Radio frequency engineering

Editor: J. Miles

CAS - CERN Accelerator School : Radio Frequency Engineering
8 - 16 May 2000 , Seeheim, Germany

REVIEW OF THEORY (I, II, III)

T. Weiland, M. Krasilnikov, R. Schuhmann, A. Skarlatos, M. Wilke
Darmstadt Technical University, Darmstadt, Germany

Abstract
After an introduction of Maxwell's equations in their most general form, an overview of the application of these equations to different types of field problems is given. The scalar and the vector wave equations are derived and solved for plane waves and guided waves in hollow waveguides. Some basic properties of modes in such guides and in cavities are discussed.

1 BASIC THEORY OF ELECTROMAGNETIC FIELDS

1.1 Maxwell's equations

The theory of electromagnetic fields is mainly involved with the solution of the set of *Maxwell's Equations* formulated by J.C. Maxwell in 1873 [1], which summarize all macroscopic phenomena of electricity. The solutions of these equations, adapted to various kinds of special problems, cover a wide range of applications. Following many other textbooks, Maxwell's equations are introduced axiomatically at the beginning of these notes. For the vector quantities

$$\vec{E}, \vec{H} \quad \text{(electric and magnetic field)} , \tag{1}$$

$$\vec{D}, \vec{B} \quad \text{(electric and magnetic flux density)} , \tag{2}$$

$$\vec{J} \quad \text{(electric current density)} , \tag{3}$$

and the scalar field

$$\rho \quad \text{(electric charge density)} \tag{4}$$

their integral form reads

$\forall A, V \subset \Omega$ and $\forall t \in R$:

$$\oint_{\partial A} \vec{E}(\vec{r}, t) \cdot \mathrm{d}\vec{s} = -\frac{d}{dt} \int_A \vec{B}(\vec{r}, t) \cdot \mathrm{d}\vec{A} , \tag{5}$$

$$\oint_{\partial A} \vec{H}(\vec{r}, t) \cdot \mathrm{d}\vec{s} = \int_A \left(\frac{\partial \vec{D}(\vec{r}, t)}{\partial t} + \vec{J}(\vec{r}, t) \right) \cdot \mathrm{d}\vec{A} , \tag{6}$$

$$\oint_{\partial V} \vec{D}(\vec{r}, t) \cdot \mathrm{d}\vec{A} = \int \rho(\vec{r}, t) \, dV , \tag{7}$$

$$\oint_{\partial V} \vec{B}(\vec{r}, t) \cdot \mathrm{d}\vec{A} = 0 . \tag{8}$$

$A, V \subset \Omega$ are arbitrary faces or volumina in the underlying problem space $\Omega \subset R^3$, and $\partial A, \partial V$ are their boundaries. Equation (5) is called Faraday's law, Eq. (6) is Ampere's law.

By use of the theorems of Gauss and Stokes we can derive the differential form of Maxwell's equations:

$$\text{curl}\,\vec{E}(\vec{r},t) = -\frac{\partial\vec{B}(\vec{r},t)}{\partial t}\,, \tag{9}$$

$$\text{curl}\,\vec{H}(\vec{r},t) = \frac{\partial\vec{D}(\vec{r},t)}{\partial t} + \vec{J}(\vec{r},t)\,, \tag{10}$$

$$\text{div}\,\vec{D}(\vec{r},t) = \rho(\vec{r},t)\,, \tag{11}$$

$$\text{div}\,\vec{B}(\vec{r},t) = 0\,. \tag{12}$$

1.1.1 Some generalizations of Maxwell's equations

With a discrete set of point charges we have

$$\int \rho(\vec{r},t)dV = \sum_i Q_i(t) = Q(t)\,, \tag{13}$$

where the summation includes all point charges inside the integration volume. Together with Eq. (7) this leads to Gauss' law of electrostatics:

$$\oint_{\partial V} \vec{D}\cdot\mathrm{d}\vec{A} = Q\,. \tag{14}$$

In Eqs. (6) and (10)

$$\frac{\partial\vec{D}(\vec{r},t)}{\partial t} + \vec{J}(\vec{r},t) = \vec{J}_{tot}(\vec{r},t) \tag{15}$$

defines the total electric current density $\vec{J}_{tot}$, including the displacement current density $\partial\vec{D}/\partial t$ and the electric current density $\vec{J}$. The electric current density itself may consist of three parts: the conduction current density $\vec{J}_c$, the convection current density $\vec{J}_{cv}$, and the impressed current density $\vec{J}_i$:

$$\vec{J}(\vec{r},t) = \vec{J}_c(\vec{r},t) + \vec{J}_{cv}(\vec{r},t) + \vec{J}_i(\vec{r},t)\,. \tag{16}$$

The conduction current density $\vec{J}_c$ is only a function of the electric field $\vec{E}$. In isotropic and linear conductors with the conductivity κ it is given by Ohm's law:

$$\vec{J}_c(\vec{r},t) = \kappa\,\vec{E}(\vec{r},t)\,. \tag{17}$$

The convection current density $\vec{J}_{cv}$ covers free charges accelerated by the Lorentz force $\vec{F} = q(\vec{E}+\vec{v}\times\vec{B})$:

$$\vec{J}_{cv}(\vec{r},t) = \rho(\vec{r},t)\,\vec{v}(\vec{r},t)\,. \tag{18}$$

The impressed current density $\vec{J}_i$ is assumed to be an externally enforced motion of charges, which is independent of all field forces (source term of electromagnetic fields).

The generalized form of Faraday's law is then

$$\oint_{\partial A} \vec{H}(\vec{r},t)\cdot\mathrm{d}\vec{s} = \int_A \left(\frac{\partial\vec{D}(\vec{r},t)}{\partial t} + \kappa\,\vec{E}(\vec{r},t) + \rho(\vec{r},t)\,\vec{v}(\vec{r},t) + \vec{J}_i(\vec{r},t)\right)\cdot\mathrm{d}\vec{A}\,. \tag{19}$$

In Ampere's law, Eq. (5), $\frac{d}{dt}$ denotes the total time derivative of the flux $\int\vec{B}\cdot\mathrm{d}\vec{A}$. This flux may change in time either by changing the magnetic flux density $\vec{B}$, or by changing the size or position of the face A. From vector analysis we get for faces moving with the velocity $\vec{v}$:

$$\oint_{\partial A} \vec{E}(\vec{r},t)\cdot\mathrm{d}\vec{s} = -\int_A \frac{\partial\vec{B}(\vec{r},t)}{\partial t}\cdot\mathrm{d}\vec{A} + \oint_{\partial A}(\vec{v}\times\vec{B}(\vec{r},t))\cdot\mathrm{d}\vec{s}\,. \tag{20}$$

1.1.2 Time-harmonic fields

In many technical applications sinusoidal field quantities occur that can be described as the real part of complex quantities:

$$f(t) = \mathrm{Re}\{\underline{f}_\omega \, e^{i\omega t}\} = |\underline{f}_\omega| \, \cos(\omega t + \varphi) \; , \tag{21}$$

where ω is the angular frequency, φ the phase angle, and $\underline{f}_\omega$ the complex amplitude. For vector quantities, e.g. the electric field vector, we can write

$$\vec{E}(\vec{r}, t) = \mathrm{Re}\{\underline{\vec{E}}_\omega(\vec{r}) \, e^{i\omega t}\} \; . \tag{22}$$

The vector complex amplitude $\underline{\vec{E}}_\omega(\vec{r})$ is also referred to as phasor. It is important to note that these complex amplitudes themselves are not physical quantities, but that the fields have to be derived by multiplying by $e^{i\omega t}$ and taking the real part.

The advantage of the complex notation is that all time derivatives can be replaced by simple multiplications with $i\omega$:

$$\frac{\partial \vec{E}(\vec{r}, t)}{\partial t} = \mathrm{Re}\{i\omega \, \underline{\vec{E}}_\omega(\vec{r}) \, e^{i\omega t}\} \; . \tag{23}$$

Maxwell's equations in stationary media can then be written as (indices ω are omitted):

$$\oint\limits_{\partial A} \underline{\vec{E}}(\vec{r}) \cdot \mathrm{d}\vec{s} \;\; = \;\; -\int\limits_{A} i\omega \underline{\vec{B}}(\vec{r}) \cdot \mathrm{d}\vec{A} \; , \tag{24}$$

$$\oint\limits_{\partial A} \underline{\vec{H}}(\vec{r}) \cdot \mathrm{d}\vec{s} \;\; = \;\; \int\limits_{A} (i\omega \underline{\vec{D}}(\vec{r}) + \underline{\vec{J}}(\vec{r})) \cdot \mathrm{d}\vec{A} \; , \tag{25}$$

$$\oint\limits_{\partial V} \underline{\vec{D}}(\vec{r}) \cdot \mathrm{d}\vec{A} \;\; = \;\; 0 \; , \tag{26}$$

$$\oint\limits_{\partial V} \underline{\vec{B}}(\vec{r}) \cdot \mathrm{d}\vec{A} \;\; = \;\; 0 \; . \tag{27}$$

Usually, the absence of free charges is assumed in this time-harmonic case (equivalent to fields in steady-state).

1.2 Fields inside matter

To complete the system of Maxwell's equations, the induction quantities $\vec{D}$ and $\vec{B}$ have to be linked to the field quantities $\vec{E}$ and $\vec{H}$. In their most general form, the constitutive equations are

$$\vec{D} \;\; = \;\; \varepsilon_0 \vec{E} + \vec{P} \; , \tag{28}$$

$$\vec{B} \;\; = \;\; \mu_0 \vec{H} + \mu_0 \vec{M} \; , \tag{29}$$

introducing the electric polarization $\vec{P}$ (also referred to as electric dipole moment density), and the magnetic polarization $\mu_0 \vec{M}$. The vector $\vec{M}$ is called magnetization. The constants

$$\varepsilon_0 \approx 8.854 \cdot 10^{-12} \frac{\mathrm{As}}{\mathrm{Vm}}, \quad \mu_0 = 4\pi \cdot 10^{-7} \frac{\mathrm{Vs}}{\mathrm{Am}} \tag{30}$$

are the permittivity and the permeability of vacuum, respectively.

The polarization quantities describe the influence of matter on electromagnetic fields. For linear materials without hysteresis and without permanent polarization they are proportional to the field vectors. The stationary relations are then given by

$$\vec{P} \;\; = \;\; \varepsilon_0 \chi_e \vec{E} \; , \tag{31}$$

$$\mu_0 \vec{M} \;\; = \;\; \mu_0 \chi_m \vec{H} \; , \tag{32}$$

with the electric and magnetic susceptibility χ_e, χ_m. For anisotropic materials the polarization vectors may have different orientations to the corresponding field vectors, which can be expressed by χ_e or χ_m becoming tensors.

1.2.1 Dielectric material characteristics

The properties of dielectric materials are based on the interaction of polarization charges (electrons or ions, in contrast to free charges in conductors) with external electromagnetic fields.

In the absence of external fields, atoms or molecules may or may not have electric dipole moments, but if they do, these are randomly oriented.[1] In the presence of a field, the atoms become polarized (or their permanent moments tend to align with the field). The dipole moment of a single polarization charge can be modelled by a pair of two point charges $\pm q$ and is expressed by

$$\vec{p} = q\vec{x}\,, \tag{33}$$

where $\vec{x}$ is the distance between the charges (see Fig. 1).

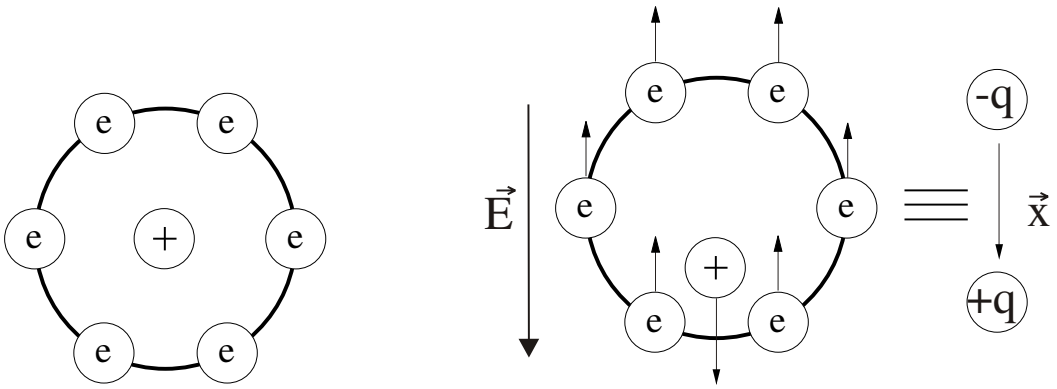


Fig. 1: Modelling of the displacement of atomic polarization charges by an electric dipole

For a set of N dipoles per unit volume with identical orientation, the macroscopic properties can be summarized in the polarization

$$\vec{P} = N\vec{p} = Nq\vec{x}\,. \tag{34}$$

Generally the polarization $\vec{P}$ is a non-linear function of the electric field $\vec{E}$, and can be expanded in terms of a power series. For many technical applications, it is sufficient to take only the linear terms into account. This leads to the susceptibility χ_e in Eq. (31), and for the flux density to

$$\begin{aligned} \vec{D} &= \varepsilon_0\vec{E} + \varepsilon_0\chi_e\vec{E} = \varepsilon_0(1+\chi_e)\vec{E} \\ &= \varepsilon_0\varepsilon_r\vec{E} \end{aligned} \tag{35}$$

with the relative permittivity ε_r (isotropic, stationary case). This linearization is also very common, if small signals are superimposed on a steady value of $\vec{E}$.

For non-stationary fields the displacement of polarization charges is a dynamic process that can be modelled by several types of differential equations. In the Debye model the first order equation (for relaxation processes)

$$\frac{\partial}{\partial t}P(t) = \frac{1}{\tau}[(\varepsilon_s - \varepsilon_\infty)\varepsilon_0\, E(t) - P(t)] \tag{36}$$

[1] Except for electrets, which have a permanent polarization.

leads to the complex permittivity in the frequency domain

$$\underline{\varepsilon}_r(\omega) = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1 + i\omega\tau} \,. \tag{37}$$

Resonant effects (e.g. for electronic polarization) can be described by the Lorentz model using a second order differential equation. The frequency dependence on the relative permittivity as a result of a superposition of several polarization mechanisms is shown in Fig. 2.

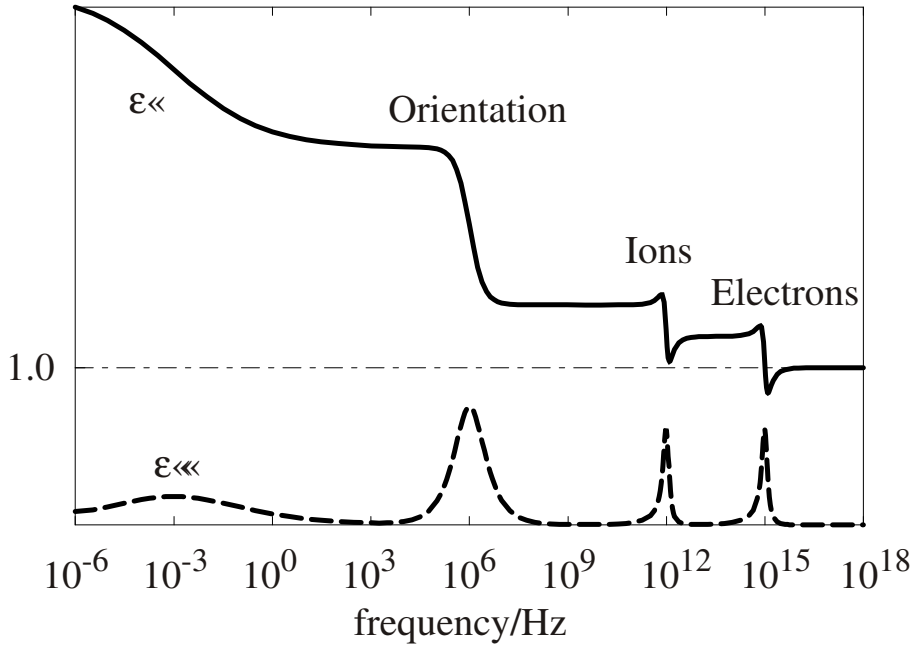


Fig. 2: Typical frequency dependence of the complex permittivity $\underline{\varepsilon}(\omega) = \varepsilon' - i\varepsilon''$ as a result of different types of polarization: orientation of permanent molecular dipoles (Debye model) and polarization of ions and electrons (Lorentz model).

A complex permittivity can also be used in conductors with a non-zero conductivity κ. Summarizing the current terms on the right-hand side of Faraday's law, Eq. (25), by

$$i\omega\underline{\vec{D}} + \underline{\vec{J}} = (i\omega\varepsilon + \kappa)\underline{\vec{E}} + \underline{\vec{J}}_i = i\omega(\varepsilon + \frac{\kappa}{i\omega})\underline{\vec{E}} + \underline{\vec{J}}_i \,, \tag{38}$$

the complex permittivity is defined as

$$\underline{\varepsilon} = \varepsilon + \frac{\kappa}{i\omega} \,. \tag{39}$$

Common notations for the complex permittivity are

$$\underline{\varepsilon} = \varepsilon' - i\varepsilon'' = \varepsilon' \,(1 - i\tan\delta_\varepsilon) \qquad (\tan\delta_\varepsilon = \varepsilon''/\varepsilon') \,, \tag{40}$$

where δ_ε is referred to as the electric loss angle of the medium.

In dielectrics we have $\tan\delta_\varepsilon \ll 1$ at most technical frequencies, and thus $\tan\delta_\varepsilon \approx \delta_\varepsilon$. On the contrary, in good conductors (metals) we have $\tan\delta_\varepsilon \gg 1$ and $\varepsilon \approx \varepsilon_0$, i.e. $\underline{\varepsilon}$ becomes almost purely imaginary.

1.2.2 Magnetic material characteristics

By analogy to electric polarization, the magnetic properties of materials can be described by means of a magnetic polarization vector, which is based on the orientation of atomic magnetic dipoles in the

presence of external fields. Such a dipole can be modelled by a circulating atomic current i_c enclosing a circuit with the normal vector $\vec{A}$. The corresponding magnetic dipole moment is defined by

$$\vec{\mu} = i_c \vec{A} \,. \tag{41}$$

For N dipoles per unit volume with identical orientation, the (macroscopic) magnetization is given by

$$\vec{M} = N\vec{\mu} = N i_c \vec{A} \,. \tag{42}$$

Referring to the dependence of the polarization mechanism on the external magnetic field, magnetic materials are classified in diamagnetic, paramagnetic, and ferromagnetic substances. The polarization in diamagnetic and paramagnetic media is approximately proportional to the external field according to Eq. (32), and the relative permeability

$$\mu_r = 1 + \chi_m \tag{43}$$

generally differs only slightly from one (e.g. $\mu_r \approx 0.999$ for diamagnetic and $\mu_r \approx 1.001$ for paramagnetic media).

In ferromagnetic substances the relation between external field and magnetic polarization is nonlinear and often depends on the preparation history of the material, leading to a hysteresis loop, as shown in Fig. 3.

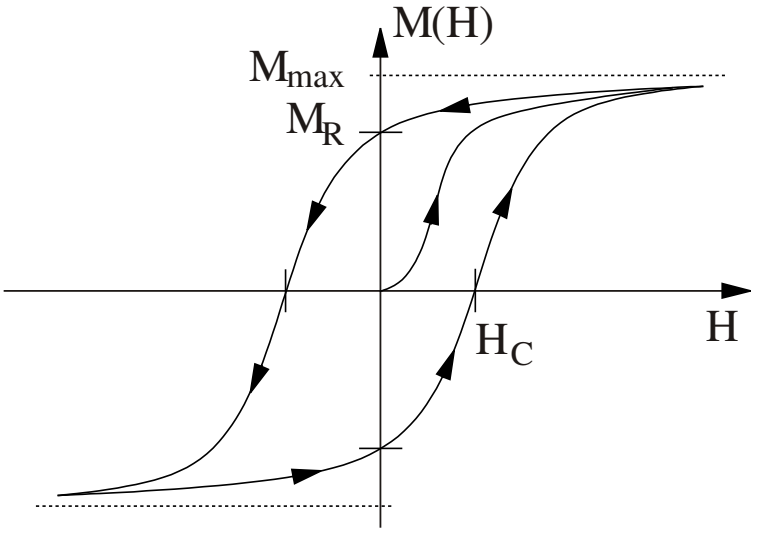


Fig. 3: Hysteresis loop of a ferromagnetic material including initial magnetization curve. H_C, M_R, and M_{max} are the coercive force, the remanence magnetization, and the saturation magnetization, respectively.

Referring to such curves, various kinds of permeability quantities can be defined. The normal permeability μ is given by the quotient B/H at the tip of the loop, and tables or graphs of μ as a function of H_{max} or B_{max} are frequently given in the literature. The differential permeability is defined as the derivative of B with respect to H, and can be as high as 10^6 for high-permeability materials. Most untreated ferromagnetic materials have a linear relation between $\vec{B}$ and $\vec{H}$ for small fields. Typical values of the initial permeability range from 10 to 10^4.

Like electric polarization, magnetization is also a dynamic process. Therefore the frequency dependence of the permeability has to be taken into account for RF fields. In the frequency domain this leads to the complex permittivity

$$\underline{\mu} = \mu' - i\mu'' = \mu' \, (1 - i \tan \delta_\mu) \tag{44}$$

with the magnetic loss angle δ_μ.

1.3 Applications of Maxwell's equations

The solutions of Maxwell's equations can be classified by several types of electromagnetic field, which can be handled by partly simplified equations: electrostatics, magnetostatics, stationary fields, quasi-stationary fields, harmonic fields, and general time-varying fields.

1.3.1 Static and stationary fields

The first simplification is based on the assumption that all field quantities are independent of time ($\partial/\partial t \equiv 0$). In this case we obtain the following system:

$$\oint_{\partial A} \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{s} \;=\; 0 \;, \tag{45}$$

$$\oint_{\partial A} \vec{H}(\vec{r}) \cdot \mathrm{d}\vec{s} \;=\; \int_{A} \vec{J}(\vec{r},t) \cdot \mathrm{d}\vec{A} \;, \tag{46}$$

$$\oint_{\partial V} \vec{D}(\vec{r}) \cdot \mathrm{d}\vec{A} \;=\; \int \rho(\vec{r})\, dV \;, \tag{47}$$

$$\oint_{\partial V} \vec{B}(\vec{r}) \cdot \mathrm{d}\vec{A} \;=\; 0 \;. \tag{48}$$

Obviously the electric-field vectors $\vec{E}$ and $\vec{D}$ and the magnetic vectors $\vec{H}$ and $\vec{B}$ do not depend on each other in current-free regions ($\vec{J} = 0$). In regions with current densities ($\vec{J} \neq 0$), they represent the sources of the magnetic fields, but there is no influence of the magnetic fields back on the electric quantities. Thus, we have two separated types of field problem: electrostatic fields and magnetic fields of stationary currents.

Electrostatic fields are governed by the equations (in differential form)

$$\mathrm{curl}\, \vec{E}(\vec{r}) \;=\; 0 \;, \tag{49}$$

$$\mathrm{div}\, \vec{D}(\vec{r}) \;=\; \rho(\vec{r}) \;. \tag{50}$$

A common approach is to fulfil Eq. (49) implicitly by defining the electric field as (negative) gradient of a scalar potential $\Phi(\vec{r})$:

$$\vec{E}(\vec{r}) = -\mathrm{grad}\, \Phi(\vec{r}) \quad \Rightarrow \quad \mathrm{curl}\, \vec{E}(\vec{r}) = -\mathrm{curl}\ \mathrm{grad}\ \Phi(\vec{r}) \equiv 0 \;. \tag{51}$$

This leads to the general equation for the scalar potential function

$$\mathrm{div}\ \varepsilon(\vec{r})\ \mathrm{grad}\ \Phi(\vec{r}) = -\rho(\vec{r}) \;. \tag{52}$$

In homogeneous regions (constant permittivity ε) we get Poisson's equation

$$\mathrm{div}\ \mathrm{grad}\, \Phi(\vec{r}) = -\frac{\rho(\vec{r})}{\varepsilon} \;. \tag{53}$$

Finally, in regions where there is no charge density, the potential satisfies Laplace's equation

$$\mathrm{div}\, \mathrm{grad}\, \Phi(\vec{r}) = 0 \;. \tag{54}$$

For the various kinds of boundary value problems emerging from these types of equations, a wide range of specialized algorithms are available [2, 3]. As an example, Fig. 4 shows the potential as well as the electric field of a two-dimensional problem.

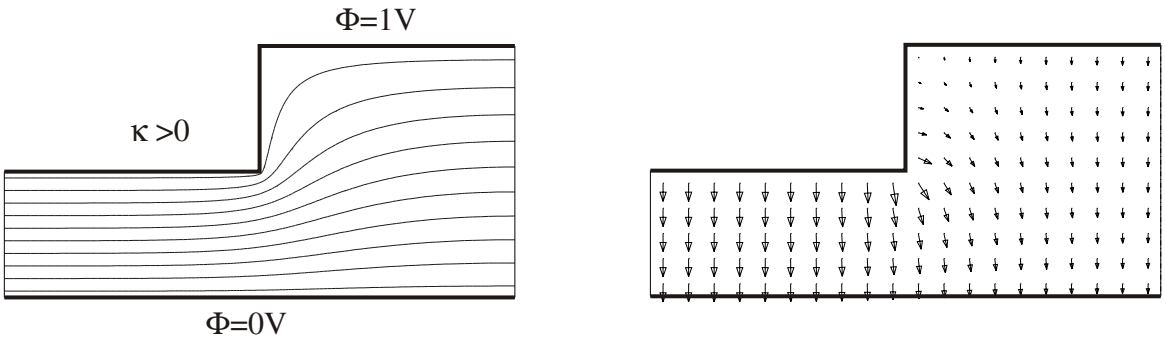


Fig. 4: Solution of an electrostatic boundary value problem: equipotential lines and electric field

An equation of the same type has to be solved for stationary current densities inside conductors. From

$$\begin{aligned} \operatorname{curl} \vec{E}(\vec{r}) &= 0, && (55)\\ \operatorname{div}\left(\kappa(\vec{r})\vec{E}(\vec{r})\right) = \operatorname{div} \vec{J}(\vec{r}) &= \operatorname{div}\operatorname{curl} \vec{H}(\vec{r}) = 0 && (56) \end{aligned}$$

we obtain the equation

$$\operatorname{div} \kappa(\vec{r}) \operatorname{grad} \Phi(\vec{r}) = 0\ . \tag{57}$$

An example of a stationary current density field $\vec{J}(\vec{r}) = -\kappa(\vec{r}) \operatorname{grad} \Phi(\vec{r})$ is shown in Fig. 5.

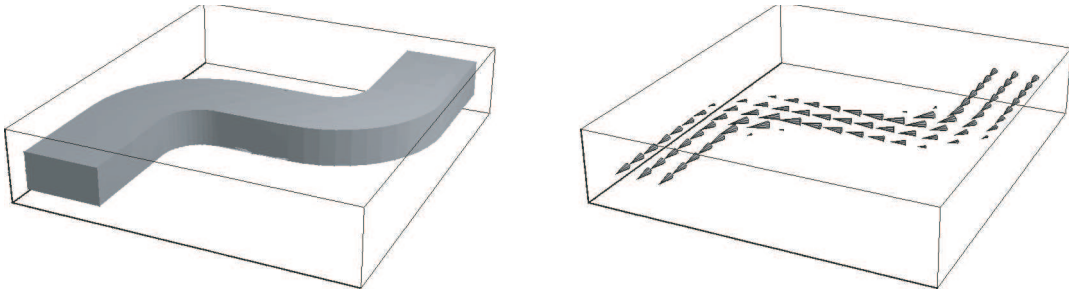


Fig. 5: Bent conductor and stationary current density

Conduction current densities as well as impressed currents are ultimately the sources for the magnetic fields of stationary currents (magnetostatic fields). Maxwell's equations now reduce to

$$\begin{aligned} \operatorname{curl} \vec{H}(\vec{r}) &= \vec{J}(\vec{r})\ , && (58)\\ \operatorname{div} \vec{B}(\vec{r}) &= 0\ . && (59) \end{aligned}$$

Because of the inhomogeneity on the right-hand side of Eq. (58), the approach with a scalar potential cannot be used here. Instead, a vector potential can be applied to fulfil Eq. (59) implicitly:

$$\vec{B}(\vec{r}) = \operatorname{curl} \vec{A}(\vec{r}) \quad \Rightarrow \quad \operatorname{div} \vec{B}(\vec{r}) = \operatorname{div}\operatorname{curl} \vec{A}(\vec{r}) \equiv 0\ . \tag{60}$$

Now the equation to solve for the potential function is

$$\operatorname{curl}\ \mu(\vec{r})^{-1} \operatorname{curl} \vec{A}(\vec{r}) = \vec{J}(\vec{r})\ . \tag{61}$$

For specialized approaches we again refer to the literature [2, 3]. As an example, Fig. 6 shows the distribution of the magnetic field $\vec{H}$ and the magnetic flux density $\vec{B}$ of a C-shaped magnet driven by two current coils.

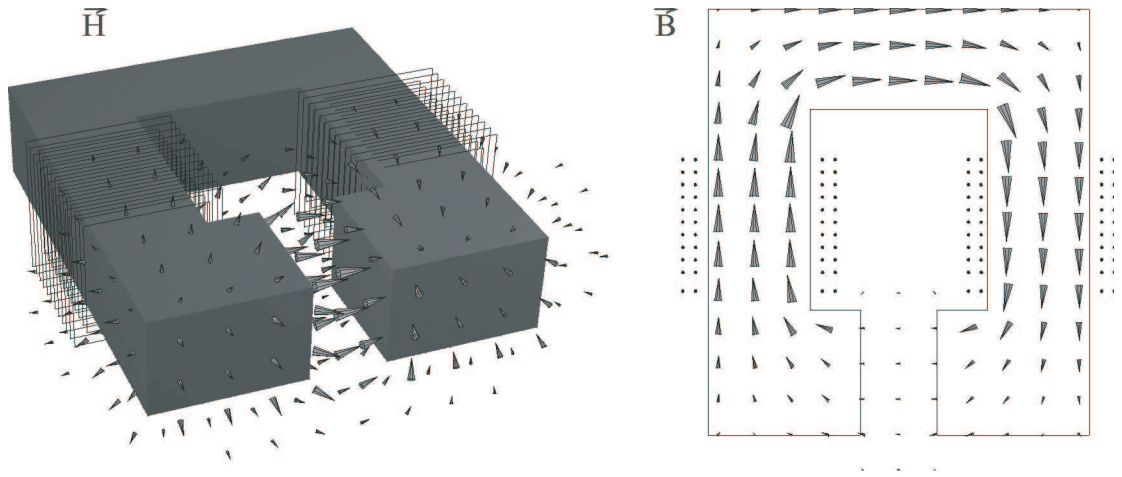


Fig. 6: Magnetic field $\vec{H}$ and magnetic flux density $\vec{B}$ of a C-shaped magnet driven by two current coils

1.3.2 Quasi-stationary fields

For quasi-stationary fields a certain time dependence is allowed, but all fields are assumed to be slowly varying. In this case, (one of) the time derivatives in Maxwell's equations can be approximated to zero.

As a first problem type, neglecting the displacement current in Eq. (10), we get magneto-quasistatic fields (see Fig. 7), with the equations

$$\operatorname{curl} \vec{E} = -\frac{\partial \vec{B}}{\partial t} , \tag{62}$$

$$\operatorname{curl} \vec{H} = \vec{J}_i + \kappa \vec{E} . \tag{63}$$

To validate the assumption $\partial \vec{D}/\partial t \approx 0$, an electric field $\vec{E}_0(\vec{r}, t)$ is considered, which is switched on during the time period T. With $\varepsilon(\vec{r}) = \varepsilon_0$ the displacement current can be approximated by

$$\left| \frac{\partial \vec{D}(\vec{r}, t)}{\partial t} \right| \approx \frac{\varepsilon_0 |\vec{E}_0(\vec{r}, t)|}{T} . \tag{64}$$

For the conduction current in a conducting material (conductivity κ) we have

$$|\vec{J}(\vec{r}, t)| = \kappa |\vec{E}_0(\vec{r}, t)| . \tag{65}$$

Therefore neglecting the displacement current is valid for

$$\frac{\varepsilon_0}{\kappa T} \ll 1 , \tag{66}$$

which is usually the case for slowly varying fields in good conductors; e.g. in copper with $T = 1$ ms:

$$\frac{\varepsilon}{\kappa T} \approx \frac{8.85 \cdot 10^{-12} As/Vm}{5.8 \cdot 10^{7} (\Omega m)^{-1} \; 10^{-3}\, s} \approx 1.5 \cdot 10^{-16} \ll 1 .$$

In the non-conducting regions of the problem domain (e.g. the surrounding air) the quasi-static approximation $|\partial \vec{D}_{air}/\partial t| \ll |\vec{J}_{air}|$ never holds, as we have $\vec{J}_{air} = 0$. However, if we are only interested in the fields inside the conductor, a quasi-static approach is permissible if the maximum value of the displacement current is much smaller than the conducting currents:

$$\max_{\vec{r} \in \Omega} \left| \frac{\partial \vec{D}(\vec{r})}{\partial t} \right| \ll \max_{\vec{r} \in \Omega} |\vec{J}_\kappa(\vec{r})| . \tag{67}$$

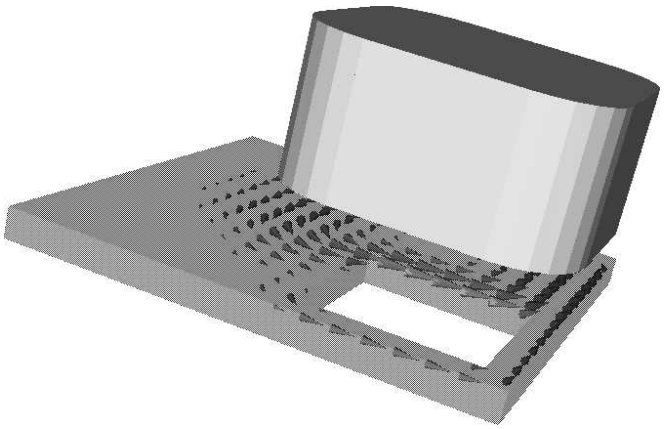


Fig. 7: Magneto-quasistatic fields: eddy currents in a copper plate with a hole induced by a coil excitation at 50 Hz

For some applications the time derivative of the magnetic flux density in Eq. (9) can also be neglected, leading to (in frequency domain)

$$\operatorname{curl} \vec{E} = 0 , \tag{68}$$

$$\operatorname{curl} \vec{H} = i\omega \vec{D} + \kappa \vec{E} . \tag{69}$$

Similar to electrostatics, such electro-quasistatic fields can be expressed in terms of a complex-valued scalar potential $\underline{\Phi}$:

$$\underline{\vec{E}} = -\operatorname{grad} \underline{\Phi} \quad \Rightarrow \quad \operatorname{curl} \underline{\vec{E}} = -\operatorname{curl} \operatorname{grad} \underline{\Phi} \equiv 0 \tag{70}$$

$$\operatorname{div} (i\omega\varepsilon + \kappa) \operatorname{grad} \underline{\Phi} = 0 . \tag{71}$$

An example, the electro-quasistatic field in an insulator contaminated by several water drops, is shown in Fig. 8.

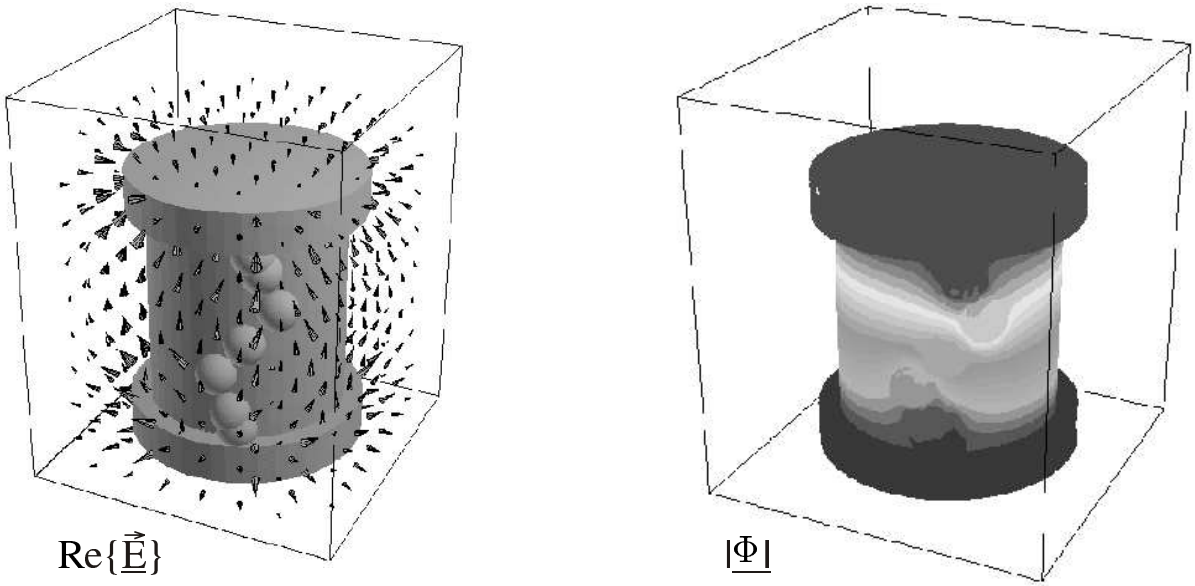


Fig. 8: High-voltage insulator contaminated by water drops at 50 Hz: real part of the electric phasor $\underline{\vec{E}}$ and absolute value of the complex potential $\underline{\Phi}$

1.3.3 Quickly varying fields

If the time derivatives of the fields cannot be neglected, the full set of Maxwell's equations have to be taken into account. In most analytical approaches — and also in the following parts of these notes — the time-harmonic formulation Eqs. (24)–(27) including the complex phasors is applied. However, especially in numerical field simulations, as well as in some measurement techniques, the usage of transient fields is also quite common.

1.4 Fields at interfaces between different media

Before we can solve electromagnetic boundary value problems, we must establish the boundary conditions satisfied by the field and flux vectors at the interface between two media with different material properties.

1.4.1 Boundary conditions for electric fields

To derive the boundary condition for the electric field $\vec{E}$ we consider the interface region shown in Fig. 9 and evaluate Faraday's law, Eq. (5), for the small circuit C_1. Assuming a very small size of C_1 we can

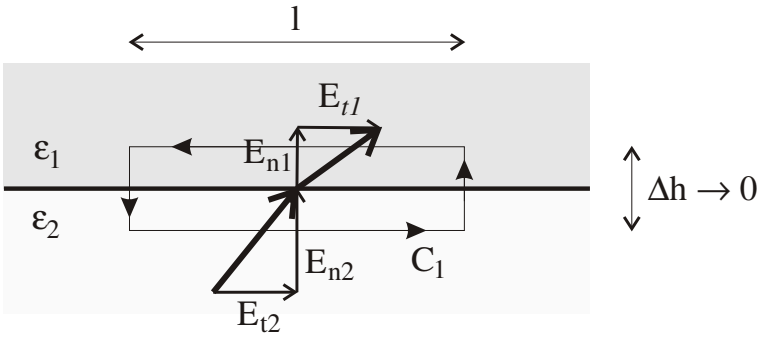


Fig. 9: Derivation of the continuity condition of the electric field. The vector fields $\vec{E}_1$ and $\vec{E}_2$ on either side of the interface are decomposed in their tangential and normal components.

approximate the fields in each of the media to be constant. We decompose them into their tangential (index 't') and their normal (index 'n') components and get

$$(E_{t2} - E_{t1}) \cdot l = \dot{B}_\perp \cdot l \, \Delta h \ , \tag{72}$$

where $B_\perp$ denotes the magnetic flux component normal to the circuit. If Δh tends to zero (keeping l fixed), the right-hand side vanishes, as the time derivative of the magnetic flux density has to be finite. This leads to the continuity of the tangential electric fields

$$E_{t2} = E_{t1} \qquad \text{or} \qquad \vec{n} \times (\vec{E}_1 - \vec{E}_2) = 0 \ , \tag{73}$$

where $\vec{n}$ is the normal vector of the interface plane.

For the normal components we consider the configuration in Fig. 10, where again the top and bottom surfaces of the cylinder are so small that constant fields can be assumed. Evaluating the surface integral on the left-hand side of Gauss' law, Eq. (7), we get

$$\oint \vec{D} \cdot \mathrm{d}\vec{A} \quad \xrightarrow{\Delta h \to 0} \quad (D_{n1} - D_{n2}) \cdot \Delta A \ , \tag{74}$$

where D_{n1} and D_{n2} are the normal components of the electric flux density on both sides of the interface. The right-hand side — the integral of the charge density inside the cylinder — vanishes for $\Delta h \to 0$, unless there is a non-zero surface charge density ρ_s at the interface plane. Therefore the boundary condition for the normal electric components reads as

$$D_{n1} - D_{n2} = \rho_s \qquad \text{or} \qquad \vec{n} \cdot (\vec{D}_1 - \vec{D}_2) = \rho_s \ . \tag{75}$$

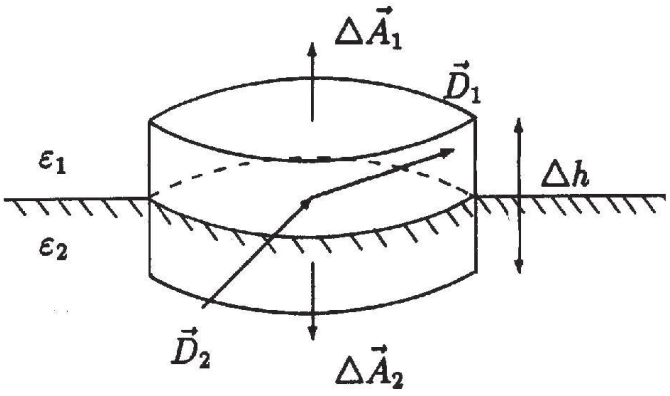


Fig. 10: Derivation of the continuity condition of the electric flux density

1.4.2 Boundary conditions for magnetic fields

Analogous configurations can be considered. For the magnetic fields, as there are no magnetic charges, the evaluation of the surface integral of the magnetic flux density (similar to Fig. 10) leads to the continuity of the normal magnetic flux density

$$\vec{n} \cdot (\vec{B}_1 - \vec{B}_2) = 0 \; . \tag{76}$$

If we evaluate Faraday's law for the circuit shown in Fig. 11, the right-hand side is given as the total current through the face ΔA:

$$\int_{\Delta A} (\vec{J} + \frac{\partial \vec{D}}{\partial t}) \cdot \mathrm{d}\vec{A} \; . \tag{77}$$

As the time derivative of the electric flux has to be finite, the integral tends to zero for $\Delta h \to 0$, unless there is an electric surface current J_s along the interface. Therefore the boundary condition for the tangential magnetic components reads as

$$H_{t2} - H_{t1} = J_s \qquad \text{or} \qquad \vec{n} \times (\vec{H}_2 - \vec{H}_1) = \vec{J}_s \; . \tag{78}$$

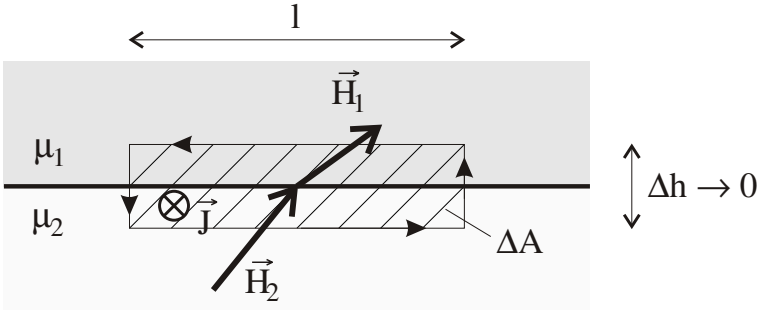


Fig. 11: Derivation of the continuity condition of the magnetic field

It can easily be shown that for non-stationary fields ($d/dt \neq 0$) it is always sufficient to impose the continuity of the tangential components of $\vec{E}$ and $\vec{H}$. This implicitly entails the continuity of the normal magnetic flux density; eventually the discontinuity of the normal components of $\vec{D}$ will yield the free electric surface charges on the interface.

1.4.3 Infinitely conducting walls

For many practical calculations the idealized model of infinitely conducting materials (also referred to as Perfect Electric Conductors, PEC) plays an important role. From $\vec{J} = \kappa\vec{E}$ we have $\vec{E} = 0$ inside the conductor, and from Ampere's law this implies $\partial\vec{B}/\partial t = 0$ and $\vec{B} = 0$ for non-stationary fields. The boundary conditions for the fields at the outer surface of the conductor then read as

$$\vec{n} \times \vec{E} = 0 \,, \tag{79}$$
$$\vec{n} \times \vec{H} = \vec{J}_s \,, \tag{80}$$
$$\vec{n} \cdot \vec{D} = \rho_s \,, \tag{81}$$
$$\vec{n} \cdot \vec{B} = 0 \,. \tag{82}$$

1.5 Electric and magnetic energy

1.5.1 Electric energy density

The energy stored in an electromagnetic field must be identical to the energy that was required for establishing this field. To derive an expression for the energy of an electrostatic field, we consider the plate capacitor (surface ΔA, width Δs) shown in Fig. 12. The field vectors $\vec{E}$ and $\vec{D}$ inside are assumed to be homogeneous. The energy required to move a charge dQ (in the opposite direction to the electric

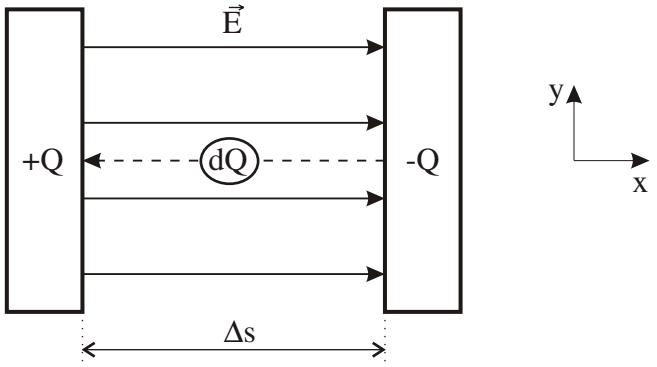


Fig. 12: Derivation of the electric energy density in a plate capacitor with surface ΔA and width Δs

field lines) from the right plate to the left amounts to

$$dW = d\vec{F} \cdot \Delta s\, \vec{e}_x = \vec{E}\, dQ \cdot \Delta s\, \vec{e}_x \,, \tag{83}$$

while the charge on the left plate increases by

$$dQ = d\vec{D} \cdot \Delta\vec{A} = d\vec{D} \cdot \Delta A\, \vec{e}_x \,. \tag{84}$$

The total energy of the field inside the capacitor volume $\Delta V = \Delta A\, \Delta s$ is given by

$$\Delta W = \int dW = \int \vec{E}(d\vec{D} \cdot \Delta A\, \vec{e}_x) \cdot \Delta s\, \vec{e}_x = \Delta V \int \vec{E} \cdot d\vec{D} \,. \tag{85}$$

The integral term is the stored energy per volume, or the energy density

$$w_e = \int \vec{E} \cdot d\vec{D} \,. \tag{86}$$

In linear isotropic media the energy density is given by

$$w_e = \frac{1}{2}\varepsilon|\vec{E}|^2 \,. \tag{87}$$

1.5.2 Magnetic energy density

A similar derivation can be performed for the energy in magnetic fields. We get the magnetic energy density

$$w_m = \int \vec{H} \cdot d\vec{B} \ , \tag{88}$$

and for linear isotropic media the simplified expression

$$w_m = \frac{1}{2}\mu |\vec{H}|^2 \ . \tag{89}$$

1.5.3 Energy densities of time-harmonic fields

The energy density of a time-harmonic electric field

$$\vec{E}(t) = \mathrm{Re}\{\underline{\vec{E}} e^{i\omega t}\} = \vec{E}_0 \ \cos(\omega t + \varphi) \tag{90}$$

in a linear isotropic medium is given by

$$\begin{aligned} w_e(t) &= \frac{1}{2}\varepsilon |\vec{E}(t)|^2 = \frac{1}{2}\varepsilon |\vec{E}_0|^2 \cos^2(\omega t + \varphi) \\ &= \frac{1}{4}\varepsilon |\vec{E}_0|^2 \ (1 + \cos(2(\omega t + \varphi))) \ . \end{aligned} \tag{91}$$

In most cases we are interested in the time average of the energy during a period $T = 2\pi/\omega$, which amounts to

$$\overline{W}_e = \int_V \overline{w}_e \mathrm{d}V \tag{92}$$

with

$$\begin{aligned} \overline{w}_e &= \frac{1}{T} \int_{t_0}^{t_0+T} w_e(t) dt \\ &= \frac{1}{4}\varepsilon |\underline{\vec{E}}|^2 = \frac{1}{4}\varepsilon \ \underline{\vec{E}} \cdot \underline{\vec{E}}^* \ , \end{aligned} \tag{93}$$

where $\underline{\vec{E}}^*$ is the conjugate vector of $\underline{\vec{E}}$. The analogous expression for the time-averaged magnetic energy density is

$$\overline{w}_m = \frac{1}{4}\mu |\underline{\vec{H}}|^2 = \frac{1}{4}\mu \ \underline{\vec{H}} \cdot \underline{\vec{H}}^* \ . \tag{94}$$

2 ELECTROMAGNETIC WAVES

2.1 Plane-wave equation

2.1.1 The wave equation in time domain

According to Maxwell's equations, a time-varying electric field is coupled (by Ampere's law) to a magnetic field, the time variation of which is in turn coupled (by Faraday's law) to an electric field. Electromagnetic field patterns that propagate through space with a finite velocity are called electromagnetic waves.

The simplest form of an electromagnetic wave is the plane wave, propagating through free space (vacuum) or homogeneous media. For this type of wave a scalar wave equation can be derived from Maxwell's equations. At first we impose a number of restrictions:

- The medium contained in the solution domain is homogeneous, isotropic and linear.
- The medium is not conducting; $\kappa = 0$.
- There is no space charge distribution and no impressed-current density; $\rho = 0$, $\vec{J}_i = 0$.
- The fields vary only in one direction, say the x-axis; $\frac{\partial}{\partial y} = \frac{\partial}{\partial z} = 0$.

Under these assumptions Maxwell's equations can be written as

$$\begin{aligned} \text{curl}\,\vec{E}(\vec{r},t) &= -\mu\frac{\partial\vec{H}(\vec{r},t)}{\partial t}\,, && (95)\\ \text{curl}\,\vec{H}(\vec{r},t) &= \varepsilon\frac{\partial\vec{E}(\vec{r},t)}{\partial t}\,, && (96)\\ \text{div}\,\vec{E}(\vec{r},t) &= \text{div}\,\vec{H}(\vec{r},t) = 0\,. && (97) \end{aligned}$$

As we supposed $\partial/\partial y = \partial/\partial z = 0$, the curl operator can be evaluated as

$$\begin{aligned} \Big(\text{curl}\,\vec{E}(x,t)\Big)_y &= -\frac{\partial E_z(x,t)}{\partial x} = -\mu\frac{\partial H_y(x,t)}{\partial t}\,, && (98)\\ \Big(\text{curl}\,\vec{E}(x,t)\Big)_z &= \frac{\partial E_y(x,t)}{\partial x} = -\mu\frac{\partial H_z(x,t)}{\partial t}\,, && (99)\\ \Big(\text{curl}\,\vec{H}(x,t)\Big)_y &= -\frac{\partial H_z(x,t)}{\partial x} = \varepsilon\frac{\partial E_y(x,t)}{\partial t}\,, && (100)\\ \Big(\text{curl}\,\vec{H}(x,t)\Big)_z &= \frac{\partial H_y(x,t)}{\partial x} = \varepsilon\frac{\partial E_z(x,t)}{\partial t}\,. && (101) \end{aligned}$$

Obviously we get two pairs of decoupled equations: Eqs. (98) and (101) with components E_z and H_y, and Eqs. (99) and (100) with components E_y and H_z. Eliminating the terms H_z and E_y from these equations we get

$$\left[\frac{\partial^2}{\partial x^2} - \mu\varepsilon\frac{\partial^2}{\partial t^2}\right]\left\{\begin{array}{c} E_y(x,t) \\ H_z(x,t) \end{array}\right\} = 0\,. \qquad (102)$$

Both H_z and E_y satisfy the following differential equation

$$\left[\frac{\partial^2}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right]\; f(x,t) = 0\,, \qquad (103)$$

where

$$v = \frac{1}{\sqrt{\mu\varepsilon}}\,. \qquad (104)$$

Equation (103) is called the homogeneous wave equation. For the solution of this equation in the time domain we use Alembert's transformation:

$$\begin{aligned} \xi &= x - v\,t\,, && (105)\\ \eta &= x + v\,t\,, && (106) \end{aligned}$$

leading to

$$\begin{aligned} \frac{\partial f}{\partial x} &= \frac{\partial f(\xi,\eta)}{\partial\xi} + \frac{\partial f(\xi,\eta)}{\partial\eta}\,, & \frac{\partial^2 f}{\partial x^2} &= \frac{\partial^2 f(\xi,\eta)}{\partial\xi^2} + 2\,\frac{\partial^2 f(\xi,\eta)}{\partial\xi\,\partial\eta} + \frac{\partial^2 f(\xi,\eta)}{\partial\eta^2}\,,\\ \frac{\partial f}{\partial t} &= -v\frac{\partial f(\xi,\eta)}{\partial\xi} + v\frac{\partial f(\xi,\eta)}{\partial\eta}\,, & \frac{\partial^2 f}{\partial t^2} &= v^2\left[\frac{\partial^2 f(\xi,\eta)}{\partial\xi^2} - 2\,\frac{\partial^2 f(\xi,\eta)}{\partial\xi\,\partial\eta} + \frac{\partial^2 f(\xi,\eta)}{\partial\eta^2}\right]\,. \end{aligned}$$

Substituting these expressions in Eq. (103) we get the simplified differential equation

$$\frac{\partial^2 f(\xi,\eta)}{\partial\xi\,\partial\eta} = 0\ . \tag{107}$$

Its general solution is

$$\begin{aligned} f(\xi,\eta) &= F(\xi) + G(\eta)\ , & (108)\\ \Rightarrow f(x,t) &= F(x - v\,t) + G(x + v\,t)\ , & (109) \end{aligned}$$

with arbitrary functions F and G. As we can see in Eq. (109), the value of F at time instant $t = t_0$ on the plane $x = x_0$ is the same as the value at time $t = t_0 + \Delta t$ on the plane $x = x_0 + v\Delta t$. In other words the field distribution is moving towards the positive x-axis as the time t is running, and v is the velocity of the wave. In an analogous manner it can be shown that the field distribution expressed by G is moving towards the negative x-direction (see Fig. 13).

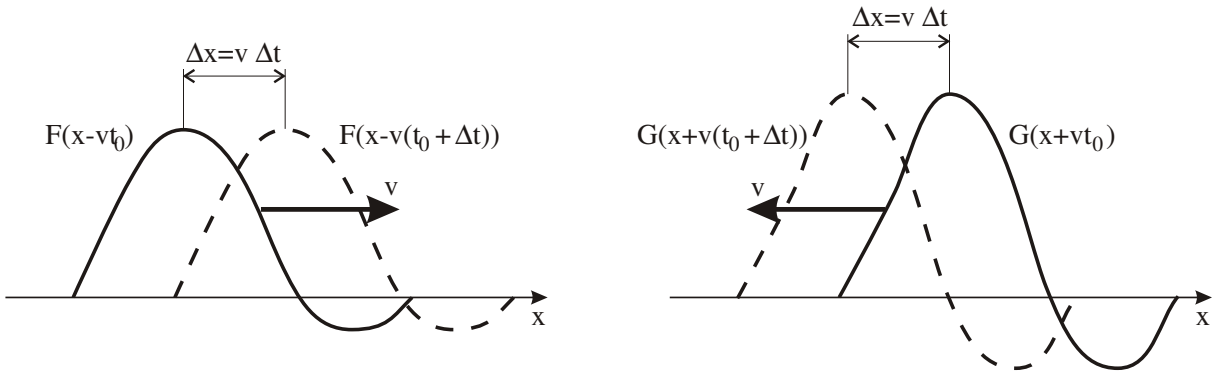


Fig. 13: Illustration of the functions F and G: spatial plots at different instants of time

Let us consider the solution for the electric field. From Eqs. (99) and (100) we get a solution for the E_y component:

$$E_y(x,t) = F(x - v\,t) + G(x + v\,t)\ . \tag{110}$$

The corresponding magnetic field can be found by evaluating Eq. (100):

$$\begin{aligned} \frac{\partial H_z}{\partial x} &= -\varepsilon\frac{\partial E_y}{\partial t} = \varepsilon\,v\,\left[F'(x - v\,t) - G'(x + v\,t)\right]\ , \\ \Rightarrow H_z &= \varepsilon\,v\,\left[F(x - v\,t) - G(x + v\,t)\right]\ . & (111) \end{aligned}$$

A second, independent solution can be obtained from Eqs. (98) and (101), having field components E_z and H_y. Both solutions are referred to as the two linear polarizations of the plane wave. In general, a superposition of both solutions results in an elliptic polarization of the wave.

In Eqs. (110) and (111), the ratio

$$Z = \frac{E_y}{H_z} = \frac{1}{\varepsilon\,v} = \sqrt{\frac{\mu}{\varepsilon}} \tag{112}$$

is referred to as the wave impedance (or intrinsic impedance) of the medium. In vacuum we have

$$Z = \sqrt{\frac{\mu_0}{\varepsilon_0}} \approx 377\,\Omega\ . \tag{113}$$

The solutions (110) and (111) represent an electromagnetic wave, whose components are normal to the direction of propagation. This kind of wave belongs to the category of plane waves, as the surfaces of equal phase and equal amplitude form planes.

In the beginning of the analysis we supposed that $\partial/\partial y = \partial/\partial z = 0$. This assumption implies that both the amplitude and the phase of the fields remain constant on the yz-plane. Thus the yz-planes are surfaces of both equal phase and equal amplitude. Waves of this kind are called homogeneous waves.

Putting it all together, the solution of the wave equation as derived in this paragraph is a homogeneous plane wave, see Fig. 14. This wave has no longitudinal field components, i.e. neither electric nor magnetic components in the direction of propagation (in our case $E_x = H_x = 0$). Therefore it belongs to the category of Transverse Electromagnetic (TEM) waves.

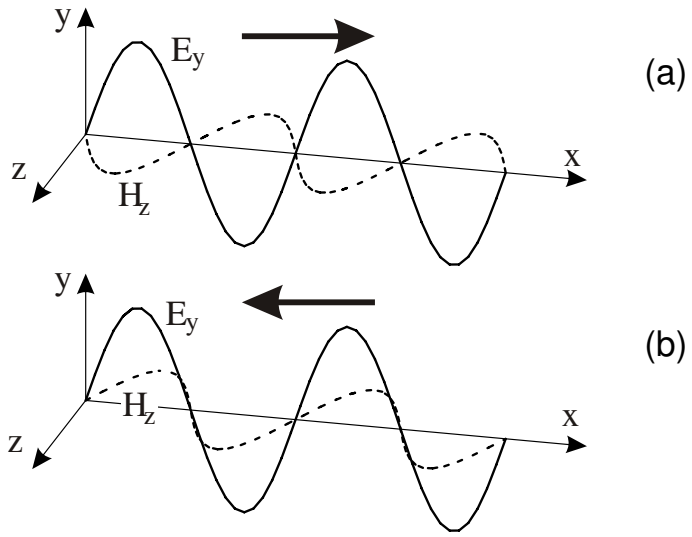


Fig. 14: Field components E_y and H_z of a homogeneous plane wave propagating in the (a) $+x$- or (b) $-x$-direction

In our analysis we assumed a wave propagation in the direction of the x-axes, which of course is not the general case. If the direction of propagation is defined by a unit vector $\vec{n}$, see Fig. 15, the solution of the wave equation is

$$f(\vec{r}, t) = f(\vec{n} \cdot \vec{r} - vt) \ . \tag{114}$$

The relation between the electric and magnetic field vectors is then given by

$$\vec{H} = \frac{1}{Z} \left(\vec{n} \times \vec{E} \right) \ . \tag{115}$$

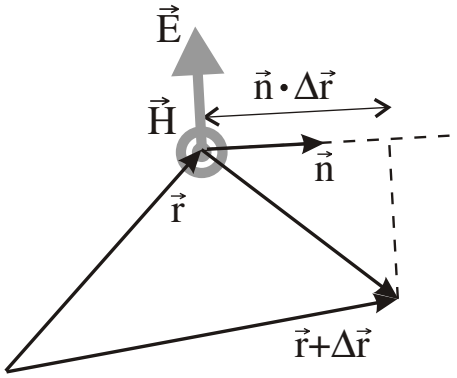


Fig. 15: Propagation of a plane wave in an arbitrary direction $\vec{n}$

2.1.2 *Time-harmonic plane waves*

In the previous section we obtained a solution of the wave equation inside an isotropic, homogeneous material for an arbitrary time variation. In many cases we are interested in problems in which the time variation is sinusoidal. In this case the complex notation for the time varying quantities can be used.

The assumptions we make here are the same as before, with the only difference being that the restriction of the non-conducting medium ($\kappa = 0$) is skipped. The rest of the assumptions remain the same: $\partial/\partial y = \partial/\partial z = 0$, ε and μ shall be constant, and no space-charge distribution is taken into account.

We define the complex propagation constant

$$\underline{\gamma} = \alpha + i\,\beta = i\underline{k} = i\omega\sqrt{\mu\underline{\varepsilon}}\,, \tag{116}$$

where the complex permittivity $\underline{\varepsilon}$ may or may not contain a conductivity term according to Eq. (39). The homogeneous wave equation (103) from time domain can then be transformed into the complex, homogeneous wave equation (without time variables)

$$\left[\frac{\partial^2}{\partial x^2} - \underline{\gamma}^2\right]\,\underline{f}(x) = 0\,. \tag{117}$$

A differential equation of this form is called a Helmholtz equation, and its solution is (for the electric field component $\underline{f} = \underline{E}_y$)

$$\underline{E}_y(x) = \underline{C}_1\,e^{-\underline{\gamma}\,x} + \underline{C}_2\,e^{+\underline{\gamma}\,x}\,, \tag{118}$$

The related magnetic field component $\underline{H}_z$ evaluates to

$$\underline{H}_z(x) = \frac{1}{\underline{Z}}\left[\underline{C}_1\,e^{-\underline{\gamma}\,x} - \underline{C}_2\,e^{+\underline{\gamma}\,x}\right]\,. \tag{119}$$

The term $\underline{Z}$ is the complex wave impedance:

$$\underline{Z} = \sqrt{\frac{\mu}{\underline{\varepsilon}}}\,. \tag{120}$$

Equations (118) and (119) describe the superposition of two independent waves propagating in the two opposite directions of the x-axis. The constant $\underline{\gamma} = \alpha + i\beta$ in Eq. (116) is called the propagation constant. From

$$\mathrm{Re}\{e^{-\underline{\gamma}x}\} = \mathrm{Re}\{e^{-\alpha x}\cdot e^{-i\beta x}\} = e^{-\alpha x}\cdot\cos(\beta x) \tag{121}$$

we find that the real part of $\underline{\gamma}$, the attenuation constant α, denotes the attenuation of the field amplitude, and the imaginary part, the phase constant β, denotes the phase shift along the direction of wave propagation (see Fig. 16). In a perfect dielectric (non-conducting) medium we get

$$\alpha = 0 \qquad \beta = \omega\sqrt{\varepsilon\mu}\,. \tag{122}$$

The time dependence of the field components in Eqs. (118) and (119) is implied in the complex notation. For example, the electric field propagating in the positive x-direction is given by

$$\begin{aligned} E_y(x,t) &= \mathrm{Re}\{\underline{C}_1\,e^{-\underline{\gamma}\,x}\,e^{i\omega\,t}\} = \mathrm{Re}\{|\underline{C}_1|\,e^{i\varphi}\;e^{-\alpha\,x - i\beta\,x + i\omega\,t}\} \\ &= |\underline{C}_1|\,e^{-\alpha\,x}\,\cos(\beta\,x - \omega\,t - \varphi) \\ &= |\underline{C}_1|\,e^{-\alpha\,x}\,\cos(\beta[x - v_p\,t] - \varphi) \qquad \text{with } v_p = \frac{\omega}{\beta}\,. \end{aligned} \tag{123}$$

The wavelength λ of a wave is defined as the spatial distance between two neighbouring planes of equal phase (see Fig. 17). From the relation $\beta\,\lambda = 2\,\pi$ we get

$$\lambda = \frac{2\,\pi}{\beta}\,. \tag{124}$$

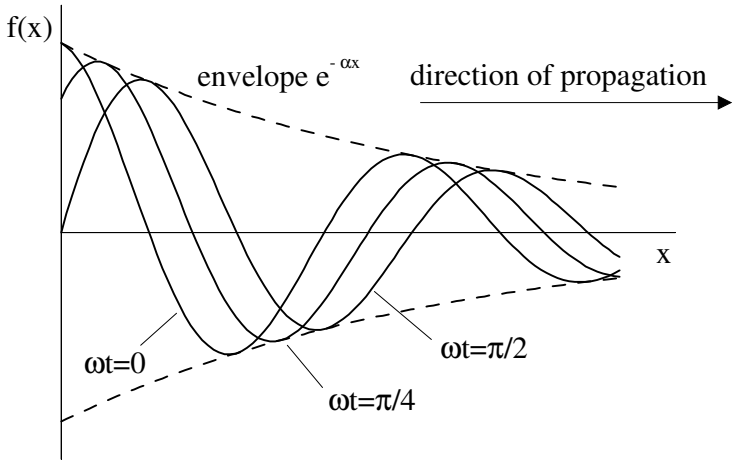


Fig. 16: Attenuated wave at several instants of time and envelope.

The period T of the wave is defined as

$$T = \frac{1}{f} = \frac{2\,\pi}{\omega} = \frac{2\,\pi}{\beta\, v_p}\,. \tag{125}$$

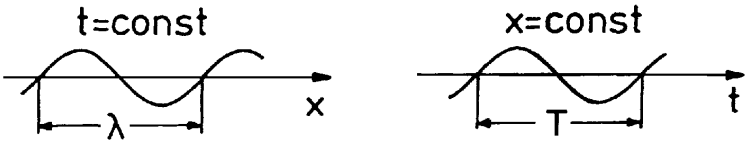


Fig. 17: Wavelength λ and period T of a harmonic plane wave

2.2 Waves in lossy matter, skin depth

In the general case of a lossy medium the propagation constant is a complex quantity, and a plane wave is attenuated in the direction of propagation according to $\exp(-\alpha x)$. As the wave impedance $\underline{Z}$ in Eq. (119) is complex too, $\underline{\vec{H}}$ is no longer in phase with $\underline{\vec{E}}$ (usually $\underline{\vec{H}}$ lags behind $\underline{\vec{E}}$). A sketch of $\underline{\vec{E}}$ and $\underline{\vec{H}}$ versus x at some instant of time would be similar to Fig. 16.

Two cases of particular interest are (i) good dielectrics (low-loss), and (ii) good conductors (high-loss). Both cases can be modelled by a complex permittivity quantity $\underline{\varepsilon} = \varepsilon' - i\varepsilon''$, where the imaginary part ε'' expresses the losses of the medium.

In the low-loss case, we have $\varepsilon'' \ll \varepsilon'$ and

$$\sqrt{\varepsilon' - i\varepsilon''} \approx \sqrt{\varepsilon'} - i\frac{\varepsilon''}{2\sqrt{\varepsilon'}} \quad \Rightarrow \quad \alpha \approx \frac{\omega\varepsilon''}{2}\sqrt{\frac{\mu}{\varepsilon'}} \qquad \beta \approx \omega\sqrt{\mu\varepsilon'} \tag{126}$$

$$|\underline{Z}| \approx \sqrt{\frac{\mu}{\varepsilon'}} \qquad \underline{Z} \approx \arctan\frac{\varepsilon''}{2\varepsilon'}\,. \tag{127}$$

Thus the attenuation is very small, and $\underline{\vec{E}}$ and $\underline{\vec{H}}$ are nearly in phase. The wave is almost the same as in the loss-free dielectric case. For example, in polystyrene a 10 MHz wave is attenuated only 0.5% per kilometre, and the phase difference between $\underline{\vec{E}}$ and $\underline{\vec{H}}$ is only $0.003°$.

In the high-loss case with $\kappa \neq 0$ we have $\underline{\varepsilon} \approx -i\varepsilon'' = -i\frac{\kappa}{\omega}$ and

$$\sqrt{\underline{\varepsilon}} \approx \sqrt{\frac{\kappa}{2\omega}}\,(1-i) \quad \Rightarrow \qquad \alpha \approx \sqrt{\frac{\omega\mu\kappa}{2}} \qquad\qquad \beta \approx \sqrt{\frac{\omega\mu\kappa}{2}} \tag{128}$$

$$|\underline{Z}| \approx \sqrt{\frac{\omega\mu}{\kappa}} \qquad\qquad \underline{Z} \approx \frac{\pi}{4}\,. \tag{129}$$

Thus the attenuation is very large, and $\underline{\vec{H}}$ lags behind $\underline{\vec{E}}$ by $45°$. The intrinsic impedance of a good conductor is extremely small at radio frequencies having a magnitude of $1.16 \cdot 10^{-3}\,\Omega$ for copper at 10 MHz. The wavelength is also very small compared to the free-space wavelength. For example, at 10 MHz the free-space wavelength is 30 m, while in copper the wavelength is only 0.131 mm. The attenuation in a good conductor is very rapid. For the 10 MHz wave mentioned above in copper the attenuation is 99.81% in 0.131 mm of travel. Thus waves do not penetrate metals very deeply. A metal acts as a shield against electromagnetic waves.

A wave starting at the surface of a good conductor and propagating inward is very quickly damped to insignificant values. The field is localized in a thin surface layer, this phenomenon being known as 'skin effect'. The distance over which a wave is attenuated to $1/e$ (36.8%) of its initial value is called the skin depth or depth of penetration δ. It is defined by $\alpha\delta = 1$, or

$$\delta = \frac{1}{\alpha} = \sqrt{\frac{2}{\omega\mu\kappa}}\,. \tag{130}$$

The skin depth is very small for good conductors at radio frequencies, for example for copper at 10 MHz it is only 0.021 mm.

2.3 Energy density, energy flow, Poynting vector

2.3.1 The Poynting vector

The space containing a time-varying electromagnetic field is a carrier of electric and magnetic energy. According to Eqs. (87) and (89) the energy density inside a linear medium (i.e. a medium with constant permittivity and permeability independent of the field intensities) can be written as

$$w = w_e + w_m = \frac{1}{2}(\vec{E}\cdot\vec{D} + \vec{H}\cdot\vec{B})\ . \tag{131}$$

Thus, the total stored energy inside the volume V is

$$W = \frac{1}{2}\int\limits_V (\vec{E}\cdot\vec{D} + \vec{H}\cdot\vec{B})\,\mathrm{d}V\ . \tag{132}$$

The time variation of the energy is given by

$$\frac{dW}{dt} = \frac{1}{2}\int\limits_V \frac{\partial}{\partial t}(\vec{E}\cdot\vec{D} + \vec{H}\cdot\vec{B})\,\mathrm{d}V\ . \tag{133}$$

A scalar product can be differentiated using the product law. So from the first term we get

$$\frac{\partial}{\partial t}(\vec{E}\cdot\vec{D}) = \vec{E}\cdot\frac{\partial\vec{D}}{\partial t} + \frac{\partial\vec{E}}{\partial t}\cdot\vec{D} = 2\vec{E}\cdot\frac{\partial\vec{D}}{\partial t}\ . \tag{134}$$

In this equation the two terms of the sum can be reduced to a single term under the additional assumption that the dielectric is isotropic, i.e. ε is a scalar value independent of the orientation of the fields. Under the same assumption for the permeability μ the time variation of the energy gives

$$\frac{dW}{dt} = \int\limits_V (\vec{E}\cdot\frac{\partial\vec{D}}{\partial t} + \vec{H}\cdot\frac{\partial\vec{B}}{\partial t})\,\mathrm{d}V\,. \tag{135}$$

Substituting Maxwell's equations

$$\frac{\partial \vec{B}}{\partial t} = -\text{curl}\,\vec{E}\,, \tag{136}$$

$$\frac{\partial \vec{D}}{\partial t} = \text{curl}\,\vec{H} - \vec{J} \tag{137}$$

in Eq. (135), we get

$$\frac{dW}{dt} = \int_V [\vec{E}\cdot(\text{curl}\,\vec{H} - \vec{J}) + \vec{H}\cdot(-\text{curl}\,\vec{E})]\,\text{d}V \tag{138}$$

$$= \int_V [\vec{E}\cdot\text{curl}\,\vec{H} - \vec{H}\cdot\text{curl}\,\vec{E} - \vec{E}\cdot\vec{J}]\,\text{d}V\,. \tag{139}$$

Using the vector identity

$$\text{div}\,(\vec{A}\times\vec{B}) = \vec{B}\cdot\text{curl}\,\vec{A} - \vec{A}\cdot\text{curl}\,\vec{B} \tag{140}$$

leads to

$$\frac{dW}{dt} = \int_V \left(-\text{div}\,(\vec{E}\times\vec{H}) - \vec{E}\cdot\vec{J}\right)\,\text{d}V\,. \tag{141}$$

Applying Gauss' law we finally get the following equation, known as Poynting's law (see also Fig. 18):

$$\boxed{\frac{dW}{dt} = -\int_{\partial V}(\vec{E}\times\vec{H})\cdot\text{d}\vec{A} - \int_V (\vec{E}\cdot\vec{J})\,\text{d}V\,.} \tag{142}$$

It states that the change of the electromagnetic energy inside a volume V can have two causes: (i) energy in the form of electromagnetic radiation is flowing through the boundary surface ∂V of the volume, (ii) an amount of the electromagnetic energy contained inside the volume is being converted into another form of energy or vice versa.

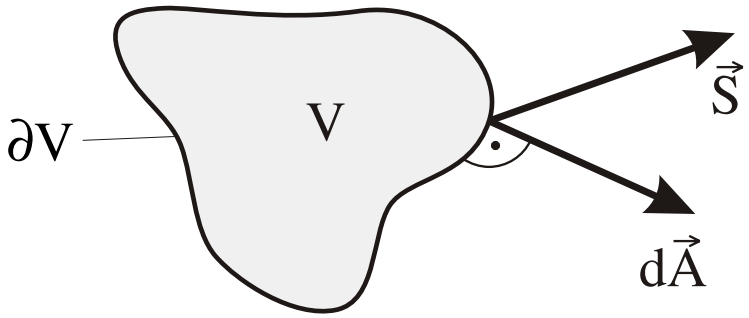


Fig. 18: Integration volume V and its boundary ∂V in Poynting's law

The Poynting vector

$$\vec{S} = \vec{E}\times\vec{H} \tag{143}$$

is defined as the cross product of the electric and magnetic field intensity. At every point of the space it points in the direction in which the electromagnetic energy flows. The integral of this vector throughout an arbitrary surface defines the amount of energy per second penetrating through this surface. The Poynting vector can thus be interpreted as a surface density of the energy flow. The terms energy flow vector or radiation vector are also commonly used.

The second term in Poynting's law describes the transformation of electromagnetic energy from or into another form of energy. In the general case the current density can be written

$$\vec{J} = \kappa \, \vec{E} + \vec{J}_i \; , \tag{144}$$

where $\vec{J}_i$ denotes the impressed-current density. Using these notations the second term of Poynting's law yields

$$\int_V (\vec{E} \cdot \vec{J}) \, \mathrm{d}V = \int_V \kappa \, |\vec{E}|^2 \, \mathrm{d}V + \int_V (\vec{E} \cdot \vec{J}_i) \, \mathrm{d}V \; . \tag{145}$$

The volume integral of $\kappa \, |\vec{E}|^2$ is always positive. Representing Joulean heat losses, it leads to a dissipation of the electromagnetic energy. The second term, the volume integral of $\vec{E} \cdot \vec{J}_i$, can also take negative values (if $\vec{E}$ and $\vec{J}_i$ are not in the same direction). In that case it can lead to an increase of the electromagnetic energy; an example is the radiation from an externally driven antenna.

2.3.2 The complex Poynting vector

When the time variation of the field is harmonic with angular frequency ω, again the complex notation of Maxwell's equations can be used:

$$\operatorname{curl} \underline{\vec{E}} = -i\omega \underline{\mu} \underline{\vec{H}} \; , \tag{146}$$

$$\operatorname{curl} \underline{\vec{H}} = i\omega \underline{\varepsilon} \, \underline{\vec{E}} \; , \tag{147}$$

with generally complex material quantities

$$\underline{\varepsilon} = \varepsilon \, (1 - i \tan \delta_\varepsilon) \quad \text{and} \quad \underline{\mu} = \mu \, (1 - i \tan \delta_\mu) \; . \tag{148}$$

To derive a modified form of Poynting's law for the phasors $\underline{\vec{E}}$ and $\underline{\vec{H}}$, Eq. (146) is left-multiplied by $\underline{\vec{H}}^*$ (the conjugate value of $\underline{\vec{H}}$):

$$\underline{\vec{H}}^* \cdot \operatorname{curl} \underline{\vec{E}} = -i\omega \underline{\mu} \underline{\vec{H}} \cdot \underline{\vec{H}}^* \; . \tag{149}$$

Left-multiplication of the conjugate form of Eq. (147) by $\underline{\vec{E}}$ yields

$$\underline{\vec{E}} \cdot \operatorname{curl} \underline{\vec{H}}^* = -i\omega \underline{\varepsilon}^* \, \underline{\vec{E}} \cdot \underline{\vec{E}}^* \; . \tag{150}$$

Using the vector identity (140) again, we get

$$\begin{aligned} \operatorname{div} (\underline{\vec{E}} \times \underline{\vec{H}}^*) &= \underline{\vec{H}}^* \cdot \operatorname{curl} \underline{\vec{E}} - \underline{\vec{E}} \cdot \operatorname{curl} \underline{\vec{H}}^* \\ &= -i\omega \underline{\mu} \underline{\vec{H}} \cdot \underline{\vec{H}}^* + i\omega \underline{\varepsilon}^* \underline{\vec{E}} \cdot \vec{E}^* \; . \end{aligned} \tag{151}$$

The Poynting vector according to its definition (143) can be written

$$\vec{S}(t) = \mathcal{R}\mathrm{e}\{\underline{\vec{\mathrm{E}}} \, \mathrm{e}^{\mathrm{i}\omega\mathrm{t}}\} \times \mathcal{R}\mathrm{e}\{\underline{\vec{\mathrm{H}}} \, \mathrm{e}^{\mathrm{i}\omega\mathrm{t}}\} \; . \tag{152}$$

In most cases we are interested in determining the time average of the energy flux density $\overline{\vec{S}}$ rather than its instantaneous value $\vec{S}(t)$. Thus we integrate Eq. (152) over a period $T = 2\pi/\omega$ and divide it by the period

$$\begin{aligned} \overline{\vec{S}} &= \frac{1}{T} \int_0^T \mathrm{Re}\{\underline{\vec{E}} \, e^{i\omega t}\} \times \mathrm{Re}\{\underline{\vec{H}} \, e^{i\omega t}\} \, \mathrm{d}t \\ &= \ldots = \mathrm{Re} \left\{ \frac{1}{2} \left(\underline{\vec{E}} \times \underline{\vec{H}}^* \right) \right\} \; . \end{aligned} \tag{153}$$

The vector

$$\underline{\vec{S}} := \frac{1}{2}\left(\underline{\vec{E}} \times \underline{\vec{H}}^*\right) \tag{154}$$

is called the complex Poynting vector. Its real part is equal to the time average of the energy flux density.

The complex form of Poynting's law is finally derived using Eq. (151):

$$\begin{aligned}
\int\limits_{\partial V} \underline{\vec{S}} \cdot \mathrm{d}\vec{A} &= -2i\omega \int\limits_V \left(\frac{\underline{\mu}|\underline{\vec{H}}|^2}{4} - \frac{\underline{\varepsilon}^*|\underline{\vec{E}}|^2}{4} \right) \mathrm{d}V \\
&= -2\omega \tan\delta_\mu \int\limits_V \frac{\mu}{4}|\underline{\vec{H}}|^2 \mathrm{d}V \; - \; 2\omega \tan\delta_\varepsilon \int\limits_V \frac{\varepsilon}{4}|\underline{\vec{E}}|^2 \mathrm{d}V \\
&\quad -2i\omega\Big(\int\limits_V \frac{\mu}{4}|\underline{\vec{H}}|^2 \mathrm{d}V - \int\limits_V \frac{\varepsilon}{4}|\underline{\vec{E}}|^2 \mathrm{d}V \Big) \\
&= -N_W - iN_B = -N_W - 2i\omega(W_m - W_e) \; .
\end{aligned} \tag{155}$$

The active power N_W represents the time-averaged Joulean heat produced inside the integration volume V, and the reactive power N_B represents the difference between the time-averaged magnetic and electric energy stored inside the volume.

For good conductors (metal) we have $W_e \ll W_m$ and

$$\begin{aligned}
\int\limits_{\partial V} \underline{\vec{S}} \cdot \mathrm{d}\vec{A} &\approx -\int\limits_V \frac{1}{2}\kappa|\vec{E}|^2 \mathrm{d}V - i2\omega W_m \\
&= -N_W - iN_B \; .
\end{aligned} \tag{156}$$

This relationship enables us to calculate the resistance R_W and the internal inductance L_i of the conductor, especially in high frequencies. Defining complex current and voltage quantities as

$$\underline{I}^* = \oint \underline{\vec{H}}^* \cdot \mathrm{d}\vec{s} \tag{157}$$

and

$$\begin{aligned}
\underline{U} &= \int\limits_1^2 \underline{\vec{E}} \cdot \mathrm{d}\vec{l} \\
&= \underline{I}\,(R_W + i\omega L_i) = \underline{I}\,\underline{Z} \, ,
\end{aligned} \tag{158}$$

N_W and N_B are given by

$$\begin{aligned}
N_W &= \frac{1}{2}|\underline{I}|^2 R_W \; , \\
N_B &= \frac{1}{2}|\underline{I}|^2 \omega L_i \; .
\end{aligned} \tag{159}$$

2.3.3 Phase, group and energy velocity

The velocity at which an equiphase surface travels is called the phase velocity of the wave. An equiphase plane $x_p(t)$ of a plane wave is defined by

$$\omega t - \beta x_p(t) = \text{constant} \; , \tag{160}$$

as we can see in Eq. (123). To derive the phase velocity of the wave, we differentiate this equation,

$$\omega - \beta \frac{dx_p}{dt} = 0 \; , \tag{161}$$

and obtain

$$v_p = \frac{\omega}{\beta} = \lambda\, f \ . \tag{162}$$

In a non-conducting medium v_p is given by

$$v_p = \frac{1}{\sqrt{\varepsilon\mu}} \ . \tag{163}$$

As we know from wireless communications for the transmission of signals, wave trains rather than single monochromatic waves are used. In this case the velocity at which the signal travels is not the phase velocity but the group velocity v_g. The simplest case is one wave train consisting of two different sinusoidal waves of the same amplitude, but with slightly different frequencies $\omega_1 = \omega$ and $\omega_2 = \omega + \Delta\omega$ where $\Delta\omega \ll \omega$:

$$f(x,t) = \hat{f}\ \sin(\omega\, t - \beta\, x) + \hat{f}\ \sin([\omega + \Delta\omega]\, t - [\beta + \Delta\beta]\, x) \ . \tag{164}$$

The superposition of the two waves creates a beat (see Fig. 19) with an envelope propagating without any significant change in shape at the group velocity v_p.

Assuming $\Delta\omega \ll \omega$ and $\Delta\beta \ll \beta$, the resulting wave is given by

$$f(x,t) = 2\ \hat{f}\ \underbrace{\sin(\omega\, t - \beta\, x)}_{\text{carrier}}\ \underbrace{\cos(\frac{1}{2}[\Delta\omega\, t - \Delta\beta\, x])}_{\text{signal information}} \ . \tag{165}$$

While the high-frequency component (carrier) travels at the phase velocity

$$v_p = \frac{\omega}{\beta} \ , \tag{166}$$

the envelope (signal) travels with the group velocity

$$v_g = \frac{\Delta\omega}{\Delta\beta} \to \frac{d\omega}{d\beta} \ . \tag{167}$$

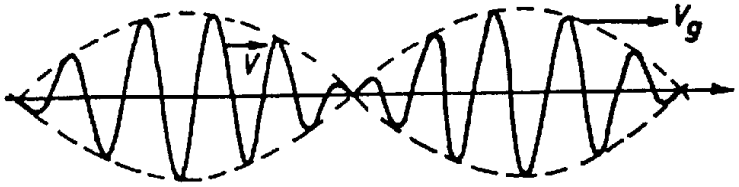


Fig. 19: The propagation of the envelope (signals carrier) of a wave train consisting of two waves with slightly differing frequencies

The energy stored in the electromagnetic fields of a wave propagates through space at a velocity called energy velocity v_E. To derive (one component of) this velocity we consider the small volume in Fig. 20.

The energy penetrating through the face ΔA_x during the time Δt is given by

$$N_x\, \Delta t = \Delta A_x\ S_x\ \Delta t \ , \tag{168}$$

where S_x is the x-component of the Poynting vector $\vec{S} = \vec{E} \times \vec{H}$. It is equivalent to the energy stored in the volume $\Delta A_x\, \Delta x$, with $\Delta x = v_{E,x}\, \Delta t$:

$$W = W'\, \Delta x = w\ \Delta A\, \Delta x \ , \tag{169}$$

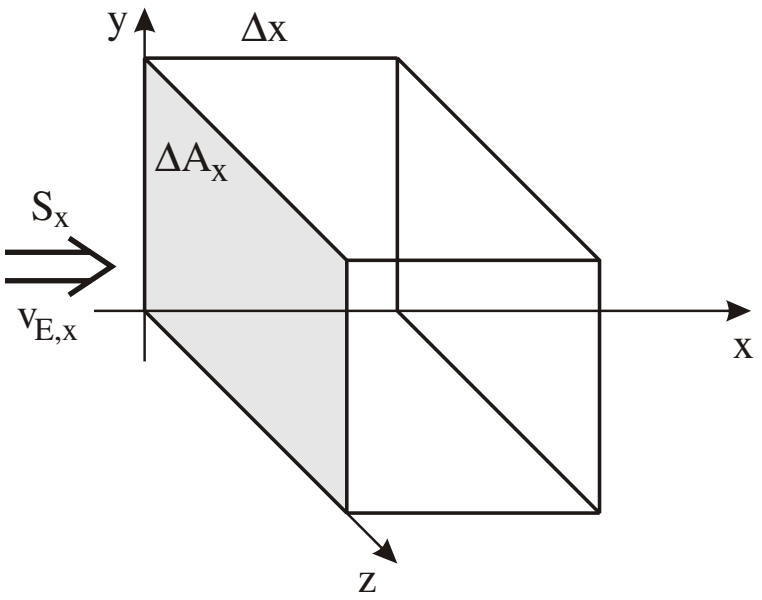


Fig. 20: Derivation of the x-component $v_{E,x}$ of the energy velocity of an electromagnetic wave

where the energy density w includes both the electric w_e and the magnetic w_m energy density.

For the energy velocity we obtain

$$N_x \, \Delta t = W' \, \Delta x \qquad \Rightarrow \qquad v_{E,x} = \frac{N_x}{W'} = \frac{S_x}{w} \; . \tag{170}$$

In the general case we get the velocity field

$$\vec{v}_E(\vec{r},t) = \frac{\vec{S}(\vec{r},t)}{w(\vec{r},t)} \; . \tag{171}$$

In most time-harmonic applications we are interested in the time-averaged quantity

$$\vec{v}_{E,\omega}(\vec{r}) = \frac{\mathrm{Re}\{\underline{\vec{S}}(\vec{r})\}}{\overline{w}(\vec{r})} \; . \tag{172}$$

For time-harmonic plane waves propagating in the x-direction in a loss-free medium,

$$\underline{\vec{E}} = \underline{E}_y \, \vec{e}_y \; e^{-i\beta x} \; , \qquad \underline{\vec{H}} = \frac{\underline{E}_y}{Z} \, \vec{e}_z \; e^{-i\beta x} \; , \tag{173}$$

we have

$$\underline{\vec{S}} \;=\; \frac{1}{2}(\underline{\vec{E}} \times \underline{\vec{H}}^*) = \frac{|\underline{E}_y|^2}{2Z} \, \vec{e}_x \; , \tag{174}$$

$$\underline{w} \;=\; \frac{1}{4}\varepsilon|\underline{E}_y|^2 + \frac{1}{4}\mu\frac{|\underline{E}_y|^2}{Z^2} = \frac{1}{2}\varepsilon|\underline{E}_y|^2 \; . \tag{175}$$

For the energy velocity we obtain

$$v_{E,\omega} = \frac{|\underline{E}_y|^2}{Z\varepsilon|\underline{E}_y|^2} = \frac{1}{\sqrt{\varepsilon\mu}} = c \; . \tag{176}$$

We find that in this case the energy velocity is identical to the phase velocity v_p.

For monochromatic plane waves all velocity definitions lead to the same value, i.e. we have $v_E = v_g = v_p$. This result, however, is *not* valid for general waves. The group velocity as well as the phase velocity may be larger than c, and in some cases the group velocity can even be negative. The energy velocity, however, is always bounded by $v_E \leq c$, ensuring the principle of causality. In waveguides we usually have

$$v_g = v_E < c \quad \text{and} \quad v_p > c \,. \tag{177}$$

3 BASIC THEORY OF WAVEGUIDES

3.1 Rectangular waveguides with perfectly conducting walls

3.1.1 The wave equation for waveguides (vector potential)

In this section the wave propagation in a rectangular waveguide is investigated. The guide should possess perfectly conducting walls and should be positioned (see Fig. 21) so the wave propagates in the $\pm z$-direction. The medium in the guide should be homogeneous, isotropic and linear. Assuming time-

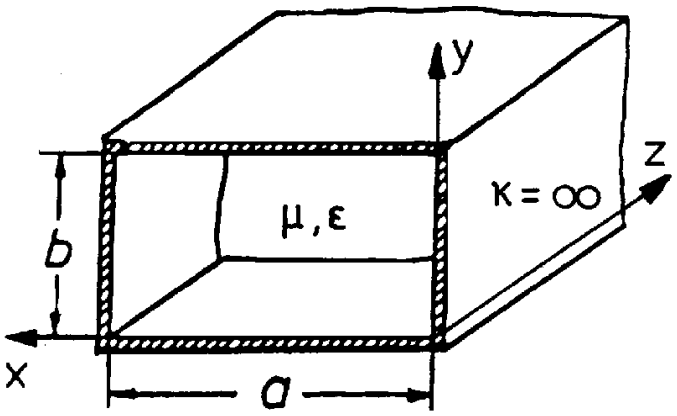


Fig. 21: Position of the rectangular waveguide in a Cartesian coordinate system

harmonic field quantities ($\sim e^{i\omega t}$) and, furthermore, the absence of space charges, Maxwell's equations result in

$$\operatorname{curl} \underline{\vec{E}} = -i\omega\mu \underline{\vec{H}} \,, \tag{178}$$

$$\operatorname{curl} \underline{\vec{H}} = i\omega\varepsilon \underline{\vec{E}} \,, \tag{179}$$

$$\operatorname{div} \underline{\vec{E}} = 0 \,, \tag{180}$$

$$\operatorname{div} \underline{\vec{H}} = 0 \,. \tag{181}$$

It can be shown that two fundamental forms of five-component waves can exist in a waveguide. On the one hand, there are E waves with one vanishing component of the magnetic field (e.g. $\underline{H}_z = 0, \underline{E}_z \neq 0$); on the other hand, there are H waves with one vanishing component of the electric field (e.g. $\underline{E}_z = 0, \underline{H}_z \neq 0$). Hence, Maxwell's equations lead to two different classes of solutions. The general solution is obtained by a superposition of E and H waves.

To solve Eq. (178)–(181), a vector potential $\vec{A}$ is introduced. Hereby two approaches are possible:

$$\underline{\vec{E}} := \operatorname{curl} \underline{\vec{A}}^H \tag{182}$$

$$\underline{\vec{H}} := \operatorname{curl} \underline{\vec{A}}^E \,. \tag{183}$$

The choice of the indices becomes more obvious later on.

This approach yields with Eq. (178):

$$\begin{aligned} \operatorname{curl} \underline{\vec{E}} &= \operatorname{curl}\operatorname{curl} \underline{\vec{A}}^H \\ &= \operatorname{grad}\operatorname{div} \underline{\vec{A}}^H - \Delta \underline{\vec{A}}^H \\ &= -i\omega\mu\, \underline{\vec{H}}. \end{aligned} \tag{184}$$

Insertion in Eq. (179) yields:

$$\begin{aligned} \operatorname{curl} \underline{\vec{H}} &= i\omega\varepsilon \operatorname{curl} \underline{\vec{A}}^H \\ \Rightarrow\ \underline{\vec{H}} &= i\omega\varepsilon\, \underline{\vec{A}}^H + \operatorname{grad} \underline{\Phi}^H, \end{aligned} \tag{185}$$

where $\operatorname{grad} \underline{\Phi}^H$ is an undefined constant of integration.
Inserting the last relation in Eq. (184) leads to:

$$\operatorname{grad}\operatorname{div} \underline{\vec{A}}^H - \Delta \underline{\vec{A}}^H = \omega^2\mu\varepsilon \underline{\vec{A}}^H - i\omega\mu \operatorname{grad} \underline{\Phi}^H .$$

According to the Lorentz convention $\underline{\Phi}^H$ is chosen as follows:

$$\operatorname{div} \underline{\vec{A}}^H = -i\omega\mu\, \underline{\Phi}^H . \tag{186}$$

This approach yields with Eq. (179):

$$\begin{aligned} \operatorname{curl} \underline{\vec{H}} &= \operatorname{curl}\operatorname{curl} \underline{\vec{A}}^E \\ &= \operatorname{grad}\operatorname{div} \underline{\vec{A}}^E - \Delta \underline{\vec{A}}^E \\ &= i\omega\varepsilon\, \underline{\vec{E}}. \end{aligned} \tag{187}$$

Insertion in Eq. (178) yields:

$$\begin{aligned} \operatorname{curl} \underline{\vec{E}} &= -i\omega\mu \operatorname{curl} \underline{\vec{A}}^E \\ \Rightarrow\ \underline{\vec{E}} &= -i\omega\mu\, \underline{\vec{A}}^E - \operatorname{grad} \underline{\Phi}^E , \end{aligned} \tag{188}$$

where $\operatorname{grad} \underline{\Phi}^E$ is an undefined constant of integration.
Inserting the last relation in Eq. (187) leads to:

$$\operatorname{grad}\operatorname{div} \underline{\vec{A}}^E - \Delta \underline{\vec{A}}^E = \omega^2\mu\varepsilon \underline{\vec{A}}^E - i\omega\varepsilon \operatorname{grad} \underline{\Phi}^E .$$

According to the Lorentz convention $\underline{\Phi}^E$ is chosen as follows:

$$\operatorname{div} \underline{\vec{A}}^E = -i\omega\varepsilon\, \underline{\Phi}^E . \tag{189}$$

For both cases the same coordinate system independent vector wave equation is obtained:

$$\boxed{\Delta \underline{\vec{A}} + k^2 \underline{\vec{A}} = 0} \tag{190}$$

with the wavenumber k:

$$\boxed{k^2 = \omega^2\mu\varepsilon\ .} \tag{191}$$

3.1.2 Solutions of the wave equation

In Cartesian coordinates the vector potential may have the components A_x, A_y, and A_z. It can be shown, however, that a vector potential with only one component is sufficient to describe all wave solutions:

$$\underline{\vec{A}}(x,y,z) = \underline{A}_z(x,y,z)\, \vec{e}_z\ . \tag{192}$$

Thus from the vector wave equation (190) the scalar wave equation, the so-called Helmholtz equation, can be obtained:

$$\frac{\partial^2 \underline{A}_z}{\partial x^2} + \frac{\partial^2 \underline{A}_z}{\partial y^2} + \frac{\partial^2 \underline{A}_z}{\partial z^2} + k^2 \underline{A}_z = 0\ . \tag{193}$$

Taking a Bernoulli ansatz in a product form,

$$\underline{A}_z(x,y,z) = \underline{f}(x)\, \underline{g}(y)\, \underline{h}(z)\ , \tag{194}$$

Eq. (193) yields

$$\underbrace{\frac{1}{\underline{f}} \frac{d^2 \underline{f}}{d x^2}}_{=-k_x^2} + \underbrace{\frac{1}{\underline{g}} \frac{d^2 \underline{g}}{d y^2}}_{=-k_y^2} + \underbrace{\frac{1}{\underline{h}} \frac{d^2 \underline{h}}{d z^2}}_{=-k_z^2} + k^2 = 0\ . \tag{195}$$

To fulfil this differential equation for all x, y, z, the addends must be constants (real or complex) and must satisfy the dispersion relation

$$k_x^2 + k_y^2 + k_z^2 = k^2. \tag{196}$$

Appropriate solution functions for a waveguide extending in the z-direction are given by

$$\begin{aligned} \underline{f}(x) &= \underline{A}_1 \cos k_x x + \underline{A}_2 \sin k_x x \,, & (197)\\ \underline{g}(y) &= \underline{B}_1 \cos k_y y + \underline{B}_2 \sin k_y y \,, & (198)\\ \underline{h}(z) &= \underline{C}_1 \, e^{+ik_z z} + \underline{C}_2 \, e^{-ik_z z} \,. & (199) \end{aligned}$$

Introducing the notation

$$\left\{ \begin{array}{c} \cos k_x x \\ \sin k_x x \end{array} \right\} = \underline{A}_1 \cos k_x x + \underline{A}_2 \sin k_x x \tag{200}$$

the general solution for the vector potential is

$$\underline{A}_z(x,y,z) = \left\{ \begin{array}{c} \cos k_x x \\ \sin k_x x \end{array} \right\} \left\{ \begin{array}{c} \cos k_y y \\ \sin k_y y \end{array} \right\} \left\{ \begin{array}{c} e^{+ik_z z} \\ e^{-ik_z z} \end{array} \right\} . \tag{201}$$

Equation (201) is the solution for the vector potential for both H and E waves. The field components derived from this solution, however, differ depending on whether ansatz (182) or Eq. (183) is chosen:

a) H **waves**:

$$\underline{\vec{A}} = \underline{\vec{A}}^H \;\Rightarrow\; \underline{\vec{E}} = \operatorname{curl} \underline{\vec{A}} \tag{202}$$

Using this approach, H waves with $\underline{H}_z \neq 0$ are received. As these waves only have transversal components of the electric field, they are also referred to as TE waves.

The electric-field components of H waves are implied by the approach above, while the magnetic-field components can be derived from the electric field using Faraday's law Eq. (178):

$$\underline{\vec{H}} = -\frac{1}{i\omega\mu} \operatorname{curl} \underline{\vec{E}} = -\frac{1}{i\omega\mu} \operatorname{curl} \operatorname{curl} \underline{\vec{A}} \,. \tag{203}$$

In the next equations, first the general derivation of all electric- and magnetic-field components is given, and then — in the last column — the explicit expressions assuming the vector potential in Eq. (201). Applying the upper sign of the exponential function, a wave propagation in the $-z$-direction, and otherwise in the $+z$-direction, is described.

$$\begin{aligned}
\underline{E}_x &= +\frac{\partial \underline{A}_z}{\partial y} &&= \underline{C}\, k_y \left\{ \begin{array}{c} \sin k_x x \\ \cos k_x x \end{array} \right\} \left\{ \begin{array}{c} \cos k_y y \\ -\sin k_y y \end{array} \right\} e^{\pm ik_z z} \,, \\
\underline{E}_y &= -\frac{\partial \underline{A}_z}{\partial x} &&= \underline{C}\, k_x \left\{ \begin{array}{c} -\cos k_x x \\ \sin k_x x \end{array} \right\} \left\{ \begin{array}{c} \sin k_y y \\ \cos k_y y \end{array} \right\} e^{\pm ik_z z} \,, \\
\underline{H}_x &= -\frac{1}{i\omega\mu} \frac{\partial^2 \underline{A}_z}{\partial x \, \partial z} &&= \mp \underline{C} \frac{k_x k_z}{\omega\mu} \left\{ \begin{array}{c} \cos k_x x \\ -\sin k_x x \end{array} \right\} \left\{ \begin{array}{c} \sin k_y y \\ \cos k_y y \end{array} \right\} e^{\pm ik_z z} \,, \qquad (204) \\
\underline{H}_y &= -\frac{1}{i\omega\mu} \frac{\partial^2 \underline{A}_z}{\partial y \, \partial z} &&= \mp \underline{C} \frac{k_y k_z}{\omega\mu} \left\{ \begin{array}{c} \sin k_x x \\ \cos k_x x \end{array} \right\} \left\{ \begin{array}{c} \cos k_y y \\ -\sin k_y y \end{array} \right\} e^{\pm ik_z z} \,, \\
\underline{H}_z &= -\frac{k^2 - k_z^2}{i\omega\mu} \underline{A}_z &&= -\underline{C} \frac{k^2 - k_z^2}{i\omega\mu} \left\{ \begin{array}{c} \sin k_x x \\ \cos k_x x \end{array} \right\} \left\{ \begin{array}{c} \sin k_y y \\ \cos k_y y \end{array} \right\} e^{\pm ik_z z} \,.
\end{aligned}$$

b) *E* **waves**:

$$\underline{\vec{A}} = \underline{\vec{A}}^E \;\Rightarrow\; \underline{\vec{H}} = \text{curl}\,\underline{\vec{A}}\ . \tag{205}$$

Using this approach, E waves with $\underline{E}_z \neq 0$ are obtained, which only have transversal components of the magnetic field (TM waves). Here the magnetic-field components are implied by the potential approach above, while the electric components can be derived from the magnetic field using Ampere's law, Eq. (179):

$$\underline{\vec{E}} = \frac{1}{i\omega\varepsilon}\text{curl}\,\underline{\vec{H}} = \frac{1}{i\omega\varepsilon}\text{curl curl}\,\underline{\vec{A}}\ . \tag{206}$$

The five field components of the wave read as

$$\begin{aligned}
\underline{H}_x &= +\frac{\partial \underline{A}_z}{\partial y} = \underline{C}\,k_y \begin{Bmatrix} \sin k_x x \\ \cos k_x x \end{Bmatrix} \begin{Bmatrix} \cos k_y y \\ -\sin k_y y \end{Bmatrix} e^{\pm ik_z z}\ , \\
\underline{H}_y &= -\frac{\partial \underline{A}_z}{\partial x} = \underline{C}\,k_x \begin{Bmatrix} -\cos k_x x \\ \sin k_x x \end{Bmatrix} \begin{Bmatrix} \sin k_y y \\ \cos k_y y \end{Bmatrix} e^{\pm ik_z z}\ , \\
\underline{E}_x &= \frac{1}{i\omega\varepsilon}\frac{\partial^2 \underline{A}_z}{\partial x\,\partial z} = \pm\underline{C}\frac{k_x k_z}{\omega\varepsilon} \begin{Bmatrix} \cos k_x x \\ -\sin k_x x \end{Bmatrix} \begin{Bmatrix} \sin k_y y \\ \cos k_y y \end{Bmatrix} e^{\pm ik_z z}\ , \\
\underline{E}_y &= \frac{1}{i\omega\varepsilon}\frac{\partial^2 \underline{A}_z}{\partial y\,\partial z} = \pm\underline{C}\frac{k_y k_z}{\omega\varepsilon} \begin{Bmatrix} \sin k_x x \\ \cos k_x x \end{Bmatrix} \begin{Bmatrix} \cos k_y y \\ -\sin k_y y \end{Bmatrix} e^{\pm ik_z z}\ , \\
\underline{E}_z &= \frac{k^2 - k_z^2}{i\omega\varepsilon}\underline{A}_z = \underline{C}\,\frac{k^2 - k_z^2}{i\omega\varepsilon} \begin{Bmatrix} \sin k_x x \\ \cos k_x x \end{Bmatrix} \begin{Bmatrix} \sin k_y y \\ \cos k_y y \end{Bmatrix} e^{\pm ik_z z}\ .
\end{aligned} \tag{207}$$

The superposition of all H_z and E_z waves yields the general solution of Eqs. (178–181). Instead of the vector potential in the z-direction, a potential in the x- or y-direction can be chosen, also leading to the general solution. However, the approach of a potential in the direction of propagation is usually preferred in a rectangular waveguide.

3.1.3 Classification of waveguide modes

So far, for the solution of the vector wave equation the specific shape of the waveguide has not been exploited. This is done in the next step, imposing boundary conditions for the electric-field components. At the (perfectly conducting) walls of the rectangular guide in Fig. 21 the tangential components must vanish, leading to (for all z):

1) $\underline{E}_x(y=0) = \underline{E}_x(y=b) = 0 \qquad \forall x \in [0,a]\ ,$
2) $\underline{E}_y(x=0) = \underline{E}_y(x=a) = 0 \qquad \forall y \in [0,b]\ ,$
3) $\underline{E}_z(y=0) = \underline{E}_z(y=b) = 0 \qquad \forall x \in [0,a]\ ,$
$\underline{E}_z(x=0) = \underline{E}_z(x=a) = 0 \qquad \forall y \in [0,b]\ .$

A TE wave with $\underline{E}_x \sim \partial \underline{A}_z^H/\partial y$ and $\underline{E}_y \sim \partial \underline{A}_z^H/\partial x$ satisfies the first and second boundary conditions only if all sine-terms vanish in Eq. (201), i.e.

$$\underline{A}_z^H \sim \cos k_x x\ \cos k_y y\ .$$

This leads to (for all z)

$$\begin{aligned}
\underline{E}_x &\sim \cos k_x x\ \sin k_y y\ , \\
\underline{E}_y &\sim \sin k_x x\ \cos k_y y\ .
\end{aligned}$$

The boundary conditions, moreover, require

$$\sin k_x a = 0 \;\text{ and }\; \sin k_y b = 0 \;,$$

which can be satisfied by the eigenvalues

$$k_x = \frac{m\,\pi}{a} \;\text{ and }\; k_y = \frac{n\,\pi}{b} \;, \tag{208}$$

where $m, n = 0, 1, 2, \ldots$ except $m = n = 0$ (this would lead to a trivial solution with all fields equal to zero).

In the TM case the third boundary condition requires vanishing cosine-terms because of $\underline{E}_z \sim \underline{A}_z^E$, i.e.

$$\underline{A}_z^E \sim \sin k_x x \;\sin k_y y \;.$$

An additional requirement is

$$\sin k_x a = 0 \;\text{ and }\; \sin k_y b = 0 \;,$$

leading to the *same* eigenvalues as in the TE case

$$k_x = \frac{m\,\pi}{a} \;\text{ and }\; k_y = \frac{n\,\pi}{b} \;,$$

but now $m, n = 1, 2, ...$ excluding the trivial solution.

In both approaches the introduction of boundary conditions leads to an infinite number of discrete modes, which are distinguished by the parameters m and n. These modes are referred to by the notation

$$\mathrm{TE}_{z\,mn} \quad \text{and} \quad \mathrm{TM}_{z\,mn} \tag{209}$$

and have the following field components:

a) $\mathrm{TE}_{z\,mn}$ modes (H_z waves):

$$\underline{E}_x = -\underline{C}^H \,\frac{n\,\pi}{b}\cos(\frac{m\,\pi}{a}x)\;\sin(\frac{n\,\pi}{b}y)\;e^{\pm i k_z z} \;, \tag{210}$$

$$\underline{E}_y = \underline{C}^H \,\frac{m\,\pi}{a}\sin(\frac{m\,\pi}{a}x)\;\cos(\frac{n\,\pi}{b}y)\;e^{\pm i k_z z} \;, \tag{211}$$

$$\underline{E}_z = 0 \;, \tag{212}$$

$$\underline{H}_x = \pm\frac{k_z}{\omega\mu}\,\underline{E}_y \;, \tag{213}$$

$$\underline{H}_y = \mp\frac{k_z}{\omega\mu}\,\underline{E}_x \;, \tag{214}$$

$$\underline{H}_z = -\underline{C}^H \,\frac{k^2 - k_z^2}{i\omega\mu}\cos(\frac{m\,\pi}{a}x)\;\cos(\frac{n\,\pi}{b}y)\;e^{\pm i k_z z} \;. \tag{215}$$

b) $\mathrm{TM}_{z\,mn}$ modes (E_z waves):

$$\underline{H}_x = \underline{C}^E \,\frac{n\,\pi}{b}\sin(\frac{m\,\pi}{a}x)\;\cos(\frac{n\,\pi}{b}y)\;e^{\pm i k_z z} \;, \tag{216}$$

$$\underline{H}_y = -\underline{C}^E \,\frac{m\,\pi}{a}\cos(\frac{m\,\pi}{a}x)\;\sin(\frac{n\,\pi}{b}y)\;e^{\pm i k_z z} \;, \tag{217}$$

$$\underline{H}_z = 0 \;, \tag{218}$$

$$\underline{E}_x = \mp\frac{k_z}{\omega\varepsilon}\,\underline{H}_y \;, \tag{219}$$

$$\underline{E}_y = \pm\frac{k_z}{\omega\varepsilon}\,\underline{H}_x \;, \tag{220}$$

$$\underline{E}_z = \underline{C}^E \,\frac{k^2 - k_z^2}{i\omega\varepsilon}\sin(\frac{m\,\pi}{a}x)\;\sin(\frac{n\,\pi}{b}y)\;e^{\pm i k_z z} \;. \tag{221}$$

In a rectangular waveguide the wave impedance is defined by the quotient of the transversal components:

$$\underline{Z}_W = \frac{\underline{E}_T}{\underline{H}_T} \; . \tag{222}$$

Considering Eqs. (213) and (214) for TE waves or Eqs. (219) and (220) for TM waves, the wave impedance is

$$\underline{Z}_W = \begin{cases} \underline{Z}^H = \dfrac{\omega\mu}{k_z} & \text{TE modes} \\ \underline{Z}^E = \dfrac{k_z}{\omega\varepsilon} & \text{TM modes} \; . \end{cases} \tag{223}$$

The propagation characteristics of the waveguide modes are determined by the longitudinal wavenumber k_z, and depend on the transversal wavenumbers (eigenvalues) k_x and k_y according to the dispersion relation (196):

$$\begin{aligned} k_x^2 + k_y^2 + k_z^2 &= k^2 = \omega^2\mu\varepsilon \\ \Rightarrow \quad k_z^2 &= \omega^2\mu\varepsilon - (k_x^2 + k_y^2) = \frac{\omega^2}{c^2} - (k_x^2 + k_y^2) \; . \end{aligned} \tag{224}$$

For frequencies with $(\omega/c)^2 > k_x^2 + k_y^2$ the squared wavenumber k_z^2 is positive and we get a propagating wave:

$$\mathrm{Re}\{e^{ik_z z}\} = \cos(k_z z) \tag{225}$$

with a real k_z. For lower frequencies with $(\omega/c)^2 < k_x^2 + k_y^2$ the wavenumber becomes imaginary, $k_z = i\alpha$, and an exponential attenuation in the z-direction occurs:

$$\mathrm{Re}\{e^{ik_z z}\} = e^{-\alpha z} \; . \tag{226}$$

The limit between both cases defines the cutoff frequency of the waveguide mode:

$$f_c = \frac{c}{2\pi}\sqrt{k_x^2 + k_y^2} \; . \tag{227}$$

With $k_x = \frac{m\pi}{a}$ and $k_y = \frac{n\pi}{b}$ in the rectangular waveguide we get

$$f_{c\,mn} = \frac{1}{\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m}{2\,a}\right)^2 + \left(\frac{n}{2\,b}\right)^2} . \tag{228}$$

Below this frequency the corresponding mode cannot propagate in the guide.

At frequencies above the cutoff frequency $f_{c\,mn}$, waves propagate corresponding to $e^{\pm ik_z z}$ in the positive or negative z-direction at a frequency f along the guide with the wavelength

$$\lambda_G = \frac{\lambda}{\sqrt{1 - (\frac{\lambda}{\lambda_c})^2}} \; , \tag{229}$$

where λ is the free-space and λ_c is the cutoff wavelength.

All modes in a waveguide can be sorted by referring to their cutoff frequency. In guides with $a > b$ (see Fig. 21) the mode with the lowest cutoff frequency is the TE_{10} mode. For example, for $a = 2 \cdot b$ it is the only propagating mode in the frequency range $f \in [f_{c,10},\, 2 \cdot f_{c,10}]$, and therefore the most important working mode in rectangular guides. The field pattern of this mode is shown in Fig. 22. The TM mode with the lowest cutoff frequency is the $E_{z\,11}$ mode.

As an example we consider a waveguide with

$$a = 6 \text{ cm}, \quad b = 3 \text{ cm} \; , \tag{230}$$

allowing modes with the following cutoff frequencies according to Eq. (228):

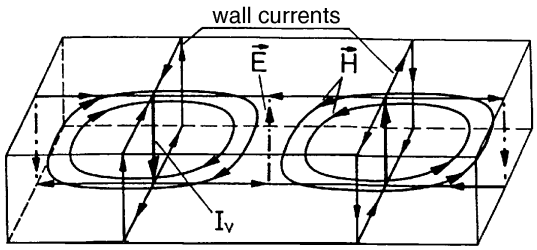


Fig. 22: $H_{z\,10}$-wave in a rectangular waveguide

Type	m	n	f_c/GHz
TE	1	0	2.498
TE	0	1	4.997
TE	2	0	4.997
TE,TM	1	1	5.586
TE,TM	2	1	7.066
			...

Figure 23 shows the dispersion curves $k_z^2(f^2)$ (left) and $|k_z(f)|$ (right) of these modes. For frequencies below cutoff, k_z^2 becomes negative and k_z imaginary. For frequencies 2.498 GHz $< f <$ 4.997 GHz the TE_{10} mode is the only propagating mode in this waveguide.

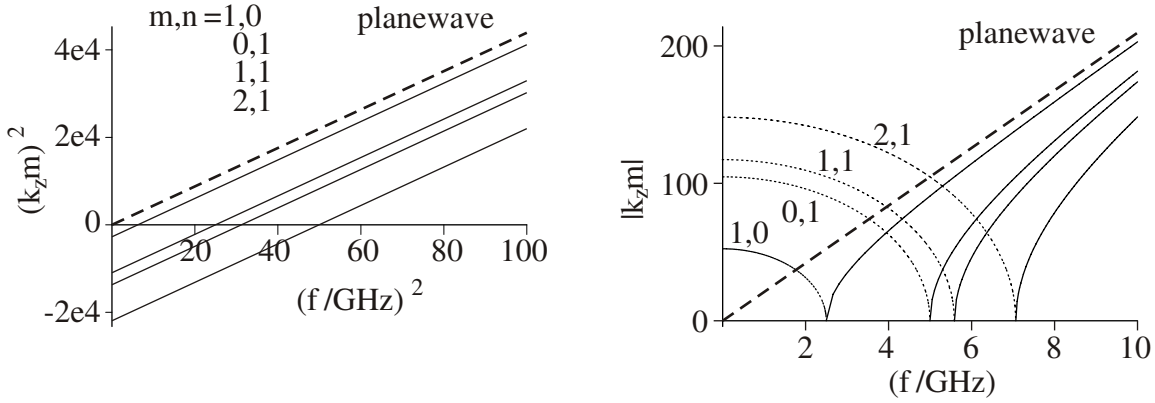


Fig. 23: Dispersion curves in a rectangular waveguide with $a = 6$ cm and $b = 3$ cm

Since $a = 2 \cdot b$ in this waveguide, the modes TE_{01} and TE_{20} are degenerate, i.e. they have identical cutoff frequencies and identical dispersion curves.

3.2 Other waveguides

3.2.1 Circular waveguides

By analogy with rectangular waveguides the approach

$$\vec{E} = \text{curl}\, A_z^H \vec{e}_z \quad (H \text{ waves}) \tag{231}$$

or

$$\vec{H} = \text{curl}\, A_z^E \vec{e}_z \quad (E \text{ waves}) \tag{232}$$

is chosen. Substituting into Maxwell's equations and enforcing the Lorentz gauge, the vector wave equation (190) and, with $\vec{A} = A\,\vec{e}_z$, the Helmholtz equation (193) are obtained.

Using the Δ operator for cylindrical coordinates in Eq. (193) yields

$$\frac{\partial^2 A}{\partial r^2} + \frac{1}{r}\frac{\partial A}{\partial r} + \frac{1}{r^2}\frac{\partial^2 A}{\partial \varphi^2} + \frac{\partial^2 A}{\partial z^2} + k^2 A = 0 \; . \tag{233}$$

Again an ansatz in product form,

$$A = f(r)\, g(\varphi)\, h(z) \; , \tag{234}$$

is used. Inserting this ansatz (234) into the Helmholtz equation (193) and dividing by A yields

$$\frac{1}{f}\frac{d^2 f}{dr^2} + \frac{1}{r}\frac{1}{f}\frac{df}{dr} + \frac{1}{r^2}\underbrace{\frac{1}{g}\frac{d^2 g}{d\varphi^2}}_{=-\mu^2} + \underbrace{\frac{1}{h}\frac{d^2 h}{dz^2}}_{=-k_z^2} + k^2 = 0 \; . \tag{235}$$

The solutions for $g(\varphi)$ and $h(z)$ are obtained using the approaches

$$\frac{1}{g}\frac{d^2 g}{d\varphi^2} = -\mu^2 \; , \tag{236}$$

$$\frac{1}{h}\frac{d^2 h}{dz^2} = -k_z^2 \; , \tag{237}$$

and for $\mu \neq 0$ and $k_z \neq 0$ they can be written as

$$g(\varphi) = \left\{ \begin{array}{c} \sin \\ \cos \end{array} \mu\varphi \right\} \quad \text{or} \quad \left\{ e^{\pm i\mu\varphi} \right\} \tag{238}$$

and

$$h(z) = \left\{ \begin{array}{c} \sin \\ \cos \end{array} k_z z \right\} \quad \text{or} \quad \left\{ e^{\pm i k_z z} \right\} \; . \tag{239}$$

To obtain the function f, Eqs. (236) and (237) are inserted into the Helmholtz equation (193) and are multiplied by f:

$$\frac{d^2 f}{dr^2} + \frac{1}{r}\frac{df}{dr} + \left(k^2 - k_z^2 - \frac{\mu^2}{r^2} \right) f = 0 \; . \tag{240}$$

This Bessel differential equation has the general solution

$$f(r) = \mathrm{Z}_\mu\left(r\sqrt{k^2 - k_z^2} \right) \; . \tag{241}$$

Consequently the solution of the Helmholtz equation for the vector potential A in cylindrical coordinates reads as

$$A = \mathrm{Z}_\mu(Kr) e^{\pm i\mu\varphi} e^{\pm i k_z z} \qquad \text{or} \quad A = \mathrm{Z}_\mu(Kr) \left\{ \begin{array}{c} \sin \\ \cos \end{array} \mu\varphi \right\} \left\{ \begin{array}{c} \sin \\ \cos \end{array} k_z z \right\} \tag{242}$$

with the dispersion equation

$$k^2 = K^2 + k_z^2 \; . \tag{243}$$

Z_μ denotes the *cylindrical functions*, which consist of Bessel and Neumann functions (see Fig. 24):

$$\mathrm{Z}_\mu(Kr) = A\mathrm{J}_\mu(Kr) + B\mathrm{N}_\mu(Kr). \tag{244}$$

A combination of the Bessel and Neumann functions yields the Hankel functions of the first and second kind

$$\mathrm{H}_\mu^{(1)}(Kr) = \mathrm{J}_\mu(Kr) + i\mathrm{N}_\mu(Kr), \tag{245}$$

$$\mathrm{H}_\mu^{(2)}(Kr) = \mathrm{J}_\mu(Kr) - i\mathrm{N}_\mu(Kr), \tag{246}$$

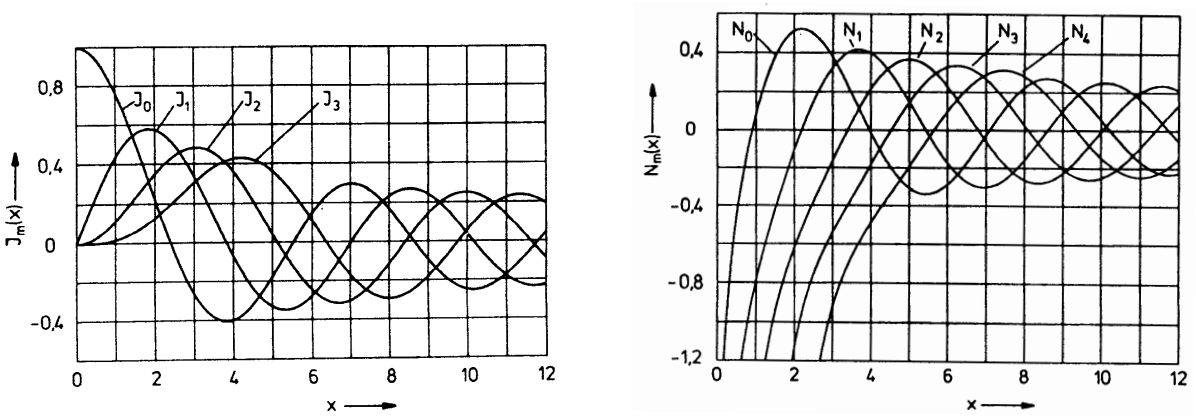


Fig. 24: The Bessel functions $\mathrm{J}_m(x)$ and Neumann functions $\mathrm{N}_m(x)$ with integer order m

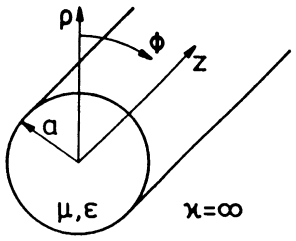


Fig. 25: Circular waveguide

which specify a wave propagating in the opposite direction to the Bessel and Neumann functions describing a standing wave in the radial direction.

To calculate the wave types in a cylindrical waveguide with a perfectly conducting wall the general solution is specialized. Because of the azimuthal periodicity of the configuration shown in Fig. 25, the condition

$$A(\varphi) = A(\varphi + 2\pi) \tag{247}$$

must hold. This leads to the relation

$$e^{\pm i\mu\varphi} = e^{\pm i\mu(\varphi+\pi)} \,, \tag{248}$$

which is satisfied by integer μ:

$$\mu = m = 0, 1, 2, 3, \ldots \tag{249}$$

The φ-dependency is described by a linear combination of sine- and cosine-functions. Because of the rotational symmetry of the configuration, the origin could be chosen such that only a sine- or a cosine-function is sufficient to describe the φ-dependency. Because the sine-function is permanently zero if $m = 0$, the cosine-function is chosen. Thus the vector potential is given by

$$A \sim \cos(m\varphi) \,. \tag{250}$$

The electromagnetic field must be finite for $r \to 0$ in the case of both wave forms, so the solution can only consist of Bessel functions (see Fig. 24):

$$A \sim \mathrm{J}_m(Kr) \,. \tag{251}$$

Consequently, for the vector potential of an electromagnetic wave propagating in the positive z-direction along a circular waveguide we have

$$\boxed{A = C\mathrm{J}_m(Kr)\ \cos m\varphi\, e^{-ik_z z} \quad \text{, where} \quad k_z = \sqrt{k^2 - K^2}\,.} \tag{252}$$

The field components for $E_{z\,mn}$ waves (TM waves) follow from the approach $\vec{H} = \text{curl}\,\vec{A}$ and Eq. (252):

$$\begin{aligned}
H_r &= \frac{1}{r}\frac{\partial A_z}{\partial \varphi} &&= -\underline{C}^E \frac{m}{r}\,\mathrm{J}_m(Kr)\,\sin(m\varphi)\,e^{-ik_z z}\,,\\
H_\varphi &= -\frac{\partial A_z}{\partial r} &&= -\underline{C}^E K\,\mathrm{J}_m'(Kr)\,\cos(m\varphi)\,e^{-ik_z z}\,,\\
E_r &= -\frac{1}{i\omega\varepsilon}\frac{\partial H_\varphi}{\partial z} &&= -\underline{C}^E \frac{k_z}{\omega\varepsilon} K\,\mathrm{J}_m'(Kr)\,\cos(m\varphi)\,e^{-ik_z z}\,,\\
E_\varphi &= \frac{1}{i\omega\varepsilon}\frac{\partial H_r}{\partial z} &&= \underline{C}^E \frac{k_z}{\omega\varepsilon}\frac{m}{r}\,\mathrm{J}_m(Kr)\,\sin(m\varphi)\,e^{-ik_z z}\,,\\
E_z &= \frac{K^2}{i\omega\varepsilon} A_z &&= \underline{C}^E \frac{K^2}{i\omega\varepsilon}\,\mathrm{J}_m(Kr)\,\cos(m\varphi)\,e^{-ik_z z}\,,
\end{aligned} \tag{253}$$

where Z_m' denotes the differentiation

$$\mathrm{Z}_m' = \frac{d\mathrm{Z}_m(Kr)}{d(Kr)}\,. \tag{254}$$

The boundary condition of vanishing E_φ and E_z at $r = a$ requires

$$\mathrm{J}_m(Kr) = 0 \quad \text{if} \quad r = a\,, \tag{255}$$

because E_φ and E_z are proportional to $\mathrm{J}_m(Kr)$. Thus the eigenvalue K is

$$K = \frac{j_{mn}}{a}\,, \tag{256}$$

where j_{mn} is the nth root of J_m. Numerical values for the roots can be found, for example, in Ref. [4]. With the eigenvalue K from Eq. (256), the cutoff frequency ω_c in a cylindrical waveguide is

$$\omega_{c\,mn} = \frac{j_{mn}}{a} c\,. \tag{257}$$

The field components for $H_{z\,mn}$ waves (TE waves) follow from the approach $\vec{E} = \text{curl}\,\vec{A}$ and again Eq. (252):

$$\begin{aligned}
E_r &= \frac{1}{r}\frac{\partial A_z}{\partial \varphi} &&= -\underline{C}^H \frac{m}{r}\,\mathrm{J}_m(Kr)\,\sin(m\varphi)\,e^{-ik_z z}\,,\\
E_\varphi &= -\frac{\partial A_z}{\partial r} &&= -\underline{C}^H K\,\mathrm{J}_m'(Kr)\,\cos(m\varphi)\,e^{-ik_z z}\,,\\
H_r &= \frac{1}{i\omega\mu}\frac{\partial E_\varphi}{\partial z} &&= \underline{C}^H \frac{k_z}{\omega\mu} K\,\mathrm{J}_m'(Kr)\,\cos(m\varphi)\,e^{-ik_z z}\\
H_\varphi &= -\frac{1}{i\omega\mu}\frac{\partial E_r}{\partial z} &&= -\underline{C}^H \frac{k_z}{\omega\mu}\frac{m}{r}\,\mathrm{J}_m(Kr)\,\sin(m\varphi)\,e^{-ik_z z}\,,\\
H_z &= -\frac{K^2}{i\omega\mu} A_z &&= -\underline{C}^H \frac{K^2}{i\omega\mu}\,\mathrm{J}_m(Kr)\,\cos(m\varphi)\,e^{-ik_z z}\,.
\end{aligned} \tag{258}$$

Because of the proportionality of E_φ to J_m', in this case the boundary condition yields the eigenvalue

$$K = \frac{j_{mn}'}{a}\,, \tag{259}$$

where j_{mn}' is the nth root of J_m'. With this eigenvalue the cutoff frequency for a $H_{z\,mn}$ mode becomes

$$\omega_{c\,mn} = \frac{j_{mn}'}{a} c\,. \tag{260}$$

A comparison of the numerical values for the cutoff frequencies (257) and (260) in a waveguide with radius $a = 3$ cm yields the lowest cutoff frequency for the TE_{11} mode:

Type	m	n	f_c/GHz
TE	1	1	2.928
TM	0	1	3.825
TE	2	1	4.858
TE	0	1	6.094
TM	1	1	6.094
TE	3	1	6.682
			...

Again, some cutoff frequencies for E and H waves are identical, e.g. the TE_{01} and TM_{11} modes are degenerate. The field pattern of the TE_{11} mode is shown in Fig. 26. Note that there exist two TE_{11} modes with different polarization of the electric field.

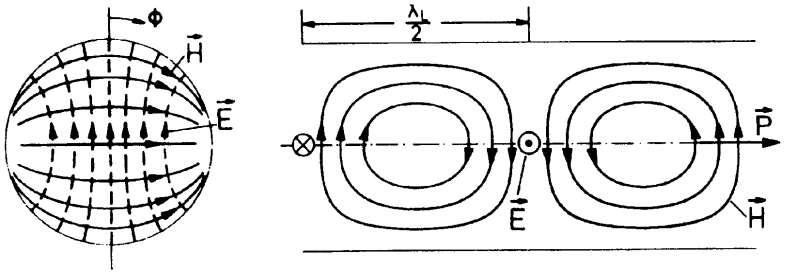


Fig. 26: Field lines of the $H_{z\,11}$ mode in a circular waveguide

3.2.2 Coaxial waveguides

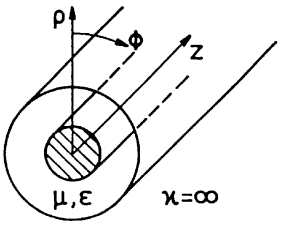


Fig. 27: A coaxial waveguide

In a coaxial waveguide, shown in Fig. 27 the z-axis is free of fields. Hence, for the general solution the cylindrical functions must contain Bessel functions as well as Neumann functions:

$$\mathrm{Z}_m(Kr) = A\mathrm{J}_m(Kr) + B\mathrm{N}_m(Kr) \; . \tag{261}$$

To determine the eigenvalue K the boundary conditions

$$E_z, E_\varphi = 0 \quad \text{if} \quad r = r_i, r_a \tag{262}$$

are imposed, leading to

$$\mathrm{Z}_m(Kr) = 0 \quad \text{if} \quad r = r_i, r_a \qquad (E \text{ waves}) \; , \tag{263}$$

$$\mathrm{Z}'_m(Kr) = 0 \quad \text{if} \quad r = r_i, r_a \qquad (H \text{ waves}) \; . \tag{264}$$

With these equations an eigenvalue equation

$$\frac{\mathrm{N}_m(Kr_i)}{\mathrm{J}_m(Kr_i)} = \frac{\mathrm{N}_m(Kr_a)}{\mathrm{J}_m(Kr_a)} \quad \text{with} \quad E \text{ waves} , \tag{265}$$

$$\frac{\mathrm{N}'_m(Kr_i)}{\mathrm{J}'_m(Kr_i)} = \frac{\mathrm{N}'_m(Kr_a)}{\mathrm{J}'_m(Kr_a)} \quad \text{with} \quad H \text{ waves} \tag{266}$$

can be established, which usually has to be evaluated numerically. Figure 28 shows the field lines of the $H_{z\,11}$ mode, which has the lowest cutoff frequency (except for the TEM waves in the next paragraph).

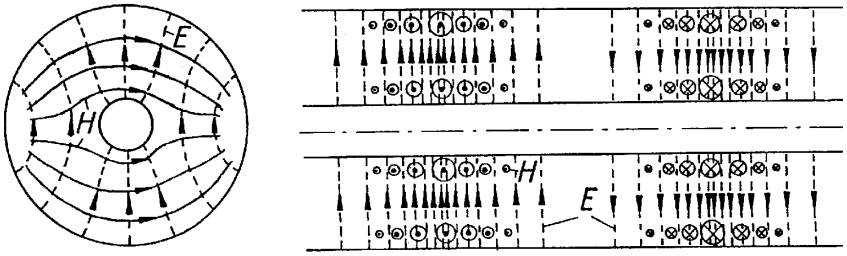


Fig. 28: Field lines of the $H_{z\,11}$ mode in a coaxial waveguide

Now we examine whether a wave exists in a coaxial waveguide with a cutoff frequency of zero. For the cutoff frequency, $k_z = 0$ applies and hence

$$\omega_c = \frac{1}{\mu\varepsilon} K \stackrel{!}{=} 0 ,$$

i.e. for a cutoff frequency of zero the eigenvalue K becomes

$$K = 0 . \tag{267}$$

The calculation of this wave starts with Eq. (240), reduced by $K = 0$ and $\mu = 0$ to

$$\frac{d^2 f}{dr^2} + \frac{1}{r}\frac{df}{dr} = 0 , \tag{268}$$

which has the solution

$$f = C_1 \ln \frac{r}{C_2} . \tag{269}$$

The vector potential can be written

$$A = C_1 \ln \frac{r}{C_2}\, e^{-ikz} , \tag{270}$$

where from $K = 0$ we have $k_z = k$. Note that this approach is not possible in a circular waveguide, because the ln-function tends toward infinity for $r \to 0$.

The field components are calculated using the approach $\vec{H} = \text{curl}\, \vec{A}$:

$$\begin{aligned} E_r &= \frac{1}{r} E_0\, e^{-ikz} , \\ H_\varphi &= \frac{1}{Z} E_r = \frac{1}{Z}\frac{1}{r} E_0\, e^{-ikz} , \\ E_\varphi &= E_z = H_r = H_z = 0 , \end{aligned} \tag{271}$$

where $Z = \sqrt{\frac{\mu}{\varepsilon}}$. (Generally the approach $\vec{E} = \operatorname{curl} \vec{A}$ would also be possible, but this would lead to all-zero field components when imposing the boundary conditions.)

Solution (271) is called a TEM wave because it only consists of transversal field components. An electrostatic field develops between the two separate conductors, resulting in a cutoff frequency $\omega_c = 0$. This cutoff frequency allows the flow of a direct current, which requires at least two separate conductors.

3.3 Attenuation of modes in waveguides

3.3.1 The power-loss method

In the previous sections the walls of the waveguides were assumed to be Perfectly Electric Conducting (PEC). Under this condition and above their cutoff frequency the waves propagate without attenuation, and the phase constant is a real number $k_z = \beta$.

In practical applications the walls only have a finite conductivity, leading to a complex propagation constant $k_z = \beta - j\alpha$ for each wave in a waveguide, and to attenuated waves even above cutoff. As an exact calculation of the losses usually is very costly, we introduce some approximations.

A common approach is the power-loss method. Here we assume that if the conductivity of the walls is sufficiently high, the electromagnetic fields can be approximated by the fields of a guide with PEC boundaries. From these fields an approximation of the power of the wall losses per unit length can be derived.

The power loss P' per unit length of a wave propagating in the $+z$-direction is given by

$$P' = -\frac{dN}{dz} \,, \tag{272}$$

where N is the power flowing through the cross section A of the guide. In the presence of losses, the field vectors decrease in the z-direction like

$$|\underline{E}, \underline{H}| \sim e^{-\alpha z} \,. \tag{273}$$

As the power through the cross section is proportional to the product of the electric and the magnetic field, we obtain

$$N(z) = N_0 \, e^{-2\alpha z} \,. \tag{274}$$

From Eq. (272) we get

$$P' = -\frac{dN}{dz} = 2\alpha N \tag{275}$$

and finally

$$\alpha = \frac{1}{2}\frac{P'}{N} \,. \tag{276}$$

In the power-loss method, the transported power N as well as the power loss P' per unit length are derived from the fields of the loss-free guide. For N we integrate the Poynting vector over the cross section A:

$$N = \frac{1}{2}\mathrm{Re}\left\{ \int_A (\underline{\vec{E}} \times \underline{\vec{H}}^*) \cdot \mathrm{d}\vec{A} \right\} \,. \tag{277}$$

For the derivation of the power loss in the wall, we consider the surface current in the loss-free case:

$$\vec{J}_s = \vec{n} \times \vec{H} \quad \Rightarrow \quad |\vec{J}_s| = H_{tan} \tag{278}$$

according to Eq. (80), where H_{tan} is the tangential magnetic field at the wall. With a finite conductivity the currents in the wall (normal direction x) decrease like

$$J(x) \sim e^{-(1-i)\frac{x}{\delta}} \tag{279}$$

(cf. formulas (128)–(130) for a good conductor), where

$$\delta = \sqrt{\frac{2}{\omega\kappa\mu}} \tag{280}$$

is the skin depth of the waveguide wall. The assumption

$$\int_0^\infty J(x)dx \approx H_{tan} \tag{281}$$

finally leads to

$$P' = \ldots = \frac{1}{2} R_m \int_{\partial A} |\underline{H}_{tan}|^2 \, ds \qquad \text{with} \quad R_m = \frac{1}{\delta\kappa} = \sqrt{\frac{\mu\omega}{2\kappa}} \,, \tag{282}$$

where the integration has to be performed along the boundary ∂A of the cross section of the guide.

3.3.2 Applicability of the power-loss method

The power-loss method is only applicable if the wall currents themselves do not excite other wave types in the guide. This may happen, however, if new field components arise by the transition from infinite to finite conductivity.

Unfortunately this is the case for all modes in rectangular waveguides except for the TE_{m0}- and TE_{0n}-modes (as their TM counterparts do not exist). In circular waveguides, however, the method is applicable to all TE and TM modes.

Criteria for the applicability of the power-loss method are summarized in Fig. 29.

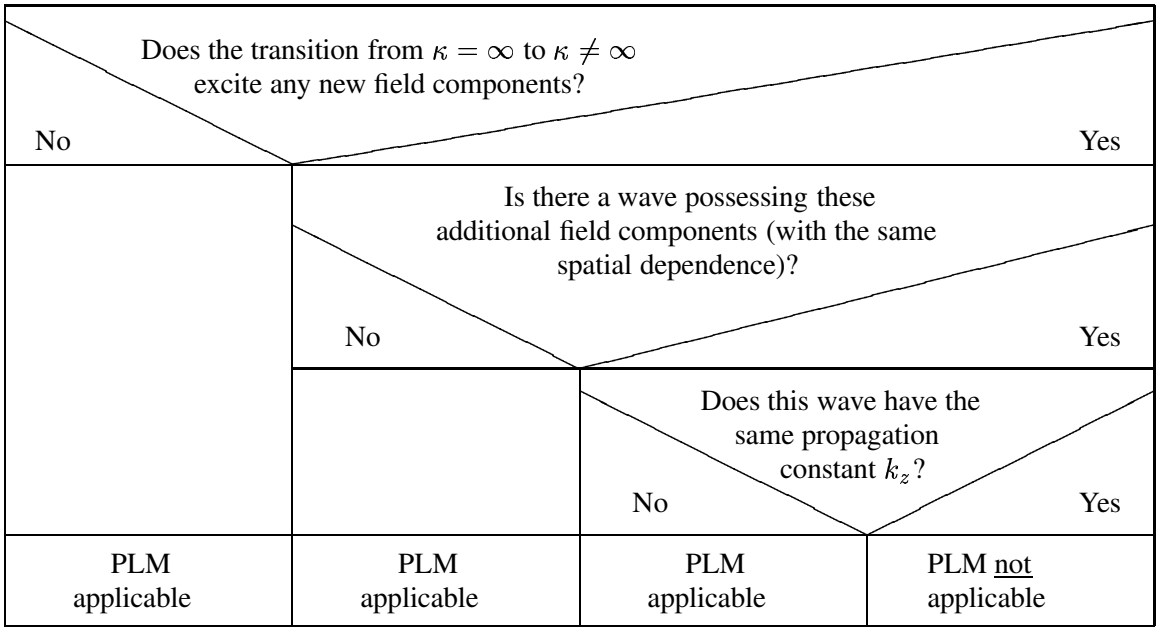


Fig. 29: Criteria for the applicability of the Power-Loss Method (PLM)

4 RESONANT CAVITIES

4.1 Solution of Maxwell's equations in cavities

In this section we consider the electromagnetic field inside resonant cavities, and again start with the assumption of PEC walls. Analogously to the previous derivation, an approach with a complex vector potential $\underline{\vec{A}}$ can be chosen, with

$$\underline{\vec{E}} = \text{curl}\, \underline{\vec{A}} \quad \text{or} \quad \underline{\vec{H}} = \text{curl}\, \underline{\vec{A}} \,.$$

From Maxwell's equations we obtain the wave equation (190)

$$\Delta \underline{\vec{A}} + k^2 \underline{\vec{A}} = 0 \,.$$

An analytical solution of this equation is only possible for a few cases, particularly if the surfaces bounding the cavity are coordinate surfaces (allowing the separation of variables). Among such cavities one should point out rectangular resonators (in a Cartesian coordinate system), and cylindrical and coaxial resonators (in a cylindrical coordinate system). In practice, cavities shaped close to cylindrical and rectangular resonators are the most frequent design.

A simple approach for these types of cavities is to interpret them as waveguides with both ends closed. This leads to a classification of the eigenmodes of the cavity, which is similar to the waveguide one:

- TM modes (E modes), characterized by a non-zero longitudinal component of the electric field and by purely transverse magnetic fields,
- TE modes (H modes), characterized by a non-zero longitudinal component of the magnetic field and by purely transverse electric fields.

Note that one distinguished coordinate direction ('longitudinal') is introduced by this classification. In the field of accelerator cavities this is usually the beam direction.

4.2 Rectangular cavities with perfectly conducting walls

If we superpose two waves in a loss-free waveguide, which have identical mode patterns $\underline{\vec{E}}(x, y)$ but propagate in opposite directions ($\pm z$), we obtain a standing wave:

$$\underline{\vec{E}}(x, y) \cdot (e^{-ik_z z} + e^{+ik_z z}) = \underline{\vec{E}}(x, y) \cdot 2\, \cos(k_z z) \,. \tag{283}$$

For TE waves, with $\underline{\vec{E}}(x, y)$ having only transversal components, the electric field vanishes for all

$$k_z z = \pi/2 + n\pi \qquad \Rightarrow \quad \cos(k_z z) = 0 \qquad (n \in Z) \,. \tag{284}$$

(A similar condition can also be found for the transversal components of TM waves.) Thus, if we insert perfectly conducting walls at a distance

$$\Delta z_p = L = \frac{p\pi}{k_z} \qquad (p \in N) \,, \tag{285}$$

the fields do not change and build the electromagnetic fields of a corresponding mode in a cavity with length L (see Fig. 30).

As the propagating constant k_z of the waveguide modes depends on the frequency of the fields, the condition (285) cannot be fulfilled at arbitrary frequencies if the geometry of the cavity (the length Δz) is kept fixed. Instead of the dispersion relation (224) of the waveguide (with a continuous relation between k_z and ω), we obtain the equation

$$\begin{aligned} k_x^2 + k_y^2 + k_z^2 &= k^2 = \left(\frac{\omega}{c}\right)^2 \\ \Rightarrow \quad \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{p\pi}{\Delta z}\right)^2 &= \left(\frac{\omega}{c}\right)^2 \,, \end{aligned} \tag{286}$$

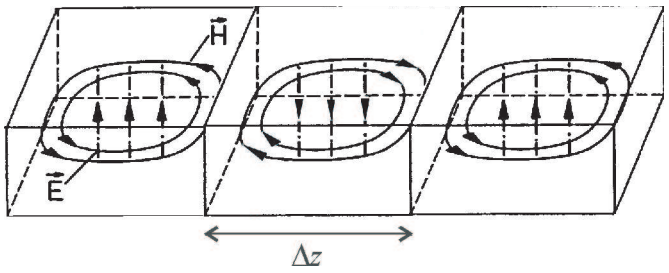


Fig. 30: Derivation of cavity modes: the fields of a standing wave (here, $H_{z,10}$ mode) are not changed if PEC walls are inserted

which defines a discrete (but infinite) set of eigensolutions in a cavity, including a discrete set of eigenfrequencies or resonance frequencies

$$f_{mnp} = \frac{c}{2\pi}\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{p\pi}{L}\right)^2} \,. \tag{287}$$

For the field components in the cavity we obtain the following expressions, first for TM modes:

$$\underline{E}_x = \underline{C}\,\frac{mp\pi^2}{aL}\frac{1}{k}\frac{1}{\kappa_e}\cos\left(\frac{m\,\pi}{a}x\right)\,\sin\left(\frac{n\,\pi}{b}y\right)\,\sin\left(\frac{p\,\pi}{L}z\right)\,, \tag{288}$$

$$\underline{E}_y = \underline{C}\,\frac{np\pi^2}{bL}\frac{1}{k}\frac{1}{\kappa_e}\sin\left(\frac{m\,\pi}{a}x\right)\,\cos\left(\frac{n\,\pi}{b}y\right)\,\sin\left(\frac{p\,\pi}{L}z\right)\,, \tag{289}$$

$$\underline{E}_z = \underline{C}\,\frac{\kappa_e}{k}\sin\left(\frac{m\,\pi}{a}x\right)\,\sin\left(\frac{n\,\pi}{b}y\right)\,\cos\left(\frac{p\,\pi}{L}z\right)\,, \tag{290}$$

$$\underline{H}_x = -\underline{C}\,\frac{1}{\kappa_e}\frac{n\pi}{b}\sin\left(\frac{m\,\pi}{a}x\right)\,\cos\left(\frac{n\,\pi}{b}y\right)\,\cos\left(\frac{p\,\pi}{L}z\right)\,, \tag{291}$$

$$\underline{H}_y = \underline{C}\,\frac{1}{\kappa_e}\frac{m\pi}{a}\cos\left(\frac{m\,\pi}{a}x\right)\,\sin\left(\frac{n\,\pi}{b}y\right)\,\cos\left(\frac{p\,\pi}{L}z\right)\,, \tag{292}$$

$$\underline{H}_z = 0\,. \tag{293}$$

And for TE cavity modes:

$$\underline{E}_x = -\underline{C}\,\frac{1}{\kappa_h}\frac{n\pi}{b}\cos\left(\frac{m\,\pi}{a}x\right)\,\sin\left(\frac{n\,\pi}{b}y\right)\,\sin\left(\frac{p\,\pi}{L}z\right)\,, \tag{294}$$

$$\underline{E}_y = \underline{C}\,\frac{1}{\kappa_h}\frac{m\pi}{a}\sin\left(\frac{m\,\pi}{a}x\right)\,\cos\left(\frac{n\,\pi}{b}y\right)\,\sin\left(\frac{p\,\pi}{L}z\right)\,, \tag{295}$$

$$\underline{E}_z = 0\,, \tag{296}$$

$$\underline{H}_x = -\underline{C}\,\frac{mp\pi^2}{aL}\frac{1}{\kappa_h}\frac{1}{k}\sin\left(\frac{m\,\pi}{a}x\right)\,\cos\left(\frac{n\,\pi}{b}y\right)\,\cos\left(\frac{p\,\pi}{L}z\right)\,, \tag{297}$$

$$\underline{H}_y = -\underline{C}\,\frac{np\pi^2}{bL}\frac{1}{\kappa_h}\frac{1}{k}\cos\left(\frac{m\,\pi}{a}x\right)\,\sin\left(\frac{n\,\pi}{b}y\right)\,\cos\left(\frac{p\,\pi}{L}z\right)\,, \tag{298}$$

$$\underline{H}_z = \underline{C}\,\frac{k}{\kappa_h}\cos\left(\frac{m\,\pi}{a}x\right)\,\cos\left(\frac{n\,\pi}{b}y\right)\,\sin\left(\frac{p\,\pi}{L}z\right)\,, \tag{299}$$

with the abbreviations

$$\kappa_e = \kappa_h = \sqrt{(m\pi/a)^2 + (n\pi/b)^2}\,. \tag{300}$$

As we can see from these formulas, for TE modes $p \geq 1$ is required to obtain a non-zero field solution, whereas $m = 0$ or $n = 0$ is allowed (but, like for waveguides, *not* $m = n = 0$).

For the TM case, we must have $m, n \geq 1$ (as for waveguides). However, the case $p = 0$ is also possible and leads to a three-component field ($\underline{H}_x, \underline{H}_y, \underline{E}_z$), which cannot be derived as a superposition of waveguide modes, but corresponds to the TM_{mn} waveguide mode at cutoff.

The cavity modes can be sorted by referring to their resonance frequency. When the resonator dimensions are such that $a \geq b \geq L$, the lowest resonance frequency is found to be

$$f_{110} = \frac{c}{2\pi}\sqrt{\left(\frac{\pi}{a}\right)^2 + \left(\frac{\pi}{b}\right)^2} \tag{301}$$

with $m = n = 1$ and $p = 0$, corresponding to the TM_{110} cavity mode. For a cavity with $a = 6$ cm and $b = 3$ cm the resonance frequency is $f_{110} = 5.59$ GHz.

The field distribution is illustrated in Fig. 31. We see that the electric fields are perpendicular to

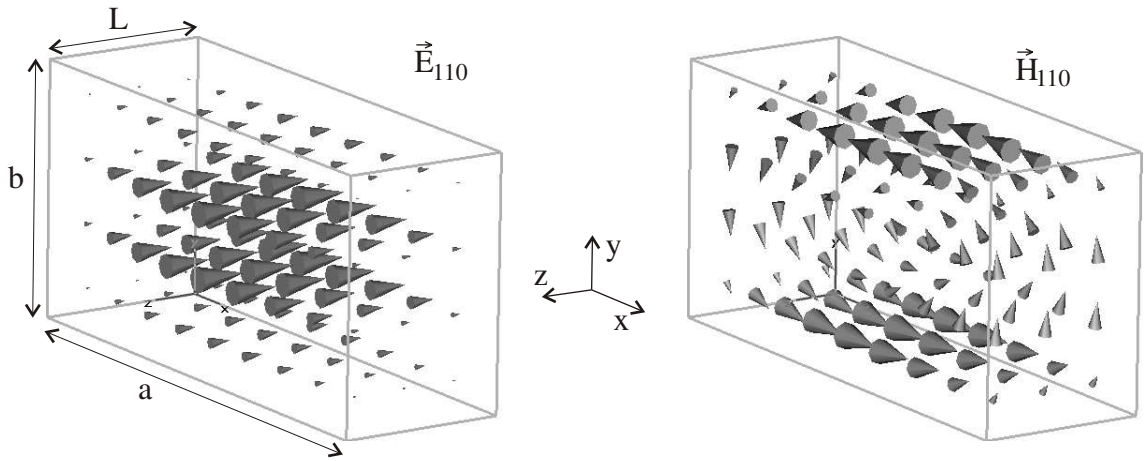


Fig. 31: Electric and magnetic field pattern of the $E_{z,110}$ mode in a rectangular cavity

the plate boundaries at $z = 0$ and $z = L$ and concentrate at the centre of the cavity so that the tangential $\vec{E}$ field vanishes at the boundaries $x = 0, a$ and $y = 0, b$. This field can also be considered as a dominant waveguide mode propagating in the y-direction and reflected at the walls $y = 0$ and $y = b$ to form a standing wave. If the labels of the coordinate axes y and z are interchanged, this mode may also be called a TE_{101} mode. To have an unambiguous notation we have to refer to $E_{z,101}$.

4.3 Cylindrical cavities

A similar derivation can be performed for the fields in cylindrical cavities. Starting with the superposition of TE and TM waveguide modes forming a standing wave and fulfilling the boundary conditions at two electric walls, we again obtain discrete sets of cavity eigenmodes.

For the resonance frequencies in a cylindrical cavity with radius R and length L we get the TM solutions

$$f_{mnp}^{(\mathrm{TM})} = \frac{c}{2\pi}\sqrt{\left(\frac{j_{mn}}{R}\right)^2 + \left(\frac{p\pi}{L}\right)^2}\,, \tag{302}$$

where j_{mn} is the nth root of the Bessel function J_m. For non-zero fields $n \geq 1$ is required, whereas $m = 0$ and $p = 0$ are allowed.

The eigenfrequencies of the TE solutions are

$$f_{mnp}^{(\mathrm{TE})} = \frac{c}{2\pi}\sqrt{\left(\frac{j'_{mn}}{R}\right)^2 + \left(\frac{p\pi}{L}\right)^2}\,, \tag{303}$$

where now j'_{mn} is the nth root of J'_m, and $m, n, p \geq 1$ is required.

The fundamental TM mode in a cylindrical cavity is the TM_{010} mode, which corresponds to the waveguide mode TM_{01} at cutoff (see Fig. 32). Its resonance frequency is

$$f_{010}^{(\mathrm{TM})} = \frac{c\, 2.405}{2\pi\, R}\ , \tag{304}$$

or $f_{010}^{(\mathrm{TM})} = 3.825$ GHz for a cavity with radius $R = 3$ cm.

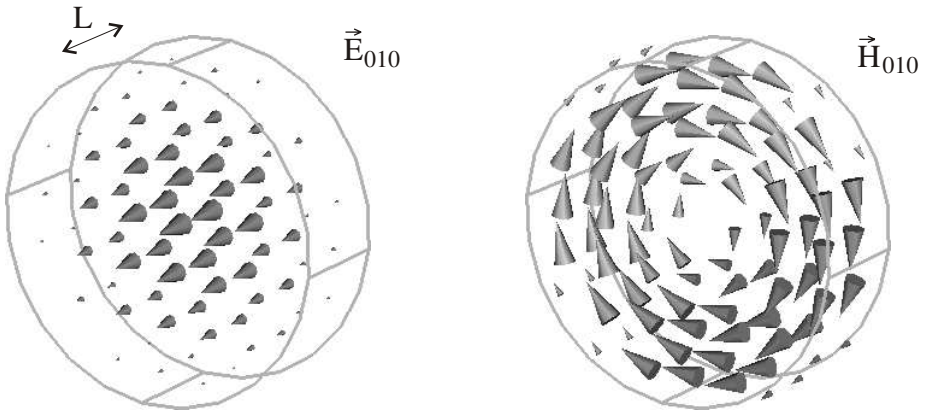


Fig. 32: Electric and magnetic field pattern of the $E_{z,010}$ mode in a cylindrical cavity

The lowest TE mode is the TE_{111} mode. Its resonance frequency is

$$f_{111}^{(\mathrm{TE})} = \frac{c}{2\pi}\sqrt{\left(\frac{1.841}{R}\right)^2 + \left(\frac{\pi}{L}\right)^2}\ , \tag{305}$$

or $f_{111}^{(\mathrm{TE})} = 8.046$ GHz for a cavity with radius $R = 3$ cm and length $L = 2$ cm.

For small L, the TM_{010} mode is the fundamental mode in a cylindrical cavity. For large L ($L > 2.03R$), however, we have $f_{111}^{(\mathrm{TE})} < f_{010}^{(\mathrm{TM})}$, and the TE_{111} mode is the fundamental. Because the frequency of this mode depends on the ratio R/L, it is possible to provide easy tuning by making the separation of the end faces adjustable.

Most cavities used in practice have a more complicated geometry, making analytical solutions more difficult or even not accessible at all. For such cavities, the eigensolutions (fields and resonance frequencies) can only be computed using numerical methods.

4.4 Losses in walls with finite conductivity, Q values

In the preceding section we neglected all sources of energy losses in a cavity, like non-perfectly electric conducting walls or conducting materials (e.g. a non-perfect vacuum) inside the cavity. Under this assumption we obtained a discrete set of eigensolutions with eigenfrequencies ω_i, corresponding to electric and magnetic fields oscillating at one specific frequency without any attenuation.

In the presence of losses, however, the cavity modes no longer have a sharp delta function singularity on the frequency axes (at their resonance frequency), but rather a narrow band of frequencies occurs around the eigenfrequency. In time domain with no excitations, this is equivalent to free attenuated oscillations with a time dependency such as

$$\vec{E}, \vec{H} \sim e^{i\omega t - \alpha t}\ . \tag{306}$$

A measure of the sharpness of response of the cavity to external excitation is the quality factor Q of a cavity mode, defined as the ratio of the time-averaged energy W stored in the cavity to the energy loss per cycle:

$$Q_0 = \omega \frac{W}{P_d} \,, \tag{307}$$

where P_d is the dissipated power in the cavity. From Eq. (306) and

$$W \sim e^{-2\alpha t} \quad \Rightarrow \quad P_d = -\dot{W} = 2\alpha W \tag{308}$$

we find

$$Q_0 = \frac{\omega}{2\alpha} \,. \tag{309}$$

To calculate the Q-value of a cavity mode due to wall losses, we again apply the power-loss approach (see formulas (272)–(282)): we approximate the fields in the lossy cavity by the solution of the loss-free case, and integrate the power loss of the wall currents over the surface of the cavity, and the stored energy over the volume V of the cavity:

$$W = \int\limits_V w \, dV = \frac{1}{2} \int\limits_V \left(\frac{\varepsilon}{2} |\underline{\vec{E}}|^2 + \frac{\mu}{2} |\underline{\vec{H}}|^2 \right) dV \,, \tag{310}$$

$$P_d = \int\limits_{\partial V} P'_d dA = \frac{1}{2} \int\limits_{\partial V} \sqrt{\frac{\omega\mu}{2\kappa}} \, |H_{tan}|^2 dA \,. \tag{311}$$

For rectangular and cylindrical cavities these integrals can easily be evaluated. The Q-values for some cavities with copper walls ($\kappa \approx 5e7$ S/m) are summarized in the following table:

Rectangular	a	b	L	Type	f	Q
	6 cm	3 cm	2 cm	TM_{110}	5.59 GHz	1.03e4
	6 cm	3 cm	2 cm	TE_{111}	9.35 GHz	1.18e4
	5 m	4 m	3 m	TM_{110}	69.3 MHz	6.60e4
Cylindrical		R	L			
		3 cm	2 cm	TM_{010}	3.83 GHz	1.15e4
		1 m	1 m	TM_{010}	0.11 GHz	8.10e4
		1 m	1 m	TE_{111}	0.17 GHz	9.22e4

The Q-values of cavities with a more complex geometrical shape can be evaluated from the results of a numerical field simulation.

Other sources of loss are the material filling of the cavity, surface irregularities of the cavity walls, and the coupling between other systems. They all contribute to the power dissipation P_d and therefore decrease Q_0.

4.5 The main electrodynamic characteristics of accelerator cavities

In charged particle accelerators two types of operation are common, namely Travelling Wave (TW) and Standing Wave (SW). In the TW case, the electrodynamic structure is an accelerating waveguide; in the SW case, an accelerating resonator in the form of a single cavity or a chain of coupled cavities is used [5].

The main characteristics of accelerating cavities are the resonant frequency, the quality factor (307), the cavity-generator coupling factor, and the effective shunt impedance.

In the absence of losses in the walls and the cavity-filling medium (or vacuum), the resonance frequency of any mode is a real valued quantity. For the simple form cavities, e.g. the 'pillbox' cavity (a circular cylindrical resonator), the frequency can be calculated analytically, as was shown above. In resonators of intricate geometry, these frequencies have to be determined by numerical methods or experimentally.

The cavity-generator coupling factor β_0 is the ratio of the unloaded quality factor Q_0 to the external quality factor Q_{ext} (a measure for the coupling of an excitation to the fields in the cavity):

$$\beta_0 = \frac{Q_0}{Q_{ext}} \,. \tag{312}$$

As a rule, the coupling factor of the beam-loaded cavity must be close to unity [5] (critical coupling), because then maximum power transmitted from the generator to the resonator. The power transmission factor is

$$k_t = 1 - |\Gamma_{in}|^2 = \frac{k_{t0}}{1 + a_1^2} \,, \tag{313}$$

where Γ_{in} is the reflection coefficient at the cavity input, k_{t0} is the transmission coefficient at the resonance frequency ω_0,

$$k_{t0} = \frac{4\beta_0}{(1+\beta_0)^2} \,, \tag{314}$$

and a_1 is the generalized detuning of the loaded resonator,

$$a_1 = 2\,Q_1\,\frac{\Delta\omega}{\omega_0} \,. \tag{315}$$

In the last formula, $\Delta\omega = \omega - \omega_0$ is the frequency detuning, and Q_1 is the loaded quality factor given by

$$\frac{1}{Q_1} = \frac{1}{Q_0} + \frac{1}{Q_{ext}} \,. \tag{316}$$

The effective shunt impedance in the accelerating cavity $R_{sh,eff}$ relates the equivalent voltage U between two points in the cavity over a given path to the power P_d dissipated in the cavity walls:[2]

$$R_{sh,eff} = \frac{U^2}{2\,P_d} \,, \tag{317}$$

where

$$U = \left| \int_{z_1}^{z_2} \underline{E}_z(z)\, \exp(ik_z z) dz \right| \,. \tag{318}$$

$\underline{E}_z(z)$ is the distribution of the longitudinal component of the electric field over the cavity axis, and z_1 and z_2 are the points on the cavity axis such that $z_2 - z_1 = L$ is the cavity length.

The effective shunt impedance, Eq. (317), takes into account the field variation in the cavity during the passage of accelerated particles. The transit time factor is given by

$$T = \frac{\left| \int_0^L \underline{E}_z(z)\, \exp(ik_z z) dz \right|}{\int_0^L |\underline{E}_z(z)| dz} \tag{319}$$

[2] Note that often the definition $R = U^2/P_d$ is used, and thus the shunt impedance in other papers may be two times higher.

with

$$R_{sh,eff} = R_{sh} T^2 \ , \tag{320}$$

where the shunt impedance of the accelerating cavity is

$$R_{sh} = \frac{U_0^2}{2\,P_d} = \frac{1}{2\,P_d} \left(\int_0^L |\underline{E}_z(z)|^2 dz \right)^2 \ . \tag{321}$$

The effective shunt impedance per unit length $r_{sh,eff}$ is given as follows:

$$r_{sh,eff} = \frac{1}{2\,P_d L} \left| \int_0^L \underline{E}_z(z)\ \exp(ik_z z) dz \right|^2 \ . \tag{322}$$

Using the expressions for shunt impedance and quality factor we may write the characteristic impedance as

$$\frac{R_{sh}}{Q} = \frac{1}{2\,\omega W} \left(\int_0^L |\underline{E}_z(z)|^2 dz \right)^2 \ . \tag{323}$$

or

$$\frac{R_{sh,eff}}{Q} = \frac{1}{2\,\omega W} \left| \int_0^L \underline{E}_z(z)\ \exp(ik_z z) dz \right|^2 \ . \tag{324}$$

The ratio R_{sh}/Q depends on the cavity geometry and does not depend on the cavity's material and quality of element matching. This ratio is sometimes called the characteristic impedance.

REFERENCES

[1] J.C. Maxwell, *A Treatise on Electricity and Magnetism* (Oxford University Press, London, 1873).

[2] K. Simony, *Theoretische Elektrotechnik* (VEB Deutscher Verlag der Wissenschaften, Berlin, 1973).

[3] J.D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1962).

[4] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (Dover Publications, New York, 1965).

[5] N.P. Sobenin and B.V. Zverev, *Electrodynamic Characteristics of Accelerating Cavities* (MEPIUP, London and Moscow, 1999).

HIGH-FREQUENCY NON-FERRITE CAVITIES

J. Le Duff
IN2P3–CNRS and Paris-Sud University, Orsay, France

Abstract
High-frequency, high-Q cavities are mostly used in electron accelerators. The electrons radiate and lose part of their energy, therefore they need relatively high RF voltages to remain on their stable orbit. Single cell and multicell cavities operating either in a standing wave mode or a travelling wave mode are presented.

1 ACCELERATORS: THE NEED FOR ELECTRIC FIELDS

Consider a rectangular coordinate system moving along the particle trajectory 's' (Fig. 1),

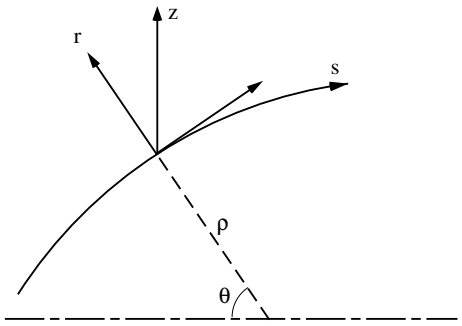


Fig. 1: Curvilinear coordinate system along the particle trajectory

and assume that the fields have single components:

$$\begin{array}{ll} \vec{E} & \rightarrow E_\theta \\ \vec{B} & \rightarrow B_z \end{array} \tag{1}$$

The Newton-Lorentz force,

$$\frac{d\vec{p}}{dt} = e\vec{E} + e\vec{v} \times \vec{B} \ , \tag{2}$$

can be expanded as follows:

$$\frac{d(mv_\theta)}{dt}.\vec{u}_\theta - m\frac{v_\theta^2}{\rho}.\vec{u}_r = eE_\theta.\vec{u}_\theta + ev_\theta B_z.\vec{u}_r \ , \tag{3}$$

where ρ is the local radius of curvature of the trajectory, and $\vec{u}_\theta$ and $\vec{u}_r$ are the unit vectors—tangent and perpendicular, respectively—to the trajectory. By identification one gets:

$$\begin{aligned} \frac{dp_\theta}{dt} &= eE_\theta \\ p_\theta/e &= B_z\rho \ , \end{aligned} \tag{4}$$

showing that the electric field provides energy and momentum while the magnetic field bends the particle trajectory. This result can be generalized to more complicated field patterns.

2 WHY RADIO FREQUENCY (RF) ELECTRIC FIELD ?

Assume a chain of electrodes (Fig. 2) fed by a high-voltage (HV) electrostatic generator.

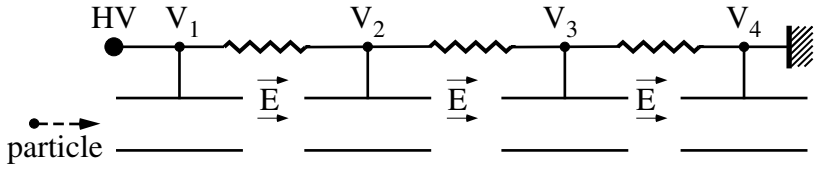


Fig. 2: Electrostatic acceleration

Since the DC voltage is shared between the electrodes in order to provide an electric field in each gap, the generator voltage is the sum of the partial voltages:

$$V_{total} = \sum_i V_i \ . \tag{5}$$

Such a system will hence be limited by electrical breakdown at the HV terminal (as will the energy of the electrostatic accelerator).

If we assume now that the electrodes are fed from an RF generator as shown on Fig. 3, the electric field has reverse polarities in consecutive gaps.

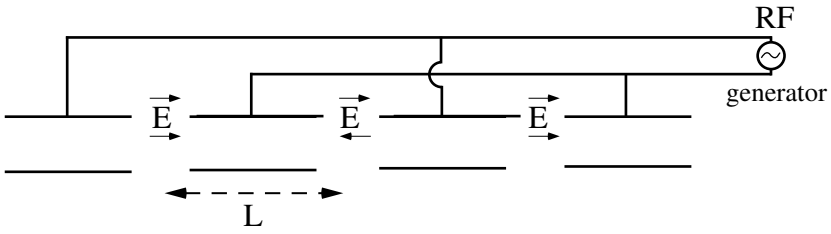


Fig. 3: Radio frequency (RF) acceleration

If the synchronism condition

$$L = \frac{vT_{RF}}{2} \tag{6}$$

is fulfilled, where v is the particle velocity, L the distance between consecutive gaps and T_{RF} the RF period, then for a given voltage V_{RF} all the gaps will accelerate the particle and provide energy. If the particle velocity increases the drift tube length grows between successive gaps, and at very high velocities it is worth considering higher RF frequencies.

3 ENERGY GAIN

In relativistic dynamics the total energy and momentum satisfy

$$\begin{aligned} E &= E_0 + W \\ E^2 &= E_0^2 + p^2c^2 \ , \end{aligned} \tag{7}$$

where E_0 is the rest energy and W the kinetic energy. Differentiating these expressions gives

$$dE = dW = vdp \ . \tag{8}$$

From the Lorentz force

$$\frac{dp}{dt} = eE_z \ , \tag{9}$$

where E_z is the electric field along the linear particle trajectory[1]

$$dW = eE_z dz \tag{10}$$

one gets:

$$W = e \int E_z dz = \mathrm{eV} \tag{11}$$

where in practice V would represent the gap voltage.

4 ELECTRIC FIELD FROM ELECTROMAGNETIC WAVES

As mentioned above, higher kinetic energies require higher operating frequencies to limit the drift tube length. However, according to the accelerating concept described above, the current flowing on the surface of the drift tubes will generate a radiated power in the free space, which will increase linearly with frequency. Therefore it is advisable to enclose the gap in a cavity that will hold the electromagnetic (EM) energy (Fig. 4).

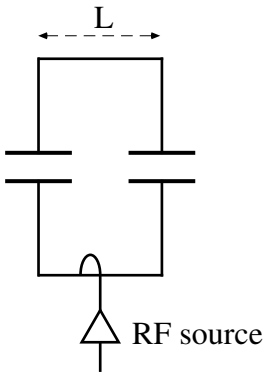


Fig. 4: Single-gap cavity

Such single-gap cavities, resonating at the operating RF frequency, can be powered and phased independently to ensure proper synchronism with respect to the particle velocity. The resonant standing

[1] In a linear accelerator z often refers to the longitudinal axis or direction of particle motion, while in a circular accelerator z can represent the vertical axis perpendicular to the plane of curvature.

wave (SW) mode is a transverse magnetic (TM) mode with a longitudinal electric field component, necessary for acceleration.

Efficient variants of the above scheme consist of placing the cavities adjacent to each other, either in a π-mode (Fig. 5(a)) or a 2π-mode (Fig. 5(b)).

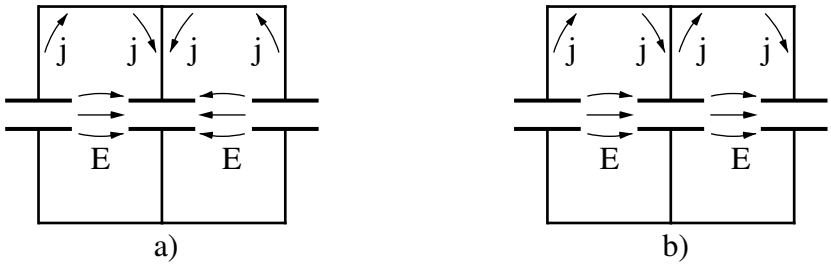


Fig. 5: Multigap cavities (E = electric field; j = wall current)

These modes can be obtained by suitable cavity geometry design leading, respectively, to the synchronism conditions

$$L = v\frac{T_{RF}}{2} \quad (\pi\text{-mode}) \qquad\qquad L = vT_{RF} \quad (2\pi\text{-mode}) \ ,$$

where L represents the distance between the gap centres.

5 THE PILL-BOX CAVITY

The most simple accelerating cavity is the pill-box cavity, which has cylindrical symmetry of radius a, and a relatively short active length l (Fig. 6)

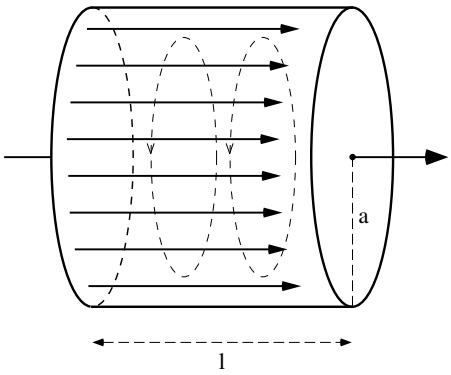


Fig. 6: Pill-box cavity with TM_{010} mode
$(\longrightarrow \ E_z \quad - - \rightarrow H_\theta)$

From Maxwell's equations

$$\begin{aligned} \nabla.\mathbf{E} &= 0 \\ \nabla.\mathbf{H} &= 0 \\ \nabla \times \mathbf{E} &= -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \\ \nabla \times \mathbf{H} &= \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \ , \end{aligned} \tag{12}$$

one can derive the wave equation

$$\nabla^2 \mathbf{A} - \epsilon_0 \mu_0 \frac{\partial^2 \mathbf{A}}{\partial t^2} = 0 \tag{13}$$

where **A** represents either the electric field or the magnetic field vector.

The lowest frequency mode is the TM_{010} with only two components for the EM field, provided $\ell < 2a$

$$\left.\begin{array}{l} E_z = J_0(kr) \\ H_\theta = -\frac{j}{Z_0} J_1(kr) \end{array}\right\} e^{j\omega t} \ , \tag{14}$$

with

$$Z_0 = \left(\frac{\mu_0}{\epsilon_0}\right)^{1/2} = 377\ \Omega$$

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{c} \ .$$

where λ is the free space wavelength and $\omega/2\pi$ is the RF frequency.

The amplitudes of the field components as function of the radial position are shown in Fig. 7. Since the longitudinal electric field must vanish at $r = a$, one gets $\lambda = 2.62a$ for the free space wavelength.

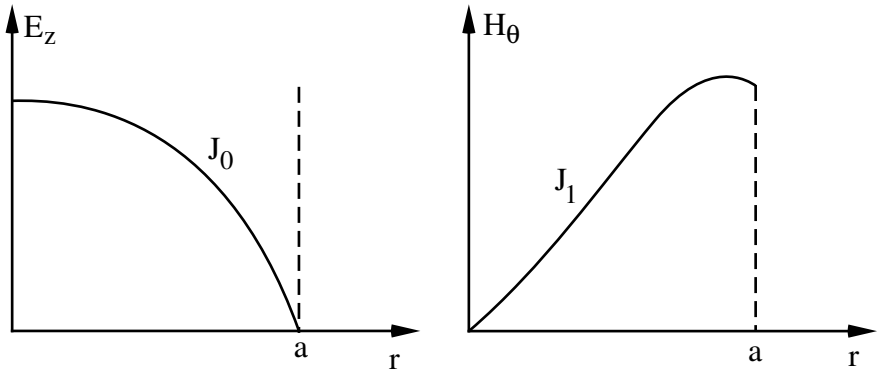


Fig. 7: Field amplitude with radial position

6 TRANSIT TIME FACTOR

In a pill-box cavity the longitudinal field amplitude is constant along the axis (Fig. 8), provided the beam tube is small enough.

$$E_z = E_0 \cos\omega t \ = \ \frac{V}{g} \cos\omega t \tag{15}$$

where V is the gap voltage and g the gap length.

Consider a particle passing through the middle of the gap at $t = 0$, such that its position is $z = vt$. The total energy gained by the particle is

$$\begin{aligned} \Delta W &= \int_{-g/2}^{g/2} e\frac{V}{g}\cos\omega\frac{z}{v}dz \\ \Delta W &= eV\frac{sin\frac{\theta}{2}}{\frac{\theta}{2}} = eVT \ , \end{aligned} \tag{16}$$

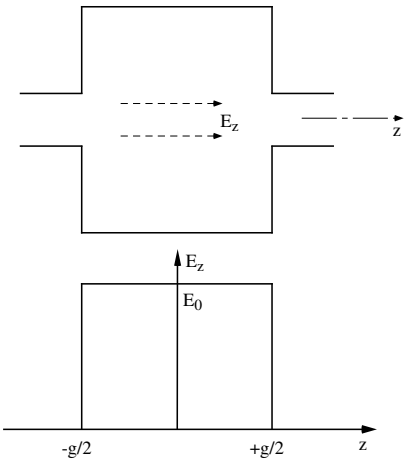


Fig. 8: Cavity gap and approximate field amplitude

where $\theta = \frac{\omega g}{v}$ is the transit angle and T is the transit time factor ($0 < T < 1$). For example, in the case of a 2π-mode structure with $g = \frac{L}{z}$ (see Fig. 3), one has $\theta = \pi$ and T = 0.637.

In the more general case where the electric field varies along the axis

$$E_z = E_z(z,t) \ , \tag{17}$$

the energy gain can be expressed as

$$\Delta W = e\mathrm{Re}\int_0^g E_0(z)e^{j\omega t}dz \tag{18}$$

$$\omega t = \omega\frac{z}{v} - \psi_p \ , \tag{19}$$

where ψ_p is the initial phase position of the particle. One can write

$$\Delta W = e\mathrm{Re}\left\{e^{-j\psi_p}\int_0^g E_0(z)e^{j\omega\frac{z}{v}}dz\right\} \tag{20}$$

or, expanding the complex integral in phase and amplitude

$$\Delta W = e\mathrm{Re}\left\{e^{-j\psi_p}e^{j\psi_i}|\int_0^g E_0(z)e^{j\omega\frac{z}{v}}dz|\right\} \ . \tag{21}$$

Introducing $\phi = \psi_p - \psi_i$, the total energy gain becomes

$$\Delta W = e|\int_0^g E_0(z)e^{j\omega\frac{z}{v}}dz|cos\phi \tag{22}$$

while the transit time factor T, which corresponds to the maximum energy a particle can gain, becomes:

$$T = \frac{|\int E_0(z)e^{j\omega\frac{z}{v}}dz|}{\int E_0(z)dz} \ . \tag{23}$$

This approach is particularly useful in case of multigap cavities, and also when using certain computer codes to design a cavity.

7 SHUNT IMPEDANCE

The shunt impedance R_s is a figure of merit that relates the accelerating voltage V in the gap to the power P_d dissipated in the cavity walls:

$$P_d = \frac{V^2}{R_s} \ . \tag{24}$$

An uncorrected shunt impedance Z is also defined through the peak voltage (the integral of the field envelope along the cavity axis) and then:

$$R_s = ZT^2 \ . \tag{25}$$

It is also usual, especially in the case of travelling wave accelerating sections, to introduce the power lost per unit length

$$\frac{dP_d}{dz} = -\frac{E_z^2}{r} \tag{26}$$

where E_z is the electric field along the axis and r the shunt impedance per unit length. The rate of power lost depends on i_w, the wall current (related to H), and on r_w the wall resistance per unit length,

$$\frac{dP_d}{dz} \propto i_w^2 r_w \ . \tag{27}$$

The wall resistance per unit length is equal to the resistivity divided by the area in which the current flows

$$r_w = \rho / 2\pi a \delta \ , \tag{28}$$

where a is the cavity radius

$$a \propto 1/\omega \ , \tag{29}$$

δ the skin depth

$$\delta = (2\rho/\omega\mu)^{1/2} \ , \tag{30}$$

and μ is the wall permeability of the material. The axial electric field is such that

$$E_z \propto i_w / a \ . \tag{31}$$

Combining the previous expressions shows that

$$r \propto \omega^{1/2} \ , \tag{32}$$

hence power efficiency will favour higher frequencies.

8 QUALITY FACTOR AND STORED ENERGY

The quality factor Q of a cavity is defined as

$$Q = \frac{\omega W_s}{P_d} \ , \tag{33}$$

where W_s is the stored energy in the cavity volume. Another quantity of interest is the ratio

$$\frac{R_s}{Q} = \frac{V^2}{\omega W_s} \ , \tag{34}$$

which only depends on the cavity geometry, and is directly accessible to measurement.

The stored energy in the cavity volume is

$$W_s = \frac{\mu}{2} \int\limits_V |H|^2 dV = \frac{\epsilon}{2} \int\limits_V |E|^2 dV \ , \tag{35}$$

while the power loss in the walls is

$$P_d = \frac{1}{2} \int\limits_S R_w |H|^2 dS \ . \tag{36}$$

The surface resistance R_w for a layer of unit area and width δ (skin depth) is

$$R_w = \frac{1}{\sigma \delta} \tag{37}$$

$$\delta = \frac{1}{\sqrt{\pi \mu \sigma f}} \tag{38}$$

with σ the conductivity and f the frequency. Hence:

$$P_d = \frac{\pi \mu \delta f}{2} \int\limits_S |H|^2 dS \ , \tag{39}$$

and finally:

$$Q = \frac{2}{\delta} \frac{\int\limits_V |H|^2 dV}{\int\limits_S |H|^2 dS} = \frac{2}{\delta} K \frac{V}{S} \ , \tag{40}$$

where K is a form factor for a given cavity geometry.

Considering the TM_{010} mode in a pill-box cavity,

$$\int\limits_V H_\theta^2 dV = \ell \int\limits_0^a J_1^2(kr) 2\pi r dr \tag{41}$$

$$\int\limits_S H_\theta^2 dS = 2 \int\limits_0^a J_1^2(kr) 2\pi r dr + 2\pi a \ell J_1^2(ka) \ , \tag{42}$$

one gets

$$\frac{1}{Q} = \frac{\delta}{\ell} \frac{\int_0^a J_1^2(kr) r dr + \frac{a\ell}{2} J_1^2(ka)}{\int_0^a J_1^2(kr) r dr} \quad . \tag{43}$$

From the relation

$$\int_0^a J_1^2(kr) r dr = \frac{a^2}{2} J_1^2(ka) \tag{44}$$

one gets

$$Q = \frac{\ell}{\delta} \frac{a}{a+\ell} \quad \propto \omega^{-1/2} \quad . \tag{45}$$

For a pill-box cavity, at 3 GHz, one gets from Section 5 radius $a = 3.8$ cm. With $\delta = 10^{-6}$ m (copper) and $\ell = 5$ cm one gets $Q = 21590$. This shows that a simple pill-box cavity, with no lossy material in the volume, has a rather high Q value. An increase in the Q value can be obtained by proper shaping of the volume to minimize losses for a given stored energy. Following the same procedure one also gets

$$\frac{r}{Q} = \frac{V^2}{\omega W_s \ell} = 2.58 \mu f \quad \propto \omega \tag{46}$$

with

$$r = \frac{R_s}{\ell} \quad . \tag{47}$$

9 FILLING TIME

It is shown from the definition of Q that the energy is dissipated at a rate proportional to the stored energy

$$P_d = -\frac{dW_s}{dt} = \frac{\omega}{Q} W_s \quad . \tag{48}$$

If the cavity is initially filled with electromagnetic energy, that stored energy will decay as follows:

$$W_s = W_{so} e^{-\frac{t}{\tau}} \quad , \tag{49}$$

where

$$\tau = \frac{Q}{\omega} \tag{50}$$

Since the stored energy is proportional to the square of the electric field, then the field will decay with a time constant 2τ. If the cavity is fed by an external RF source, then the stored energy will build up like

$$W_s = W_{so} [1 - e^{-\frac{t}{2\tau}}]^2 \quad , \tag{51}$$

and τ is called the filling time of the cavity. If the cavity is coupled to an external load, Q should be replaced by the loaded Q, Q_L, to take account of the additional losses.

10 CAVITY EQUIVALENT CIRCUIT

Solutions of Maxwell's equations show that when the cavity resonates on a given mode, such as the TM_{010}, the time average energy stored in the electric field equals the time average energy stored in the magnetic field:

$$W_{se} = W_{sm} \tag{52}$$

$$\frac{\epsilon}{4}\int_V |E|^2 dV = \frac{\mu}{4}\int_V |H|^2 dV \ , \tag{53}$$

and within an RF period the energy oscillates between electric and magnetic. This is the case for a lumped RLC parallel circuit at resonance (Fig. 9).

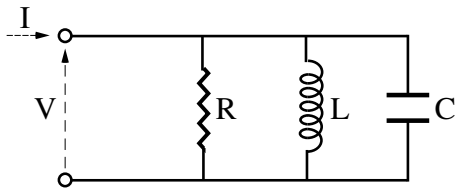


Fig. 9: Resonant circuit

The time average energy stored in the electric field (capacitor) is

$$W_{se} = \frac{1}{4} C V V^* \tag{54}$$

while the time average energy stored in the magnetic field (inductance) is

$$W_{sm} = \frac{1}{4} L I_L I_L^* \ , \tag{55}$$

where I_L is the current in the inductance L.

At resonance

$$\begin{aligned} \omega_0 &= (LC)^{-1/2} \\ V &= \omega_0 L I_L \end{aligned} \tag{56}$$

and the time average energy stored in the system is

$$W_s = W_{se} + W_{sm} = \frac{1}{2} C V V^* \ . \tag{57}$$

Since the power loss in the 'equivalent' resistance is

$$P_d = \frac{1}{2} G V V^* \quad \text{with } G = R^{-1} \tag{58}$$

the quality factor of the circuit is

$$Q = \frac{\omega_0 C}{G} = \omega_0 R C = \frac{R}{\omega_0 L} \ . \tag{59}$$

From the knowledge of ω_0, Q and R/Q, it is possible to build up the equivalent resonant circuit of a cavity.

11 INPUT IMPEDANCE

The impedance of the resonant circuit, as seen from the input, is

$$Z_{in} = \left(\frac{1}{R} + \frac{1}{j\omega L} + j\omega C \right)^{-1} \tag{60}$$

with

$$\omega = \omega_0 + \Delta\omega \ . \tag{61}$$

Expanding for small values of $\Delta\omega/\omega_0$ gives

$$Z_{in} = \frac{\omega_0^2 RL}{\omega_0^2 L + j2R\Delta\omega} \tag{62}$$

or

$$Z_{in} = \frac{R}{1 + j2Q\frac{\Delta\omega}{\omega_0}} \ . \tag{63}$$

If $Q = \omega_0/(2\Delta\omega)$ then $\Delta\omega$ corresponds to $0.707|Z_{in}|_{max}$, where $|Z_{in}|_{max} = R$. The relative bandwidth (BW) is defined as $2\Delta\omega/\omega_0$ so one can write

$$Q = 1/BW \ . \tag{64}$$

If R represents the losses in the resonant circuit then Q is the unloaded Q. If additional losses come from the coupling to an external load, represented by R_L in parallel, the quality factor becomes

$$\begin{aligned} Q_L &= \frac{R_t}{\omega_0 L} \\ R_t &= \frac{RR_L}{R+R_L} \ . \end{aligned} \tag{65}$$

Introducing the external Q, Q_e

$$Q_e = \frac{R_L}{\omega_0 L} \ , \tag{66}$$

one gets

$$\frac{1}{Q_L} = \frac{1}{Q} + \frac{1}{Q_e} \ . \tag{67}$$

12 MULTICELL STANDING WAVE CAVITIES

If the accelerator requires high accelerating voltages (such as electron linacs and storage rings) it is often more efficient to use multicell (coupled together) cavities rather than many single-cell cavities (e.g. pill-box cavities) powered separately.

Consider again a two-gap system enclosed in two metallic boxes and powered independently with proper phasing to provide a 2π phase difference between gaps (Fig. 10).

Since the circulating currents compensate each other in the common wall, this wall is useless in terms of boundary conditions for solving Maxwell equations. Hence a variant of this system consists of placing the drift tubes in a single resonant tank. The corresponding accelerating section is well-known as the Alvarez structure (Fig. 11) and is still used in proton linacs.

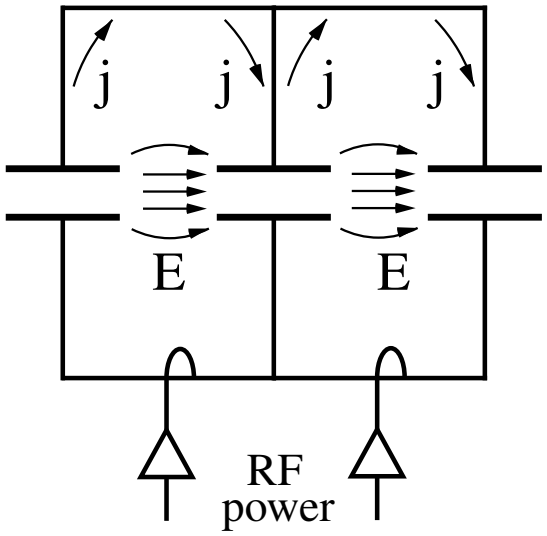


Fig. 10: 2π-mode in a two-cell cavity

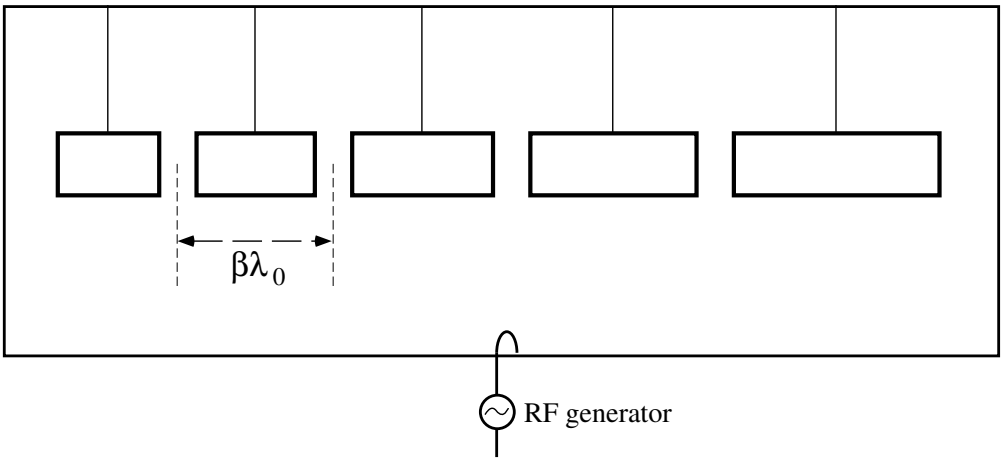


Fig. 11: Alvarez structure

There are other types of multicell standing wave (SW) structures used in modern accelerators, such as the side-coupled structure shown in Fig. 12, where the coupling between cells is reinforced using resonant cavities.

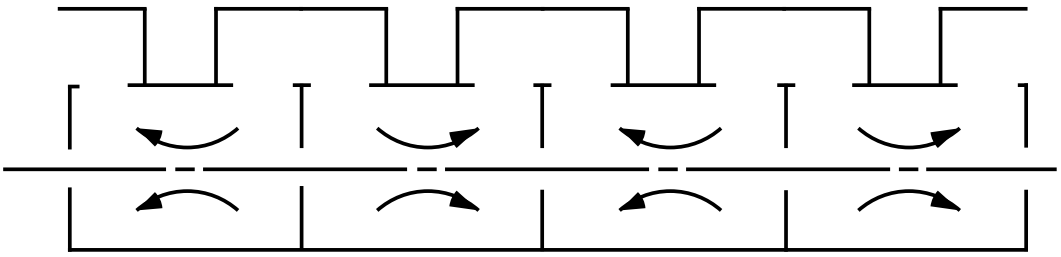


Fig. 12: Side-coupled structure

13 TRAVELLING-WAVE STRUCTURES

For ultra-relativistic particles (electrons, positrons) travelling-wave (TW) accelerating structures are preferred to SW accelerating structures because:

- a particle can travel on the crest of the wave;
- there is no need to care about the transit time factor;
- there is higher shunt impedance (forward wave only dissipates); and
- there is continuous acceleration (no drift).

However, proper acceleration with a TW cavity requires the phase velocity v_p of the wave to be equal to the particle velocity v ($\simeq c$).

Since standard waveguides (e.g. cylindrical) have $v_p > c$, the trick of lowering the phase velocity consists of loading the guide with iris. This has led to the concept of 'iris-loaded structures' (Fig. 13).

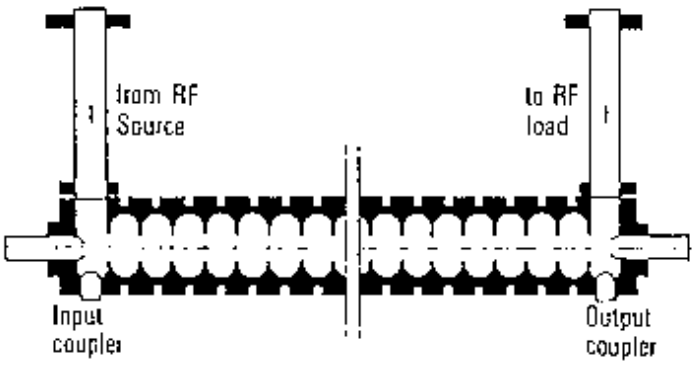


Fig. 13: Travelling-wave accelerating structure

Most electron linacs use these structures, operating on a $\frac{\pi}{2}$ or $\frac{2\pi}{3}$ mode, and they can be as long as 7 m.

14 THE TM_{01} MODE

In a cylindrical waveguide the simplest mode with a longitudinal electric field that can propagate is the TM_{01} mode:

$$\left.\begin{array}{rcl} E_z &=& E_0 J_0(k_c r) e^{-j\beta z} \\ E_r &=& j\frac{\beta}{k_c} E_0 J_1(k_c r) e^{-j\beta z} \\ H_\theta &=& \frac{1}{Z_0} j\frac{k}{k_c} E_0 J_1(k_c r) e^{-j\beta z} \end{array}\right\} e^{j\omega t} \quad , \tag{68}$$

where $Z_0 = 377\ \Omega$ is the vacuum impedance and β is the propagation factor of the wave travelling in the + z direction and satisfying the relation

$$\beta^2 = k^2 - k_c^2 \tag{69}$$

with

$$\begin{array}{rcccl} k &=& \frac{2\pi}{\lambda} &=& \frac{\omega}{c} \\ k_c &=& \frac{2\pi}{\lambda_c} &=& \frac{\omega_c}{c} \quad , \end{array} \tag{70}$$

where λ is the free space wavelength and ω_c the cut-off frequency. Since E_z must vanish at $r = a$, where a is the inner radius of the cylindrical waveguide, one has:

$$\begin{aligned} J_0(k_c a) &= 0 \\ k_c a &= 2.4 \end{aligned} \tag{71}$$

In order for the wave to propagate, β must be real and positive, hence

$$\begin{aligned} k^2 &> k_c^2 \\ \omega &> \omega_c \ , \end{aligned} \tag{72}$$

The velocity at which the wave propagates is just

$$v_p = \frac{\omega}{\beta} \tag{73}$$

and the guided wave length, λ_g, is such that

$$\beta = \frac{2\pi}{\lambda_g} = \frac{\omega}{v_p} \ . \tag{74}$$

15 PHASE VELOCITY AND GROUP VELOCITY

For propagating waves one has the relation

$$\frac{\omega^2}{v_p^2} = \frac{\omega^2}{c^2} - \frac{\omega_c^2}{c^2} \ , \tag{75}$$

showing that in a cylindrical waveguide

$$v_p > c \ . \tag{76}$$

In the Brillouin diagram this is represented by an hyperbola (Fig. 14).

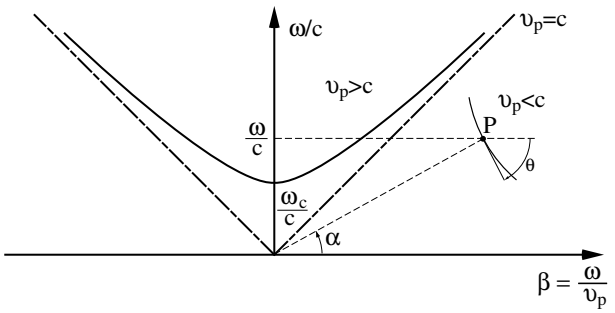


Fig. 14: Brillouin diagram

To lower the phase velocity the waveguide can be loaded with equally spaced disks (ultra-relativistic particles). In the Brillouin diagram a slow wave structure operates below the 45° line. If P(ω, β) represents the operating point, then

$$\mathrm{tg}\alpha = \frac{v_p}{c} \tag{77}$$

$$\mathrm{tg}\theta = \frac{d\left(\frac{\omega}{c}\right)}{d\left(\frac{\omega}{v_p}\right)} = \frac{1}{c}\frac{d\omega}{d\beta} = \frac{1}{c} v_g \ , \tag{78}$$

where $v_g = \left(\frac{d\beta}{d\omega}\right)^{-1}$ is called the group velocity.

16 ENERGY FLOW VELOCITY

The average power that flows through a transverse cross section of a waveguide is given by the integral of the Poynting vector

$$P = \frac{1}{2}\mathcal{R}e\int_S (E_T \times H_T)dS \ . \tag{79}$$

For a TM mode the relation between the transverse field components is

$$\frac{E_T}{H_T} = Z_0\frac{\lambda}{\lambda_g} \tag{80}$$

hence

$$P = \frac{1}{2Z_0}\frac{k}{\beta}\int_S |E_T|^2 dS \ . \tag{81}$$

The energy stored in the magnetic field (purely transverse) per unit length is:

$$w_{sm} = \frac{\mu}{4}\int_S |H_T|^2 dS = \frac{\mu}{4}\frac{1}{Z_0^2}\frac{k^2}{\beta^2}\int_S |E_T|^2 dS \tag{82}$$

The energy stored in the electric field per unit length is equal to that of the magnetic field. Hence the total energy stored per unit length is

$$w_s = w_{se} + w_{sm} = 2w_{sm} \ . \tag{83}$$

The velocity of the energy flow is then

$$v_e = \frac{P}{w_s} = \frac{1}{\mu}Z_0\frac{\beta}{k} = \frac{\beta}{k}c \ . \tag{84}$$

Since

$$v_g = \left(\frac{d\beta}{d\omega}\right)^{-1} = \left[\frac{d\left(\frac{\omega^2}{c^2} - k_c^2\right)^{1/2}}{d\omega}\right]^{-1} \tag{85}$$

one finally gets

$$v_g = \frac{\beta c^2}{\omega} = \frac{\beta}{k}c = v_e \ , \tag{86}$$

showing that in the particular case discussed the group velocity corresponds to the energy flow velocity.

17 SPACE HARMONICS IN LOADED WAVEGUIDES

In an infinite periodic structure (Fig. 15) the wave equation must satisfy the periodic boundary conditions of the disk arrangement (equally spaced for ultra-relativistic particles). Floquet's theorem suggests solutions of the form

$$\begin{aligned} E(r,\theta,z) &= e^{-\gamma z}E_1(r,\theta,z) \\ H(r,\theta,z) &= e^{-\gamma z}H_1(r,\theta,z) \ , \end{aligned} \tag{87}$$

where E_1 and H_1 are periodic functions of z, with period d:

$$\begin{aligned} E_1(r,\theta,z+d) &= E_1(r,\theta,z) \\ H_1(r,\theta,z+d) &= H_1(r,\theta,z) \ . \end{aligned} \tag{88}$$

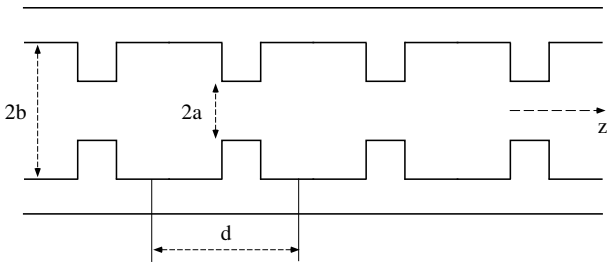


Fig. 15: Periodic iris loaded waveguide

The Fourier expansion of the field is

$$E_1(r,\theta,z) = \sum_{n=-\infty}^{+\infty} E_{1n}(r,\theta) e^{-j2n\pi z/d} \tag{89}$$

and since for lossless structures $\gamma = j\beta_0$ (with β_0 real), one can write

$$E(r,\theta,z) = \sum_{n=-\infty}^{+\infty} E_{1n}(r,\theta) e^{-j\beta_n z} \tag{90}$$

with

$$\beta_n = \beta_0 + 2n\pi/d \ . \tag{91}$$

The wave can be looked at as the sum of 'space harmonics', each harmonic having a different phase velocity,

$$v_{pn} = \frac{\omega}{\beta_0 + \frac{2\pi n}{d}} \ , \tag{92}$$

but the same group velocity

$$v_{gn} = \left(\frac{d\beta_n}{d\omega}\right)^{-1} = \left(\frac{d\beta_0}{d\omega}\right)^{-1} = v_{g0} \ . \tag{93}$$

Since the same value of the group velocity repeats with a distance $2\pi/d$ on the scale of the Brillouin diagram (see Fig. 16), the dispersion curve passes though minima and maxima. The waveguide is of the pass-band type.

In practice, with particle velocity c, the operating point P_0 should be on the 45° line. Most commonly $\pi/2$ and $2\pi/3$ modes are used, which correspond to $d = \lambda/4$ or $d = \lambda/3$.

A periodically loaded waveguide can be represented by a chain of coupled resonant circuits (Fig. 17). Each cell is represented by an equivalent RLC circuit while the electric coupling (iris) is represented by an additional capacitor C′. The analysis of such a network, in the absence of losses ($R = 0$), leads to the dispersion relation

$$\omega^2 = \omega_{\pi/2}^2 (1 - K\cos\beta d) \ , \tag{94}$$

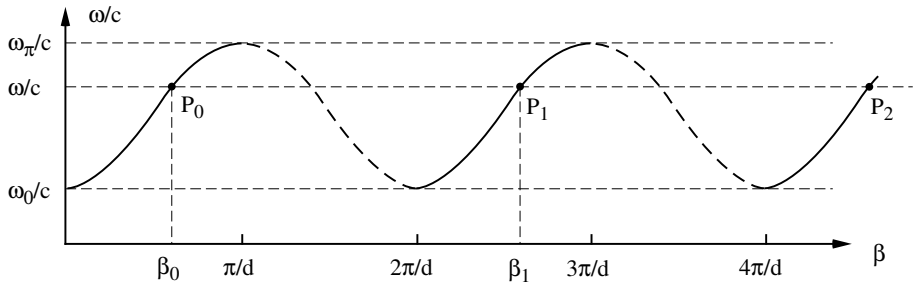


Fig. 16: Brillouin diagram for a slow wave structure

where $K < 1$ is the coupling factor.

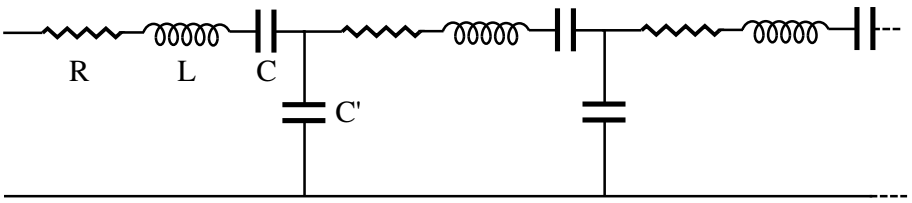


Fig. 17: Coupled resonant circuits

18 DEFLECTING CAVITIES

For the accelerating TM modes considered abhttp://mostinfo.net/hlam/test/ove, only the E_z component was non-zero on axis. However, higher order modes can show non-zero transverse components on axis as well, which can then deflect particles.

a) Standing-wave cavities

Consider the TM_{110} mode with components (θ, r, z):

$$\begin{aligned} E_z &= E_0 J_1(kr)\cos\theta \approx E_0 kr\cos\theta \quad (\text{near axis}) \\ H_r &= -j\tfrac{E_0}{Z_0}\tfrac{J_1(kr)}{kr}\sin\theta \approx -j\tfrac{E_0}{Z_0}\sin\theta \quad (\text{near axis}) \\ H_\theta &= -j\tfrac{E_0}{Z_0}J_1'(kr)\cos\theta \approx 0 \quad (\text{near axis}) \ . \end{aligned} \tag{95}$$

Since $r\cos\theta = x$ and if one assumes $\sin\theta = 1$ then H_r becomes H_y and

$$\begin{aligned} E_z &= \left(\tfrac{\partial E_z}{\partial x}\right) x \quad (\text{zero on axis}) \\ H_y &= -j\tfrac{E_0}{Z_0} = -\tfrac{j}{Z_0 k}\left(\tfrac{\partial E_z}{\partial x}\right) \quad (\neq 0 \text{ on axis}) \ . \end{aligned} \tag{96}$$

Consider now a relativistic electron ($v \leq c$) traversing the cavity of length L on the axis. It gets a horizontal impulse from the H_y component,

$$\Delta p_x = -e\int_0^L (v_z \mu H_y)\frac{dz}{v_z} = j\frac{e}{\omega}\int_0^L \left(\frac{\partial E_z}{\partial x}\right) dz \ , \tag{97}$$

showing that a SW TM mode can deflect relativistic particles travelling on the axis. This transverse kick is related to the transverse gradient of the longitudinal electric field.

b) Travelling-wave cavities

Consider now a TW structure with transverse field components and a particle travelling parallel to the axis with $v \leq c$:

$$\begin{aligned} \mathcal{E}_\perp &= \mathbf{E}_\perp(x,y)e^{j(\omega t-\beta z)} \\ \mathcal{H}_\perp &= \mathbf{H}_\perp(x,y)e^{j(\omega t-\beta z)} \ , \end{aligned} \tag{98}$$

where the phase velocity is $v_p = \omega/\beta$. The Newton Lorentz force perpendicular to the axis is

$$\mathbf{F}_\perp = e[\mathbf{E}_\perp + v\mu(\mathbf{u}\times\mathbf{H}_\perp)]e^{j(\omega t-\beta z+\phi_0)} \ , \tag{99}$$

where $\mathbf{u}$ is the unit vector along the axis. Analysis of Maxwell solutions for TM modes gives the identity

$$\mu(\mathbf{u}\times\mathbf{H}_\perp) = -\frac{1}{v_p}\mathbf{E}_\perp + \frac{j}{\omega}\nabla_\perp \mathrm{E}_z \ , \tag{100}$$

hence the force becomes

$$\mathbf{F}_\perp = e\left[\left(1-\frac{v}{v_p}\right)\mathbf{E}_\perp + j\frac{v}{\omega}\nabla_\perp \mathrm{E}_z\right]e^{j(\omega t-\beta z+\phi_0)} \ . \tag{101}$$

If the particle is synchronous with the wave, $v = v_p$, and since $\omega t = \omega\frac{z}{v} = \beta z$ one gets:

$$\mathbf{F}_\perp = \frac{e}{\beta}\nabla_\perp E_z \ , \tag{102}$$

with the particle on the crest of the wave. In principle there will be no transverse deflection from the TM_{01} mode. In practice, however, input and output couplers can give a local field asymmetry that can provide locally a non-zero $\nabla_\perp E_z$. Higher order modes can deflect particles and as a matter of fact they can be used as RF beam separators.

The relation between the transverse deflection and the transerve gradient of the longitudinal electric field is often referred as the Panofsky–Wenzel theorem. [2]

BIBLIOGRAPHY

M.S. Livingston and J.P. Blewett, *Particle Accelerators*, McGraw-Hill Book Company, New York, 1962.

E. Persico, E. Ferrari and S.E. Segré, *Principles of Particle Accelerators*, W.A. Benjamin, New York, 1968.

P. Lapostolle and A. Septier (Eds.), *Linear Accelerators*, North Holland Publishing Company, Amsterdam, 1970.

R.E. Collin, *Foundations for Microwave Engineering*, International-Student Edition, McGraw-Hill Book Company, New York, 1966.

M. Puglisi, Conventional RF cavity design, CERN Accelerator School on RF Engineering for Particle Accelerators, CERN 92-03 (June 1992).

[2] W.K.H. Panofsky, W.A. Wenzel, *Review of Scientic Instrument*, **27**, (1956) 967.

BASIC CONCEPTS I AND II

H. Henke
Berlin Technical University, Berlin, Germany

Abstract
This is as an introduction to microwave techniques. It does not treat active devices nor special materials used in microwave components, but it deals with the fundamentals of transmission lines, microwave networks, and cavity resonators. It also introduces some basic measurements closely related to this material.

1 MICROWAVE ENGINEERING

The term microwave refers to high-frequency signals with short wavelengths. Because of the short wavelengths standard circuit theory can no longer be used, since the size of components is comparable to the wavelength. Standard circuit theory is a special case of Maxwell's theory, where the components are small in relation to the wavelength. Microwave components are distributed elements, where the fields vary significantly over the physical length of the device. In general, they have to be computed by applying Maxwell's equations and thus the mathematical complexity arises. Microwave engineering tries to reduce the complexity and expresses solutions found by field theory in terms of circuit theory.

The foundations of electromagnetic theory were laid down in Maxwell's theory [1] and proven experimentally by Heinrich Hertz in the years 1887–1891. After a long period of development, the birth of microwave engineering is often attributed to the time of World War II when the Radiation Laboratory was established at MIT, USA, to develop radar theory and technique. Brilliant scientists, including N. Marcuvitz, I.I. Rabi, J.S. Schwinger, H.A. Bethe, E.M. Purcell, C.G. Montgomery, and R.H. Dicke gathered to develop the field of microwaves. Their work is summarized in the 28-volume Radiation Laboratory Series of books [2].

Passive microwave components are transmission lines, filters, couplers, junctions, antennas, ferrite devices and others. Active devices include tubes and solid-state devices, and are used for sources, detectors, amplifiers, mixers, and so on. This paper is an introduction to microwave engineering, treating passive components.

2 TRANSMISSION LINES

Transmission-line theory normally refers to cylindrical (constant cross section) waveguides that support TEM modes or quasi TEM modes, i.e. modes with no (or negligible) longitudinal field components. The field patterns are equal or close to static field distributions and propagate with the velocity of light. These are two-wire lines, coaxial lines, parallel-plate lines, striplines, co-planar striplines and so on. However, much of what we learn in the following can be extended, under certain restrictions, to other waveguides (TM, TE, and hybrid lines) and is therefore of general importance.

2.1 Transmission-line equations

The wave propagation on TEM lines can be calculated, in a formal way, from Maxwell's equations. Here, however, we will used a more intuitive approach and derive it from a short piece of line of length dz, which is modelled as a lumped-element circuit. As an example we choose a two-wire line.

The current in the wire causes a magnetic field around the axis of the wire and experiences a resistance in transporting the electrons. Therefore, the equivalent circuit has a series inductance $L'\mathrm{d}z$ and a series resistance $R'\mathrm{d}z$ (the prime indicates a quantity per unit length), Fig. 1. Likewise, the voltage

between the conductors includes surface charges on the conductors and a leakage current between the conductors due to dielectric losses. These effects are represented by a shunt capitance $C'\mathrm{d}z$ and a shunt conductance $G'\mathrm{d}z$.

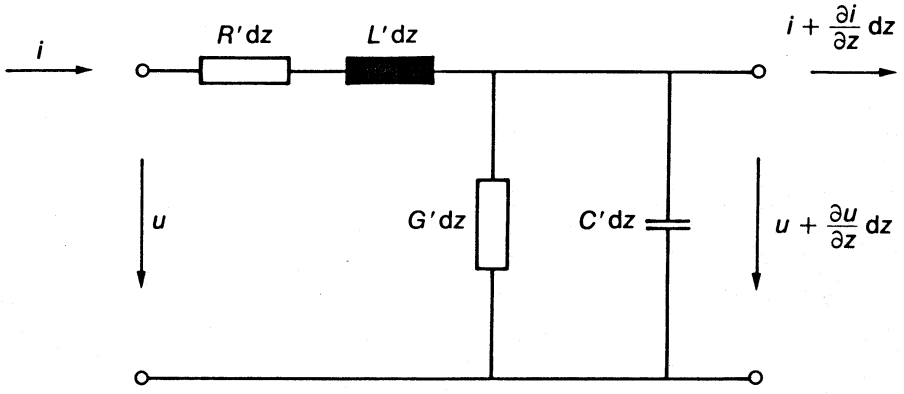


Fig. 1: Equivalent circuit of a piece of line of length $\mathrm{d}z$

Applying Kirchhoff's voltage law

$$u - R'\,\mathrm{d}z\,i - L'\,\mathrm{d}z\,\frac{\partial i}{\partial t} - u - \frac{\partial u}{\partial z}\,\mathrm{d}z = 0 \tag{1}$$

and Kirchhoff's current law

$$i - G'\,\mathrm{d}z\,u - C'\,\mathrm{d}z\,\frac{\partial u}{\partial t} - i - \frac{\partial i}{\partial z}\,\mathrm{d}z = 0\,, \tag{2}$$

where we have neglected second-order terms, yields the time-domain form of the transmission-line equations

$$\frac{\partial u}{\partial z} = -R'\,i - L'\,\frac{\partial i}{\partial t}\,, \tag{3}$$

$$\frac{\partial i}{\partial z} = -G'\,u - C'\,\frac{\partial u}{\partial t}\,. \tag{4}$$

The solution of Eqs. (3) and (4) plays an important role in the transmission of steps and impulses. Here, we are interested in the sinusoidal steady-state condition with a time dependence $\exp(j\omega t)$, and the equations simplify to

$$\frac{\partial u}{\partial z} = -(R' + j\,\omega\,L')\,i\,, \tag{5}$$

$$\frac{\partial i}{\partial z} = -(G' + j\,\omega\,C')\,u\,. \tag{6}$$

To solve Eqs. (5) and (6) we differentiate (5) and substitute (6):

$$\frac{\mathrm{d}^2 u}{\mathrm{d}z^2} = \gamma^2\,u \tag{7}$$

with the complex propagation constant

$$\gamma = \sqrt{(R' + j\,\omega\,L')(G' + j\,\omega\,C')} = \alpha + j\,\beta\,. \tag{8}$$

The solution of Eq. (7) are travelling waves

$$u = u_0^+\, e^{j\omega t - \gamma z} + u_0^-\, e^{j\omega t + \gamma z} \tag{9}$$

and

$$i = \frac{u_0^+}{Z_0}\, e^{j\omega t + \gamma z} - \frac{u_0^-}{Z_0}\, e^{j\omega t + \gamma z}\,, \tag{10}$$

where we used Eq. (5) to find i, and Z_0 is the **characteristic impedance**

$$Z_0 = \sqrt{\frac{R' + j\,\omega\,L'}{G' + j\,\omega\,C'}}\,. \tag{11}$$

The term Z_0 describes the ratio of the voltage and current amplitudes, either for the forward- or backward-travelling wave.

The complex propagation constant (8) consists of the real part

$$\alpha = \sqrt{\frac{1}{2}\,(R'\,G' - \omega^2\,L'\,C') + \frac{1}{2}\,\sqrt{(R'^2 + \omega^2\,L'^2)(G'^2 + \omega^2\,C'^2)}}\,, \tag{12}$$

which is the **attenuation constant**, and the imaginary part

$$\beta = \sqrt{-\frac{1}{2}\,(R'\,G' - \omega^2\,L'\,C') + \frac{1}{2}\,\sqrt{(R'^2 + \omega^2\,L'^2)(G'^2 + \omega^2\,C'^2)}}\,, \tag{13}$$

which is the **phase constant**. As an example, the forward-travelling wave is shown in Fig. 2. It decays with α along z and has a wavelength

$$\lambda = 2\,\pi/\beta\,. \tag{14}$$

The phase of each travelling wave is

$$\Phi = \omega\,t \mp \beta\,z\,, \tag{15}$$

which after differentiation with respect to t determines the phase velocity:

$$v_{ph} = \frac{dz}{dt} = \pm\frac{\omega}{\beta}\,. \tag{16}$$

2.2 Terminated lines

If one excites a wave at the input end of a semi-infinite line there will be only one wave travelling away from the input. The second wave will not be excited (if it were, an infinitely strong source would have to be present at the infinitely remote end of the line). The pattern of the wave is given in Fig. 2.

To treat lines of finite length l we express u, i in Eqs. (9) and (10) for instance, by their terminal values (here and in the following we drop the time dependence)

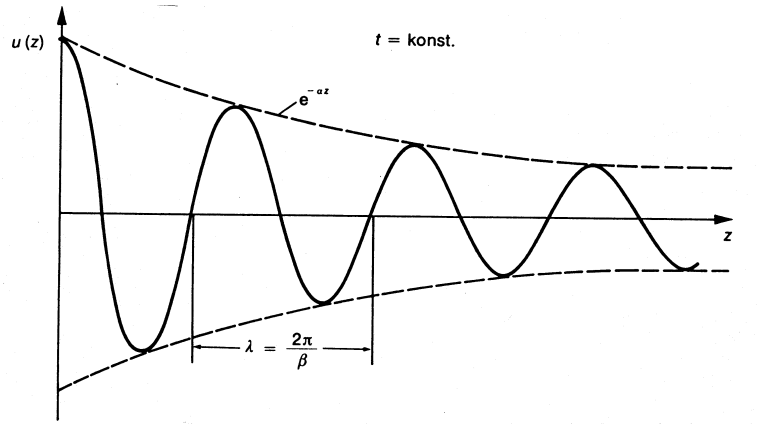


Fig. 2: Voltage of the forward-travelling wave along the line at a fixed time instant

$$u(l) = u_e = u_0^+ e^{-\gamma l} + u_0^- e^{\gamma l}\,, \tag{17}$$

$$Z_0\, i(l) = Z_0\, i_e = u_0^+ e^{-\gamma l} - u_0^- e^{\gamma l}\,, \tag{18}$$

yielding

$$u = \frac{1}{2}\,(u_e + Z_0\, i_e)\; e^{\gamma(l-z)} + \frac{1}{2}\,(u_e - Z_0\, i_e)\; e^{-\gamma(l-z)}\,, \tag{19}$$

$$Z_0\, i = \frac{1}{2}\,(u_e + Z_0\, i_e)\; e^{\gamma(l-z)} - \frac{1}{2}\,(u_e - Z_0\, i_e)\; e^{-\gamma(l-z)}\,. \tag{20}$$

First, let us consider a matched line where the terminating impedance Z_e equals the line impedance Z_0. Then,

$$Z_e = \frac{u_e}{i_e}\, Z_0 \tag{21}$$

and the backward-travelling wave vanishes. Similar to the infinitely long line, there is only one forward-travelling wave. At position $z = l$, this wave is completely absorbed by the load impedance. The line is **matched**. This is an important situation because an unwanted reflection of the wave is suppressed. Furthermore, since Z_0 is essentially real the matched situation provides an optimal transfer of power to the load.

In the general case of an arbitrary load impedance a part of the forward-travelling wave will be reflected (Fig. 3).

At the end of the line u and i are connected by $u_e = Z_e\, i_e$ and we obtain for Eq. (19)

$$u = \frac{1}{2}\, u_e\; (1 + Z_0/Z_e)\; e^{\gamma(l-z)} + \frac{1}{2}\, u_e\; (1 - Z_0/Z_e)\; e^{-\gamma(l-z)}\,, \tag{22}$$

i.e. the ratio of the backward to forward voltage wave at the end of the line is

$$r_e = \frac{Z_e - Z_0}{Z_e + Z_0}\,. \tag{23}$$

This ratio is called the **reflection coefficient**. If the line is terminated by its impedance, $(Z_e = Z_0)$, then $r = 0$; if it is short circuited $(Z_e = 0)$, $r = -1$; and in the case of an open circuit $(Z_e = \infty)$, $r = +1$.

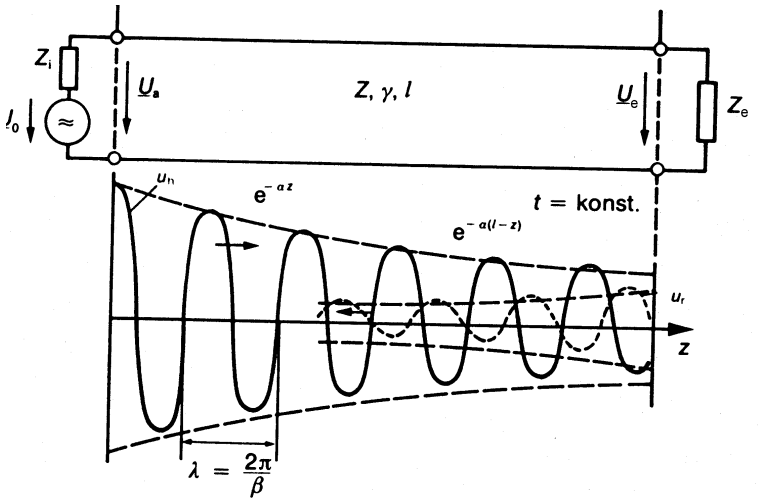


Fig. 3: Line excited by an a.c. voltage and terminated with an arbitrary load

In the last two cases the wave is fully reflected with a cancellation of the voltage or the current for short and open circuits, respectively.

In the case of an arbitrary termination, with the reflection coefficient Eq. (23), the voltage along the line is made up by the two waves Eq. (22). They can be described by complex phasors:

$$u^+(z) = u^+(l)\, e^{\alpha(l-z)}\, e^{j\beta(l-z)}\,, \qquad u^+(l) = \frac{1}{2}\, u_e\, (1 + Z_0/Z_e)\,, \tag{24}$$

$$u^-(z) = u^-(l)\, e^{-\alpha(l-z)}\, e^{-j\beta(l-z)}\,, \qquad u^-(l) = r\, u^+(l)\,. \tag{25}$$

Going from the line end to the beginning, the phasor of the forward wave increases and rotates counter-clockwise, while the phasor of the backward wave decreases and rotates clockwise. The projection onto the real axis of the vector sum of both phasors is the real voltage.

2.3 Terminated lossless line

Typical transmission lines have small losses and for short lengths one can assume $\alpha l \ll 1$, then $\gamma \approx j\,\beta$ and the voltage and current waves can be written as

$$u(z) = u^+(l)\, \left[e^{j\beta(l-z)} + r\, e^{-j\beta(l-z)}\right]\,, \qquad u^+(l) = \frac{u_e}{2}\, (1 + Z_0/Z_e)\,, \tag{26}$$

$$Z_0\, i(z) = u^+(l)\, \left[e^{j\beta(l-z)} - r\, e^{-j\beta(l-z)}\right]\,. \tag{27}$$

With the distance ζ from the line end, $z = l - \zeta$, and $r = |r|\, \exp(j\vartheta)$, the voltage magnitude is

$$|u(z)| = \left|u^+(l)\right|\, \left|1 + |r|\, e^{j(\vartheta - 2\beta\zeta)}\right|\,, \tag{28}$$

i.e. it oscillates between maximum values

$$|u|_{\max} = \left|u^+(l)\right|\, |1 + |r||\, \tag{29}$$

at positions

$$\zeta_{\max} = (\vartheta - n\,2\,\pi)\,/2\,\beta\,, \qquad n = 0, 1, 2... \tag{30}$$

and minimum values

$$|u|_{\min} = \left|u^{+}(l)\right|\,|1 - |r||\tag{31}$$

at positions

$$\zeta_{\min} = (\vartheta - (2\,n + 1)\,\pi)\,/2\,\beta\,. \tag{32}$$

As $|r|$ increases, the ratio of $|u|_{\max}$ to $|u|_{\min}$ increases, so a measure of the mismatch of the line is the **standing wave ratio**, defined as

$$SWR = \frac{|u|_{\max}}{|u|_{\min}} = \frac{1 + |r|}{1 - |r|}\,. \tag{33}$$

The *SWR* is a real number between 1 and ∞, where 1 implies a matched line and ∞ refers to an open or shortened line. From Eqs. (29) to (32) it follows that the distance between subsequent maxima or minima is $\pi/\beta = \lambda/2$, while the distance between a maximum and a minimum is $\pi/2\beta = \lambda/4$.

While the voltage amplitude is oscillatory with position on the line, the time-average power flow is constant. From Eqs. (26) and (27) follow

$$\begin{aligned} P_{av} &= \frac{1}{2}\,\mathrm{Re}\left\{u(z)\,i^{*}(z)\right\} = \frac{1}{2Z_0}\,\left|u^{+}(l)\right|^{2}\,\mathrm{Re}\left\{1 - r^{*}\,e^{j2\beta(l-z)} + r\,e^{-j2\beta(l-z)} - |r|^{2}\right\} \\ &= \frac{|u^{+}(l)|^{2}}{2Z_0}\,\mathrm{Re}\left\{1 - |r|^{2} + 2\,j\,\mathrm{Im}\left(r\,e^{-j2\beta(l-z)}\right)\right\} \\ &= \frac{1}{2Z_0}\,\left|u^{+}(l)\right|^{2}\,\left(1 - |r|^{2}\right)\,, \end{aligned} \tag{34}$$

which shows that the average power flow is constant along the line and that the power delivered to the load impedance is equal to the incident power $|u^{+}(l)|^{2}/2Z_0$ minus the reflected power $|ru^{+}(l)|^{2}/2Z_0$.

As shown above, the voltage is oscillating on the line and the real power flow is constant. It can therefore be concluded that the impedance seen looking into the line must vary with position. At a distance $\zeta = 1 - z$ from the load the **input impedance** follows from Eqs. (26) and (27) and Eq. (23) as

$$\begin{aligned} Z_{in} &= \frac{u(\zeta)}{i(\zeta)} \\ &= Z_0\,\frac{\exp(j\beta\zeta) + r\,\exp(-j\beta\zeta)}{\exp(j\beta\zeta) - r\,\exp(-j\beta\zeta)} \\ &= Z_0\,\frac{Z_e\,\cos\beta\zeta + j\,Z_0\,\sin\beta\zeta}{j\,Z_e\,\sin\beta\zeta + Z_0\,\cos\beta\zeta} \\ &= Z_0\,\frac{Z_e + j\,Z_0\,\tan\beta\zeta}{Z_e + Z_0\,\tan\beta\zeta}\,. \end{aligned} \tag{35}$$

This important result gives the input impedance of a length ζ of a transmission line with an arbitrary terminating impedance.

2.4 Impedance transformation

On a lossy line the voltage and current waves follow Eqs. (26) and (27) with $j\beta$ replaced by γ. Then, the input impedance corresponding to Eq. (35) is

$$Z_{in} = Z_0 \, \frac{Z_e + Z_0 \, \tanh \gamma\zeta}{Z_0 + Z_e \, \tanh \gamma\zeta} \, . \tag{36}$$

We write

$$\tanh \gamma\zeta = \frac{\tanh \alpha\zeta + j \, \tan \beta\zeta}{1 + j \, \tanh \alpha\zeta \cdot \tan \beta\zeta} \, . \tag{37}$$

$$Z_e/Z_0 = \tanh(x + j \, y) \, , \tag{38}$$

and obtain for Eq. (36)

$$\frac{Z_{in}}{Z_0} = \tanh \, [x + \alpha\zeta + j \, (y + \beta\zeta)] \, . \tag{39}$$

Six special cases are of interest:

1. The **long lossy line**, where $\alpha\zeta \gg 1$, then

$$\tanh \alpha\zeta \approx 1, \qquad \tanh \gamma\zeta \approx 1 \tag{40}$$

and

$$Z_{in} \approx Z_0 \, . \tag{41}$$

The input impedance is independent of the terminating impedance and equals the line impedance.

2. A **small mismatch** with $Z_e = Z_0(1 + x)$, $|x| \ll 1$, then

$$\frac{Z_{in}}{Z_0} = \frac{1 + x/(1 + \tanh \gamma\zeta)}{1 + x \, \tanh \gamma\zeta/(1 + \tanh \gamma\zeta)} \approx 1 + x \, \frac{1 - \tanh \gamma\zeta}{1 + \tanh \gamma\zeta} = 1 + x \, e^{-2\gamma\zeta} \, , \tag{42}$$

i.e. the smaller the mismatch x and the larger the attenuation $\alpha\zeta$, is the better the input impedance approximates the line impedance. The deviation depends on $\beta\zeta$, but is always smaller than $|x| \, \exp(-2\alpha\zeta)$. This is a good estimate for how strong a small mismatch appears at the input.

3. Line terminated in a short circuit, $Z_e = 0$, then

$$Z^s_{in} = Z_0 \, \tanh \gamma\zeta \tag{43}$$

4. Line terminated in an open circuit, $Z_e = \infty$, then

$$Z^s_{in} = Z_0 \, \coth \gamma\zeta \, . \tag{44}$$

From Eqs. (43) and (44) follows

$$Z_0 = \sqrt{Z^0_{in} \, Z^s_{in}} \tag{45}$$

$$\tanh \gamma\zeta = \sqrt{Z^s_{in}/Z^0_{in}} \tag{46}$$

or

$$e^{2\gamma\zeta} = \left(1 + \sqrt{Z^s_{in}/Z^0_{in}}\right) / \left(1 - \sqrt{Z^s_{in}/Z^0_{in}}\right) , \tag{47}$$

and line impedance as well as propagation constant can be determined from measuring the input impedance of the open and shorted line.

Particularly useful for impedance transformation are lossless lines (or short lines with small losses). They transform impedances with no losses.

5. The $\lambda/4$-impedance inverter
   A line of length $\zeta = \lambda/4$ has $\beta\zeta = \pi/2$ and the input impedance (35) is

$$Z_{in} = Z_0^2/Z_e \, . \tag{48}$$

   A real load $Z_e = R_e$ can be transformed into any real impedance $Z_{in} = R_{in}$ by choosing $Z_0 = \sqrt{R_e \, R_{in}}$. However, since β depends on ω, the exact transformation is obtained only at the right frequency. For a broad band transformation several cascaded $\lambda/4$ inverters with different Z_0 are used.

6. The $\lambda/2$ transformer
   If the line has a length $\zeta = \lambda/2$, then $\beta\zeta = \pi$ and the input impedance (48) equals the termination

$$Z_{in} = Z_e \, . \tag{49}$$

2.5 The Smith chart

The Smith chart was developed by P. Smith at the Bell Telephone Laboratories in 1939. It is a graphical aid to solve transmission line problems.

We express the complex ratio of terminating impedance to line impedance by

$$\frac{Z_e}{Z_0} = z = x + j\, y \tag{50}$$

and write the reflection coefficient (23) as a function of z:

$$r = r_r + j\, r_j = \frac{z-1}{z+1} \, . \tag{51}$$

Then, the Smith chart maps the complex z-plane into the polar plot of the voltage reflection coefficient, Fig. 4(a). We invert Eq. (51),

$$z = \frac{1+r}{1-r} \, ,$$

and split it into real and imaginary parts:

$$x = \frac{1 - r_r^2 - r_i^2}{(1-r_r)^2 + r_i^2} \, , \qquad y = \frac{2\, r_i}{(1-r_r)^2 + r_i^2} \, .$$

The last two equations can be rearranged such that they represent two sets of circles in the r-plane:

$$\left(r_r - \frac{x}{1+x}\right)^2 + r_i^2 = \frac{1}{(1+x)^2} \, , \tag{52}$$

$$(r_r - 1)^2 + \left(r_i - \frac{1}{y}\right)^2 = \frac{1}{y^2} \, . \tag{53}$$

Equation (52) defines resistance circles for x = const. and Eq. (53) defines reactance circles for y = const. All resistance circles have centres on the horizontal $r_i = 0$ axis, and pass through the point $r = 1$. The centres of the reactance circles lie on the vertical $r_r = 1$ line, and the circles pass also through the point $r = 1$.

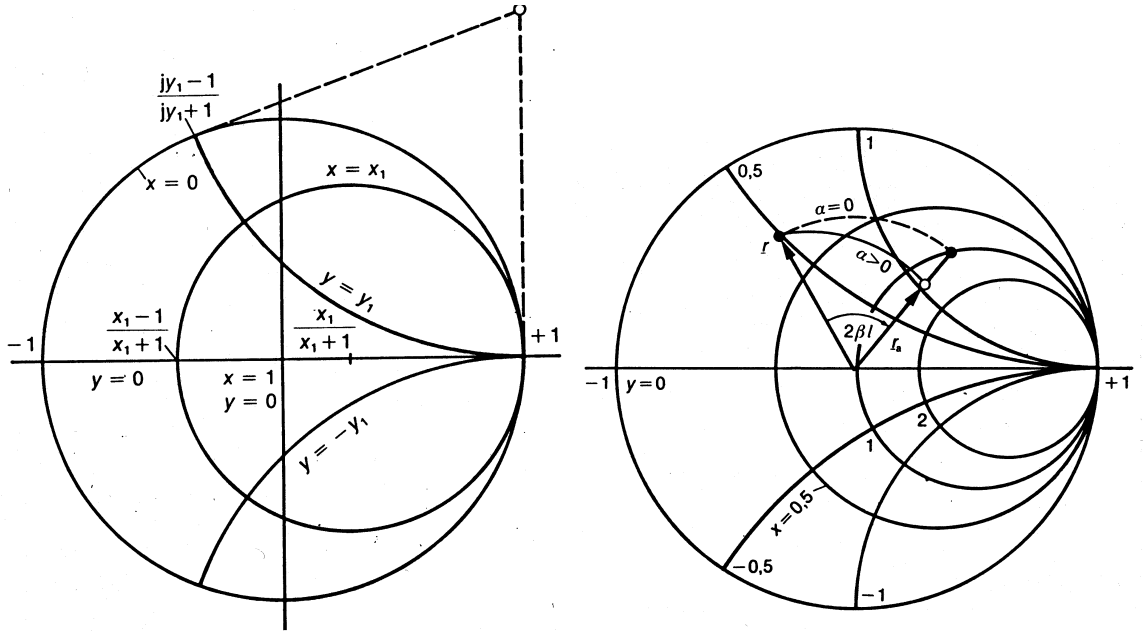


Fig. 4: (a) Construction of a Smith chart; (b) Impedance transformation in the Smith chart

The Smith chart allows impedance transformations in a convenient way. Following the arguments in Section 2.2 and defining the reflection coefficient at any position $z = 1 - \zeta$ as the ratio of backward to forward voltage wave, we obtain from Eqs. (22) and (23)

$$r(\zeta) = \frac{Z_e - Z_0}{Z_e + Z_0} e^{-2\gamma\zeta} = r_e^{-2\alpha\zeta} e^{-j\,2\beta\zeta} . \tag{54}$$

Then, we find for the impedance at position ζ

$$Z_{in}(\zeta) = \frac{u(\zeta)}{i(\zeta)} = Z_0 \frac{1 + r(\zeta)}{1 - r(\zeta)} \tag{55}$$

or after inversion

$$r(\zeta) = \frac{Z_{in}(\zeta)/Z_0 - 1}{Z_{in}(\zeta)/Z_0 + 1} . \tag{56}$$

The relations (55) and (56), together with the Smith chart (Fig. 4(b)), now allow the input impedance to be determined when the terminating impedance z is given. We find z in the chart and read the corresponding reflection coefficient $r(0)$. Corresponding to Eq. (55) we rotate r by $-2\beta\zeta$ and demagnify it by $\exp(-2\alpha\zeta)$. The readings of the new point determine $Z_{in}(\zeta)/Z_0$.

3 MICROWAVE NETWORKS

Microwave networks consist of different elements connected by lines. An element may have several 'ports' (connections with lines), Fig. 5. At each port we use the voltages and currents we have defined previously. The relations between the port quantities define the element.

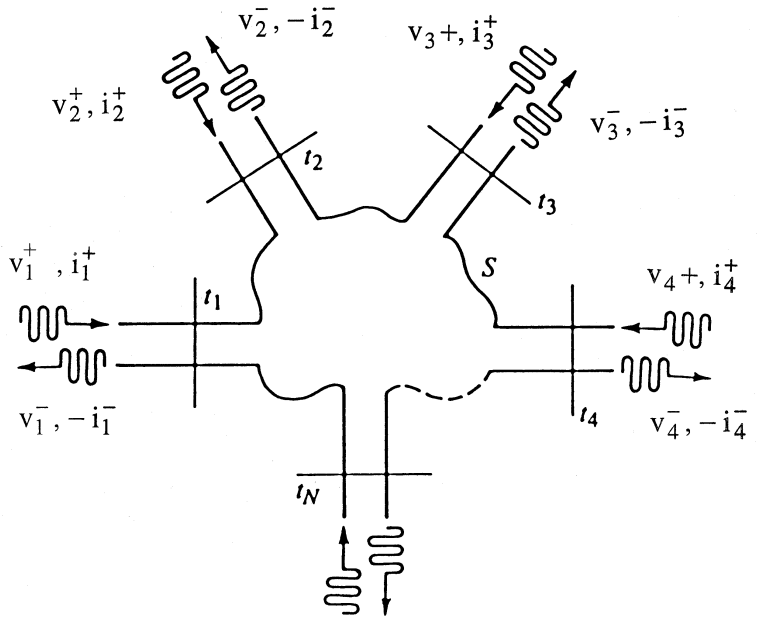


Fig. 5: An N-port microwave network

3.1 Impedance and admittance matrices

Let us consider an N-port network as shown in Fig. 5. At the nth port, in a well-defined terminal plane t_n, the voltage and current are given by

$$u_n = u_n^+ + u_n^- , \qquad i_n = i_n^+ - i_n^- . \tag{57}$$

The terminal plane t_n is important in providing a phase reference for the voltage and current phasors. Then, assuming linear networks, the relation between the voltages and currents at the different parts can be expressed by the **impedance matrix**

$$\mathbf{U} = \mathbf{Z}\,\mathbf{I} , \quad \mathbf{Z} = \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1N} \\ Z_{21} & Z_{22} & \dots & \vdots \\ \vdots & & & \vdots \\ Z_{N1} & Z_{N2} & \dots & Z_{NN} \end{bmatrix} , \quad \mathbf{U} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} , \quad \mathbf{I} = \begin{bmatrix} i_1 \\ i_2 \\ \vdots \\ i_N \end{bmatrix} , \tag{58}$$

or **admittance bmatrix**

$$\mathbf{I} = \mathbf{Y}\,\mathbf{U} , \qquad \mathbf{Y} = \mathbf{Z}^{-1} . \tag{59}$$

The elements of, for instance, the impedance matrix can be found as

$$Z_{ij} = \frac{u_i}{i_j} , \quad \text{if} \quad i_k = 0 \quad \text{for} \quad k \neq j , \tag{60}$$

that is by driving port j with i_j, while all other ports are open, and measuring the open-circuit voltage at port i. Thus, Z_{ii} is the input impedance of port i when all other ports are open-circuited, and Z_{ii} is the transfer impedance between ports i and j when all other ports are open-circuited.

In general, each Z_{ii} and Y_{ii} may be complex, and an N-port network will have $2N^2$ independent quantities. In practice, however, many networks are reciprocal or lossless, or both. If the network is

reciprocal the impedance and admittance matrices are symmetric. If the network is lossless, the elements Z_{ii}, Y_{ii} are purely imaginary.

A network is **reciprocal** if it does not contain active devices or non-reciprocal material such as ferrites or plasmas. Reciprocity is a consequence of Maxwell's equations (Lorentz reciprocity theorem) and can be shown easily [3]. Here, we will only state the consequences. Let us shorten all ports but the first and second. Then, a current i_1' applied to port 1 will create a voltage u_2' at port 2. Conversely, a current i_2'' at port 2 will create a voltage u_1'' at port 1. Reciprocity means that the transfer impedances are equal,

$$\frac{u_2'}{i_1'} = \frac{u_1''}{i_2''}\,, \quad or \quad Z_{21} = Z_{12}\,,$$

and since this is valid for any two ports the impedance matrix $\mathbf{Z}$ is symmetric. Its inverse, the admittance matrix $\mathbf{Y}$ is also symmetric.

In **lossless networks** the averaged real power delivered to the network must be zero. We start with all ports but the ith open and obtain

$$P_{av} = \frac{1}{2}\,\mathrm{Re}\{u_i\, i_i^*\} = \frac{1}{2}\,\mathrm{Re}\{Z_{ii}\}\,|i_1|^2 = 0$$

or $\mathrm{Re}\{Z_{ii}\} = 0$. Next, we drive the ith and jth port with i_i and i_j,

$$\begin{aligned} P_{av} &= \frac{1}{2}\,\mathrm{Re}\{u_i\, i_i^* + u_j\, i_j^*\} = \frac{1}{2}\,\mathrm{Re}\{Z_{ii}\,|i_i^2| + Z_{jj}\,|i_j|^2 + Z_{ij}\, i_j\, i_i^* + Z_{ji}\, i_i\, i_j^*\} \\ &= \frac{1}{2}\,\mathrm{Re}\{Z_{ij}\, i_j\, i_i^* + Z_{ji}\, i_i\, i_j^*\} = \frac{1}{2}\,\mathrm{Re}\{Z_{ij}(i_i\, i_j^* + i_i^*\, i_j)\} = 0\,, \end{aligned}$$

and since $i_i\, i_j^* + i_i^*\, i_j$ is real it follows that $\mathrm{Re}\{Z_{ij} = 0\}$. Here, we have used $Z_{ij} = Z_{ji}$. That means the elements of the impedance and admittance matrix are purely imaginary for lossless and reciprocal networks.

Often, in practice, the network is a symmetric 2-port network and the impedance matrix reduces to

$$\mathbf{Z} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}. \tag{61}$$

One possible equivalent network is the T-junction shown in Fig. 6(a). The elements are easily found by applying open circuits:

$$\begin{aligned} Z_{11} &= \left.\frac{u_1}{i_1}\right|_{i_2=0} = Z_a + Z_c\,, \\ Z_{22} &= \left.\frac{u_2}{i_2}\right|_{i_1=0} = Z_b + Z_c\,, \\ Z_{12} &= \left.\frac{u_1}{i_2}\right|_{i_1=0} = Z_c\,. \end{aligned} \tag{62}$$

The equivalent network for the admittance matrix is normally a π network, Fig. 6(b). The elements can be found by short circuits:

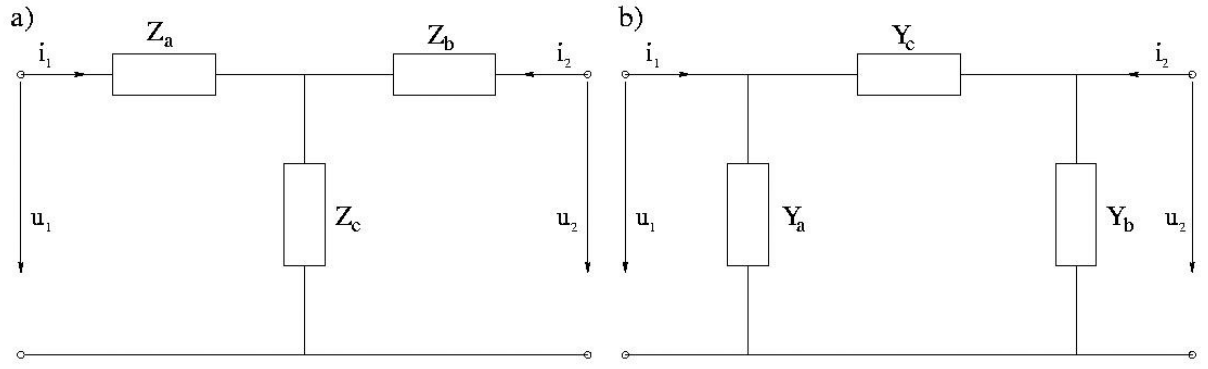


Fig. 6: (a) Equivalent T network with a symmetric impedance matrix; (b) Equivalent π network for a symmetric admittance matrix

$$\begin{aligned} Y_{11} &= \left.\frac{i_1}{u_1}\right|_{u_2=0} = Y_a + Y_c\,, \\ Y_{22} &= \left.\frac{i_2}{u_2}\right|_{u_1=0} = Y_b + Y_c\,, \\ Y_{12} &= \left.\frac{i_1}{u_2}\right|_{u_1=0} = -Y_c\,. \end{aligned} \tag{63}$$

3.2 Scattering matrix

The description of microwave networks by voltages and currents is not always the best choice. This has to do with problems of measuring voltages and currents at high frequencies but also with difficulties in defining voltages and currents for non-TEM lines (waveguides). Therefore, the description of networks by impedance and admittance matrices becomes somewhat an abstraction. A representation more in accordance with direct measurements is given by the scattering matrix $\mathbf{S}$. The scattering matrix also provides a complete description of the network as seen at its N ports. It relates the incoming waves to the reflected waves.

Consider the N-port of Fig. 5, where u_n^+ and u_n^- are the amplitudes of the incoming and reflected waves, respectively. Referring to Eqs. (9) and (10) and choosing the port plane t_n as a reference plane we can normalize the amplitudes of the voltage and current waves,

$$a_n = u_n^+/\sqrt{Z_0}\,, \quad b_n = u_n^-/\sqrt{Z_0}\,, \tag{64}$$

such that the averaged real power delivered into port n is

$$\begin{aligned} P_{av} &= \frac{1}{2}\,\mathrm{Re}\{u_n\, i_n^*\} = \frac{1}{2}\,\mathrm{Re}\{\sqrt{Z_0}\,(a_n + b_n)\,\frac{1}{Z_0}\,(a_n^* - b_n^*)\} \\ &= \frac{1}{2}\,\mathrm{Re}\{|a_n|^2 - |b_n|^2 + b_n\,a_n^* - a_n\,b_n^*\} = \frac{1}{2}\,\left(|a_n|^2 - |b_n|^2\right)\,, \end{aligned} \tag{65}$$

i.e. the incoming power is given by $|a_n|^2/2$ and the reflected power by $|b_n|^2/2$.

The relation between the wave amplitudes a_i, b_i is given by a system of N linear equations

$$\mathbf{b} = \mathbf{S}\,\mathbf{a}\,, \quad \mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & ... & S_{1N} \\ S_{21} & S_{22} & ... & S_{2N} \\ \vdots & & & \\ S_{N1} & S_{N2} & ... & S_{NN} \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}, \tag{66}$$

and the system matrix is the **scattering matrix**. The elements are found by

$$S_{ij} = \frac{b_i}{a_j}\,, \quad if \quad a_k = 0 \quad for \quad k \neq j\,. \tag{67}$$

That is, S_{ii} is the reflection coefficient at port i when all other ports are matched, and S_{ij} is the transmission coefficient for a wave coming in at port j and going out at port i when all ports are matched.

For networks satisfying reciprocity, the scattering matrix is symmetric, as are the impedance and admittance matrices.

If the network is loss-free, then the real power coming out of the ports must equal the real power delivered to the ports

$$\frac{1}{2}\,\mathbf{b}^t\,\mathbf{b}^* = \frac{1}{2}\,\mathbf{a}^t\,\mathbf{a}^*\,,$$

which we write as

$$(\mathbf{S}\,\mathbf{a})^t(\mathbf{S}\,\mathbf{a})^* = \mathbf{a}^t\,\mathbf{S}^t\,S^*\,a^* = \mathbf{a}^t\,a^* = \mathbf{a}^t\,1a^*\,.$$

Thus

$$\mathbf{S}^t\mathbf{S}^* = 1\,,$$

and the transpose is the conjugate of the inverse matrix

$$\mathbf{S}^t = (\mathbf{S}^{-1})^*\,, \tag{68}$$

which is called a **unitary matrix**. We can write these relations as sums,

$$\sum_{k=1}^{N} S_{ki}\,S_{ki}^* = 1\,, \tag{69}$$

$$\sum_{k=1}^{N} S_{ki}\,S_{kj}^* = 0\,, \quad i \neq j\,, \quad i,j = 1,2,...\,N\,, \tag{70}$$

stating that the columns of the scattering matrix are orthonormal.

Equations (69) and (70) reduce the number of independent quantities but they also impose restrictions on what can or cannot be done with loss-free junctions. Consider for instance the loss-free power divider or power combiner of Fig. 7. If sources are introduced at ports 1 and 2 with the combined power obtained at 3, one might wish to have $S_{12} = S_{21} = 0$ in order to eliminate direct interaction between the sources. But Eq. (70) with $i = 1$, $j = 2$ gives

$$S_{11}\,S_{12}^* + S_{21}\,S_{22}^* + S_{31}\,S_{32}^* = 0\,,$$

requiring either S_{31} or S_{32} to be equal to zero and thus eliminating one of the two desired couplings. The junction will act as a power combiner, but the sources will interact, and if they are not identical in magnitude and phase, one source will feed power to the other.

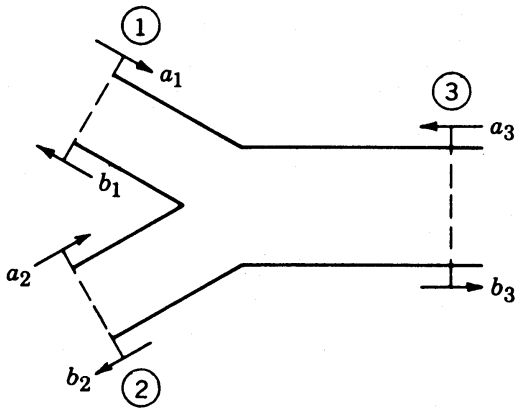


Fig. 7: A Y-junction power combiner

In dealing with cascaded circuits, the scattering formalism is not convenient. For that purpose we restrict ourselves to 2-port circuits and rearrange the scattering matrix such that the amplitudes a_2, b_2 are the independent and b_1, a_1 the dependent variables:

$$\begin{bmatrix} b_1 \\ a_1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} ; \tag{71}$$

with

$$T_{11} = S_{12} - S_{11}\, S_{22}/S_{21}\,, \quad T_{12} = S_{11}/S_{21}\,, \quad T_{21} = -S_{22}/S_{21}\,, \quad T_{22} = 1/S_{21}\,.$$

Now, the output wave b_2 of a first circuit is the input wave a_1' of a subsequent circuit, and the input a_2 is the output b_1', see Fig. 8. Simple multiplication of the transfer matrices gives the overall matrix

$$\begin{bmatrix} b_1 \\ a_1 \end{bmatrix} = T \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}, \quad \begin{bmatrix} b_1' \\ a_1' \end{bmatrix} = T' \begin{bmatrix} a_2' \\ b_2' \end{bmatrix} ;$$

and since

$$\begin{bmatrix} a_2 \\ b_2 \end{bmatrix} = \begin{bmatrix} b_1' \\ a_1' \end{bmatrix},$$

$$\begin{bmatrix} b_1 \\ a_1 \end{bmatrix} = T\,T' \begin{bmatrix} a_2' \\ b_2' \end{bmatrix} . \tag{72}$$

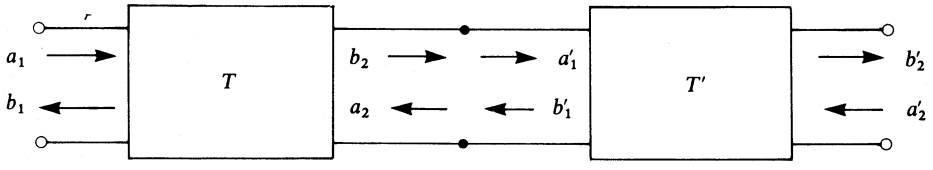


Fig. 8: Cascade connection of two 2-port networks

3.3 Some common microwave elements

T-junction power dividers, Fig. 9, are simple 3-port networks that can be used for power division or combining. As mentioned above, such a network cannot be lossless, reciprocal and matched at all ports. If it is lossless and reciprocal the scattering matrix must be symmetric and Eqs. (69) and (70) hold. In case of the H-plane T, for instance, we can further assume that the waves going out from the ports 1 and 2 are symmetric if port 3 is driven, i.e. $S_{13} = S_{23}$, and that port 3 is matched, $S_{33} = 0$. Then Eqs. (69) and (70) reduce to

$$
\begin{aligned}
&|S_{11}|^2 + |S_{12}|^2 + |S_{13}|^2 = 1\,,\\
&|S_{12}|^2 + |S_{22}|^2 + |S_{13}|^2 = 1\,,\\
&2\,|S_{13}|^2 = 1\,,\\
&S_{11}\,S_{12}^* * S_{12}\,S_{22}^* + |S_{13}|^2 = 0\,,\\
&S_{11}\,S_{13}^* * S_{12}\,S_{13}^* = 0\,,\\
&S_{12}\,S_{13}^* * S_{22}\,S_{13}^* = 0\,,
\end{aligned}
$$

which gives

$$
S_{11} = S_{22} = -S_{12}, \quad |S_{13}| = 1/\sqrt{2}, \quad |S_{11}| = 1/2
$$

and, after choosing the reference planes such that the elements are real,

$$
\mathbf{S}_H = \frac{1}{2}\begin{bmatrix} 1 & -1 & \sqrt{2} \\ -1 & 1 & \sqrt{2} \\ \sqrt{2} & \sqrt{2} & 0 \end{bmatrix}. \tag{73}
$$

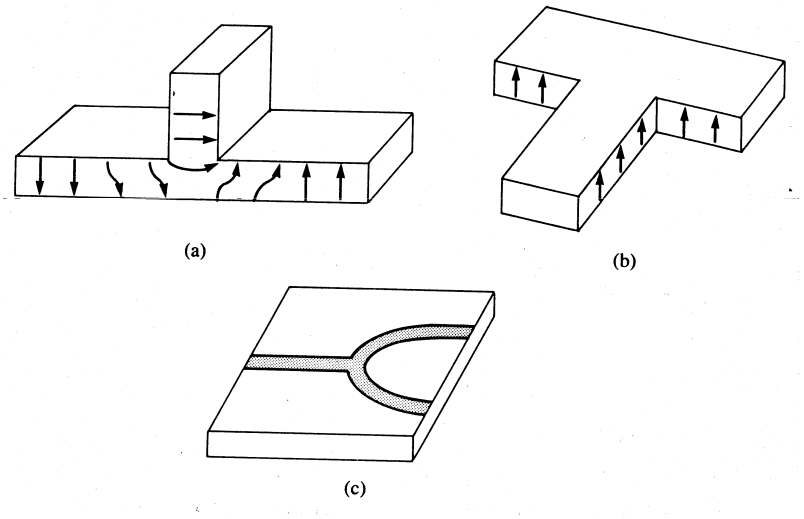


Fig. 9: Various T-junction dividers: (a) E-plane waveguide; (b) H-plane waveguide; (c) Microstrip

If the 3-port network is non-reciprocal, $S_{ij} \neq S_{ji}$, then the condition of input matching at all ports, $S_{ii} = 0$, can be satisfied. We further assume a lossfree device and Eqs. (69) and (70) become

$$
\begin{aligned}
|S_{21}|^2 + |S_{31}|^2 &= 1\,, \qquad & S_{31}\,S_{32}^* &= 0\,;\\
|S_{12}|^2 + |S_{32}|^2 &= 1\,, \qquad & S_{21}\,S_{23}^* &= 0\,;\\
|S_{13}|^2 + |S_{23}|^2 &= 1\,, \qquad & S_{12}\,S_{13}^* &= 0\,.
\end{aligned}
$$

The equations can be satisfied in one of two ways

$$
S_{12} = S_{23} = S_{31} = 0\,, \qquad |S_{21}| = |S_{13}| = |S_{32}| = 1
$$

or

$$
S_{21} = S_{13} = S_{32} = 0\,, \qquad |S_{31}| = |S_{12}| = |S_{23}| = 1\,.
$$

The result is that $S_{ij} \neq S_{ji}$ for $i \neq j$, i.e. the device is non-reciprocal. The two solutions are shown in Figs. 10(a) and (b). The first solution is a device where the power coming in at port 1 exits at port 2, and power coming in at port 2 will exit at 3 and so on. Such a device is called a circulator and allows the separation of a source at port i from reflections coming from a load at port $i+1$. In the second solution the direction of power flow is changed. Circulators employ generally anisotropic materials, such as ferrites, and can be realized in very different ways. Two solutions are shown in Figs. 10(c) and (d).

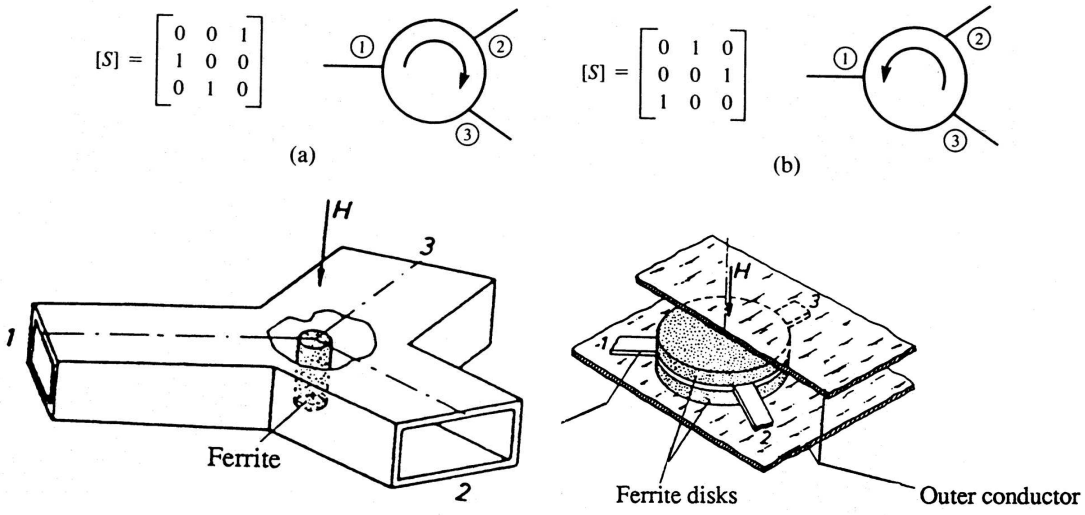


Fig. 10: Different circulators. (a) Clockwise circulation; (b) counter-clockwise circulation; (c) waveguide circulator; (d) stripline circulator.

Two examples of 4-port networks are directional couplers and hybrids, see Fig. 11. We assume a reciprocal loss-free network with all four ports matched, $S_{ii} = 0$. We further assume isolation between the ports 1–4 and 2–3, $S_{14} = S_{23} = 0$. Then, the scattering matrix reduces to

$$
\mathbf{S} = \begin{bmatrix} 0 & S_{12} & S_{13} & 0 \\ S_{12} & 0 & 0 & S_{24} \\ S_{13} & 0 & 0 & S_{34} \\ 0 & S_{24} & S_{34} & 0 \end{bmatrix} \tag{74}
$$

with four unknowns. Equations (69) and (70) yield for $i = j$

$$|S_{12}|^2 + |S_{13}|^2 = 1\,, |S_{12}|^2 + |S_{24}|^2 = 1\,, |S_{13}|^2 + |S_{34}|^2 = 1\,, |S_{24}|^2 + |S_{34}|^2 = 1\,, \qquad (75)$$

implying $|S_{13}| = |S_{24}|$ and $|S_{13}| = |S_{24}|$. Furthermore, we select the terminal plane t_2 with respect to t_1 so that S_{12} is real and positive, and similarly 4 with respect to 3 so that S_{34} is real and positive, i.e.

$$S_{12} = S_{34} = \alpha\,.$$

From Eqs. (69) and (70) with $i \neq j$ remains

$$S_{12}\,S_{24}^* + S_{13}\,S_{34}^* = 0\,,$$
$$S_{12}\,S_{13}^* + S_{24}\,S_{34}^* = 0\,,$$

requiring $S_{13} = -S_{24}^*$. S_{13} and S_{24} can only differ in phase, therefore

$$S_{13} = \beta\, e^{j\varphi}\,, \qquad S_{24} = \beta\, e^{j\vartheta}$$

and the relation between phases is

$$\varphi + \vartheta = \pi + 2\,n\,\pi\,.$$

Ignoring integer multiples of 2π, there are two choices:

- symmetrical coupler, $\varphi = \vartheta = \pi/2$,
- antisymmetrical coupler, $\varphi = 0$, $\vartheta = \pi$.

The related scattering matrices are

$$\mathbf{S}_s = \begin{bmatrix} 0 & \alpha & j\,\beta & 0 \\ \alpha & 0 & 0 & j\,\beta \\ j\,\beta & 0 & 0 & \alpha \\ 0 & j\,\beta & \alpha & 0 \end{bmatrix}, \quad \mathbf{S}_{as} = \begin{bmatrix} 0 & \alpha & \beta & 0 \\ \alpha & 0 & 0 & -\beta \\ \beta & 0 & 0 & \alpha \\ 0 & -\beta & \alpha & 0 \end{bmatrix}. \qquad (76)$$

Note that the two couplers differ only in the choice of reference planes. Also, the constants α, β are not independent because of Eq. (75),

$$\alpha^2 + \beta^2 = 1;$$

therefore, an ideal directional coupler has only one degree of freedom. The coupling mechanism is either hole coupling, Fig. 11, where the strength is adjusted by the number and the size of the holes, or coupled transmission lines, where the distance between lines determines the strength.

Hybrid couplers are special cases of directional couplers, where the coupling factor is 3 dB, i.e. $\alpha = \beta = 1/\sqrt{2}$. There are two types of hybrid:

- Quadrature hybrid with 90° phase shift between ports 2 and 3 when fed at port 1. It is a symmetrical coupler, see Fig. 12.
- Magic-T or rat-race hybrid with 180° phase difference between ports 2 and 3 when fed at port 4. It is an antisymmetrical coupler, see Fig. 12.

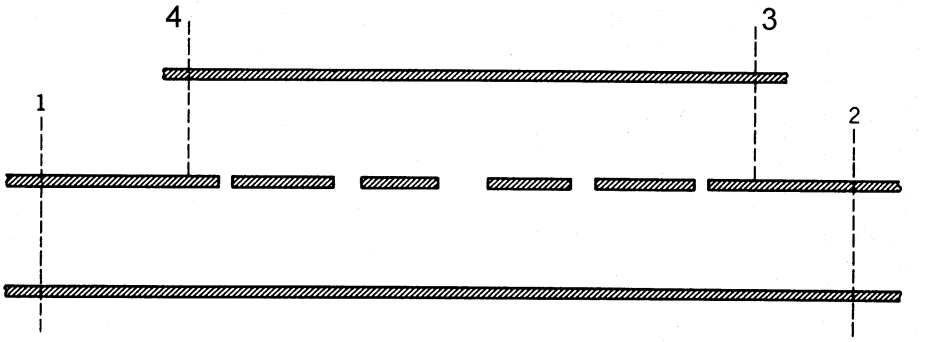


Fig. 11: Hole-coupled directional coupler

Directional couplers are usually characterized by three quantities:

$$\begin{array}{lrcl} \text{coupling} & C & = & 10\,\log P_1/P_3 = -20\,\log\beta\,dB, \\ \text{directivity} & D & = & 10\,\log P_3/P_4 = 20\,\log\dfrac{\beta}{|S_{14}|}\,dB, \\ \text{isolation} & I & = & 10\,\log P_1/P_4 = -20\,\log\,|S_{14}|\,dB. \end{array} \tag{77}$$

C indicates the fraction of the input power which is coupled to the output. D is a measure of the coupler's ability to isolate forward and backward waves, as is I. These quantities are related as follows:

$$I = D + C\,.$$

An ideal coupler [Eq. (76)] has infinite directivity and isolation.

4 CAVITY RESONATORS

Cavity resonators have an infinite number of oscillatory modes. Each mode is characterized by its resonant frequency, the stored electromagnetic energy and the losses dissipated in the walls and eventually also radiated into external circuits. Near resonance, a mode can be modelled by an equivalent lumped-element resonant circuit.

4.1 Lumped-element resonant circuits

Lumped-element resonant circuits can be series or parallel circuits (Fig. 13).

Series circuit

The input impedance is

$$Z_{in} = R + j\,\omega\,L + \frac{1}{j\,\omega\,C} \tag{78}$$

and the complex power delivered to the circuit is

$$\begin{aligned} P_{in} &= \frac{1}{2}\,u\,i^* = \frac{1}{2}\,Z_{in}\,|\imath|^2 = \frac{1}{2}\,|\imath|^2\,\left(R + j\,\omega\,L + \frac{1}{j\,\omega\,C}\right) \\ &= P_{loss} + 2\,j\,\omega\,(W_m - W_e)\,, \end{aligned} \tag{79}$$

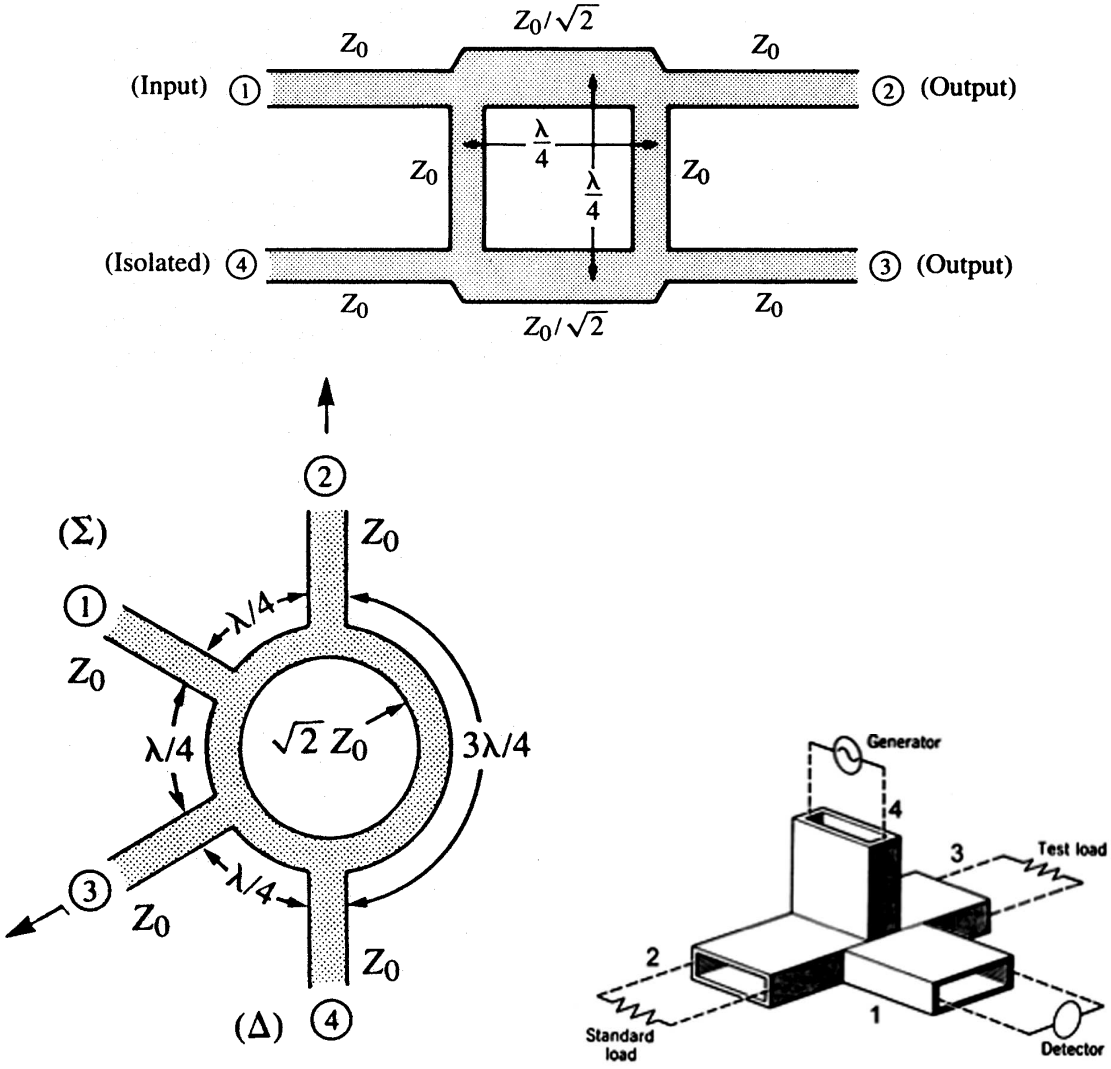


Fig. 12: Top: 90° branch-line coupler; bottom left: 180° ring or rat-race hybrid; bottom right: Magic-T

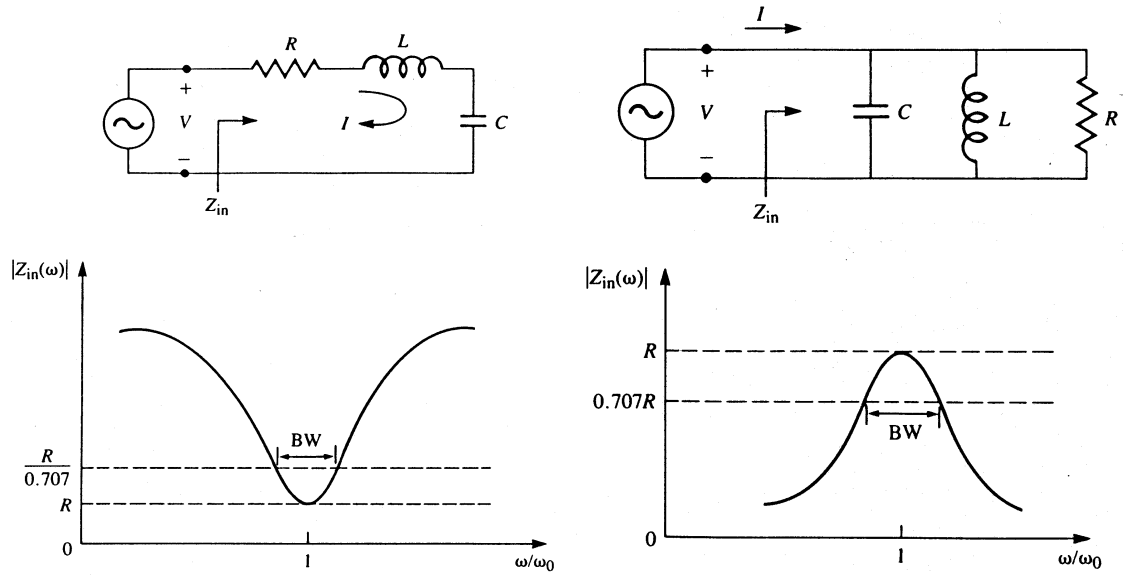


Fig. 13: Left: Series RLC resonator and its response; right: parallel RLC resonator and its response

where

$$P_{loss} = \frac{1}{2} R\,|i|^2 \tag{80}$$

is the power dissipated in the resistor, and

$$W_m = \frac{1}{4} L\,|i|^2, \qquad W_e = \frac{1}{4}\frac{1}{\omega^2 C}\,|i|^2 \tag{81}$$

are the average magnetic energy stored in the inductor and the average electric energy stored in the capacitor, respectively. At the **resonance frequency**

$$\omega_0 = \frac{1}{\sqrt{L\,C}}$$

the stored electric and magnetic energies are equal, $W_e = W_m$, and the input impedance is $Z_{in} = R$.

A figure of merit of the losses in the circuit is the **quality factor**

$$Q_0 = \omega_0 \frac{W_e + W_m}{P_{loss}} = \frac{\omega_0 L}{R} = \frac{1}{\omega_0 RC} \tag{82}$$

which shows that Q_0 increases if R decreases.

Close to resonance $\omega = \omega_0 + \Delta\omega$, the input impedance (78) can be written as

$$\begin{aligned} Z_{in} &= R + j\,\omega\,L\left[1 - \frac{1}{\omega^2 LC}\right] = R + j\,\omega\,L\,\frac{\omega^2 - \omega_0^2}{\omega^2} \\ &\approx R + j\,2\,L\,\Delta\omega = R + j\,2\,R\,Q_0\,\frac{\Delta\omega}{\omega_0}\,. \end{aligned} \tag{83}$$

From Eq. (83) we find the **bandwidth B** of the resonator. If $|Z_{in}|^2 = 2R^2$ the real power delivered to the circuit (79) is one half of that delivered at resonance. The corresponding $\Delta\omega/\omega_0$ defines the half-bandwidth:

$$\left| R + j\,2\,R\,Q_0\,\frac{\Delta\omega}{\omega_0} \right|^2 = R^2\,\left(1 + 4\,Q_0^2\,\left(\frac{\Delta\omega}{\omega_0}\right)^2\right) = 2\,R^2\,,$$

$$2\,\Delta\omega/\omega_0 = B = 1/Q_0\,. \tag{84}$$

Alternatively to Eq. (83), a resonator with losses can be treated as a lossless resonator whose resonance frequency ω_0 has been replaced by a complex resonant frequency:

$$\omega_0 \to \omega_0\,\left(1 + \frac{j}{2Q_0}\right)\,. \tag{85}$$

The input impedance of a lossless resonator becomes, after substitution of Eq. (85),

$$\begin{aligned} Z_{in} &= j\,2\,L\,(\omega - \omega_0) = j\,2\,L\,\left(\omega - \omega_0 - j\,\frac{\omega_0}{2Q_0}\right) \\ &= \omega_0\,\frac{L}{Q_0} + j\,2\,L\,(\omega - \omega_0) = R + j\,2\,L\,\Delta\omega\,, \end{aligned}$$

which is identical to Eq. (83).

Parallel circuit

For a parallel circuit the input impedance is

$$Z_{in} = \left(\frac{1}{R} + \frac{1}{j\omega L} + j\,\omega\,C\right)^{-1}\,. \tag{86}$$

$$P_{in} = \frac{1}{2}\,u\,i^* = \frac{1}{2}\,|u|^2/Z_{in}^* = \frac{1}{2}\,|u|^2\,\left(\frac{1}{R} - \frac{1}{j\omega L} - j\,\omega\,C\right)\,, \tag{87}$$

$$P_{loss} = \frac{1}{2}\,|u|^2/R \tag{88}$$

$$W_e = \frac{1}{4}\,C\,|u|^2\,, \qquad W_m = \frac{1}{4}\,\frac{1}{\omega^2 L}\,|u|^2\,, \tag{89}$$

$$Q_0 = \omega_0\,\frac{W_e + W_m}{P_{loss}} = \frac{R}{\omega_0 L} = \omega_0\,R\,C\,, \tag{90}$$

near resonance, $\omega = \omega_0 + \Delta\omega$, the input impedance is given by

$$\begin{aligned} Z_{in} &\approx \left(\frac{1}{R} + \frac{1}{j\,\omega_0 L}\,\left(1 - \frac{\Delta\omega}{\omega_0}\right) + j\,\omega_0\,C + j\,\Delta\omega\,C\right)^{-1} \\ &\approx \left(\frac{1}{R} + j\,\Delta\omega\,\left(\frac{1}{\omega_0 L} + C\right)\right)^{-1} = \left(\frac{1}{R} + j\,2\,C\,\Delta\omega\right)^{-1} \\ &\approx \frac{R}{1 + j\,2\,Q_0\,\Delta\omega/\omega_0}\,, \end{aligned} \tag{91}$$

$$B = 1/Q_0\,. \tag{92}$$

Loaded Q

The above-defined quality factor is characteristic of the circuit itself, i.e. in the absence of any loading effects, and is called **unloaded** Q. In practice, however, a circuit is inevitably connected to other circuits which will cause additional losses, thus lowering the overall, **loaded** Q_l. For a series circuit the external load resistor R_l adds in series with R, and from (82) follows

$$\frac{1}{Q_l} = \omega_0 \, C \, (R + R_l) = \frac{1}{Q_0} + \frac{1}{Q_{ext}} \tag{93}$$

with $Q_{ext} = 1/\omega_0 \, R_l \, C$.

For a parallel circuit the load resistor is parallel to R and one obtains from Eq. (90)

$$\frac{1}{Q_l} = \frac{1}{\omega_0 \, c} \left(\frac{1}{R} + \frac{1}{R_l} \right) = \frac{1}{Q_0} + \frac{1}{Q_{ext}} \, , \tag{94}$$

with $Q_{ext} = \omega_0 \, R_l \, C$.

4.2 Time response of resonators

In a free-running resonator the dissipated power must equal the rate of change of the stored energy:

$$P_{loss} = -\frac{\mathrm{d}}{\mathrm{d}t} \, (W_e + W_m) = -\frac{\mathrm{d}W}{\mathrm{d}t} \, ,$$

which becomes, with the definition of Q in Eq. (82) or (90),

$$\frac{\mathrm{d}W}{\mathrm{d}t} = -\frac{\omega_0}{Q} \, W \, . \tag{95}$$

Thus, the energy decays exponentially

$$W(t) = W_0 \, e^{-2t/T_f} \, , \tag{96}$$

where

$$T_f = 2 \, Q/\omega_0 \tag{97}$$

is the filling time. Since $W \sim u^2$ or i^2, the voltage or current in the circuit decays exactly with T_f.

As an example, let us consider the parallel circuit in Fig. 13. The source may be an ideal current source, with infinite internal resistance, and delivers a short (compared to the resonance wavelength) pulse of total charge $q = \int i \, dt$. The pulse charges up the capacitor instantaneously (with no current through the inductance) to a voltage $u_0 = q/C$ and a stored energy $W_0 = q^2/2C$. Now, the resonator starts ringing with its resonance frequency $\omega_0 = 1/\sqrt{LC}$ and the envelope of the voltage decays with T_f, Eq. (97).

The other extreme is driving the resonator in steady state with a current $i = i_0 \exp(j \, \omega \, t)$. Then, away from resonance, e.g. $\omega \ll \omega_0$, the input impedance (86) is

$$
\begin{aligned}
Z_{in} &= R\left(1 - j\,\frac{R}{\omega L} + j\,\omega\,R\,C\right)^{-1} \\
&= R\left(1 - j\,Q_0\,\frac{\omega_0}{\omega} + j\,Q_0\,\frac{\omega}{\omega_0}\right)^{-1} \\
&\approx R/(1 - j\,Q_0\,\omega_0/\omega) \\
&\approx j\,\frac{R}{Q_0}\,\frac{\omega}{\omega_0} << j\,R\,,
\end{aligned}
\tag{98}
$$

and the voltage is small and inductive. Near resonance, the impedance is given by Eq. (91) and the voltage is

$$
u = \frac{R\,i_0}{1 + j\,2\,Q_0\,\Delta\omega/\omega_0}\,e^{j\omega t}\,, \tag{99}
$$

i.e. it is $R\,i_0$ at resonance and decays fast off-resonance (for large-Q circuits). At ω_0 it changes from a positive to a negative phase. The full behaviour is shown in Fig. 14.

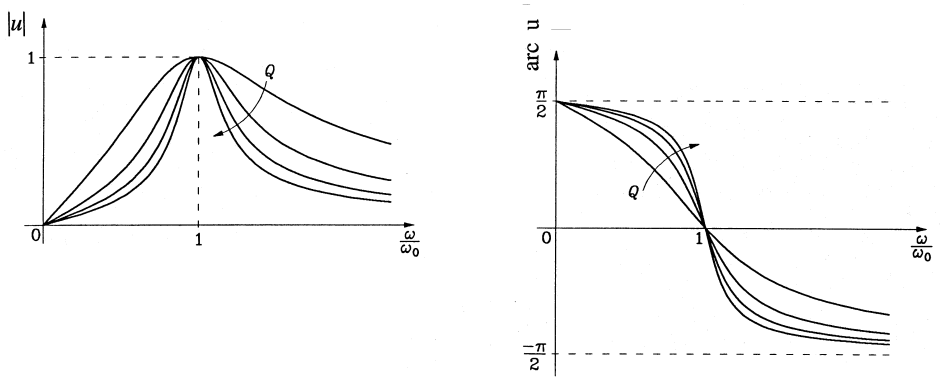


Fig. 14: Magnitude and phase of the voltage across a parallel resonance circuit

Equation (99) shows a phase difference $\Delta\varphi$ between the driving current and the resonator voltage of

$$
\tan\,\Delta\varphi = -2\,Q_0\,\frac{\Delta\omega}{\omega_0}\,. \tag{100}
$$

This phase difference is zero at resonance and $\mp 45^\circ$ for $\Delta\omega = \pm B/2$.

As a last example we consider the switching on of a harmonic current with frequency ω_0:

$$
i(t) = \begin{cases} 0 & for\ t \le 0 \\ i_0\,\sin\omega_0\,t & for\ t > 0\,. \end{cases} \tag{101}
$$

The differential equation of the circuit is

$$
i(t) = \frac{1}{R}\,u(t) + \frac{1}{L}\int u(t)\,\mathrm{d}t + C\,\frac{\mathrm{d}u(t)}{\mathrm{d}t}\,. \tag{102}
$$

We apply the Laplace transform and obtain

$$\frac{1}{C}\,\frac{\omega_0}{s^2+\omega_0^2}\,i_0 = \frac{2}{T_f}\,U(s) + \omega_0^2\,\frac{1}{s}\,U(s) + s\;U(s)$$

or, after solving for $U(s)$,

$$U(s) = \frac{s}{s^2+\omega_0^2}\cdot\frac{1}{s^2+2s/T_f+\omega_0^2}\cdot\frac{\omega_0}{C}\,i_0\,.$$

After decomposing the right side into partial fractions,

$$U(s) = \left[\frac{\omega_0}{s^2+\omega_0^2} - \frac{\omega_0}{s^2+2s/T_f+\omega_0^2}\right] R\,i_0\,,$$

the inverse Laplace transform gives

$$u(t) = \left\{\sin\omega_0 t - \frac{\exp(-t/T_f)}{\sqrt{1-1/4Q_0^2}}\,\sin\left(\omega_0\sqrt{1-1/4Q_0^2}\,t\right)\right\} R\,i_0\,,$$

which for high Q-values can be written as

$$u(t) \approx \left\{\left(1-e^{-t/T_f}\right)\,\sin\omega_0 t + \frac{t}{4T_f}\,e^{-t/T_f}\,\cos\omega_0 t\right\} R\,i_0\,. \tag{103}$$

In deriving Eq. (103) we put $\cos(t/4T_f) \approx 1$ and $\sin(t/4T_f) \approx t/4T_f$, since these terms will vanish owing to the exponential for large values of t/T_f.

As can be seen from Eq. (103) the resulting voltage consists of a transient part, which decays exponentially, and the steady-state part, given by Eq. (99) for $\Delta\omega = 0$.

4.3 Transmission-line resonator

The most simple microwave resonator is one made of a transmission line. It is also an arrangement which allows for the study of the filling of a resonator with microwave energy in a very intuitive way.

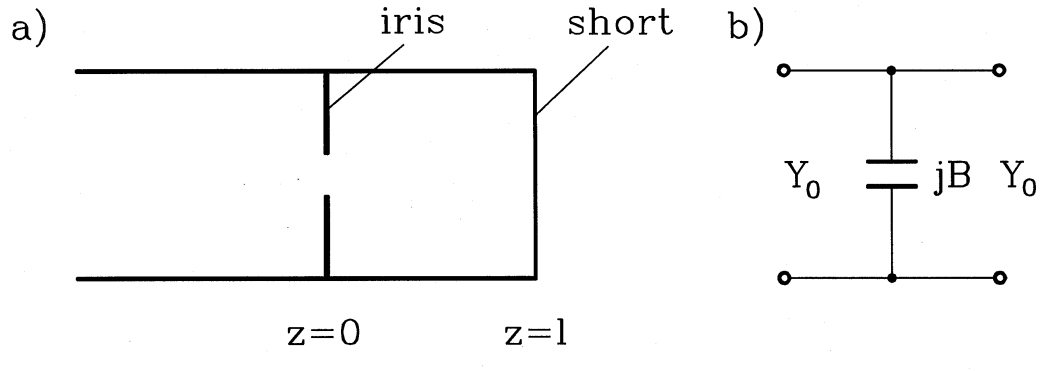


Fig. 15: (a) Shorted transmission line with iris; (b) equivalent circuit of a thin lossless iris parallel to electric fields

Let us consider a shorted transmission line with an iris a distance l away from the short, see Fig. 15(a). If the mode on the line is such that the electric fields are parallel to the iris, the iris will concentrate the electric fields and electric energy will be stored around the iris. Then, if we assume negligible losses and a thin iris, the iris can be represented as a lumped capacitance, Fig. 15(b). The value of B has to be

computed with field theoretical means and will not be derived here. The interested reader will find it in Vol. 10 of Ref. [2]. Such an iris in an infinitely long transmission line causes a reflection, Eq. (23), of

$$r = \frac{Z_e - Z_0}{Z_e + Z_0} = \frac{Y_0 - Y_e}{Y_0 + Y_e} = \frac{-j\,B}{2\,Y_0 + j\,B} \tag{104}$$

and has a transmission (see also Vol. 10 of Ref. [2]) of

$$t = 1 + r. \tag{105}$$

Now, we turn on the RF and launch a wave down the line, toward the structure formed by the iris and the short, as seen in Fig. 16. For a small iris opening one expects that most of the wave will be reflected. Some will be transmitted through the iris. The transmitted part travels to the short, is reflected, returns to the iris, where it radiates a little bit through, but most of it is reflected. With time, and if the phases are right, the fields build up in the resonator. In steady state the incoming reflected wave interferes destructively with the radiation from the cavity, and if the coupling of the iris and the dissipation in the cavity are in a 'matched' condition the reflection disappears.

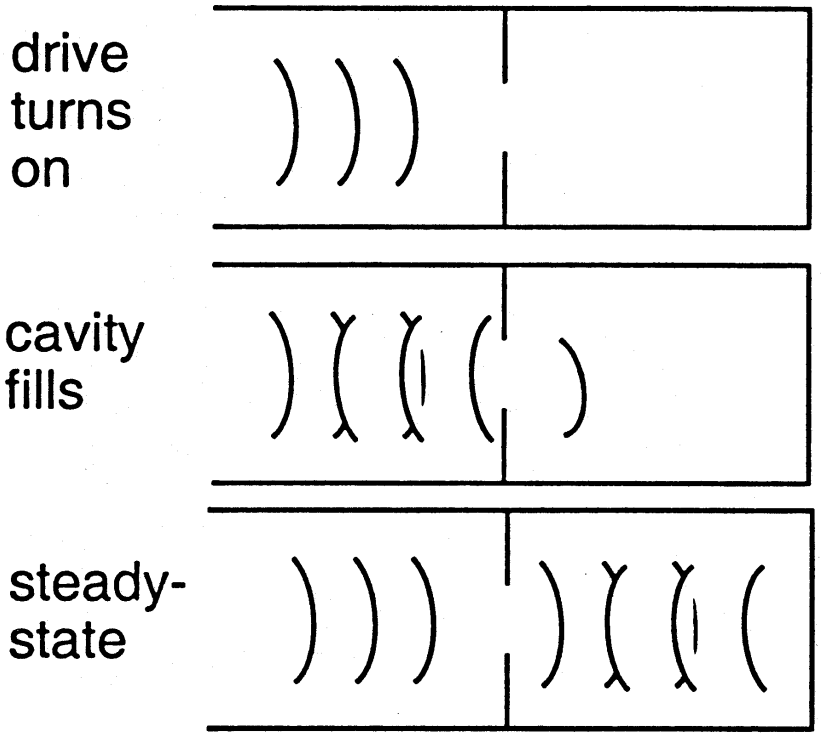


Fig. 16: Illustration of the filling of the cavity of Fig. 15

The process can be calculated in a step-by-step approach. The incident wave with complex amplitude a is reflected by the iris with value ra, while a fraction ta is transmitted. This wave travels to the short in a time $\tau = l/v_g$, where v_g is the group velocity in the unloaded line, is reflected with $r_s = -1$ and returns to the iris, where it is reflected and partially transmitted, and so on. Mathematically we find for the wave on the right side of the iris and travelling to the right

$$a_+ = t\,a - r\,e^{-2\gamma l}\,t\,a + \left(r\,e^{-2\gamma l}\right)^2\,t\,a - \left(r\,e^{-2\gamma l}\right)^3\,t\,a \pm \ldots = \frac{ta}{1 + r\,e^{-2\gamma l}}\,. \tag{106}$$

Similarly, we find for the wave on the left side of the iris and travelling to the left

$$\begin{aligned} a_- &= r\,a - e^{-2\gamma l}\,t^2\,a + r\,e^{-4\gamma l}\,t^2\,a - r^2\,e^{-6\gamma l}\,t^2\,a \pm ... \\ &= r\,a - e^{-2\gamma l}\,\left[1 - r\,e^{-2\gamma l} + \left(r\,e^{-2\gamma l}\right)^2 \mp ...\right]\,t^2\,a \\ &= \left[r - \frac{t^2\,e^{-2\gamma l}}{1 + r\,e^{-2\gamma l}}\right]\,a = \frac{r + (r^2 - t^2)e^{-2\gamma l}}{1 + r\,e^{-2\gamma l}}\,a\,. \end{aligned} \tag{107}$$

Making use of Eq. (105) the condition for zero reflection in steady state is

$$r = \frac{e^{-2\gamma l}}{1 - 2\,e^{-2\gamma l}}\,, \tag{108}$$

which, substituted into Eq. (106), yields

$$a_+ = \frac{1 + r}{1 + r\,e^{-2\gamma l}}\,a = \frac{a}{1 - e^{-2\gamma l}}\,. \tag{109}$$

The admissible parameter space which satisfies Eqs. (108) and (109) and the requirement $|r| < 1$ can be shown graphically (Fig. 17). For given losses $2\alpha l$ the dashed segment indicates the area where $|r|$ would be larger than one. Any $2\beta l$ larger than $(2\beta l)_c$ is allowed. The inverse of the magnitude of the phasor $\tilde{p}$ gives then the build-up factor $|a_+/a|$ in the resonator.

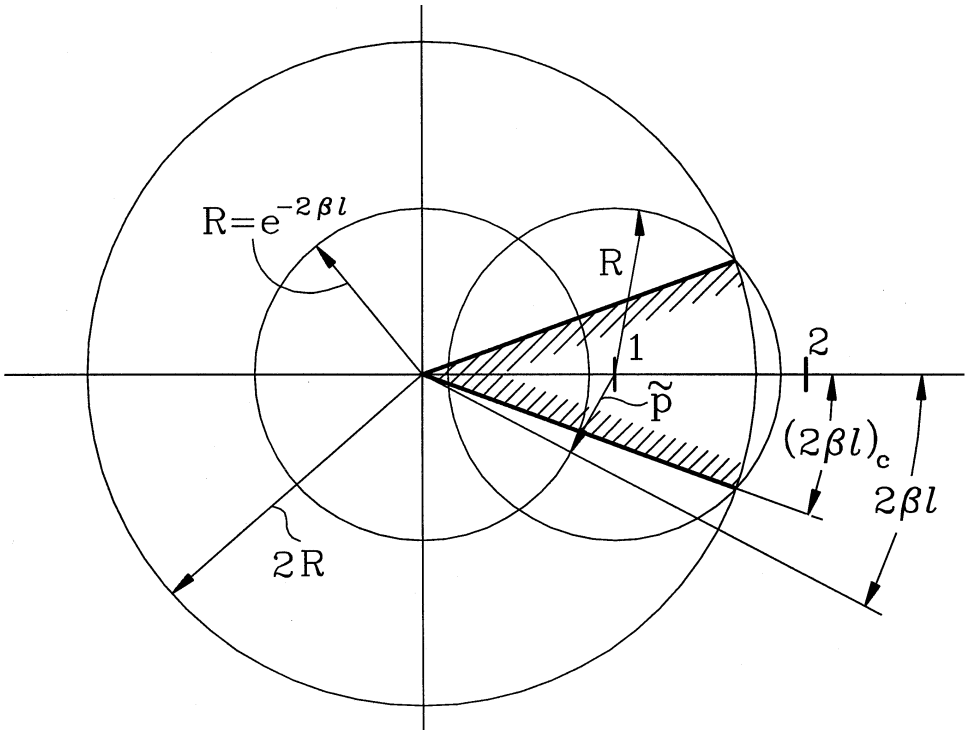


Fig. 17: Graphical solution of Eqs. (108) and (109). Any $2\beta l > (2\beta l)_c$ is admissible; admissible parameter $> (2\beta l)_c \cdot |\tilde{p}|^{-1}$ gives the field build-up in the resonator.

4.4 Cavity-resonator parameters

Cavity resonators for acceleration are basically characterized by four parameters:

- The resonant frequency ω_0.
- The Q-value.
- The R upon Q, which is a measure for providing an accelerating voltage V_0 with a certain stored energy W:

$$\frac{R}{Q} = \frac{V_0^2}{\omega_0 W} \,. \tag{110}$$

- The **shunt impedance** R, which measures the efficiency to create an accelerating voltage V_0 with a certain dissipated power P_d:

$$R = \frac{V_0^2}{P_d} \,. \tag{111}$$

Note that sometimes the definition $R = V_0^2/2P_d$ is used, where the factor 2 is inserted for similarity with a peak a.c. voltage V_0 at a resistor R.

The accelerating voltage is given by the integration of the accelerating field over a typical length L:

$$V_0 = \int_0^L E(z, t(z))\, \mathrm{d}z \,, \tag{112}$$

where $t(z)$ is the time at which a particle is at location z, i.e. $t(z) = \int \mathrm{d}z/\nu(z) + t_0$ and t_0 has to be chosen such that V_0 becomes maximum.

To illustrate Eq. (112) we take the example of a closed cylindrical cavity of length g which is driven in the TM_{010} mode, Fig. 18. The electric field is parallel to the cavity axis and is independent of z:

$$\mathbf{E} = E_0 \, \cos \omega_0 \, t \, \mathbf{e_z} \,.$$

Then, the accelerating voltage for a particle with constant speed ν is given by

$$\begin{aligned} V_0 &= \max_{t_0} \left\{ E_0 \int_0^g \cos[\omega_0 (z/\nu + t_0)] \, \mathrm{d}z \right\} \\ &= E_0 \max_{t_0} \left\{ \frac{\nu}{\omega_0} [\sin \omega_0 (g/\nu + t_0) - \sin \omega_0 t_0] \right\} , \end{aligned}$$

which is maximum for $\omega_0 t_0 = -\omega_0 g/2\nu$ and has the value

$$V_0 = E_0 \, g \, T, \qquad T = \sin\left(\frac{\omega_0 g}{2\nu}\right) / \left(\frac{\omega_0 g}{2\nu}\right) \,. \tag{113}$$

T is the **transit time factor** and takes into account the change of E with time while the particle traverses the cavity gap.

Often, instead of a single accelerating gap, there is a chain of cavities each of length L. Then, the parameters (110) and (111) are given per unit length:

$$
\begin{aligned}
R' &= \frac{E_0^2}{P_d'}\,, \\
E_0 &= V_0/L\,, \\
(R/Q)' &= \frac{E_0^2}{\omega_0 W'}\,.
\end{aligned}
\tag{114}
$$

The shunt impedance is a key parameter because it determines the accelerating voltage available from a given input power $P_{in} = P_d$, if the cavity is not loaded by a beam. The R upon Q, on the other hand, is independent of the material losses and depends only on the cavity mode and geometry. It measures how much stored energy is required for the wanted accelerating voltage.

One mode in a cavity resonator can be modelled by an equivalent circuit, as in Fig. 13. In the case of the parallel circuit, for instance, the gap voltage V_0 would be across the capacitor and the circuit parameters follow from the cavity parameters as

$$
\begin{aligned}
R_{circuit} &= \frac{1}{2}R\,, \\
C_{circuit} &= \frac{2}{\omega_0 R/Q}\,, \\
L_{circuit} &= \frac{1}{2\omega_0}\frac{R}{Q}\,.
\end{aligned}
\tag{115}
$$

Finally, let us calculate the cavity parameters for a simple example, namely for the TM_{010} mode in a closed cylindrical cavity, Fig. 18. This is a good approximation for a cavity with small beam pipes and is the most standard configuration.

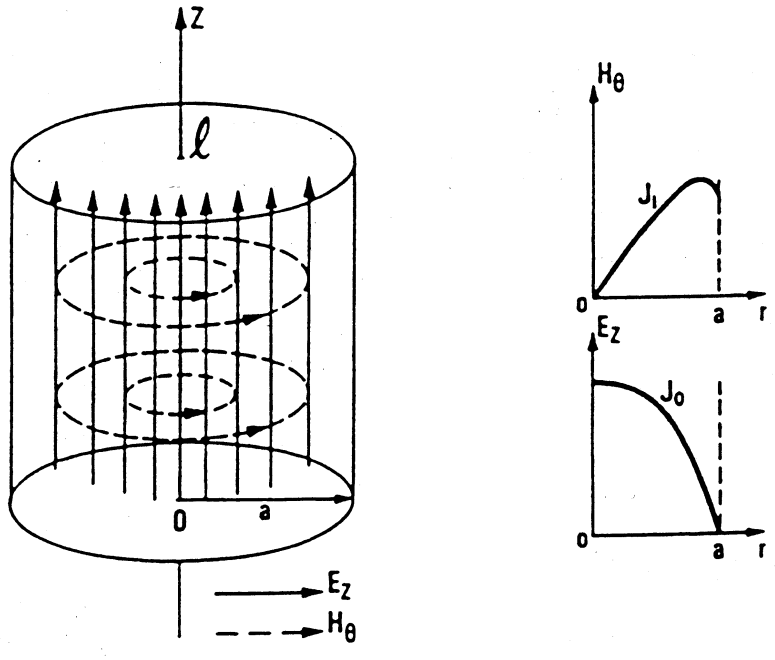


Fig. 18: TM_{010} mode in a pill-box cavity

The field components are

$$
\begin{aligned}
E_z &= E_0\, J_0\left(j_{01}\,\rho/a\right), \\
Z_0\, H_\varphi &= j\, J_1\left(j_{01}\,\rho/a\right), \\
Z_0 &= \sqrt{\mu_0/\varepsilon_0}\,,
\end{aligned}
\tag{116}
$$

where a is the cavity radius and j_{01} is the first zero of the zero-order Bessel function.

The stored energy in the cavity is

$$
\begin{aligned}
W &= W_{may} + W_{el} = 2\,W_{el} = 2\,\frac{\varepsilon_0}{4}\int_0^a\int_0^{2\pi}\int_{-g/2}^{g/2} E_0^2\, J_0^2\left(j_{01}\,\frac{\rho}{a}\right)\rho\mathrm{d}\rho\,\mathrm{d}\varphi\,\mathrm{d}z \\
&= \frac{\varepsilon_0}{2}\,2\pi g a^2\, E_0^2\int_0^1 x J_0^2\left(j_{01}\,x\right)\mathrm{d}x \\
&= \frac{\pi}{2}\,\varepsilon_0\, E_0^2\, g a^2\, J_1^2\left(j_{01}\right).
\end{aligned}
\tag{117}
$$

The losses are calculated by means of the power-loss method, that is from the wall currents of the ideal conducting cavity:

$$
\begin{aligned}
P_d &= \frac{1}{2\kappa\delta}\left\{2\int_0^a \left|H_\varphi(z=0)\right|^2\, 2\pi\rho\mathrm{d}\rho + \int_{-g/2}^{g/2}\left|H_\varphi(\rho=a)\right|^2\, 2\pi a\mathrm{d}z\right\} \\
&= \frac{\pi}{\kappa\delta}\,\frac{\varepsilon_0}{\mu_0}\,E_0^2\left\{2a^2\int_0^1 J_1^2\left(j_{01}\,x\right)x\mathrm{d}x + agJ_1^2\left(j_{01}\right)\right\} \\
&= \frac{\pi}{\kappa\delta}\left(\frac{E_0}{Z_0}\right)^2 a(a+g)\,J_1^2\left(j_{01}\right),
\end{aligned}
\tag{118}
$$

where $\delta = 2(\omega\mu_0\kappa)^{1/2}$ is the skin depth.

The cavity voltage is given in Eq. (113). Then, shunt impedance, Q-value and frequency become

$$
\begin{aligned}
R &= \frac{V^2}{P_d} = \frac{1}{\pi}\,\kappa\,\delta\,Z_0^2\,\frac{(g/a)^2}{1+g/a}\,\frac{\sin kg/2}{kg/2}\,\frac{1}{J_1^2\left(j_{01}\right)} \\
k &= \omega_0/c_0\,, \\
Q_0 &= \frac{\omega w}{P_d} = \frac{1}{\delta}\,\frac{g}{1+g/a}\,, \\
f &= \frac{j_{01}}{2\pi}\,\frac{c}{a}\,.
\end{aligned}
\tag{119}
$$

For a 3 GHz copper cavity with a gap of half a wavelength the values are:

$$
\begin{array}{lll}
j_{01} = 2.405, & J_1(j_{01}) = 0.5191 & \kappa = 58\times10^6\ \Omega^{-1}\,\mathrm{m}^{-1}\,, \\
a = 3.828\ \mathrm{cm}, & g = 5\ \mathrm{cm}, & \delta = 1.207\ \mu\mathrm{m}\ ,
\end{array}
$$

$$
T = \frac{\sin kg/2}{kg/2} = \frac{2}{\pi} \qquad \text{transit time factor}\,,
$$

$$
R' = R/g = 100\ \mathrm{M\Omega/m}, \quad Q_0 = 17963, \quad R'/Q_0 = 6.1\ \mathrm{k\Omega/m}, \quad T_0 = \frac{2Q_0}{\omega} = 1.9\ \mu\mathrm{s}.
$$

4.5 Cavity shape perturbation

Inserting small metallic objects into a cavity or slightly deforming the shape of the cavity can be treated by the perturbation technique [4], [5].

Let us designate the fields and the resonant frequency of the unperturbed cavity by E_0, H_0, ω_0 and of the perturbed cavity by E, H, ω; then Maxwell's equations are

$$\nabla \times \mathbf{E_0} = -j\,\omega_0\,\mu\,\mathbf{H_0}\,, \quad \nabla \times \mathbf{H_0} = j\,\omega_0\,\varepsilon\,\mathbf{E_0}\,, \tag{120}$$

$$\nabla \times \mathbf{E} = -j\,\omega\,\mu\,\mathbf{H}\,, \quad \nabla \times \mathbf{H} = j\,\omega\,\varepsilon\,\mathbf{E}\,, \tag{121}$$

Manipulating Eqs. (120) and (121) we find

$$\mathbf{H}\cdot(\nabla\times\mathbf{E_0^*}) - \mathbf{E_0^*}\cdot(\nabla\times\mathbf{H}) = \nabla\cdot(\mathbf{E_0^*}\times\mathbf{H}) = j\,\omega_0\,\mu\,\mathbf{H}\cdot\mathbf{H_0^*} - j\,\omega\,\varepsilon\,\mathbf{E_0^*}\cdot\mathbf{E}\,,$$
$$\mathbf{H_0^*}\cdot(\nabla\times\mathbf{E}) - \mathbf{E}\cdot(\nabla\times\mathbf{H_0^*}) = \nabla\cdot(\mathbf{E}\times\mathbf{H_0^*}) = -j\,\omega\,\mu\,\mathbf{H_0^*}\cdot\mathbf{H} + j\,\omega_0\,\varepsilon\,\mathbf{E}\cdot\mathbf{E_0^*}\,,$$

and after adding both equations and integration over the volume V of the perturbed cavity

$$\int_V \nabla\cdot(\mathbf{E}\times\mathbf{H_0^*} + \mathbf{E_0^*}\times\mathbf{H})\,dV = \oint_S (\mathbf{E}\times\mathbf{H_0^*} + \mathbf{E_0^*}\times\mathbf{H})\,d\mathbf{S}$$
$$= \oint_S (\mathbf{E_0^*}\times\mathbf{H})\cdot d\mathbf{S} = -j(\omega-\omega_0)\int_V (\varepsilon\,\mathbf{E}\cdot\mathbf{E_0^*} + \mu\,\mathbf{H}\cdot\mathbf{H_0^*})\,dV. \tag{122}$$

In deriving Eq. (122) we used Gauss' theorem and the fact that $n \times E = 0$ on the surface S of the perturbed cavity. Referring to Fig. 19, we see that $S = S_0 - \Delta S$ and write for the left side of Eq. (122)

$$\oint_S (\mathbf{E_0^*}\times\mathbf{H})\,d\mathbf{S} = \oint_{S_0} (\mathbf{E_0^*}\times\mathbf{H})\,d\mathbf{S} - \oint_{\Delta S} (\mathbf{E_0^*}\times\mathbf{H})\,d\mathbf{S} = -\int_{\Delta S} (\mathbf{E_0^*}\times\mathbf{H})\,d\mathbf{S}\,,$$

because $n \times E_0 = 0$ on S_0. Substitution into Eq. (122) gives

$$\omega - \omega_0 = -j\,\frac{\oint_{\Delta S}(\mathbf{E_0^*}\times\mathbf{H})\,d\mathbf{S}}{\int_V (\varepsilon\,\mathbf{E}\cdot\mathbf{E_0^*} + \mu\,\mathbf{H}\cdot\mathbf{H_0^*})\,dV}\,. \tag{123}$$

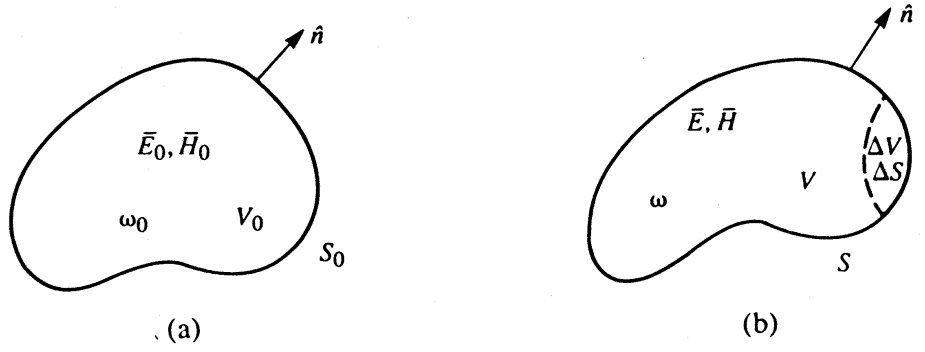


Fig. 19: Resonant cavity perturbed by a change in shape: (a) original cavity; (b) perturbed cavity

Equation (123) is the exact expression for the change in the resonant frequency. However, it is of little use since we do not know the quantities E, H of the perturbed cavity. But if the perturbation is small, E, H can be replaced by E_0, H_0 in the denominator of Eq. (123) because it is essentially the stored energy in the cavity and this will not change much. In the numerator we approximate H by H_0 and use Poynting's theorem,

$$\oint_{\Delta S} (\mathbf{E_0^*} \times \mathbf{H})\, d\mathbf{S} \approx \oint_{\Delta S} (\mathbf{E_0^*} \times \mathbf{H_0})\, d\mathbf{S} = -j\,\omega_0 \int_{\Delta V} (\varepsilon\,|\mathbf{E_0}|^2 - \mu\,|\mathbf{H_0}|^2)\, dV\,,$$

which finally gives for Eq. (123)

$$\frac{\omega - \omega_0}{\omega_0} \approx \frac{\int_{\Delta V} (\mu\,|\mathbf{H_0}|^2 - \varepsilon\,|\mathbf{E_0}|^2)\, dV}{\int_{V_0} (\mu\,|\mathbf{H_0}|^2 + \varepsilon\,|\mathbf{E_0}|^2)\, dV} = \frac{\Delta W_m - \Delta W_e}{W_m + W_e}\,. \tag{124}$$

The terms ΔW_m, ΔW_e are the changes in the stored magnetic and electric energy, respectively, and $W_m + W_e$ is the total stored energy. The result shows that the frequency may either increase or decrease depending on the location and the character of the perturbation.

The formula (124) was derived by pushing the cavity wall inwards by a small amount. It seems reasonable to suppose that introducing a small metallic object into the interior of the cavity should perturb the frequency in a similar way by an amount depending upon the local fields, and thus we could use the frequency shift to measure the field strength at an interior point. This is in fact the case. We might further suppose that we only have to perform the integration of the unperturbed fields over the volume of the perturbing object. This, however, is far from the case because the object perturbs the field in a way that is essential. In order to calculate the field perturbation we follow a procedure for a small metallic sphere as outlined in Ref. [6]. With the well-known electric field of a metallic sphere in a homogeneous electrostatic field the volume integral over the electric field was performed when changing the sphere radius from r_0 to $r_0 + dr_0$. For the total perturbation caused by the sphere of radius r_0 the resulting expression was integrated from zero to r_0. In an analogous manner the volume integral over the magnetic field was carried out. As a result form factors for the volume integrals in the numerator of Eq. (124) were found:

$$f_e = 3/2, \quad f_m = 3/4\,.$$

In general, these form factors depend on the shape and orientation and material of the perturbing object. For some geometries, like ellipsoids, they are calculated [6]; for other more complicated geometries they can be determined experimentally [7].

5 MEASUREMENTS

Measurement techniques are a vast and complicated area. Here, I present a few basic techniques directly related to the subjects treated in the previous section.

5.1 Line mismatch

An old-fashioned but instructive way to measure a line mismatch is with a slotted line, Fig. 20. A movable capacitive probe measures the voltage standing wave ratio, Eq. (33), along the mismatched line. This yields the magnitude of the reflection coefficient. We further know from Section 2.3 that the first voltage minimum occurs at a distance ζ_{min} from the load

$$2\beta\,\zeta_{\text{min}} = \vartheta - \pi$$

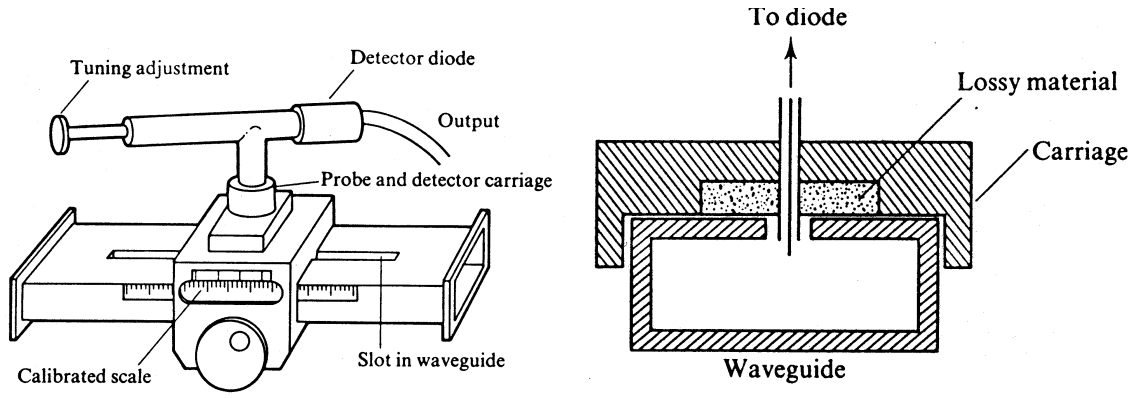


Fig. 20: Standing wave detector: left: complete device with slotted rectangular waveguide; right: probe in waveguide

which fixes the phase ϑ of the reflection coefficient.

An alternative solution for measuring the magnitude of the reflection coefficient is shown in Fig. 21. A fraction of the forward- and backward-travelling power is coupled out by two directional couplers, giving the measurements shown.

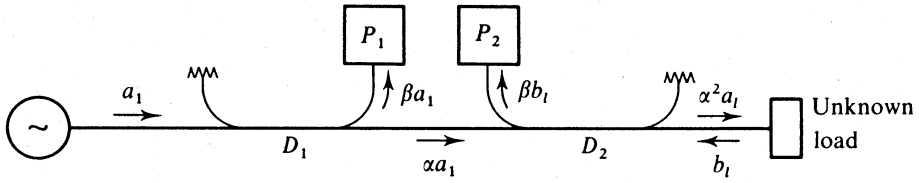


Fig. 21: Measurement of the ratio of backward (P_2) to forward (P_1) travelling power with two directional couplers D_1, D_2.

5.2 Q-value and coupling factor of a cavity [8]

Let us consider a cavity coupled to a signal generator via a piece of a transmission line and a coupling device. The source has an internal impedance Z_0 equal to the line impedance. The equivalent circuits are shown in Fig. 22. L_1 represents the self-inductance of the coupling device and M the mutual inductance between it and the cavity inductance L. The terminal plane of the coupling device is presumed to be located at some arbitrary position a–a near the cavity. The coupling device is assumed to be lossfree. The circuit can be simplified further as shown on the right of Fig. 22.

The normalized impedance at the terminal plane a–a is then

$$\begin{aligned} \frac{Z_{aa}}{Z_0} &= j\,\frac{X_1}{Z_0} + \frac{1}{Z_0}\,\frac{(\omega M)^2}{R_s + j\,(\omega L1/\omega C)} \\ &= j\,\frac{X_1}{Z_0} + \frac{\beta_1}{1 + j\,(\omega L/R_s)[1-(\omega_0/\omega)^2]} \approx j\,\frac{X_1}{Z_0} + \frac{\beta_1}{1 + j2Q_0\delta}\,, \end{aligned} \tag{125}$$

where $X_1 = \omega L_1$, $\beta_1 = (\omega M)^2/Z_0 R_s$, $\delta = (\omega - \omega_0)/\omega$.

The analysis can be simplified by shifting the reference plane to a location where the term with X_1 vanishes. Such a location is called the **detuned-short position** because a short circuit will appear if the resonator impedance is far off-resonance. We find the detuned-short position by transforming the impedance Z_{aa} with a piece of line of length l. From Eq. (35) follows

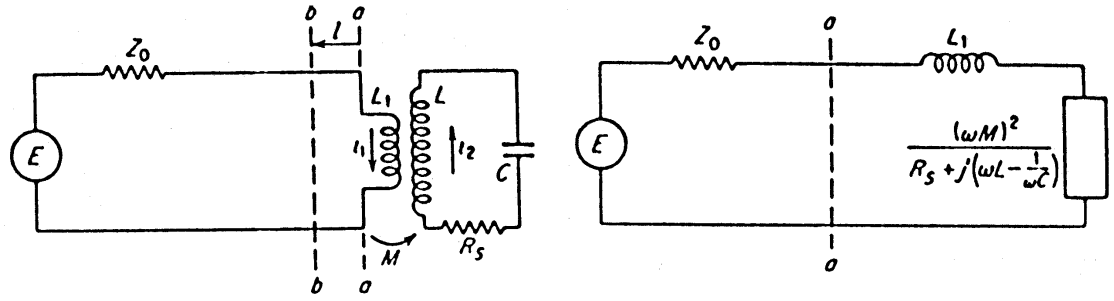


Fig. 22: Cavity coupled to a generator through a line; left: equivalent circuit; right: equivalent circuit with the impedances referred to the primary

$$\frac{Z_{bb}}{Z_0} = \frac{Z_{aa} + j\, Z_0 \tan\beta l}{Z_0 + j\, Z_{aa} \tan\beta l}\,, \tag{126}$$

which is zero at

$$\tan\beta l = j\,\frac{Z_{aa}}{Z_0} = -\frac{X_1}{Z_0} \quad \text{for} \quad \frac{Z_{aa}}{Z_0} = j\,\frac{X_1}{Z_0}\,. \tag{127}$$

Substituting Eqs. (125) and (127) into (126) gives the simple expression for the impedance at the detuned-short position:

$$\frac{Z_{bb}}{Z_0} = \frac{\beta}{1 + j2Q_0(\delta - \delta_0)}\,, \tag{128}$$

with

$$\beta = \frac{\beta_1}{1 + (X_1/Z_0)^2}\,, \qquad \delta_0 = \frac{\beta}{2Q_0}\,\frac{X_1}{Z_0}\,, \qquad \beta_1 = \frac{(\omega M)^2}{Z_0 R_s}\,.$$

Equation (128) represents the impedance of a parallel resonant circuit with a resonant impedance βZ_0. In the Smith chart, Fig. 23, it is a circle with the centre located on the real axis and touching the point $r = R/Z_0 = 0$. At resonance the circle cuts the real axis at the point $r_0 = \beta$. This determines the coupling coefficient. If

$$r_0 = \beta \begin{cases} < 1 \\ = 1 \\ > 1 \end{cases} \qquad \text{the resonator is} \begin{cases} \text{undercoupled} \\ \text{matched} \\ \text{overcoupled} \end{cases} \qquad \text{with} \quad Q_{ext} \begin{cases} > Q_0 \\ = Q_0\,. \\ < Q_0 \end{cases}$$

At certain frequencies the imaginary part of the denominator becomes equal to ± 1, then

$$\frac{Z_{bb}}{Z_0} = \frac{\beta}{1 \pm j}\,,$$

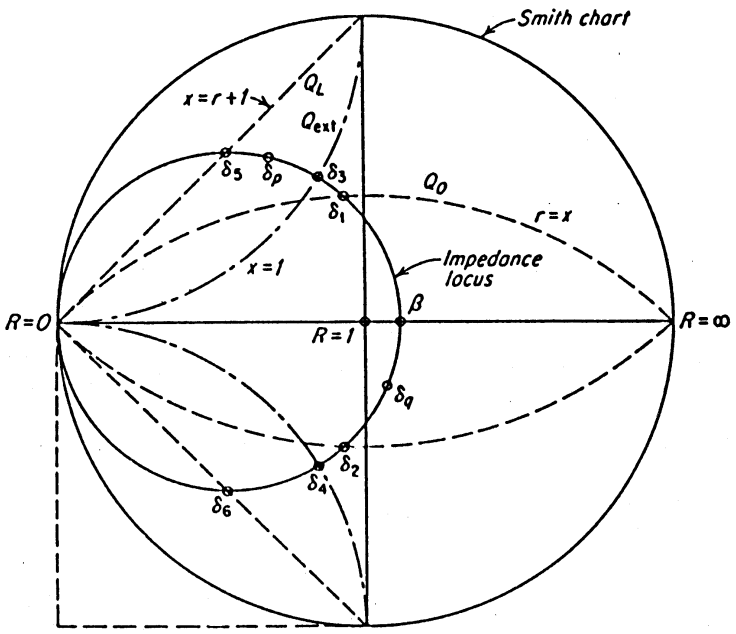


Fig. 23: Identification of the half-power points from the Smith chart. The Q_0 locus is given by $x = r$; Q_l by $x = 1 + r$; Q_{ext} by $x = 1$.

and the real part of Z_{bb} equals the imaginary part. The locus of these points is the dashed segments $(r = x)$ in Fig. 23. They cut the impedance circle at frequencies δ_1 and δ_2 and determine the unloaded Q:

$$\begin{aligned} 2Q_0\,(\delta_1 - \delta_0) &= 1 \\ 2Q_0\,(\delta_2 - \delta_0) &= -1 \end{aligned} \quad \Rightarrow \qquad Q_0 = \frac{1}{\delta_1 - \delta_2}\,. \tag{129}$$

The coupling coefficient β is also the ratio of the power radiated from the resonator into the external circuit to the power dissipated in the resonator

$$\beta = \frac{P_{ext}}{P_d} = \frac{Q_0}{Q_{ext}}\,. \tag{130}$$

Eq. (130) together with Eq. (93) gives for the loaded Q

$$\frac{1}{Q_l} = \frac{1}{Q_0} + \frac{1}{Q_{ext}} = \frac{1}{Q_0} + \frac{1}{Q_0}\,\frac{Q_0}{Q_{ext}}$$

or

$$Q_l = \frac{Q_0}{1+\beta}\,, \qquad Q_{ext} = \frac{Q_0}{\beta}\,. \tag{131}$$

We write Eq. (128) in terms of Q_l and Q_{ext},

$$\frac{Z_{bb}}{Z_0} = \frac{\beta}{1 + j2Q_l(1+\beta)(\delta - \delta_0)} = \frac{\beta}{1 + j2Q_{ext}\beta(\delta - \delta_0)}\,,$$

and see that for frequencies δ_5, δ_6 at which

$$2Q_l(\delta - \delta_0) = \pm 1 \qquad \text{or} \qquad Q_l = 1/(\delta_5 - \delta_6) \tag{132}$$

the impedance is

$$\frac{Z_{bb}}{Z_0} = \frac{\beta}{1 \pm j(1+\beta)} + \frac{1}{1/\beta \pm j(1+1/\beta)} .$$

The locus of these points is $x = 1 + r$ and is given by straight lines in Fig. 23. They cut the impedance circle at frequencies δ_5 and δ_6 and determine the loaded Q_l. Finally, for frequencies δ_3, δ_4 at which

$$2Q_{ext}(\delta - \delta_0) = \pm 1 \qquad \text{or} \qquad Q_{ext} = 1/(\delta_3 - \delta_4) \ ,$$

the impedance is

$$\frac{Z_{bb}}{Z_0} = \frac{\beta}{1 \pm j\,\beta} = \frac{1}{1/\beta \pm j} .$$

with the locus of the points given by $x = 1$ (dot–dashed line in Fig. 23). The intersections with the impedance circle determine δ_3, δ_4 and thus Q_{ext}.

5.3 R upon Q of a cavity [5], [9]

The measurement of the resonant frequency and Q-value of a cavity is relatively easy. More difficult is the measurement of the shunt impedance. We use the definition of the R upon Q, Eq. (110), for a standing wave cavity where the electric field is proportional to $\exp(j\omega t)$. Then, Eq. (110), together with Eq. (112), is equivalent to

$$\frac{R}{Q} = \frac{1}{\omega_0 W} \left| \int E_z(z)\, e^{j\omega t}\, \mathrm{d}z \right|^2 = \frac{1}{\omega_0 W} \left| \int E_z(z)\, e^{jkt}\, \mathrm{d}z \right|^2 .$$

for a relativistic, $\nu \approx c$, $k = \omega/c$, particle. This can be written as

$$\frac{R}{Q} = \frac{1}{2\pi f_0 W} \left\{ \left(\int E_z \cos kz\, dz \right)^2 + \left(\int E_z \sin kz\, dz \right)^2 \right\} . \tag{133}$$

Next, we make use of the perturbation formula Eq. (124) where we choose a trajectory with zero magnetic field, $H_0 = 0$, and where we write the integral over the perturbed electric field with the aid of a form factor K_e:

$$\frac{\Delta\omega}{\omega_0} = \frac{1}{4W} K_e\, \varepsilon\, E_0^2 . \tag{134}$$

The term K_e depends on the material, size, and shape of the perturbing object (bead). Since E_0 is a function of the position z, $\Delta\omega = \Delta\omega(z)$. Identifying E_0 in Eq. (134) as being the accelerating field E_z in Eq. (133), we get after substitution

$$\frac{R}{Q} = \frac{2}{\pi f_0 K_e \varepsilon} \left\{ \left(\int \sqrt{\frac{\Delta\omega(z)}{\omega_0}} \cos kz \, dz \right)^2 + \left(\int \sqrt{\frac{\Delta\omega}{\omega_0}} \sin kz \, dz \right)^2 \right\} . \tag{135}$$

Therefore the problem is reduced to the measurement of the frequency shifts $\Delta\omega/\omega$ as a function of the position z of the perturbing object. Instead of the frequency shift it is easier to measure the phase shift of the signal which is related to the frequency shift by (see Eq. (99))

$$\tan\varphi = 2Q_l \frac{\Delta\omega}{\omega_0} . \tag{136}$$

A set-up to measure the phase shift with a network analyser is shown in Fig. 24. The perturbing bead is pulled through the cavity by a stepping motor while the network analyser continuously measures the phase.

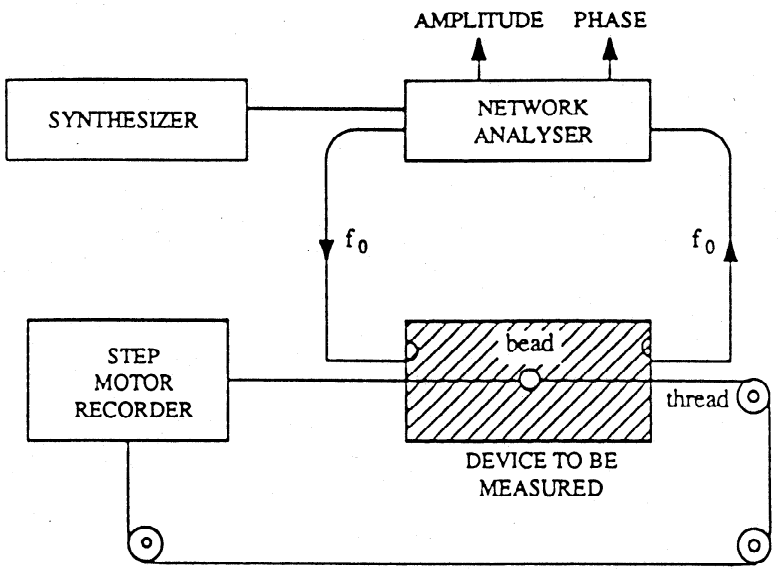


Fig. 24: Bead-pull measurement set-up for measuring the R upon Q of a cavity

REFERENCES

[1] J.C. Maxwell, *A Treatise on Electricity and Magnetism* (Dover, New York, 1954).

[2] Radiation Laboratory Series, Massachusetts Institute of Technology, McGraw-Hill, New York: CD-ROM published by Artech House, Norwood, 1999.

[3] S. Ramo, J.R. Whinnery and Th. van Duzer, *Fields and Waves in Communication Electronics* (John Wiley & Sons, New York, 1984).

[4] D.M. Pozar, *Microwave Engineering* (Addison-Wesley, Reading, MA, 1993).

[5] CERN Accelerator School, RF engineering for particle accelerators, Oxford, UK, 1991, CERN 92–03, Vol. I, 1992.

[6] L.C. Maier and J.C. Slater, *J. Appl. Phys.* **23** (1952) 68.

[7] H. Hahn and H.J. Halama, *IEEE Trans. Microwave Theory Tech.* **MTT-16** (1968) 20.

[8] E.L. Ginzton, *Microwave Measurements* (McGraw-Hill, New York, 1957).

[9] E. Karantzoulis, An overview on impedances and impedance measuring methods for accelerators. Internal report ST/M-91/1, Synchrotrone Trieste, 1991.

IMPROVEMENTS IN CAVITY CONSTRUCTION TECHNIQUES

Walter Wuensch
CERN, Geneva, Switzerland

Abstract
Two machining techniques relevant to the construction of accelerator RF cavities, optical-quality diamond turning and Electrical Discharge Machining (EDM), have undergone extensive evolution in recent years. The techniques are described and examples of their use in RF structures are presented.

1 INTRODUCTION

A comprehensive review of fabrication techniques relevant to the construction of RF cavities has already been published [1] and this report represents an appendix to it. Because the majority of the techniques described in the original publication have not changed much in recent years, they are not repeated here. This report rather concentrates on two machining techniques, optical-quality diamond turning and electrical discharge machining, where both the techniques and their application to the construction of accelerator RF cavities have evolved substantially over the last decade.

2 OPTICAL-QUALITY DIAMOND TURNING

Diamond tools have been used for some time in the fabrication of accelerator RF cavities. An example is the 3 GHz travelling wave accelerating structures of the LEP injector linac at CERN [2]. Using a standard lathe and very careful tool adjustment, 0.01 mm precision and an N3 (Ra < 0.1 μm) surface finish can be obtained in OFHC copper. The primary objective of such diamond turning has usually been to obtain as high an RF quality factor, Q, as possible.

Taking advantage of a relatively new and rapidly developing technology, RF cavities can now be fabricated to much tighter tolerances and with a greatly improved surface finish (see Figs. 1 and 2). This is possible on specialized sub-micron precision lathes using single point cutting with natural- diamond tools. The combination gives an improvement over 'classical' diamond turning of a soft material like copper of an order of magnitude in tolerance, to $\pm$ 1 μm, and over an order of magnitude in surface finish, to better than N1 ($R_a < 25$ nm). These advantages are particularly important for high-frequency accelerator applications, where the tolerances are critical for transverse beam stability and a high Q is difficult to achieve because of the high frequency and where precision assembly requires diffusion bonding techniques.

The original motivation for the development of the precision lathes that give such a performance was the direct turning of aspherical optical components and an important early application was germanium infrared lenses for military applications. The range of materials that can be diamond turned has increased and includes non-ferrous metals, crystals and polymers. Servo-controlled tool movement linked to the rotation of the lathe spindle now also allow for the creation of 'free-form' optics—simultaneously aspheric and acylindrical forms.

Application of the technology is now rapidly spreading and includes pick-up lenses for DVD and CD players, lenses for fax machines and photocopiers, infrared lenses and mirrors, vision systems for defense applications, semiconductor lithographic equipment, fibre optics and data transmission systems and switches, scanner mirrors, X-ray telescope mirrors, hard disk drives, automotive fuel delivery systems, contact lenses, lens moulds and intraocular lenses.

A number of these applications now even require sub-micron tolerances, albeit for harder materials than copper. For the machining of copper, the ubiquitous RF material, these developments at least provide some margin in the capability of the machines themselves.

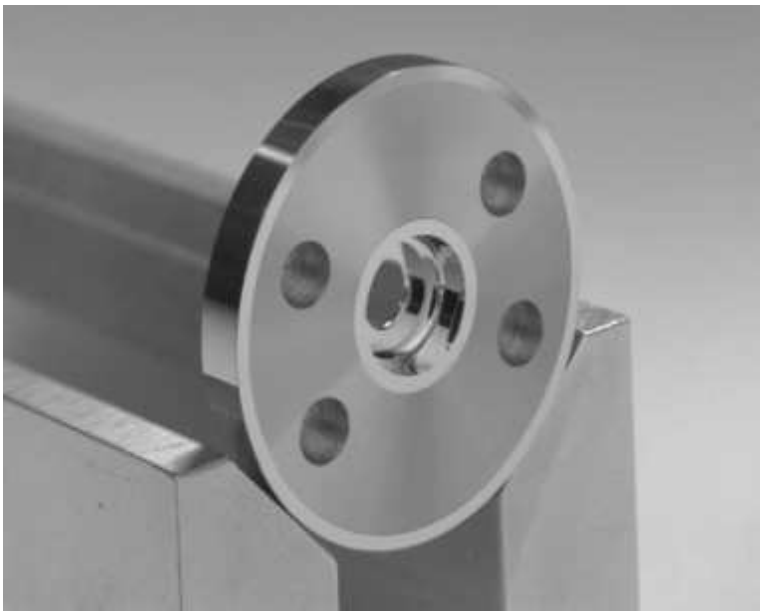


Fig. 1: Prototype diamond-turned disk of a CLIC 30 GHz accelerating structure. The outside diameter is 35 mm. The RF structure is in the 8.7 mm diameter cell and 4 mm diameter loading iris, which are the shiny features at the centre of the disk. The outer features are for bonding the disks together.

Fig. 2: Prototype diamond-turned disk of an NLC X-band accelerating structure. The four holes coupled to the accelerating cells are used for transverse wakefield suppression.

2.1 A description of the technology and the machines

There are some common characteristics among the specially designed lathes that are necessary for micron tolerances and optical (Ra values of less than 10 nm) surface finishes. The support structures of such lathes are very stiff, have low coefficients of thermal expansion and are dimensionally stable. Supports are often made from natural or synthetic granite and are vibration isolated from the ground. The individual temperature components of the machine are very tightly controlled—this is accomplished through

multiple temperature control loop systems. Hydrostatic bearings, usually either air or oil, are used for the spindle of the lathes. The rotational accuracy is in the nanometer range. Tool motion is controlled using laser interferometer feedback loops. The control loops are designed to give a very high 'stiffness'. Tool positioning resolutions are of the order of 1 nm.

In order to obtain micron tolerances, parts to be machined must be fixed to the lathe's spindle in such a way as to induce little stress. Typically a vacuum chuck is used. During the set-up of the machine the vacuum chuck is turned using a diamond tool to ensure that it is flat and to provide a precise reference surface.

Rough-machined blanks to be diamond turned are typically over-dimensioned by a few hundredths of a millimetre. Care must be taken when machining the blanks to induce a minimum of stress. Blanks are first diamond turned on a face so that the face can be fixed to the vacuum chuck with minimum stress. The process of making a flat reference surface may have to be repeated a few times as induced stresses are released. Diamond cutting forces are of the order of a few grams. Parts which can not be vacuum chucked because the flat surfaces are not large enough to provide enough chucking force, or parts with many holes can be glued to a backing plate. Gluing is not desirable for RF components, but if it is unavoidable, the glue should be cleaned from the piece as soon as possible. Another important issue to consider when designing parts for highest precision is that as many critical features be machined as possible during a single chucking. This is especially important for maintaining concentricity between internal RF features and external mechanical reference surfaces.

2.2 Examples of application in accelerator RF systems

One of the primary applications of diamond turning to accelerating RF cavities has been in the linear collider studies at CERN [3], KEK [4] and SLAC [5]. All of the studies require travelling wave accelerating structures with tolerances in the 1–10 μm range and as a consequence have pursued optical quality diamond machining. Typical parts are 30–100 mm in diameter, with minimum turned inside diameters of 4 mm. In addition to producing prototypes, the studies have shown that for mass production diamond machining can be implemented at a reasonable cost.

3 ELECTRICAL DISCHARGE MACHINING

EDM is a process in which material is removed from a workpiece by electrical discharges. The technique works for any material with electrical conductivity and includes metals, alloys, carbides, and graphite—therefore almost all of the materials used in accelerator RF cavities. There are two distinct classes of EDM: die sinking and wire cutting. In die sinking, the form of the machining reproduces the shape of an electrode. In wire cutting a wire electrode cuts a programmed contour in the workpiece. The advantages of EDM over more classical types of machining include:

- Its capacity to make complex shapes, especially those with small internal features. For example, the internal radii of a milled rectangular hole are limited to the radius of the mill. A much smaller radius, 0.1 mm, is possible with wire-cut EDM.
- The absence of cutting forces allows simplified jigging and the possibility to machine very delicate structures.
- The absence of cutting forces gives the potential for very tight tolerances.
- Certain materials such as carbides and titanium are difficult to machine classically but present no difficulty for EDM. Electrodes can be made from a material that is easy to machine.
- Simplified assembly. Since EDM can form complex inside shapes, parts can often be machined from a single block of material. To make the same inside shape with classical machining may require machining into a few blocks of material and then bonding.

The main disadvantages of EDM are that it is relatively slow (and consequently expensive) and that it is not easy to produce a good surface finish (this is even slower). For these reasons EDM is not

competitive with turning for producing the main inside RF surfaces of large and/or circularly symmetric cavities. It can, however, prove a very competitive technique for the production of auxiliary features such as power couplers, damping waveguides, cooling channels. It is, for example, ideal for producing rectangular waveguide holes inside solid material. EDM is also interesting for very high frequency (30 GHz and above) accelerating structures which have features smaller than a millimetre. EDM may also be applicable to cavities of very complex geometry such as planar and muffin-tin structures.

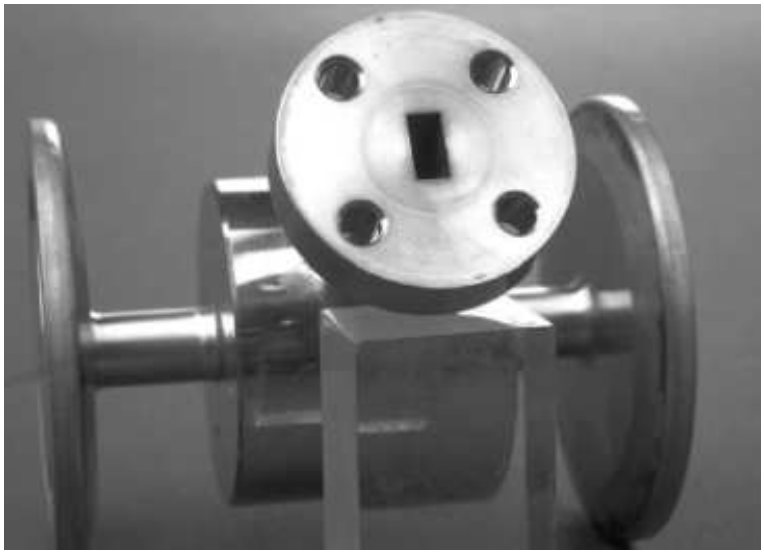


Fig. 3: Close-up view of a 30 GHz test cavity with a wire-cut output waveguide

3.1 A description of the technology

EDM removes material by creating a succession of sparks between an electrode and a workpiece via a pulsed DC voltage. Each spark makes a microscopic crater in the workpiece. The electrode and the workpiece are immersed in a dielectric bath, either water or mineral oil. Because of the submersion in a liquid bath, RF parts made with electrical discharge machining require an especially thorough cleaning. With typical bias voltages of X to Y volts and gaps of a few microns to a millimetre, sparks with peak temperatures of 8000–12 000°C are produced. The liquid dielectric is used to flush away the metallic particles produced by the sparks. Lower applied gradients produce a better surface finish but at the expense of a lower machining speed.

Die sinking electrodes are typically made from copper, graphite, or tungsten copper and wire electrodes from copper or brass. Wire electrodes have typical diameters of 0.2 to 0.1 mm. Electrode wires can be shifted longitudinally during machining, renewing the part of the wire doing the cutting, thus eliminating any effects of tool wear. This allows a very high degree of reproducibility over may parts. Die sinking electrodes are eroded somewhat during machining, but the effect is usually much less than the tool wear of mills.

3.2 Examples of application in accelerator RF systems

An example of wire cutting used in an RF application is shown in Fig. 3. It shows a 30 GHz high gradient test cavity with its output waveguide and RF and vacuum flange facing the viewer. The dimensions of the WR-28 standard waveguide are 3.56×7.11 mm. The waveguide and flange have been formed from a single block of stainless steel and the waveguide itself has been formed by wire cutting. The cut was made with a 0.2 mm diameter wire, which gives only 0.1 mm diameter radii in the waveguide corners.

This allows connection with very small impedance mismatch directly to components made from extruded waveguide tube. Because the piece was made from a single block of material, no intermediate brazing steps were required.

REFERENCES

[1] I. Wilson, Cavity Construction Techniques, Proceedings of the CERN Accelerator School of RF Engineering for Particle Accelerators, Oxford, 1991, CERN 92-03, vol. 2.

[2] D. Blechschmidt and D.J. Warner, Parameters of the LEP Injector Linacs, CERN/PS/88-07 (LPI).

[3] The CLIC Study Team, A 3 TeV e^+e^- Linear Collider Based on CLIC Technology, CERN 2000-008 and http://cern.web.cern.ch/CERN/Divisions/PS/Projects/CLIC.html .

[4] http://www-jlc.kek.jp/ .

[5] NLC Design Report, SLAC-Report 474 and
http://www-project.slac.stanford.edu/lc/nlc.html .

REVIEW OF RF POWER SOURCES FOR PARTICLE ACCELERATORS

R.G. Carter
Lancaster University, Lancaster, UK

Abstract

This paper reviews the main types of RF power amplifiers used for particle accelerators. It covers solid-state devices, tetrodes, inductive output tubes, klystrons, and gyrotrons with power outputs greater than 10 kW CW or 100 kW pulsed at frequencies from 50 MHz to 50 GHz. Factors affecting the satisfactory operation of amplifiers, including cooling, matching, and protection circuits, are discussed. The paper concludes with a summary of the state of the art for the different technologies.

1 INTRODUCTION

Many particle accelerators use high-power RF sources [1]. These sources must usually be amplifiers in order to achieve sufficient frequency and phase stability. The frequencies employed range from about 50 MHz to 50 GHz or higher. Power requirements range from 10 kW to 2 MW or more for continuous sources and up to 150 MW for pulsed sources. When sufficient power cannot be obtained from a single amplifier then the output from several amplifiers may be combined. In some cases power is supplied to a number of accelerating cavities from separate power amplifiers fed from a single low-power source. Other important factors include power conversion efficiency reliability and, in some cases, bandwidth.

Several types of power amplifier exist. The two main categories are solid-state devices and vacuum tubes. The former can only yield quite low powers and large numbers must be operated in parallel to reach even the lowest power levels required for accelerators. The vacuum tubes most commonly used to power accelerators are tetrodes and klystrons. Newer devices including the diacrode, the Inductive Output Tube (IOT), the Multiple-Beam Klystron (MBK), and the gyrotron may find applications in future. This paper sets out to provide a review of these types of amplifier together with a discussion of some of the factors affecting their successful operation.

2 SOLID-STATE AMPLIFIERS

Solid-state power amplifiers for use at high frequencies employ silicon bipolar transistors, silicon MOSFETs or silicon carbide Static Induction Transistors (SIT). An indication of the capabilities of UHF and VHF power amplifiers can be obtained by studying the state of the art of television transmitters at these frequencies [2]. The amplifiers used for accelerators are sometimes commercially available television transmitters or derived from them. The power required is obtained by operating numerous transistors in parallel. Reference [3] describes a 30 kW UHF television transmitter comprising six 6 kW power amplifiers each made up of eight 800 W modules. Each module in turn has twelve 110 W bipolar power output transistors, making a total of 576 transistors. A further seven power transistors of the same type are used in the driver stages of each module. Table 1 shows the state of the art of the principal types of transistor used in the UHF band.

Table 1: Solid-state UHF transistors: state of the art

	Si bipolar	**Si MOSFET**	**SiC SIT**
Peak power (W)	175	100	280
Mean power (W)	40	45	70
Operating voltage (V)	26 to 32	28 to 50	65
UHF gain (dB)	8 to 11	13 to 16	9 to 10
Maximum junction temperature (°C)	150	150	225

It is claimed that solid-state amplifiers are more reliable than their vacuum-tube counterparts by a factor of as much as 2.5 [3]. Because many transistors are operated in parallel, the failure of one produces a negligible drop in power output. However, the transistors are operated close to their design limits and they are known to be vulnerable to accidental overloads. Appreciable power is dissipated in the combining circuits and provision may be made to isolate a device that fails so that it does not adversely affect the performance of others nearby. For this reason further increases in the power output of solid-state amplifiers must be achieved by improvements in the transistors rather than by increasing the number used. Other advantages claimed for solid-state amplifiers are their high stability, low maintenance requirements, absence of warm-up time and low voltage operation. The supply voltage is low enough to avoid the safety problems associated with the HT supplies of tube amplifiers. The penalty which must be paid for this is the need to supply and handle very large d.c. currents. The transmitter described above requires a total current of the order of 2000 A. Such currents require large bus-bars in which there will be appreciable ohmic losses. It is not easy to deduce information about the conversion efficiencies of solid-state television transmitters from published information. This is because the figures quoted are calculated for a standard TV test signal rather than for continuous high-power operation. The transistors are operated in class AB (see Section 3.2) and it appears that the overall efficiencies of the UHF amplifiers are around 40%.

3 TETRODE AMPLIFIERS

Tetrode vacuum tubes are well established as high-power RF sources in the VHF (30–300 MHz) band. Figure 1(a) shows a schematic diagram of a tetrode. The voltages of the anode (known as the 'plate' in the USA) and the screen grid are positive with respect to the cathode whereas that of the control grid is normally negative. The current of electrons emitted from the cathode surface is controlled by the field of the control grid modified by those of the other two electrodes. The screen grid, maintained at RF ground, prevents capacitive feedback from the anode to the control grid. Figure 1(b) shows the characteristic curves of a typical tetrode. It is clear that the anode current depends strongly on the control grid voltage in a non-linear fashion. This voltage is normally negative with respect to the cathode to prevent electrons being collected on the grid with consequent problems of heat dissipation. If the anode voltage falls below that of the screen grid then any secondary electrons liberated from the anode are collected by the screen grid. For this reason the characteristic curves show kinks in that region. When tetrodes are operated as power amplifiers the anode voltage is always greater than the screen grid voltage. The curves show that the anode current varies very slowly with changes in the anode voltage.

The arrangement of a 150 kW, 30 MHz tetrode is shown in Fig. 2 [4]. The construction is coaxial with the cathode inside and the anode outside. This arrangement is necessary to enable the anode to be cooled (see Section 8.1). The output power available from such a tube is limited by the maximum current density available from the cathode and by the maximum power density that can be dissipated by the anode. The length of the anode must be much less than the free-space wavelength of the signal to be amplified to avoid variations in the signal level along it. The perimeter of the anode must likewise be much less than the free-space wavelength to avoid azimuthal higher-order modes in the space between the anode and the screen grid. The spacings between the electrodes must be small enough for the transit time of an electron from the cathode to the anode to be much less than the RF period. If attempts are

made to reduce the transit time by raising the anode voltage then there may be flashover between the electrodes. Reference [5] is the standard text on gridded tubes.

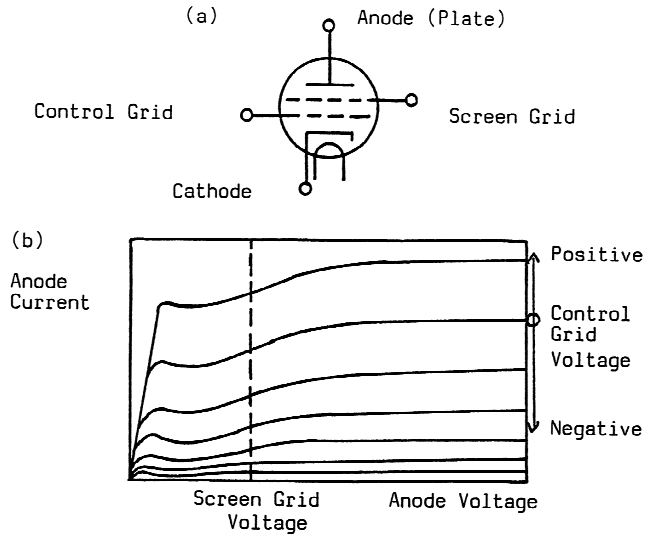


Fig. 1: Tetrode: (a) schematic diagram and (b) characteristic curves

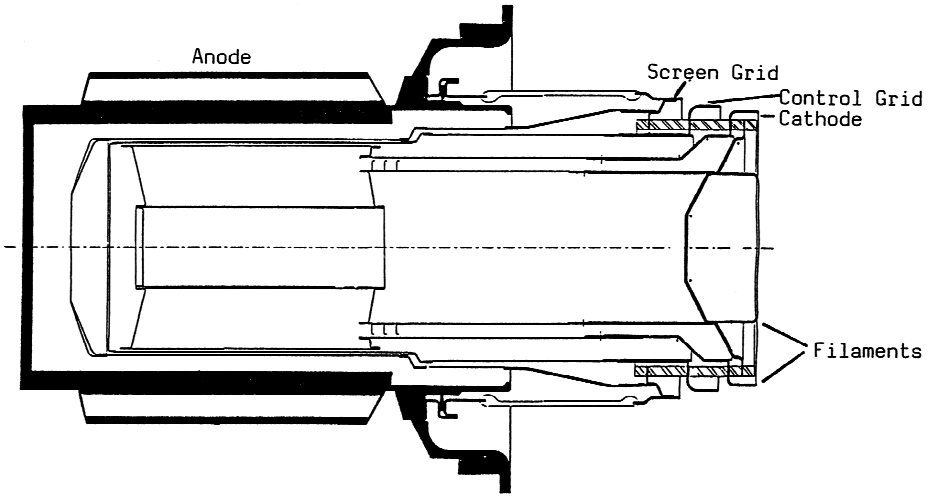


Fig. 2: Sectional view of a high-power tetrode (courtesy of Marconi Applied Technologies)

3.1 Tetrode amplifier circuits

Figure 3 shows the circuit of a grounded cathode tetrode amplifier with a tuned anode circuit. At lower frequencies a resistive anode load is used but this is unsatisfactory in the VHF band because of the effects of parasitic capacitance. The amplifiers used in accelerators are operated at a single frequency at any one time so the limited bandwidth of the tuned anode circuit is not a problem. At the resonant frequency the load in the anode circuit comprises the shunt resistance of the resonator (R_s) and the load resistance

(R_L) in parallel. If the load impedance has a reactive component then it merely detunes the resonator and can easily be compensated for. The d.c. electrode potentials are maintained by the power supplies shown and the capacitors provide d.c. blocking and RF bypass.

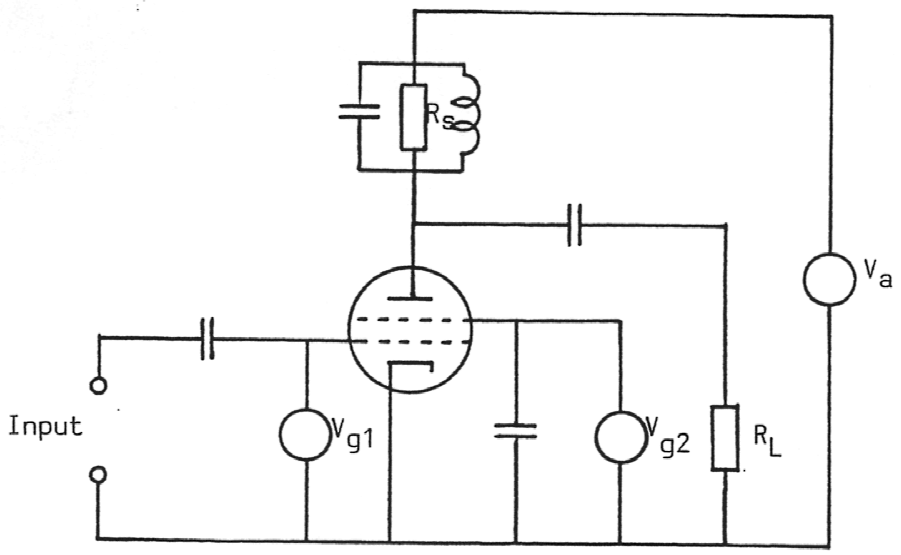


Fig. 3: Tetrode amplifier with grounded cathode

A commonly used alternative circuit is the grounded control grid circuit shown in Fig. 4. This circuit is not as simple to analyse as the previous one but it is easier to construct using tetrodes with the coaxial arrangement of electrodes shown in Fig. 2. Because both grids are at RF earth there is better isolation between the input and the output circuits. The other important difference is that in this circuit the full anode current flows in the input circuit, with the result that the input impedance and the gain are both lower than for the grounded cathode circuit.

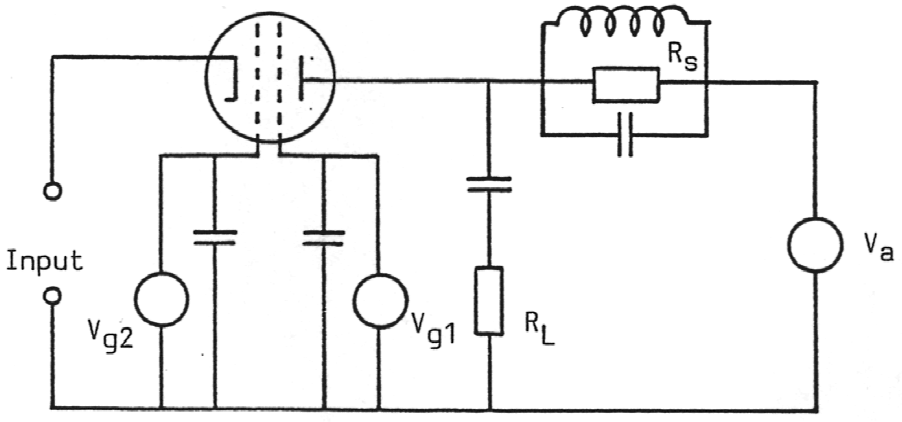


Fig. 4: Tetrode amplifier with grounded control grid

3.2 Classes of operation

Tuned amplifiers are normally operated in one of three modes, designated classes A, B, and C.

3.2.1 Class A

Figure 5(a) shows the form of the characteristic curves of a tetrode most commonly used for circuit design. The control grid voltage is plotted against the anode voltage with curves of constant anode, control grid, and screen grid currents shown. Because the anode circuit is tuned to the signal frequency the voltages on the two axes vary sinusoidally in antiphase with each other. The operation of the tube throughout the RF cycle is therefore represented by points on the straight load-line shown. In class A operation the tube conducts throughout the RF cycle, as shown in Fig. 5(b). The anode current curve is not quite a pure sinusoid because of the non-linearity of the tube characteristics. The ends of the load line are chosen so that the anode current is greater than or equal to zero at the maximum anode voltage and the control grid voltage is zero or slightly positive at the minimum anode voltage. The tube must then be biased so that in the absence of RF drive it sits at the quiescent ('Q') point shown at the centre of the load line.

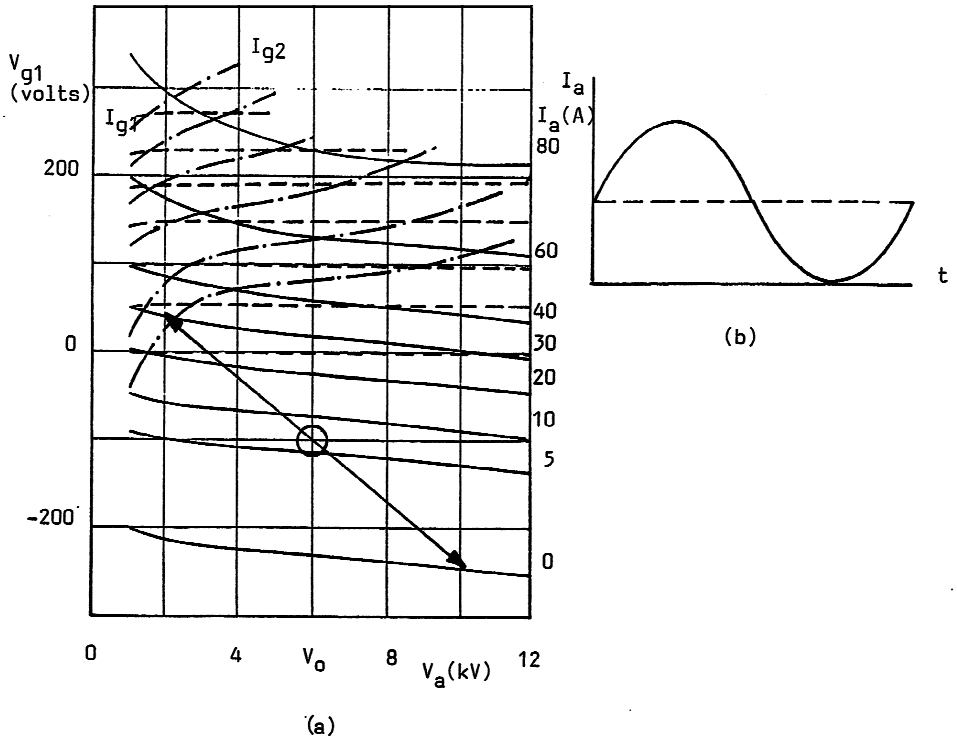


Fig. 5: Class A operation

This class of operation gives an output signal with low harmonic content, provided that the anode resonator does not have any higher-order modes that coincide with the harmonics of the signal frequency. The efficiency can be estimated easily if it is assumed that the amplifier is in linear operation with the anode current swinging from zero to I_{pk}. The mean anode current is then

$$I_0 = 0.5\, I_{\mathrm{pk}} \tag{1}$$

and the amplitude of the RF current is

$$I_1 = 0.5\, I_{\rm pk} \ . \tag{2}$$

If the minimum anode voltage is close to zero then the amplitude of the RF voltage is given by

$$V_1 = V_0 \ , \tag{3}$$

where V_0 is the anode voltage at the Q point. The d.c. power input to the tube is then

$$P_{\rm dc} = V_0\, I_0 = 0.5\, V_0\, I_{\rm pk} \tag{4}$$

and the RF power output is

$$P_{\rm rf} = 0.5\, V_1\, I_1 = 0.25\, V_0\, I_{\rm pk} \ . \tag{5}$$

From Eqs. (4) and (5) the efficiency is

$$\eta = P_{\rm rf}/P_{\rm dc} = 50\% \ . \tag{6}$$

The difference between the d.c. input and RF output powers is dissipated as heat in the anode. A more exact analysis, taking into account the non-linearity of the tube and the actual values of the voltages and currents, gives a lower figure for the efficiency. The calculation above therefore gives an upper limit for class A operation. If the tube is operated in class A then power is drawn from the supply and dissipated in the anode even when there is no RF drive and the tube is sitting at the Q point.

3.2.2 Class B

In class B operation the tube is biased so that the Q point lies on the curve of zero anode current. It then conducts only during the positive half-cycle of the control grid voltage. This is illustrated in Fig. 6. During the negative half-cycle of the control grid voltage the anode voltage swings to a voltage that is nearly twice that at the Q point because of the resonant output circuit. The anode current waveform has a higher harmonic content than in class A operation.

The estimation of the efficiency of the amplifier follows the same pattern as in the previous Section. Assuming that the tube is linear while it is conducting, the mean and RF currents can be found by Fourier analysis of the waveform in Fig. 6(b):

$$I_0 = I_{\rm pk}/\pi \tag{7}$$

and

$$I_1 = 0.5\, I_{\rm pk} \ , \tag{8}$$

while

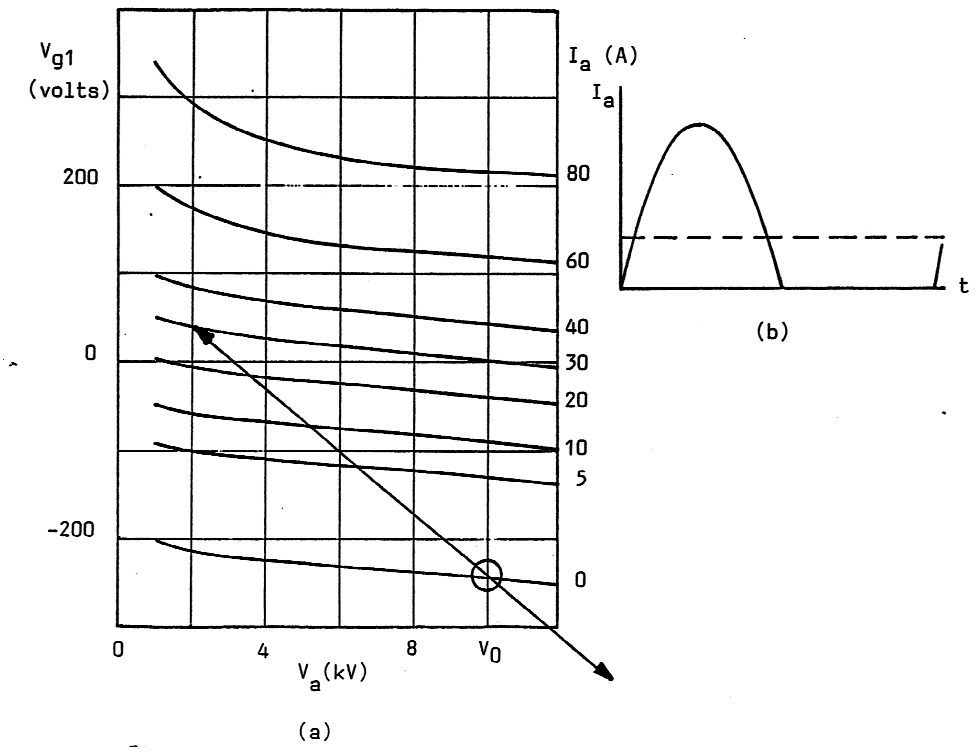


Fig. 6: Class B operation

$$V_1 = V_0 \tag{9}$$

as before. The d.c and RF powers are given by

$$P_{\rm dc} = V_0\, I_0 = V_0\, I_{\rm pk}/\pi \tag{10}$$

and

$$P_{\rm rf} = 0.5\, V_1\, I_1 = 0.25\, V_0\, I_{\rm pk} \ . \tag{11}$$

Comparison of Eqs. (10) and (11) with Eqs. (4) and (5) shows that the d.c. power is proportionately less in this case because the mean anode current is lower. The efficiency is

$$\eta = \pi/4 = 78.5\% \ . \tag{12}$$

This, again, is the upper limit of the possible efficiency of this class of operation. The example in Section 3.3 shows the effect of taking non-linearity and other imperfections into account. Devices are sometimes operated in a regime between classes A and B and this is described as class AB operation. It represents a compromise between the requirements of low harmonic content and high efficiency.

3.2.3 Class C

In class C operation the tube is biased so that it is cut off for more than half of the RF cycle, as shown in Fig. 7. If it is assumed that the tube operation is linear when it is conducting, then the efficiency can be estimated by the same method as before. For example if the angle of conduction is 90° then

$$I_0 = 0.165\, I_{\rm pk} \ , \tag{13}$$

$$I_1 = 0.31\, I_{\rm pk} \ , \tag{14}$$

$$V_1 = V_0 \tag{15}$$

and

$$P_{\rm dc} = 0.165\, V_0\, I_{\rm pk} \ . \tag{16}$$

So

$$P_{rf} = 0.5\, V_1\, I_1 = 0.155\, V_0\, I_{\rm pk} \tag{17}$$

and

$$\eta = 94\% \ . \tag{18}$$

Evidently the efficiency increases as the conduction angle decreases, but with the penalty of greater non-linearity and higher harmonic content in the output waveform. The other penalty to be paid for this increase in efficiency is a reduction in gain because the tube must be driven harder. This is not a serious penalty because the power consumed by the driver stages of the amplifier is a small proportion of the whole power consumption.

3.3 Tetrode amplifier design

The process by which a tetrode amplifier can be designed is best explained by means of an example. This is based upon a 62 kW, 200 MHz amplifier used in the CERN SPS [6]. The example was chosen because sufficient information is available about the amplifier to verify the results of the calculations. The amplifier uses a single RS2058CJ tetrode [7] operating with a d.c. anode voltage of 10 kV and 900 V screen grid bias. The design procedure described below is based upon that given in Ref. [8].

The actual amplifier is operated in class AB, but quite close to class B. For simplicity we will assume class B operation in the calculations that follow. The first stage is to estimate the probable efficiency of the amplifier. We know that this must be less than the theoretical limit given in Eq. (12) so let us try 75%. Then the d.c. power input necessary to obtain the desired output power is

$$P_{\rm in} = 62/0.75 = 83\,{\rm kW} \ . \tag{19}$$

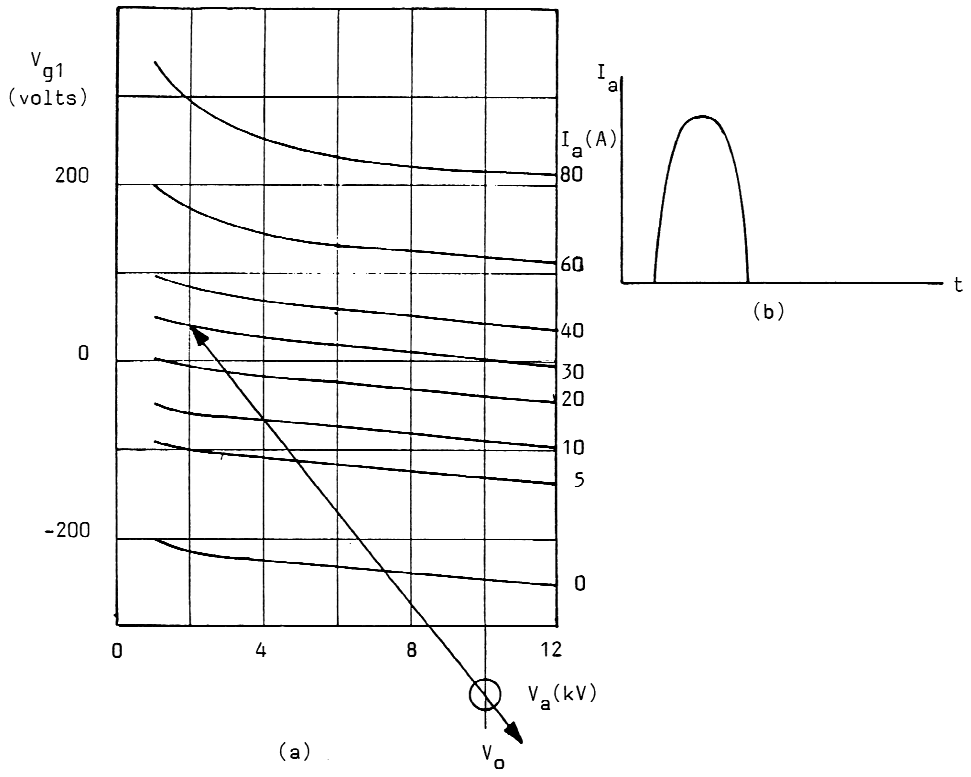


Fig. 7: Class C operation

The d.c. anode voltage was chosen to be 10 kV so the d.c. anode current is

$$I_0 = 83/10 = 8.3\,\mathrm{A} \quad . \tag{20}$$

The theoretical value of I_{pk} is given by Eq. (7) but Ref. [8] suggests that a factor of 3.5 to 4.5 should be taken as a first estimate of the ratio of the currents in place of π because of the non-linearity of the tube. If we take the factor to be 4.0 then

$$I_{\mathrm{pk}} = 40 \times 8.3 = 33\,\mathrm{A} \quad . \tag{21}$$

Next we construct the load line on the characteristic curves for the tube shown in Fig. 8 by choosing the minimum anode voltage. To ensure that the anode voltage is always greater than the screen grid voltage we select a value of 1.5 kV. The left-hand end of the load line is then fixed by this voltage and the peak anode current of 33 A, as shown in Fig. 8. We note that this requires the control grid voltage to swing slightly positive with a maximum of 60 V.

To find the anode current waveform we carry out the construction shown in the lower part of Fig. 8. The radius of the arc is equal to the length of the load line from the Q point to one end. From the arc we construct lines from which the anode current can be found at 15° phase intervals of the anode voltage. The results are shown in Table 2.

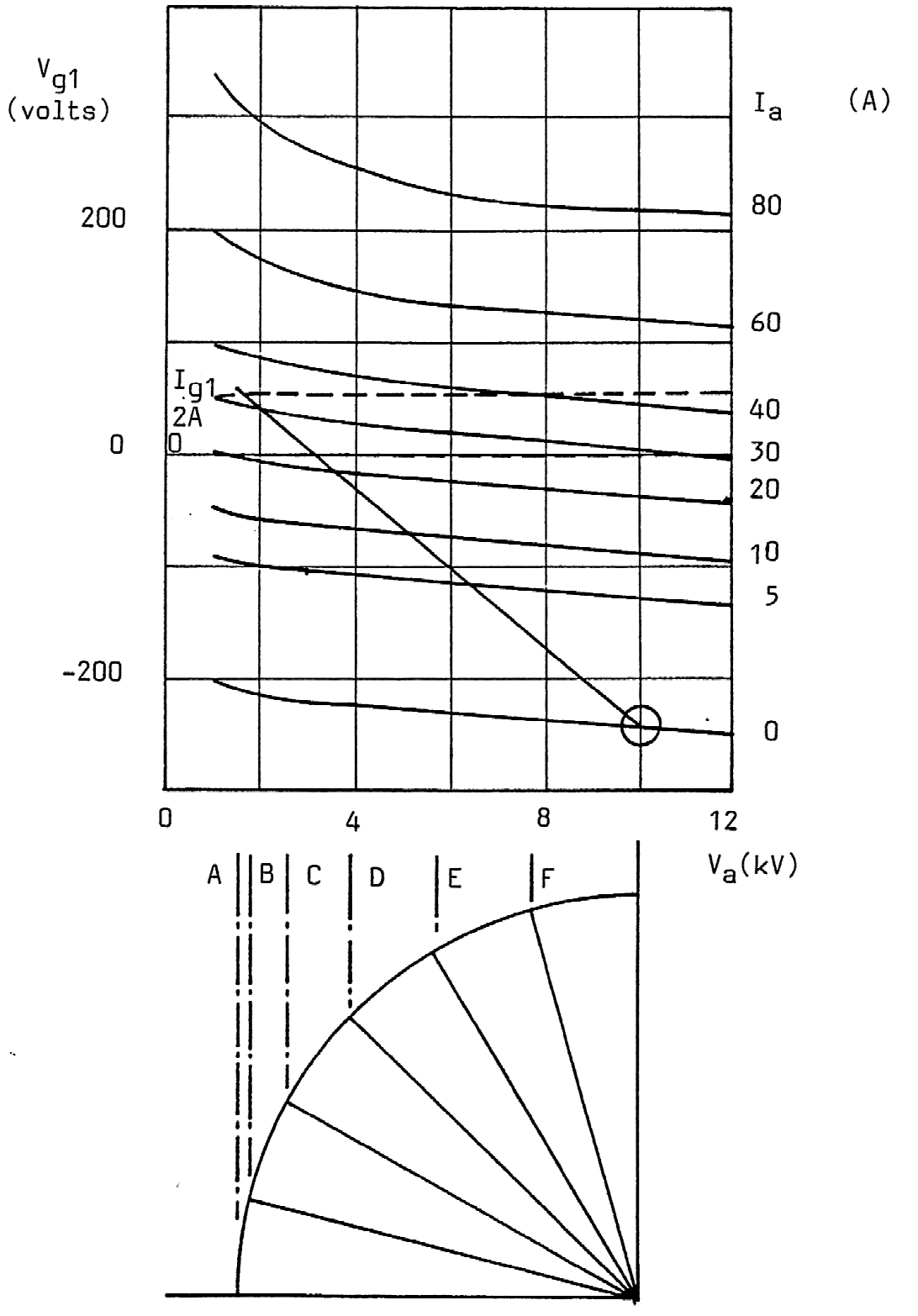


Fig. 8: RS 2058 CJ characteristic curves (courtesy of Siemens AG)

Table 2: Anode currents at 15° phase intervals taken from Fig. 8

Point	**Degrees**	I_a **(A)**
A	0	33
B	15	32
C	30	27
D	45	18
E	60	8
F	75	3

The d.c. and RF anode currents can be found by Fourier analysis of the current waveform using the numerical formulae given in Ref. [8]:

$$I_0 = (0.5\,A + B + C + D + E + F)/12 \tag{22}$$

and

$$I = (A + 1.93\,B + 1.73\,C + 1.41\,D + E + 0.52\,F)/12 \ . \tag{23}$$

When these formulae are used with the data from Table 2 the results are:

$$I_0 = 8.7\,\mathrm{A} \tag{24}$$

and

$$I_1 = 14.7\,\mathrm{A} \ . \tag{25}$$

The figures given by Eqs. (24) and (25) can be compared with the values of 8.3 A and 10.5 A given by Eqs. (20) and (7), respectively. The amplitude of the RF voltage is

$$V_1 = 10.0 - 1.5 = 8.5\,\mathrm{kV} \ . \tag{26}$$

The d.c. input power is then

$$P_{\mathrm{dc}} = 10.0 \times 8.7 = 87\,\mathrm{kW} \tag{27}$$

and the RF output power is

$$P_{\mathrm{rf}} = 0.5\,V_1\,I_1 = 62.5\,\mathrm{kW} \ , \tag{28}$$

which is very close to the desired value. The efficiency is 72%, which is slightly less than the value originally assumed. We can also calculate the input and output resistances of the tube. The output resistance is

$$R_{\text{out}} = V_1/I_1 = 5/8\ \Omega\ . \tag{29}$$

To find the input impedance we note that the amplitude of the RF control grid voltage is

$$V_{\text{in}} = 245 + 60 = 305\,\text{V}\ , \tag{30}$$

and that for grounded grid operation the RF input current is

$$I_{\text{in}} = I_1 + I_{g1} = 14.7 + 0.6 = 15.3\,\text{A}\ , \tag{31}$$

where the control grid current I_{g1}, is obtained by reading the control grid currents off Fig. 8 at 15° intervals and employing Eq. (23). Then the RF input resistance is

$$R_{\text{in}} = V_{\text{in}}/I_{\text{in}} = 20\,\Omega\ . \tag{32}$$

Finally we note that the input power is

$$P_{\text{in}} = 0.5\,V_{\text{in}}\,I_{\text{in}} = 2333\,\text{W} \tag{33}$$

and that the power gain of the amplifier is

$$\text{Gain} = 10\ \log(62.5/2.3) = 14.3\,\text{dB}\ . \tag{34}$$

Table 3 shows a comparison between the figures calculated above and those reported in Ref. [6]. The differences between the two columns of Table 3 are entirely attributable to the difference between the actual class AB operation and the class B operation assumed in the calculations.

Table 3: Comparison between actual and calculated parameters of the amplifier described in Ref. [6]

Parameter	**Actual**	**Calculated**	**Units**
V_0	10	10	kV
I_0	9.4	8.7	A
V_{g2}	900	900	V
V_{g1}	– 200	– 245	V
I_{g1}	105	600	mA
P_{out}	62	62.5	kW
P_{in}	1.8	2.3	kW
Gain	15.4	14.3	dB
η	64	72	%

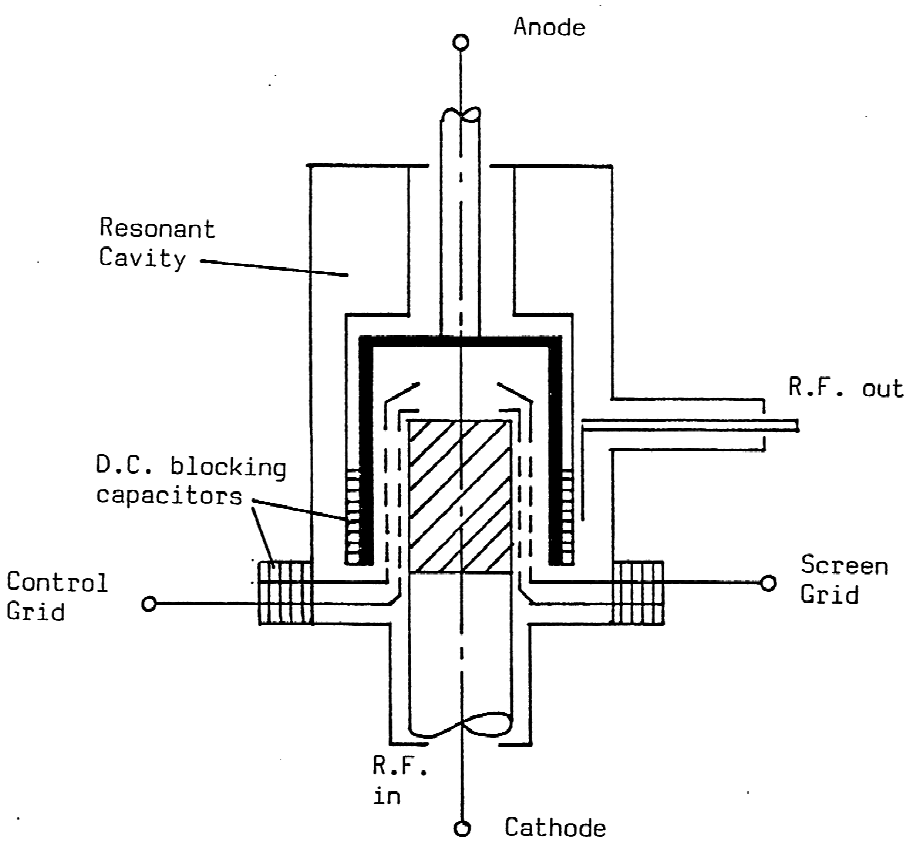


Fig. 9: Arrangement of a tetrode amplifier

3.4 Practical details

Figure 9 shows a simplified diagram of the arrangement of the tetrode amplifier for the RIKEN ring cyclotron described in Ref. [9]. The tube is operated in the grounded grid configuration with coaxial input and output circuits. The outer conductors of the coaxial lines are at ground potential and they are separated from the grids by d.c. blocking capacitors. The anode resonator is a re-entrant coaxial cavity, which is separated from the anode by a d.c. blocking capacitor. The output power is coupled through an impedance matching device to a coaxial line. The anode HT connection and water cooling pipes are brought in through the centre of the resonator.

The electrodes of the tube form coaxial lines with characteristic impedances of a few ohms. We have seen above that the input impedance of the amplifier is typically a few tens of ohms and the output impedance a few hundred ohms. Thus both the input and output lines are terminated in near open-circuit loads. The anode resonator therefore has one end open circuited and the other short circuited and it must be an odd number of quarter wavelengths long at resonance. Typically the resonator is 3/4 of a wavelength long. In this case the point at which the output coaxial line is brought into the resonator can be used to transform the impedance to provide a match. There is also a voltage node towards the lower end of the outer part of the resonator, and this can be used to bring connections through to the anode [10]. The higher-order modes of the cavity can be troublesome and it is usually necessary to damp them by the selective placing of lossy material or of coupling loops connected to external loads within the cavity [6, 9–11]. The tube heater connections must incorporate some means of decoupling from the RF circuit [6, 9].

It will be clear from what has already been said that the tube input and output are mismatched to the external connections. It is therefore necessary to devise matching networks for these connections. Figure 10(a) shows a simplified form of the input circuit of the amplifier discussed in Section 3.3. The source (normally 50 Ω) feeds the tube through a length of low impedance coaxial line. The line is terminated by the tube input impedance shunted by the capacitance between the control grid and the cathode. For the RS2058CJ tetrode this capacitance is 140 pF so that it has a reactance of 5.7 Ω at 200 MHz. This is comparable with the input resistance of 20 Ω so its effect must be allowed for. In practice the connecting coaxial line is made up of several sections having different impedances [9, 10].

Figure 10(b) shows the output circuit of the amplifier. The output resistance is shunted by the capacitance between the screen grid and the anode, which is 40 pF, with a reactance of 20 Ω at 200 MHz in this case. The tube is connected to the coaxial cavity resonator by a coaxial line and the load is connected by another line and, possibly, an impedance matching network. The effect of the anode/screen grid capacitance will normally be to tune the frequency of the cavity somewhat. Variations in the load impedance may have a similar effect, which can be compensated for by tuning the cavity. They may also require a variable matching network so that the load can be kept correctly matched to the amplifier [6, 9, 10]. If the impedance presented to the anode is too high the anode voltage swing may be excessive with the possibility of damage to the tube through internal arcing.

3.5 Operation of tetrode amplifiers in parallel

When higher powers are required than can be obtained from a single tube, it is then possible to operate several tubes in parallel. Two such systems are described in Ref. [11]. The original four 500 kW, 200 MHz power amplifiers for the CERN SPS each comprised four 125 kW tetrodes operating in parallel. Figure 11 shows the arrangement of one amplifier. The loads on the fourth arms of the 3 dB couplers normally receive no power. If one tube fails, however, they must be capable of absorbing the power from the unbalanced coupler. The amplifier also contains coaxial transfer switches (not shown in Fig. 11), which make it possible for a faulty tube to be completely removed from service. The remaining three tubes can then still deliver 310 kW to the load. A more recent design of a 500 kW amplifier for the same accelerator employs sixteen 35 kW units operated in parallel, with a seventeenth unit as the driver stage. Both types of amplifier operate at anode efficiencies greater than 55% and overall efficiencies greater than 45%.

3.6 The Diacrode®

A recent development of the tetrode is the Diacrode[1] [12, 13]. In this tube the coaxial line formed by the anode and the screen grid is extended to a short circuit, as shown in Fig. 12. The consequence of this change is that the standing wave now has a voltage antinode and a current node at the centre of the active region of the tube. The tetrode, in contrast, has a voltage antinode and current node just beyond the end of the active region, as shown in Fig. 12. Thus, for the same RF voltage difference between the anode and the screen grid, the Diacrode has a smaller reactive current flow and much smaller power dissipation in the screen grid than a tetrode of similar dimensions. This means that, compared with conventional tetrodes, Diacrodes can either double the output power at a given operating frequency or double the frequency for a given power output. The gain and efficiency of the Diacrode are the same as those of a conventional tetrode. Table 4 shows the comparison between the TH 526 tetrode and the TH 628 Diacrode operated at 200 MHz.

[1] The name Diacrode is the registered property of Thomson Tubes Electroniques.

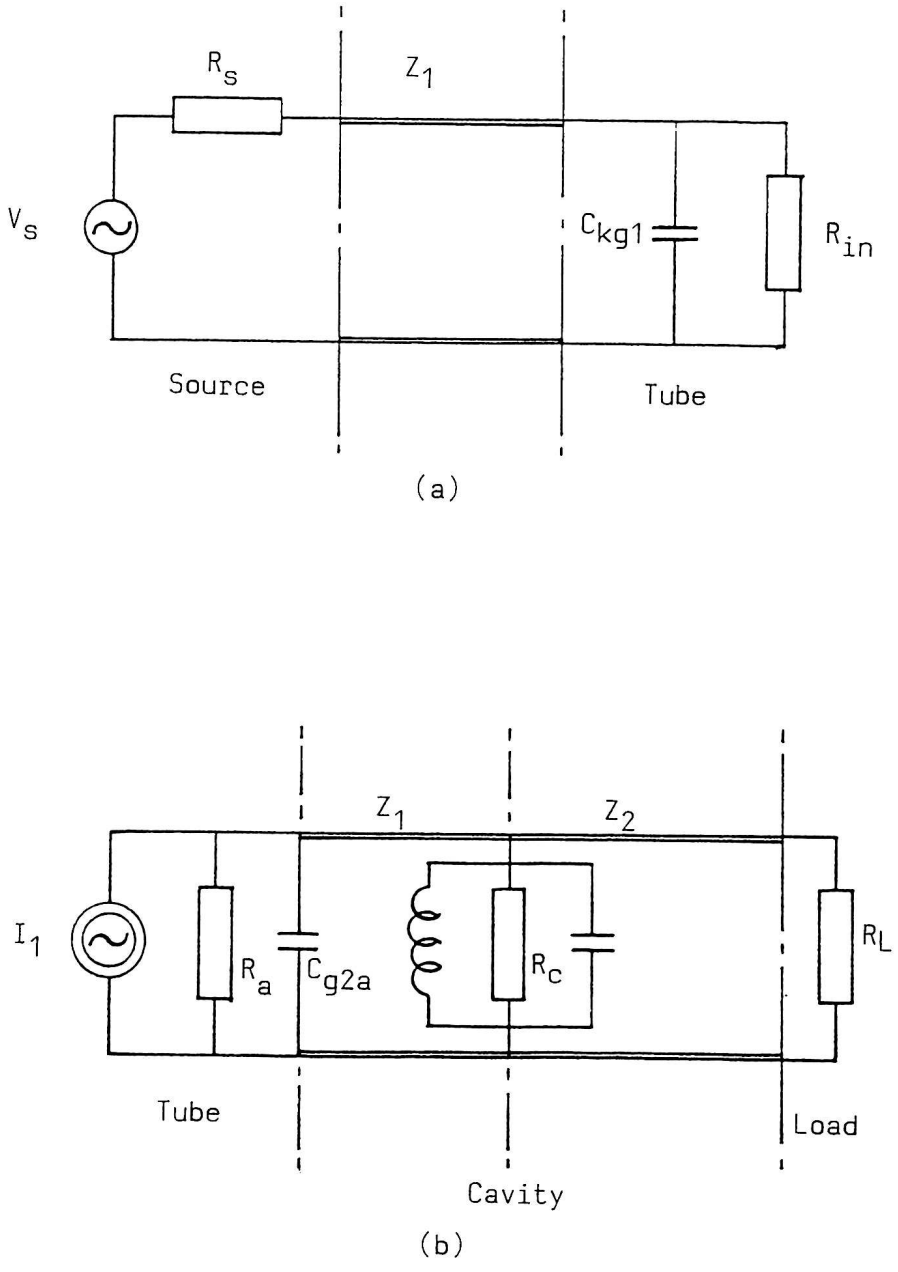


Fig. 10: Tetrode amplifier: (a) input circuit and (b) output circuit

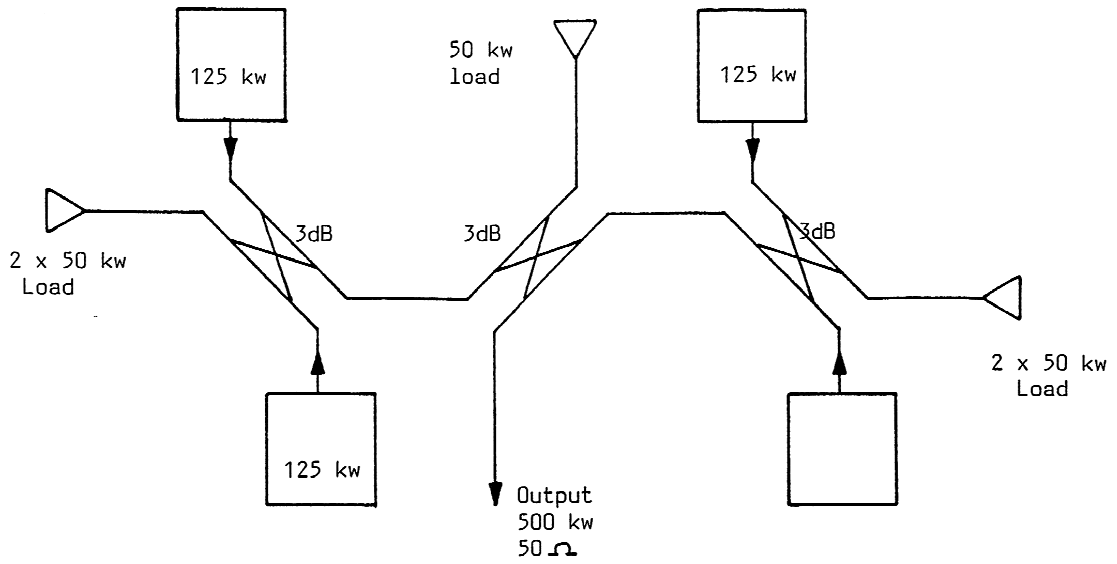


Fig. 11: Parallel operation of tetrode amplifiers

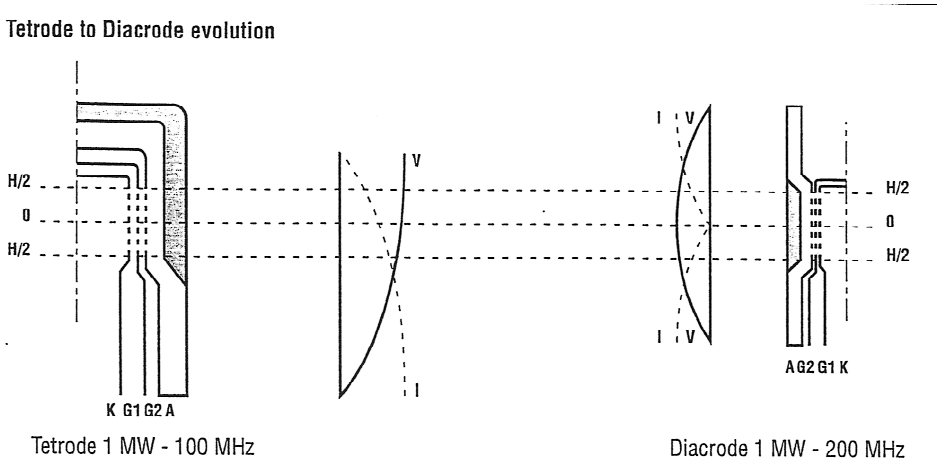


Fig. 12: Comparison between a tetrode and a Diacrode (courtesy of Thales)

Table 4: Comparison between the TH 526 tetrode and the TH 628 Diacrode

Parameter	**Units**	**TH 526**	**TH 628**
Pulse duration	ms	2.2	2.5
Peak output power	kW	1600	3000
Average output power	kW	240	600
Anode voltage	kV	24	26
Anode current	A	124	164
Peak input power	kW	64.9	122.5
Gain	dB	13.9	13.9

4 INDUCTIVE OUTPUT TUBES

The tetrode suffers from the disadvantage that the same electrode, the anode, is part of both the d.c. and the RF circuits. The output power is, therefore, limited by grid and anode dissipation. To get high power at high frequencies it is necessary to employ high velocity electrons and to have a large collection area for them. It is therefore desirable to separate the electron collector from the RF output circuit. The possibility that these two functions might be separated from each other was originally recognized by Haeff in 1939, but it was not until 1982 that a commercial version of this tube was described [14]. Haeff called his invention the Inductive Output Tube (IOT) but it is also commonly known by the proprietary name Klystrode®[2]. The basis of operation of the IOT is illustrated in Fig. 13, which shows how a current is induced in a resonant cavity by the azimuthal magnetic field of a bunch of electrons passing through it (hence the name of the tube). The bunches of electrons can be produced by a high-voltage electron gun and the spent beam can be collected on an electrode with a large surface area.

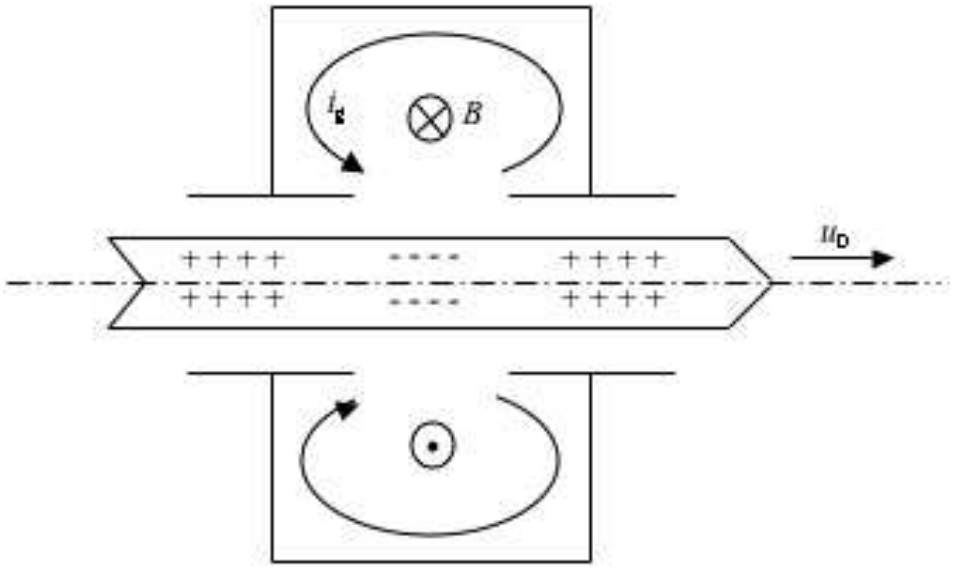


Fig. 13: Current induced in a cavity resonator by a bunched electron beam

Figure 14 shows a schematic diagram of an IOT. The electron beam is formed by a gridded, convergent-flow electron gun and confined by an axial magnetic field (not shown). The gun is biased so that no current flows except during the positive half-cycle of the RF input. The resulting bunches of electrons excite the cavity resonator in a manner analogous to class B or class C amplifier operation. The operation of this tube can be understood by comparing it with that of a tetrode, as shown in Fig. 15. In this figure the solid lines show the potentials at the peak of the current flow, and the broken lines show them when there is no current flowing. In a tetrode amplifier the anode voltage is lowest at maximum anode current so the electrons cross the tube relatively slowly. The gap between the screen grid and the anode must therefore be kept small to minimize transit time effects and the peak anode voltage is, accordingly, limited. In the IOT the electrons enter the gap with high velocity and encounter an opposing electric field which extracts energy from them and transfers it into the RF fields in the cavity resonator. Since the electron velocity is high it is possible to use a much longer output gap than in the tetrode.

Advantages of the IOT are that it does not need a d.c. blocking capacitor in the RF output circuit since the cavity is at ground potential, and it has higher isolation between input and output and a longer life than an equivalent tetrode. These advantages are offset to some extent by the need for a magnetic focusing field. The typical gain is 24 dB, which is appreciably higher than that of a tetrode; high enough

[2] The name Klystrode is the registered property of CPI Inc.

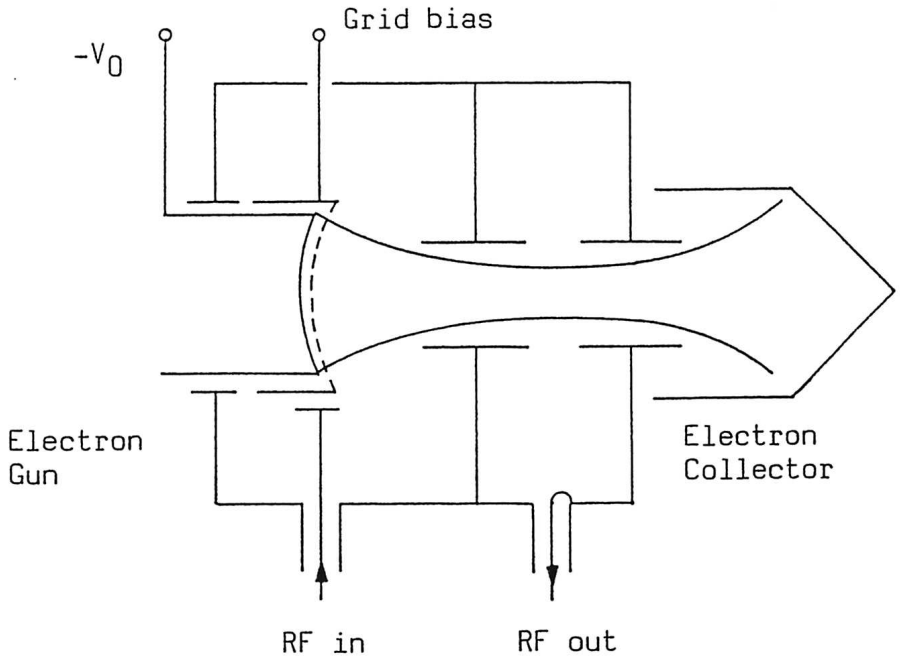


Fig. 14: Schematic diagram of an inductive output tube

in fact for a 60 kW tube to be fed by a solid-state driver stage. The IOTs designed for television service are operated in class B and have dual input and output cavities to give the necessary bandwidth. Tubes currently available give up to 60 kW of power in the UHF TV bands (470–860 MHz). In principle the efficiency can be still further enhanced by collector depression (see Section 5.3.1). A tube for use in an accelerator does not need to have the additional cavities and can operate in class C with an efficiency of around 70%. It has been suggested that IOTs could be useful from about 100 MHz to perhaps 3 GHz with CW powers up to 1 MW at the lower frequencies and some tens of kilowatts at the higher frequencies. Further information about the IOT can be found in Refs. [14] and [15].

5 KLYSTRONS

At frequencies of 300 MHz and above the power sources for accelerators are generally klystrons. The klystron extracts power from a bunched electron beam in the same manner as an IOT, but the bunches are produced by velocity modulation of the beam rather than by switching it on and off. Figure 16 shows a schematic diagram of a two-cavity klystron. Most klystrons used for accelerators have four, five, or even more cavities in order to get high gain and the highest possible efficiency. The long, high-current electron beam is confined by an axial magnetic field throughout the interaction region.

5.1 Electron bunching in klystrons

The method by which the electron beam in a klystron is velocity modulated is illustrated in Fig. 17. The beam passes through a cavity resonator, very like the output cavity of an IOT, which is excited in the TM_{010} mode so that there is an axial electric field across the gap in the drift tube. The electrons are accelerated or retarded according to the phase of the RF field in the gap as they cross it. As the beam leaves this cavity it is velocity modulated, but there is no current modulation.

Figure 18 shows the way in which the velocity modulation produces electron bunches. The motion of a number of sample electrons is displayed in a frame of reference, which is moving at the initial beam velocity. The peak accelerating phase is marked $\oplus$ and the peak retarding phase is marked O. Those

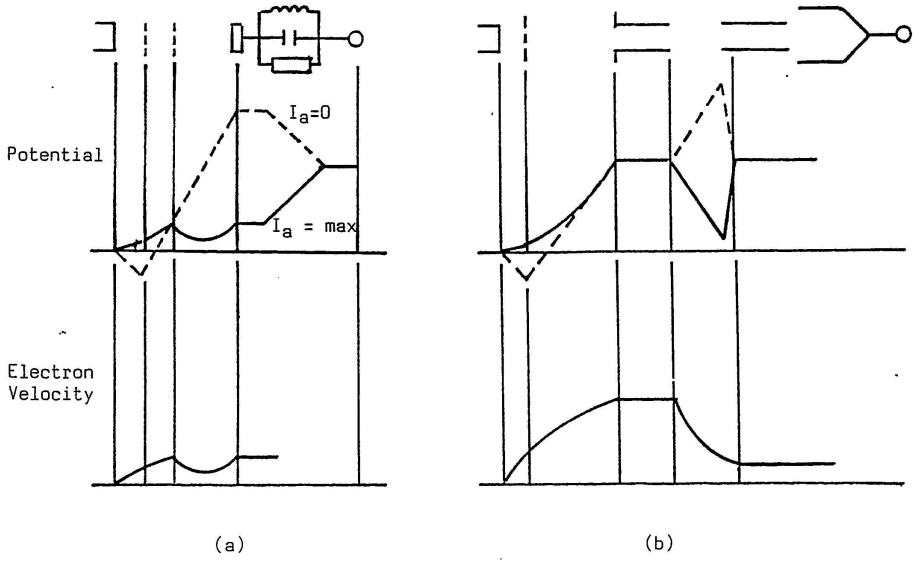


Fig. 15: Comparison between (a) tetrode and (b) IOT operation

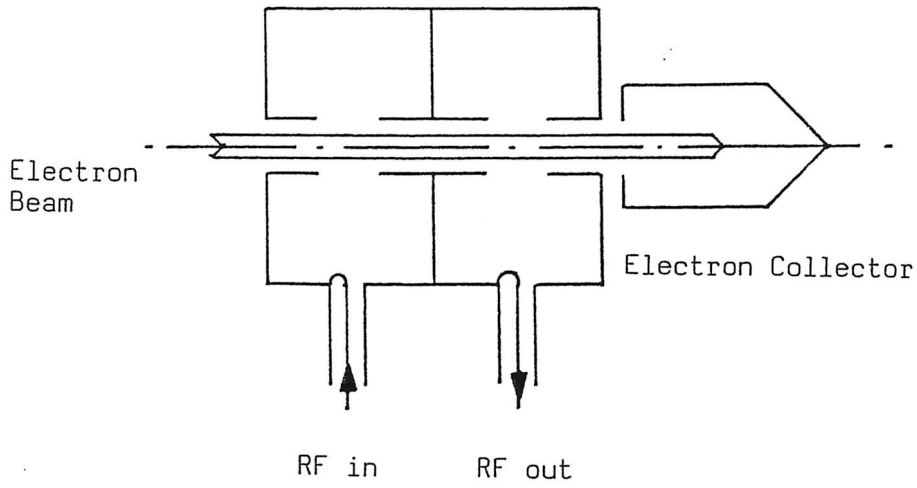


Fig. 16: Schematic diagram of a two-cavity klystron

electrons which cross the input gap at an instant when the field is zero proceed without any change in their velocities and appear as horizontal straight lines. Retarded electrons move upwards and accelerated electrons move downwards in the diagram. The velocity modulation causes the electrons to become bunched together until further bunching is prevented by the space-charge repulsion between the electrons. Under small-signal conditions the beam has current modulation but no velocity modulation at the plane of the bunch. Figure 18 is an example of an Applegate diagram. In a simple klystron (Fig. 17) a second cavity, tuned to the signal frequency, is placed at the plane of maximum bunching. This cavity presents a resistive impedance to the current induced in it by the electron beam, so the phase of the field across the gap is in anti-phase with the RF beam current. Electrons that cross the gap within $\pm$ 90° of the bunch centre are retarded and give up energy to the field of the cavity. Since more electrons cross the second gap during the retarding phase than the accelerating phase there is a net transfer of energy to the RF field of the cavity. Thus the klystron operates as an amplifier by converting some of the d.c. energy

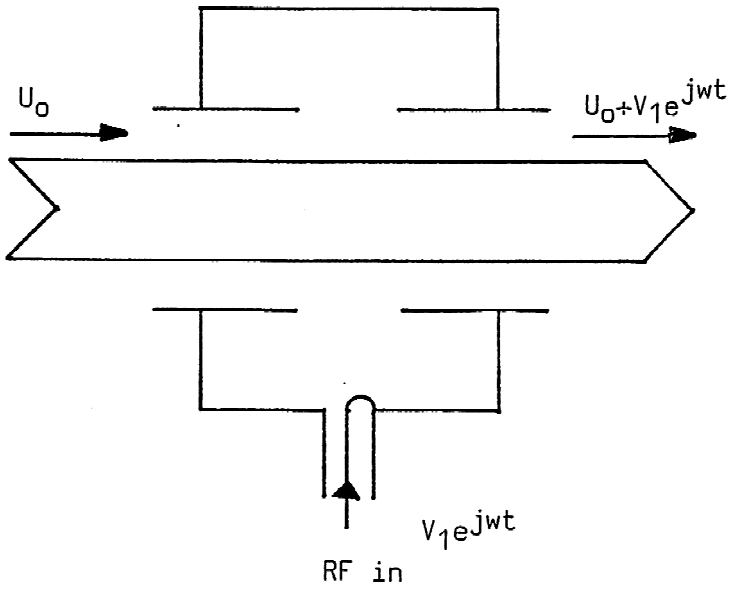


Fig. 17: Velocity modulation of an electron beam by a cavity resonator

input into RF energy in the output cavity. If the second cavity were removed the electron bunches would disperse under the influence of space-charge, only to re-form further down the tube. Under small-signal conditions this process is repeated periodically. From the point of view of an observer travelling with the mean electron velocity the electrons would appear to be executing oscillations about their mean positions at the electron plasma frequency. The plasma frequency is modified to some extent by the boundaries surrounding the beam and by the presence of the magnetic focusing field.

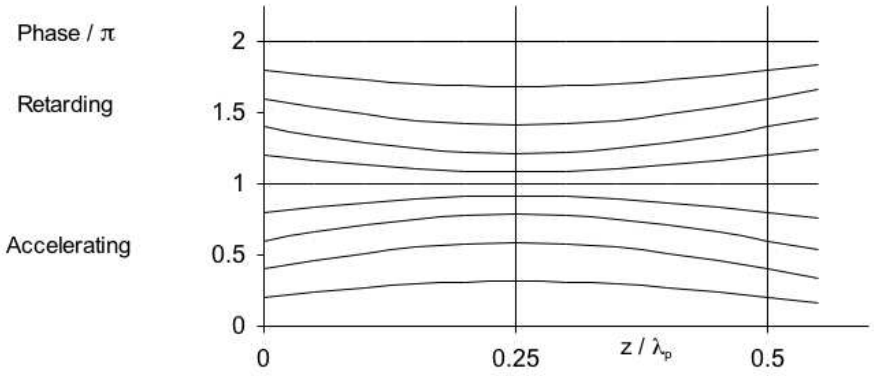


Fig. 18: Bunching of a velocity modulated electron beam

The electron plasma frequency is given by

$$\omega_p = (\eta\rho/\varepsilon_0)^{0.5} \quad , \tag{35}$$

where η is the charge-to-mass ratio of the electron and ρ is the charge density in the beam. The distance

from the input gap to the first plane at which the bunching is maximum is then a quarter of a plasma wavelength (λ_p), given by

$$\lambda_p = 2\pi u_0/\omega_p \tag{36}$$

where u_0 is the mean electron velocity.

The bunching length is independent of the input signal except at very high drive levels, when it is found that the bunching length is reduced. If attempts are made to drive the tube still harder the electron trajectories cross over each other and there is less bunching. The transfer characteristic of a klystron (Fig. 19) shows that the device is a linear amplifier at low signal levels but that the output saturates at high signal levels. A tube used in an accelerator would normally be run at, or close to, saturation to obtain the highest possible efficiency.

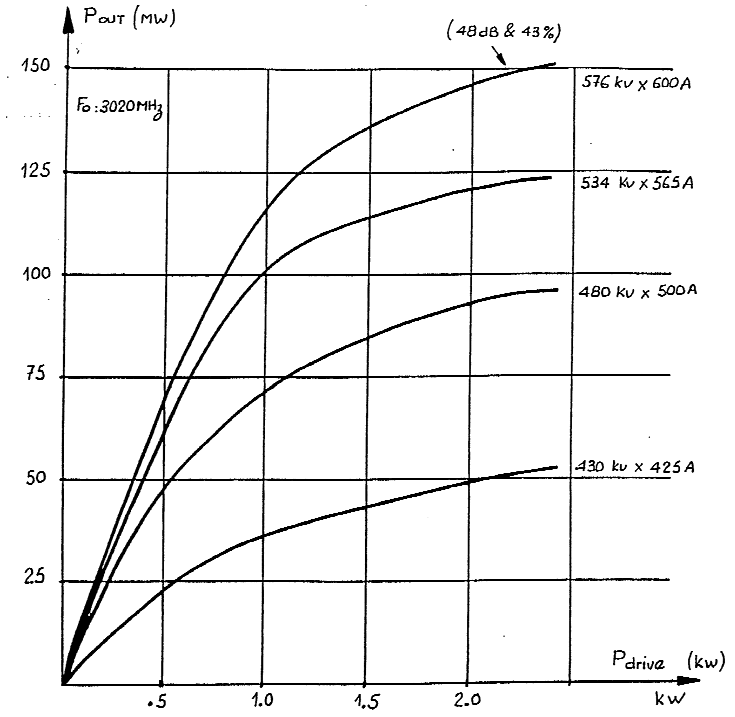


Fig. 19: Klystron transfer characteristics (courtesy of Thales)

5.1.1 The effect of bunching on efficiency

The conversion efficiency of a klystron depends upon the interaction between the bunched beam and the electric field of the output cavity. This is illustrated in Fig. 20, which shows the waveform of the output gap field (assumed to be sinusoidal) and a number of possible current waveforms [16]. To simplify the discussion we assume that the electrons in the bunch all have the same energy V_0 electron volts and that the magnitude of the output gap voltage V_g is equal to V_0, as shown in Fig. 20(a). It follows that any electron that crosses the gap when the field is at its maximum and retarding loses all its energy to the field if transit time effects are neglected.

It used to be supposed that the maximum conversion efficiency would be obtained by maximizing the amplitude of the fundamental harmonic component of the beam current. It can be shown that the ratio of the RF current i_1, to the mean current I_0, is 1.16. Figure 20(b) shows the normalized RF beam current

when the ratio is 1.0 and no higher harmonics are present. Only a small proportion of the electrons cross the gap when the field is close to its maximum retarding value, and those in the shaded region of the diagram experience an accelerating field and remove energy from the cavity. The instantaneous power flow from the beam to the cavity, normalized to the d.c beam power, is

$$P/P_0 = \frac{i_g V_g}{I_0 V_0} = \cos\theta\,[1 + \cos\theta] \quad , \tag{37}$$

where θ is the phase angle of the field. The mean power and hence the efficiency are obtained by integrating Eq. (38) over one RF cycle to give

$$\eta = \frac{1}{2\pi}\int_0^{2\pi} \cos\theta\,[1 + \cos\theta]\,d\theta = 0.5 \quad . \tag{38}$$

The theory of klystron bunching that ignores the effects of space-charge predicts the optimum current waveform shown in Fig. 20(c). In this case some electrons also cross the gap at an accelerating phase of the field. A calculation along the lines of that above shows that the maximum possible efficiency is 58%. The identical figure is given by the calculation above if i_g/I_0 is assumed to be 1.16.

Figures 20(d) and 20(e) show two possible waveforms in which all the electrons cross the gap when the field is retarding. The waveform in Fig. 20(d), which is analogous to class B operation of a tetrode, produces an efficiency of 71%, whereas that in Fig. 20(e) gives 90%. When the effects of circuit losses are considered it is evident from the efficiencies measured that high-power klystrons commonly operate under conditions close to those of Fig. 20(e) (see Section 5.3).

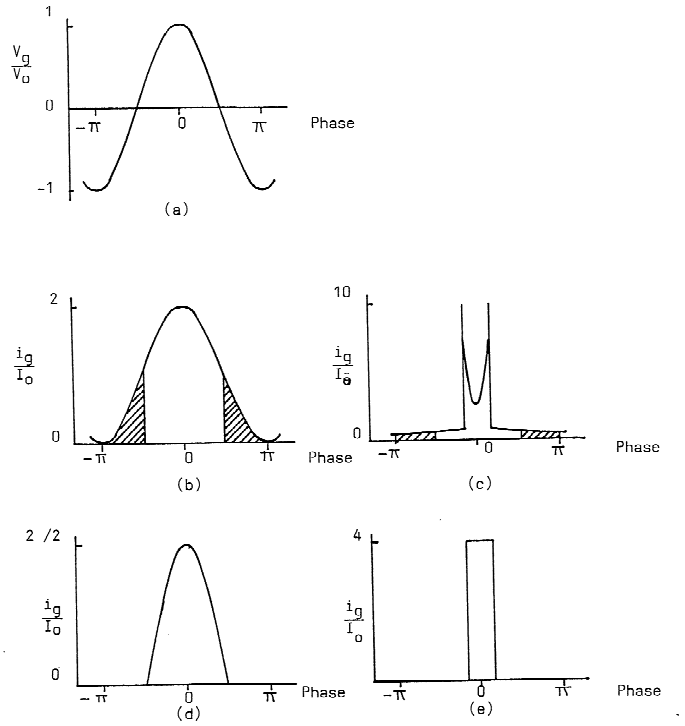


Fig. 20: Klystron output gap waveforms (courtesy of H. Bohlen)

5.1.2 Formation of bunches

Figure 18 shows that the tightness of the bunches formed by velocity modulation is limited by space-charge forces. It follows that it is easiest to obtain high efficiency with a beam in which the current density is low and the velocity is high, that is, with a low perveance ($I_0/V_0^{1.5}$). Figure 21 shows the effect of perveance on efficiency for a particular klystron.

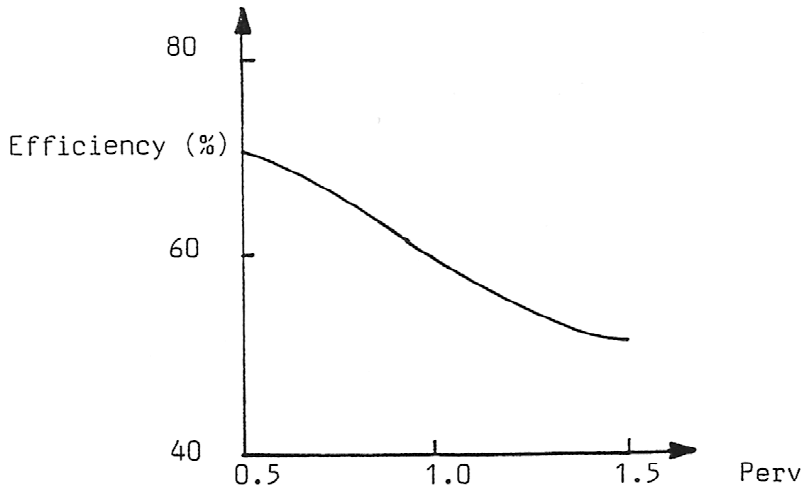


Fig. 21: Dependence of the efficiency of a klystron on the beam perveance

Figure 22 shows the Applegate diagram computed for a typical super-power klystron [17]. The tube has five cavities placed as shown, and the peak accelerating and retarding fields are marked in the same way as before. The bunching produced by the first cavity is imperceptible on the scale of this diagram but it is sufficient to excite the RF fields in the second cavity. The second cavity is tuned to a frequency above the signal frequency so that it presents an inductive impedance to the beam current. As a result the bunch centre coincides with the neutral phase of the field in the cavity and further velocity modulation is added to the beam, which produces much stronger bunching at the third cavity. The third cavity is tuned to the second harmonic of the signal frequency, as can be seen from a careful examination of the diagram. The principal purpose of this cavity is to cause the electrons lying farthest from the bunch centre to be gathered into the bunch. The use of a second harmonic cavity increases the efficiency of a klystron by at least ten percentage points. The splitting of the lines in the diagram that occurs at this plane is caused by a divergence in the behaviour of electrons in different radial layers within the electron beam. The fourth cavity is similar to the second cavity and produces still tighter bunching of the electrons. By the time they reach the final cavity nearly all the electrons are bunched into a phase range which is $\pm$ 90° with respect to the bunch centre. The output cavity is tuned to the signal frequency so that the electrons at the bunch centre experience the maximum retarding field and all electrons that lie within a phase range of $\pm$ 90° with respect to the bunch centre are also retarded. If the impedance of the output cavity is chosen correctly then a very large part of the bunched beam's kinetic energy can be converted into RF energy. Computer simulations have shown that the ratio of the fundamental beam RF current to the d.c. beam current can be as high as 1.6 to 1.7 at the output cavity.

5.2 Terminal characteristics of klystrons

The performance of a klystron is appreciably affected by variations in beam voltage, signal frequency, and output match, and we now examine these in turn.

Klystrons for use in accelerators are normally operated at or close to saturation. Figure 19 shows that the output power is then insensitive to variations in input power and, by extension, to variations in

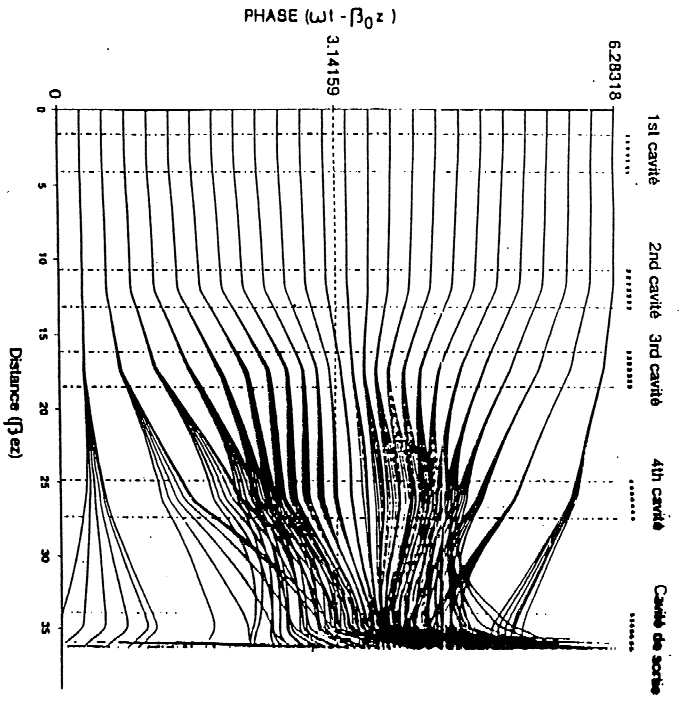


Fig. 22: Electron bunching in a high power five-cavity klystron (courtesy of Thales)

beam voltage. The effects on the phase of the output signal are more serious because of the distance from the input to the output. If the distance from the centre of the input gap to the centre of the output gap is L then the phase difference between the input and the output is

$$\phi = \omega L / u_0 \tag{39}$$

where the beam velocity is given by

$$u_0 = c\left[1 - \left(1 + (eV/m_0c^2)\right)^{-2}\right]^{0.5} \quad . \tag{40}$$

Thus if the normal beam voltage is 90 kV, the tube length 1.17 m, and the frequency 500 MHz, the sensitivity of phase to changes in the beam voltage is –5.8° per kV.

The transfer characteristic of a klystron with synchronously tuned cavities is essentially that of a resonant circuit as far as changes in frequency are concerned, namely

$$H(\omega) = R/\left[1 - jQ(\omega_0/\omega - \omega/\omega_0)\right] \quad . \tag{41}$$

The effective Q factor takes account of the combined effects of all the cavities and of any external loading. The klystron used in the example above has a bandwidth of 1 MHz, giving an effective Q factor of 500. Small changes in the centre frequency are produced by changes in the working temperature of the tube. If $\omega = \omega_0 + \delta\omega$ and if $\delta\omega$ is small then

$$\text{phase}(H(\omega)) = \arctan(-2Q\delta\omega/\omega_0) \quad , \tag{42}$$

giving a phase sensitivity of -63° per MHz. If the cavities are made from copper, whose coefficient of thermal expansion is 16×10^{-6} K^{-1}, then the sensitivity of phase to variations in temperature is 0.53 degrees/K.

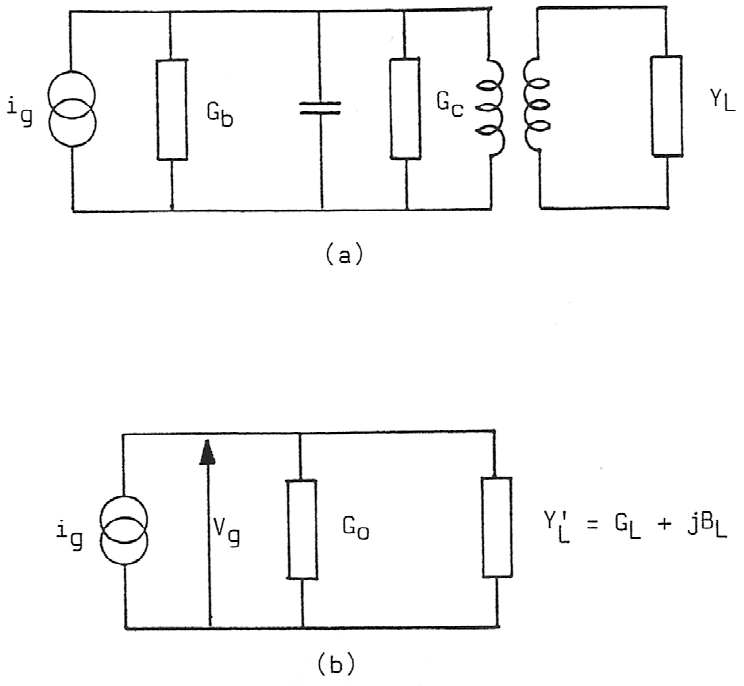


Fig. 23: Equivalent circuits of the output of a klystron

The output circuit of a klystron can be represented by the equivalent circuit shown in Fig. 23(a). The beam behaves as a current source with an impedance R_b, which is of the order of 20 kΩ. The shunt resistance of the output cavity, R_c, is in the range 50 to 200 kΩ. The transformer represents the output coupling and matching network required to match the load impedance Z_L to the output impedance of the klystron. Some klystrons incorporate arrangements for adjusting the coupling between the load and the cavity. If the cavity is tuned to the signal frequency the equivalent circuit can be simplified to that shown in Fig. 23(b). The load admittance is referred to the plane of the output gap as $Y'_L = G_L + jB_L$. When the load is correctly matched to the tube $G_L = R_b$ and $B_L = 0$. The effects of a mismatched load can be deduced from the circuit in Fig. 23(b). The gap voltage is given by

$$V_{\rm g} = i_{\rm g}/(G_0 + G_{\rm L} + jB_{\rm L}) \ . \tag{43}$$

When the load is matched

$$V_{\rm g} = V_{\rm g0} = i_{\rm g}/2G_0 \ . \tag{44}$$

Thus

$$V_{\rm g}/V_{\rm g0} = 2G_0/(G_0 + G_{\rm L} + jB_{\rm L}) = 2/(1 + g_{\rm L} + jb_{\rm L}) \ , \tag{45}$$

in which the lower case symbols represent admittances that have been normalized to that of a matched load. Then the normalized gap voltage has amplitude

$$|V_{\rm g}/V_{\rm g0}| = 2/\left[(1+g_{\rm L})^2 + b_{\rm L}^2\right]^{1/2} \tag{46}$$

and phase

$$\text{phase}\,(V_{\rm g}/V_{\rm g0}) = \arctan\left[-b_{\rm L}/(1+g_{\rm L})\right] \quad . \tag{47}$$

The power delivered to the load is

$$P_{\rm L} = 0.5\,|V_{\rm g}|^2\,G_{\rm L} \tag{48}$$

and to a matched load is

$$P_0 = 0.5\,|V_{g0}|^2\,G_0 \tag{49}$$

so that

$$P_{\rm L}/P_0 = 4\,g_{\rm L}/\left[(1+g_{\rm L})^2 + b_{\rm L}^2\right] \quad . \tag{50}$$

Contours of constant gap voltage amplitude and of constant load power can be calculated from Eqs. (44) and (48). Figure 24 shows these contours plotted on a Smith chart of normalized load admittance for a typical klystron [18]. It is important to ensure that the working point remains in the region for which the normalized gap voltage is less than unity in order to avoid the risk of voltage breakdown in the output gap. If the gap voltage becomes too high it is possible for electrons to be reflected, reducing the efficiency of the tube and providing a feedback path to the other cavities, which may cause the tube to become unstable. A further complication is provided by the effect of harmonic signals in the output cavity. Since the klystron is operated in the non-linear regime to obtain maximum efficiency it follows that the signal in the output waveguide will have harmonic components. These are incompletely understood but it is known that the reflection of harmonic signals from external components such as a circulator can cause the klystron output to behave in unexpected ways.

5.3 Typical klystrons

5.3.1 UHF television klystrons

At moderate powers in the UHF band it is possible to use klystrons designed for use in UHF television transmitters as power sources for accelerators. Tubes are available with output powers in the 10–70 kW range and gains of 30–40 dB. The beam current can usually be controlled independently of the beam voltage by the voltage applied to a separate modulating anode. The conversion efficiency is around 50% but in some modern tubes this is increased by the use of a multi-element depressed collector. This type of collector has several electrodes at different potentials between ground and cathode potentials. The spent electron beam can therefore be collected in a way that reduces both the heat dissipated in the collector and the power supplied to the tube [19]. If greater power is required than can be supplied by one tube it is possible to operate several tubes in parallel. A 450 kW, 800 MHz amplifier for the CERN SPS comprises eight modified UHF television klystrons operated in parallel.

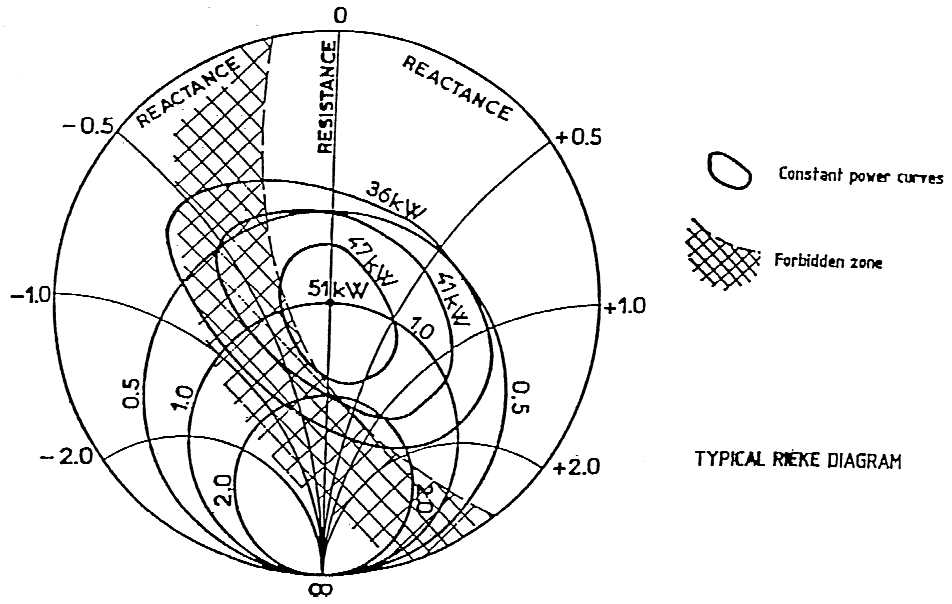


Fig. 24: Typical Rieke diagram of a klystron (courtesy of Thales)

5.3.2 Super-power klystrons

Some klystrons, commonly known as super-power klystrons, have been developed specifically for use in accelerators. Tables 5 and 6 give a summary of the characteristics of some typical tubes.

Table 5: Characteristics of typical CW super-power klystrons

Type	**TH2089B**	**K351C**	**YK1303**	**TH2103**	**Units**
Frequency	352	352	508	3700	MHz
Beam voltage	100	100	90	60	kV
Beam current	20	20	18.2	20	A
Power	1300	1300	1	0.5	kW
Efficiency	65	65	61	43	%
Gain	40	40	41	45	dB
Length	4.8	3.8	3.75	2.0	m

Table 6: Characteristics of typical pulsed super-power klystrons

Type	**TH 2153**	**SLAC**	**VKS 8333A**	**Units**
Frequency	3.0	2.87	2.998	GHz
Beam voltage	576	470		kV
Beam current	600	612		A
Power	150	150	150	MW
Pulse length	1.2	3	3	μsec
Efficiency	43	50		%
Gain	48	52		dB

5.4 The SLAC Energy Doubler

Sometimes higher pulsed powers are required than can be obtained from a single conventional klystron. One way of achieving this is the energy doubler devised at SLAC (SLED) [20]. The principle of operation of this device is illustrated by Fig. 25(a). The input to the klystron is via a 180° switched phase shifter. The output passes to a 3 dB coupler connected to a pair of cavities. If the cavity fields are initially zero then part of the output power of the klystron goes into building them up while the remainder is reflected and passes on towards the accelerator. As the cavity fields increase they re-radiate power in anti-phase with the incident signal so that the mean power transmitted to the accelerator is small. The cavities are over-coupled so that the peak re-radiated power rises to a level greater than the transmitted power. If the phase of the klystron drive is then abruptly reversed the re-radiated and transmitted signals are in phase with each other and the power transmitted to the accelerator rises steeply, as shown in Fig. 25(b). The power level then decays as the energy stored in the cavities is discharged. The process can be repeated to provide a train of very high energy pulses. The principle is similar to that of a pulse modulator with RF storage in the cavities replacing d.c. storage in capacitors. The original SLED system [20] used a pulsed 30 MW klystron to produce SLED output pulses at 125 MW. More recently a similar system has been suggested that employs a CW klystron [21].

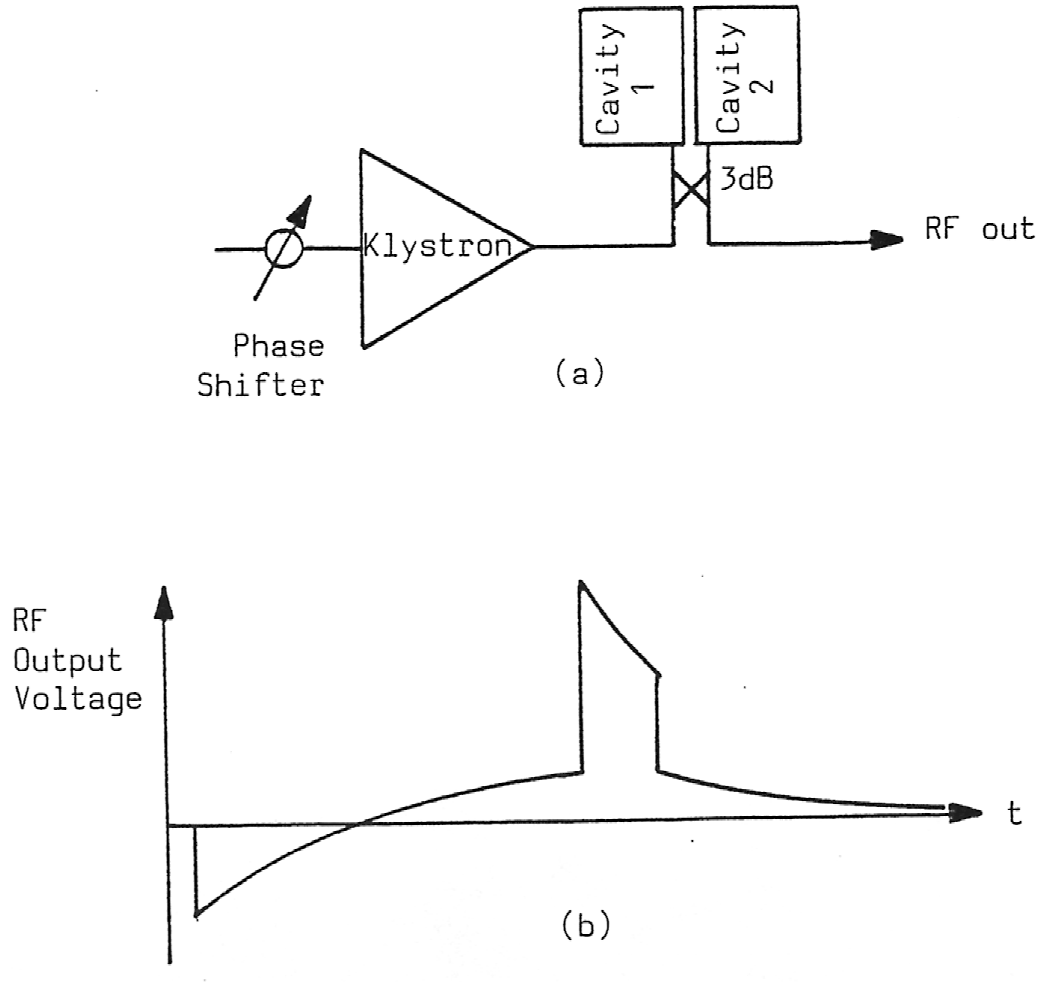


Fig. 25: The SLAC Energy Doubler (SLED): (a) circuit diagram and (b) output waveform

5.5 Multiple beam klystrons

We have seen that the efficiency of a klystron is determined by the perveance of the electron beam, so, to get high efficiency, it is necessary to use a high-voltage, low-current beam. The use of high voltages produces problems with voltage breakdown and it is therefore difficult to obtain very high power with high efficiency, as can be seen by comparing Table 6 with Table 5. One solution to this problem is to use several electron beams within the same vacuum envelope, as shown in Fig. 26. A klystron designed in this way is known as a Multiple Beam Klystron (MBK). The principle of the MBK has been known for many years [22] but, until recently, the only such tubes constructed were in the former Soviet Union, for military applications. The first MBK designed specifically for use in particle accelerators is the TTE type TH1801, whose performance is shown in Table 7 [23].

Table 7: Characteristics of a multiple beam klystron

Type	**TH 1801**	**Units**
Frequency	1300	MHz
Beam voltage	115	kV
Beam current	133	A
Number of beams	7	
Power	9.8	MW
Pulse length	1500	μs
Efficiency	64	%
Gain	47	dB

5.6 Relativistic klystrons

Some experimental work has been carried out on klystrons with very high energy beams produced by induction linear accelerators [24]. Table 8 shows the results achieved in one such experiment.

Table 8: Characteristics of the LLNL SL-4 relativistic klystron

Type	**TH 1801**	**Units**
Frequency	11.45	GHz
Beam voltage	975	kV
Beam current	625	A
Power	200	MW pk
Pulse length		μs
Efficiency	32	%
Gain	55	dB

6 GYROTRONS

An alternative type of tube for producing very high pulsed RF power at high frequencies is the gyrotron. This type of tube has been the subject of intensive developmental work mainly with a view to providing RF power for plasma heating experiments. A good summary of this work and of the development of other, novel, high-power RF sources is given in Ref. [25]. The gyrotron employs the interaction between an annular electron beam and the azimuthal electric field of a circular waveguide mode, as shown in Fig. 27. There is a strong axial magnetic field so that the electrons move in small orbits at the cyclotron frequency within the beam (as shown). The cyclotron frequency is made equal to the signal frequency.

At frequencies above 60 GHz this means that a super-conducting solenoid is needed to produce the magnetic field. It is essential to the working of the gyrotron that the electrons should have relativistic velocities. The cyclotron frequency is then given by

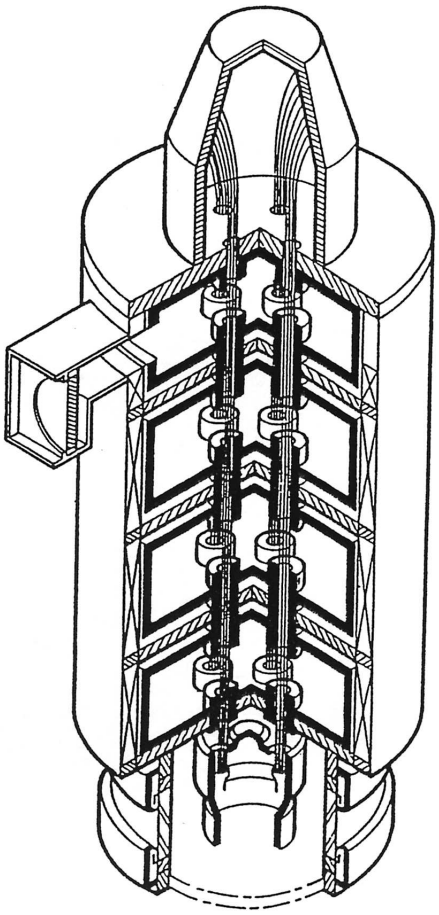


Fig. 26: Schematic diagram of a multiple-beam klystron (MBK) (courtesy of Thales)

$$\omega_c = (eB/m_0)(1 - \nu^2/c^2)^{0.5} \tag{51}$$

so that it is a function of the electron velocity. As the electron velocity increases the cyclotron frequency decreases so that the faster electrons lag behind the field. Slow electrons, similarly, lead the RF field and phase bunching occurs with a net transfer of energy to the RF field.

Commercially available gyrotrons are oscillators with the general arrangement shown in Fig. 28. The hollow electron beam is produced by a magnetron electron gun and confined by an axial magnetic field whose strength varies as shown. The interaction takes place in a section of cylindrical waveguide made resonant by the mismatches at its ends. The spent electron beam is allowed to spread sideways to be collected on the wall of the larger, cylindrical, output waveguide. The RF output power passes down this guide and through the output window. At millimetre wavelengths it is usual to use an over-moded output waveguide to avoid problems with breakdown in it. Table 9 shows the characteristics of some typical gyrotron oscillators.

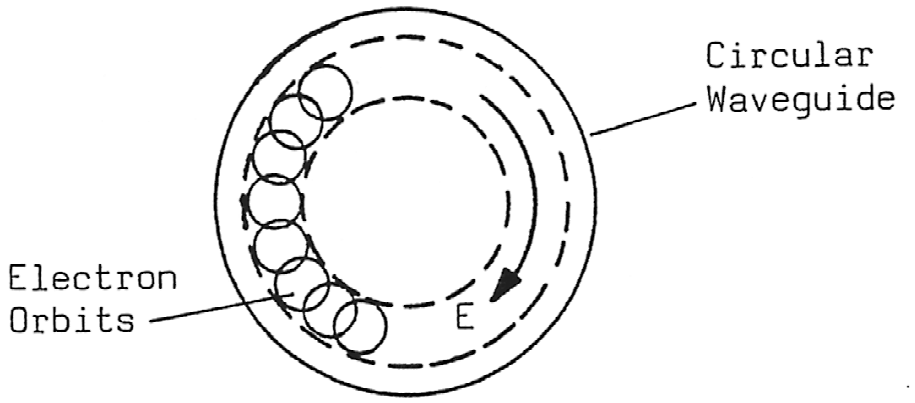


Fig. 27: Principle of operation of a gyrotron

Table 9: Characteristics of typical pulsed gyrotron oscillators

Type	**TH1504**	**TH1503B**	**Units**
Frequency	8	110	GHz
Beam Voltage	90	80	kV
Beam Current	27	15	A
Power	1	0.35	MW
Efficiency	41	29	%

Experimental gyrotron amplifiers have been made that are analogues of klystrons and travelling-wave tubes. These tubes are potentially better than klystrons for producing multi-megawatt RF pulses at frequencies of 20 GHz and above.

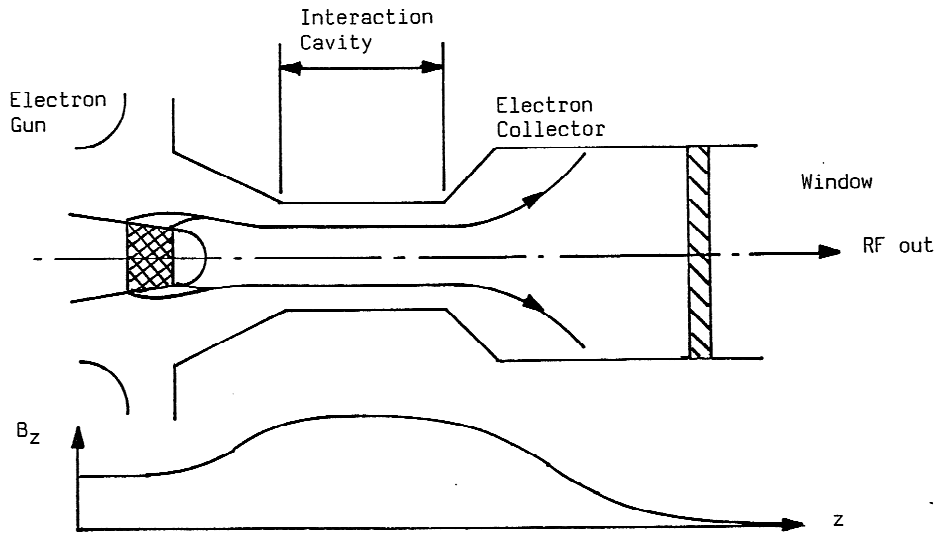


Fig. 28: Arrangement of a gyrotron oscillator

7 LIMITATIONS OF MICROWAVE TUBES

The performance of microwave tubes is limited by a number of factors, which operate in much the same way for both klystrons and gyrotrons. The chief of these are heat dissipation, voltage breakdown, output window failure, and multipactor discharges.

The RF structures and the windows of microwave tubes generally scale inversely with frequency. The maximum CW or average power that can be handled by a particular type of tube depends upon the maximum temperature that the internal surfaces can be allowed to reach. Now this temperature is independent of the frequency, so the power that can be dissipated varies inversely with the frequency. Gyrotrons can handle a higher power of the same frequency than klystrons because they have simpler structures and, if operated in a higher order mode, their structures are larger for a given frequency.

The power is also limited by the power that can be generated by an electron gun and formed into a beam. The beam diameter scales inversely with frequency and the beam current density is determined by the maximum attainable magnetic focusing field. Since the field is independent of frequency the beam current scales inversely with the square of the frequency. The beam voltage is related to the current by the gun perveance ($I/V^{1.5}$), which usually lies in the range 1.0 to 2.0 for power tubes. The maximum gun voltage is limited by the breakdown field in the gun and so varies inversely with frequency for constant perveance. These considerations suggest that the maximum power obtainable from a tube of a particular type varies as frequency to the power –2.5 to –3.0, depending upon the assumptions made. For pulsed tubes the peak power is limited by the considerations mentioned here and the mean power by those mentioned above. The efficiencies of tubes tend to fall with increasing frequency. This is partly because the RF losses increase with frequency and partly because of the design compromises that must be made at higher frequencies.

The maximum power obtainable from a pulsed tube is often determined by the power-handling capability of the output window. The output window of an external cavity klystron is in the form of a cylinder within the cavity and close to the output gap. This arrangement is limited to powers of about 70 kW. At higher power levels integral cavities are used and the power is brought out through waveguide or coaxial line windows. Very high power klystrons commonly have two windows in parallel to handle the full output power. Windows can be destroyed by excessive reflected power, by arcs in the output waveguide, by X-ray bombardment, and by the multipactor discharges described in the next Section. The basic cause of failure is overheating and it is usual to monitor the window temperature and to provide reverse power and waveguide and cavity arc detectors.

7.1 Multipactor discharge

Radio frequency vacuum breakdown can happen in a variety of ways, all described as multipactor discharges. The basic mechanism is illustrated by Fig. 29, which shows a pair of parallel metal plates in vacuum with a sinusoidally varying voltage between them. If an electron is liberated from one of the plates at a suitable phase of the RF field it will be accelerated towards the other plate and may strike it and cause secondary electron emission. If the phase of the field at the moment of impact is just 180° from that at the time when the electron left the first plate, then the secondary electrons will be accelerated back towards the first plate. These conditions make it possible for a stable discharge to be set up if the secondary electron emission coefficients of the surfaces are greater than unity. It is found that electrons liberated at a phase of the RF field lying within a certain range of the ideal one tend to reach the other electrode at a phase closer to the ideal.

The secondary emission coefficients of many materials vary with the energy of electrons with normal incidence according to the universal curve shown in Fig. 30. The constants of this curve for a number of materials used in vacuum tubes are given in Table 10.

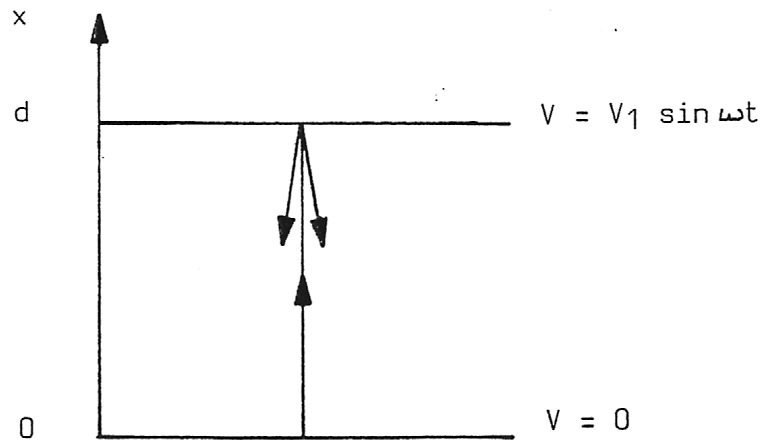


Fig. 29: Principle of the multipactor discharge

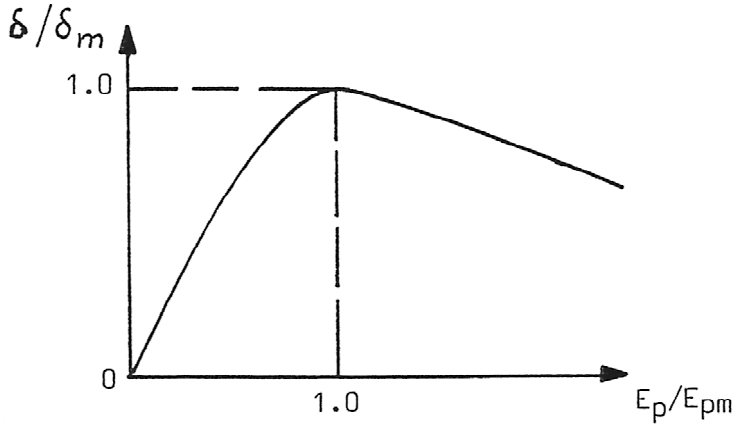


Fig. 30: Variation of secondary electron emission coefficient with primary electron energy

Table 10: Secondary electron emission coefficients of materials used in vacuum tubes

Material	δ_m	E_{pm} **(V)**
Copper	1.3	600
Platinum	1.8	800
Carbon	0.45	500
Alumina	2.35	500

A two-surface multipactor discharge can, therefore, only occur within a fairly limited range of voltages and products of frequency and electrode separation, as illustrated by Fig. 31. The limits in the vertical direction are set by the need for the secondary electron emission coefficient to be greater than unity. Table 10 shows that this happens for a small range of energies of the order of a few hundred electron volts. The limits in the horizontal direction are set by the need for the correct phase relationships. Figure 31 also shows the ranges in which higher order multipactor discharges can occur. The two-surface multipactor discharge typically involves currents of less than 1 A and voltages of a few hundred volts, so the power is moderate and the discharge is not normally destructive. It is probable that discharges of

this kind occur in most microwave power tubes and their main effect is to cause some additional loss and loading of the RF circuit.

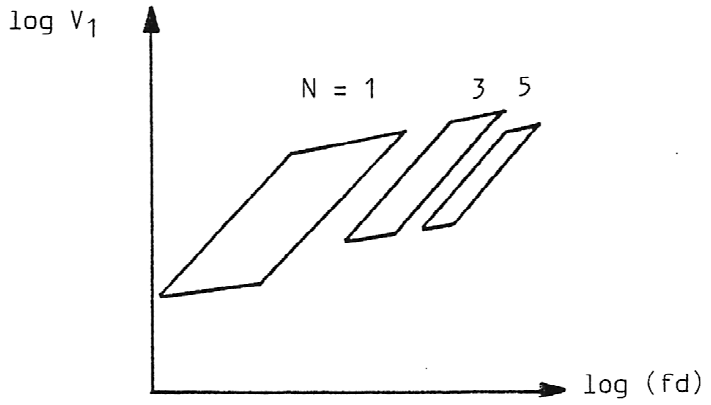


Fig. 31: Ranges in which two-surface multipactor discharges can occur

A more serious kind of multipactor discharge can occur in the presence of a magnetic field, as shown in Fig. 32. The electron trajectories are bent by the field so that the impacts made on the surfaces are oblique. This type of discharge can involve impacts on either one or two surfaces. Figure 33 shows the effects of oblique incidence on the secondary electron emission coefficient of a surface. First the peak value of δ is greater for oblique incidence than for normal incidence, and second, the range of energies over which δ is greater than unity is greatly increased. Thus crossed-field multipactor discharges can occur at much higher energies than the simple multipactor and they are, in consequence, potentially much more damaging. Because strong magnetic fields are used to focus linear-beam tubes it is quite possible for the conditions for crossed-field multipactor to exist somewhere within the tube. The manufacturer will normally have taken steps to ensure that this is not the case but, if the magnetic field around the tube is disturbed in any way (by the field of a circulator for example), then it is possible for a destructive discharge to occur.

It is also possible for multipactor discharges to occur on ceramic surfaces, with surface charge providing a static field. The local heating of a window ceramic in this way can be sufficient to cause window failure. Further information about multipactor discharges can be found in Ref. [26].

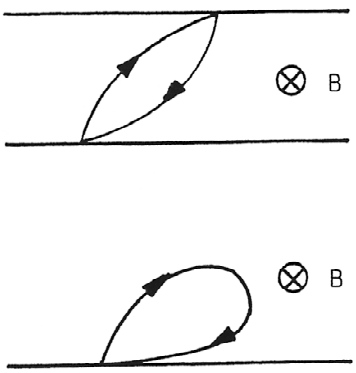


Fig. 32: Crossed-field multipactor discharges

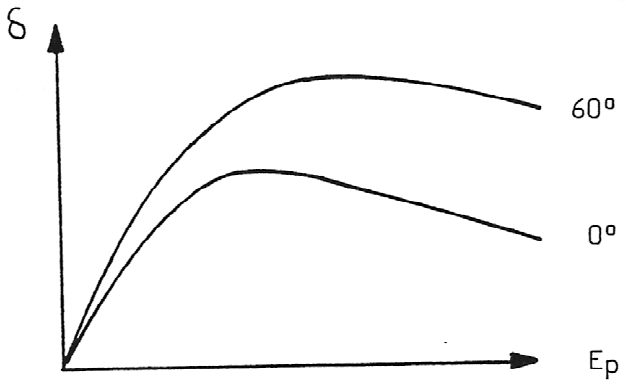


Fig. 33: Effect of the angle of incidence on secondary electron emission coefficient

8 COOLING AND PROTECTION

8.1 Cooling power tubes

The power tubes used in accelerators typically have efficiencies between 40% and 70%. It follows that a proportion of the d.c. input power is dissipated as heat within the tube. The heat to be dissipated is between 40% and 150% of the RF output power provided that the tube is never operated without RF drive. If a linear beam tube is operated without RF drive then the electron collector must be capable of dissipating the full d.c. beam power. The greater part of the heat is dissipated in the anode of a tetrode or in the collector of a linear-beam tube. These electrodes are normally cooled in one of three ways: by blown air (at low power levels); by pumped liquid (usually de-ionized water); or by vapour phase cooling. The last of these may be less familiar than the others and needs a little explanation.

The electrode to be cooled by vapour phase cooling is immersed in a bath of the liquid (normally de-ionized water), which is permitted to boil. The vapour produced is condensed in a heat exchanger, either in the cooling tank or part of an external circuit. The cooling system therefore forms a closed loop so that water purity is maintained. In all water cooling systems it is important to maintain the water purity to ensure that the electrodes cooled are neither contaminated nor corroded. Either of these effects can degrade the effectiveness of the cooling system and cause premature failure of the tube. In blown air systems careful filtering of the air is necessary for the same reasons.

It is important to remember that, in a high power tube, appreciable quantities of heat will be dissipated on parts of the tube other than the anode or collector, especially if a fault occurs during operation. It is common to provide air or water cooling for these regions too. Inadequate cooling may lead to the internal distortion or melting of the tube and its consequent destruction. Further information on the cooling of tubes is given in Refs. [7, 8].

8.2 Tube protection

Power tubes are very expensive devices and it is vital that they are properly protected when in use. The energy densities in the tubes and their power supplies are so high that it is easy for a tube to be destroyed if it is not properly protected. However, with adequate protection tubes are in fact very good at withstanding accidental overloads and may be expected to give long, reliable service.

Two kinds of protection are required. First, a series of interlocks must be provided to ensure that the tube is switched on in the correct sequence. It must be impossible to apply the anode voltage until the heater is at the correct working temperature, the cooling systems are functioning correctly, and

so forth. The exact switch-on sequence depends upon the tube type and reference must be made to the manufacturer's operating instructions. The sequence must also be maintained if the tube has to be restarted after tripping off for any reason.

The second provision is of a series of trips to ensure that power is removed from the tube in the event of a fault such as voltage breakdown, excessive reflected power, and so forth. Again, the range of parameters to be monitored and the speed with which action must be taken varies from tube to tube. Examples are coolant flow rate, coolant temperature, tube vacuum, output waveguide reverse power, and electrode over-currents. If a tube has not been used for some time it is sometimes necessary to bring it up to full power gradually to avoid repeated trips. The manufacturer's operating instructions should be consulted. If a tube trips out repeatedly it is wise to consult the manufacturer to avoid the risk of losing the tube completely by unwise action taken in ignorance of possible causes of the trouble. General information about tube protection and safe operation is given in Refs. [7, 8].

9 CONCLUSION

This paper has set out to review the main types of RF power source used in particle accelerators. In conclusion we review the state of the art of the different types of power source.

Figure 34 shows the state of the art for RF power sources in terms of their CW or mean powers as a function of frequency. Solid-state sources can only compete with tubes at the lower frequencies and power levels and even that requires massively parallel operation, as noted in Section 2. The fall-off in power output at high frequencies for each type of tube is related to the fundamental principles of its operation, as discussed in Section 7. The power achieved by klystrons at low frequencies does not generally represent a fundamental limitation but merely the maximum that has been demanded to date. For tetrodes and solid-state devices the maximum power is probably closer to the theoretical limits for those devices. In any case higher powers can be produced by parallel operation.

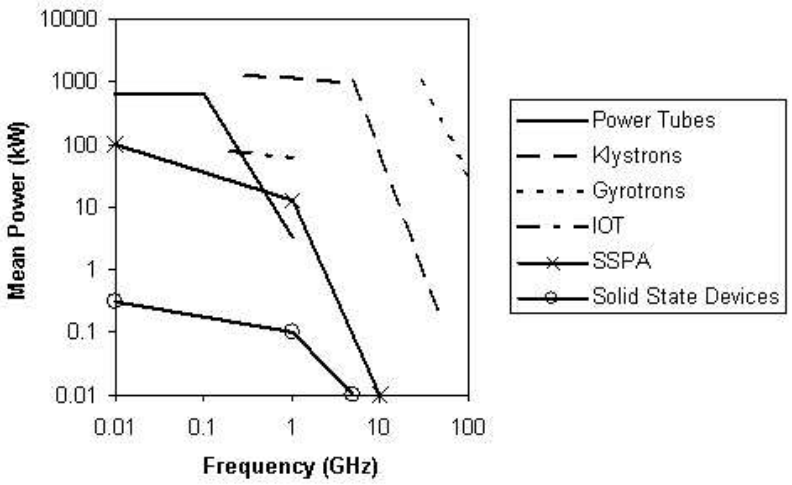


Fig. 34: Output power of state of the art of continuous wave sources

Figure 35 shows the state of the art of the devices discussed in terms of their efficiencies as a function of frequency. In general the efficiency of any kind of device falls with increasing frequency for the reasons discussed in Section 7. The overall efficiencies of linear-beam tubes designed for the lower power levels can be increased by collector depression. Super-power tubes do not have separately insulated collectors.

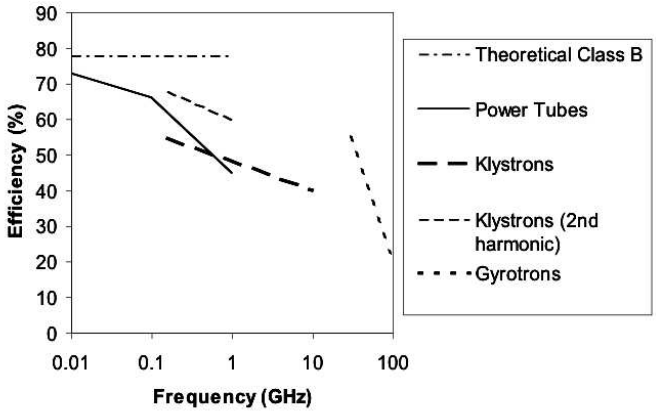


Fig. 35: Efficiencies of state-of-the-art continuous wave sources

Finally, Fig. 36 shows the state of the art for pulsed RF power sources. The figure includes conventional microwave tubes, relativistic derivatives of conventional tubes and gyrotrons. It also shows a number of other experimental devices which have not been discussed in this paper. Further information is given in Ref. [25].

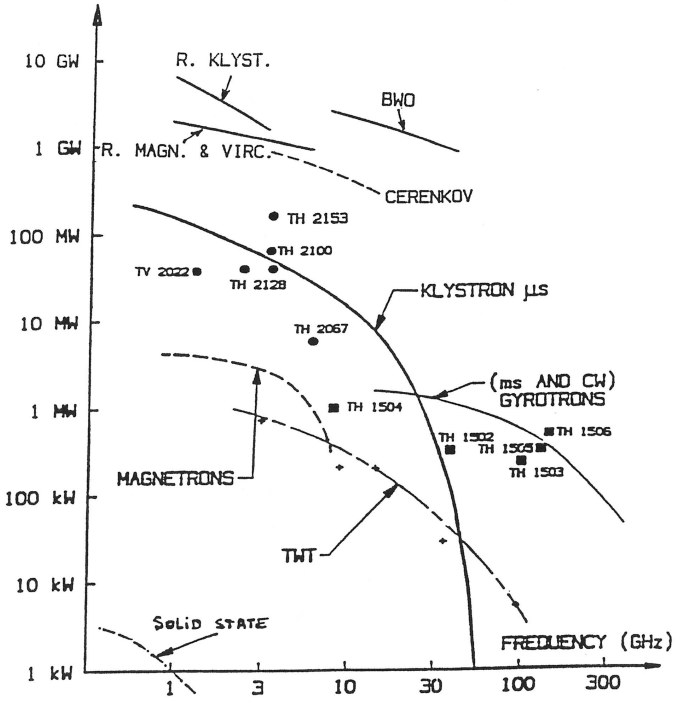


Fig. 36: Peak output power of state-of-the-art pulsed sources (Courtesy of Thales)

For further information on the theory of microwave tubes and for suggestions for background reading see Refs. [27–29].

ACKNOWLEDGEMENTS

This paper could not have been written without the generous help of a number of people. Special thanks are due to D. Carr and R. Heppinstall of Marconi Applied Technologies, H. Bohlen of CPI, H.P. Kindermann and H. Frischholz of CERN, V.P. Suller and his colleagues at the SERC Daresbury Laboratory, J. Stahl of Siemens, J. Vogel of Philips, F. Bernier and G. Faillon of Thomson Tubes Electroniques, and R.A. Rimmer of the Lawrence Berkeley Laboratory.

REFERENCES

[1] C. Zettler, RF systems for accelerators, CERN 87–10 (1987).

[2] C. Davis, J. Hawkins, and C. Einholf Jr., *IEEE Trans. Broadcasting*, **43** (1997) 252.

[3] T. Ikegami, A newly design UHF solid-state television transmitter, IEE Colloquium Digest No. 1988/16 (1988).

[4] CY1172 RF Power Tetrode data sheet, EEV Ltd., 1990.

[5] K. Spangenburg, *Vacuum Tubes*, McGraw-Hill (1948).

[6] W. Herdrich and H.P. Kindermann, RF power amplifier for the CERN SPS operating as LEP injector, Proc. Particle Accelerator Conf., Vancouver, 1985.

[7] Transmitting Tubes Data Book 1986/87, Siemens AG, 1986.

[8] Preamble–Tetrodes, EEV Ltd., 1976.

[9] T. Fujisawa et al., *Nuclear Inst. Methods Phys.Res.* **A292** (1990) 1.

[10] G. Schaffer, Components for high-power RF systems in modern accelerators, Lecture given at the 3rd Workshop New Techniques for Future Accelerators, Erice-Trapani, May 1987.

[11] W. Herdrich, H.P. Kindermann, and W. Sinclair, The RF power plant of the SPS, Proc. Particle Accelerator Conf., Santa Fe, 1983.

[12] A. Boussaton, G. Clerc, J.P. Ichac, and C. Robert, A new generation of gridded tubes for c.w. operation on new fusion magnetic machines, 1997 Symposium on Fusion Engineering, San Diego, California, October 6–10, 1997.

[13] G. Clerc, *Diacrode ®: Fundamentals and Performance*, Thomson Tubes Electroniques (1997).

[14] D.H. Preist and M.B. Shrader, The klystrode–an unusual transmitting tube with potential for UHF TV, Proc. IEEE, **70** (1982) 1318.

[15] H. Bohlen, Advanced high power microwave vacuum electron device development, Proc. 1999 Particle Accelerator Conference, New York, p. 445.

[16] H. Bohlen, CPI Inc., Private communication.

[17] G. Faillon, Technical and industrial overview of RF and microwave tubes for fusion, *Fusion Engineering and Design*, **46** (1999) 371.

[18] G. Faillon, Klystrons and related devices, Presented at the 48th Scottish Universities Summer School in Physics: Generation and Application of High Power Microwaves, St. Andrews, 1996.

[19] H. Bohlen *et al.*, Improved technological solutions for UHF power tubes, EEV Ltd., 1990.

[20] Z.D. Farkas *et al., IEEE Trans. Nucl. Sci.*, **NS-22** (1975) 1299.

[21] Z.D. Farkas, RF energy compressor, IEEE MTT Society, Int. Microwave Symposium Digest, 1980, p. 84.

[22] M.R. Boyd, R.A. Dehn, J.S. Hickey, and T.G. Mihran, *IRE Trans. Electron Devices*, **ED-9** (1962) 247.

[23] A. Beunas and G. Faillon, 10 MW/1.5ms, L-band multi-beam klystron, Proc. Conf. Displays and Vacuum Electronics, Garmisch-Partenkirchen, Germany, 1998, p. 257.

[24] Allen, RF power sources, Proc. 1988 European Particle Accelerator Conf., Rome, 1988, p. 101.

[25] V.L. Granatstein and I. Alexeff (eds.), *High-Power Microwave Sources*, Artech House (1987).

[26] J.R.M. Vaughan, *IEEE Trans. Electron Devices*, **ED-35** (1988) 1172.

[27] M.J. Smith and G. Phillips, *Power Klystrons Today*, Research Studies Press (1995).

[28] A.S. Gilmour, *Microwave Tubes*, Artech House (1986).

[29] L. Sivan, *Microwave Tube Transmitters*, Chapman and Hall (1994).

LOW-LEVEL RF SYSTEMS FOR SYNCHROTRONS
Part I: the low-intensity case

P. Baudrenghien
CERN, Geneva, Switzerland

Abstract
The low-level RF system generates the drive sent to the high-power equipment. It uses signals from the main bending magnet (B field) and from beam pick-ups (radial and longitudinal positions). It must minimize the beam losses and provide a beam with reproducible parameters (intensity, bunch length, average momentum and momentum spread) for either the next accelerator or the physicists. This presentation is the first of two: it considers the low intensity case where the voltage in the RF cavity is not influenced by the beam. The feedback loops that do not include the beam (accelerating field amplitude and cavity tuning) will not be presented.

1 BASIC NOTIONS ON ACCELERATION IN SYNCHOTRONS

1.1 The harmonic number

The RF system accelerates particles by producing a time-varying electric field in a cavity. At each turn the particle crosses the cavity. The RF frequency f_{rf} must stay locked in frequency to the revolution frequency f_{rev} otherwise the accelerating voltage will average to zero,

$$f_{rf} = h f_{rev} \ . \tag{1}$$

The integer *h* is called the *harmonic number*.

1.2 The stable phase

At each traversal of the cavity the particle receives an energy kick [1]–[3]

$$\Delta E = qV sin\phi_s \tag{2}$$

where q is the charge of the particle, V is the accelerating voltage and ϕ_s is the phase of the RF when the particle crosses the cavity (synchronous phase or stable phase). The corresponding momentum kick is

$$\Delta p = \frac{\Delta E}{2\pi R_0 f_{rev}} \ , \tag{3}$$

where $2\pi R_0$ is the machine circumference. The rate of change of momentum is thus

$$\frac{dp}{dt} = \Delta p \, f_{rev} = qV sin\phi_s \frac{1}{2\pi R_0} \ . \tag{4}$$

To stay on the centred orbit, the momentum must follow the B field [1]–[3]

$$p = q\rho B \ , \tag{5}$$

where ρ is the bending radius. During acceleration the RF voltage and stable phase will thus obey the relation

$$V sin\phi_s = 2\pi R_0 \rho \frac{dB}{dt} \ . \tag{6}$$

Once the RF voltage is defined, the above relation can be inverted to give the stable phase as a function of the RF voltage and the time derivative of the B field

$$\phi_s = \arcsin\left(2\pi R_0 \rho \frac{\frac{dB}{dt}}{V}\right) \ . \tag{7}$$

The above function is called the stable phase programme (see Fig. 1).

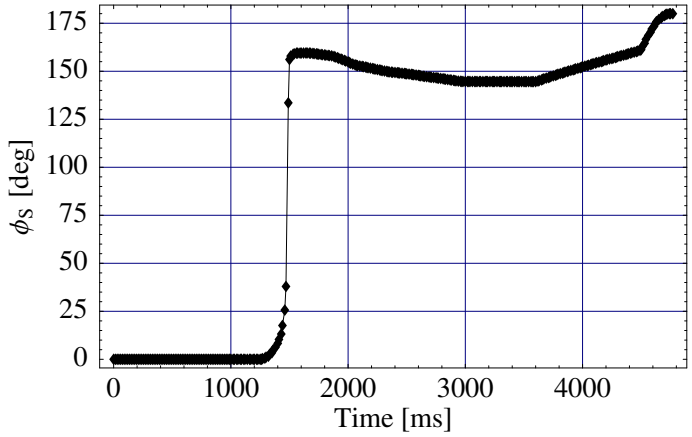


Fig. 1: Variation of the stable phase during the proton cycle in the CERN SPS (acceleration from 14 GeV to 450 GeV). Transition energy (see Section 1.4) at 23 GeV crossed around time 1500 ms with a phase jump from ϕ_s^- to $\phi_s^+ = \pi - \phi_s^-$). Courtesy of T. Bohl, CERN SL/HRF.

1.3 Synchrotron oscillations

By definition the synchronous particle is the particle that stays at the stable phase during the acceleration, i.e. the RF phase when it crosses the cavity is ϕ_s. Its momentum, p_s, is such that it stays on the centre orbit.[1] We define the synchronous RF frequency ω_s as $2\pi h$ times the revolution frequency of the synchronous particle. A bunch consists of many particles. Each particle is characterized by its coordinates in phase and momentum: let $\phi = \phi_s + \delta\phi$ be the phase of the RF when the particle crosses the cavity and let $p = p_s + \delta p$ be its momentum. We now derive the equation of motion in the presence of a small modulation of the RF frequency $\omega_{rf} = \omega_s + \delta\omega_{rf}$. First the momentum gain per turn:

$$\frac{d}{dt}\delta p = \frac{qV}{2\pi R_0}\left(sin(\phi_s + \delta\phi) - sin\phi_s\right) \approx \frac{qV\,cos\,(\phi_s)}{2\pi R_0}\delta\phi \ . \tag{8}$$

The above linear approximation is valid for small phase deviations only. A second equation relates the phase slip to the difference in beam frequency f_b and the RF frequency modulation:

$$\frac{d}{dt}\delta\phi = -2\pi\,\delta f_b + \delta\omega_{rf} \ , \tag{9}$$

where δf_b is the particle frequency deviation that can be related to the momentum deviation via the slippage factor (Eq. 10)[2]

$$\eta = -\frac{\frac{\delta f_b}{f_b}}{\frac{\delta p}{p}} = \frac{1}{\gamma_t^2} - \frac{1}{\gamma^2} \ , \tag{10}$$

[1] The subscript s refers to the synchronous particle.

[2] Some authors define η as $\frac{\frac{\delta f_b}{f_b}}{\frac{\delta p}{p}}$.

where $\gamma = E/E_0$ (E_0 being the rest energy) and γ_t is the value of γ at the transition energy. Recall that $\phi_s + \delta\phi$ is the phase of the RF when the particle crosses the cavity while $f_s + \delta f_b$ is the particle frequency. A faster particle will cross the cavity earlier, i.e. with a negative $\delta\phi$. This explains the minus sign in Eq. (9). By differentiating Eq. (9) and using Eqs. (10) and (8) we get

$$\frac{d^2}{dt^2}\delta\phi + \Omega_s^2\,\delta\phi = \frac{d\delta\omega_{rf}}{dt} \quad . \tag{11}$$

Similarly, by differentiating Eq. (8) and using Eqs. (9) and (82) we get

$$\frac{d^2}{dt^2}\delta p + \Omega_s^2\,\delta p = \frac{qV\,cos\,(\phi_s)}{2\pi R_0}\delta\omega_{rf} \quad . \tag{12}$$

Ω_s is called the synchrotron frequency

$$\Omega_s = \sqrt{-\frac{hqc}{2\pi R_0^2}\frac{\beta_s}{p_s}\eta_s V\,cos\,(\phi_s)} \quad , \tag{13}$$

where β_s is the normalized velocity v_s/c of the synchronous particle. The equation of motion is one of an *undamped resonator excited by the RF frequency noise*. Its resonant frequency Ω_s varies during the acceleration. It depends on the synchronous momentum via the parameters p_s, β_s and γ_s , and on the RF parameters V and ϕ_s. Figure 2 shows a mechanical analogy of the longitudinal motion: an object of mass m is free to move on an horizontal axis. It is subject to the force of a spring of strength k. Let $x(t)$

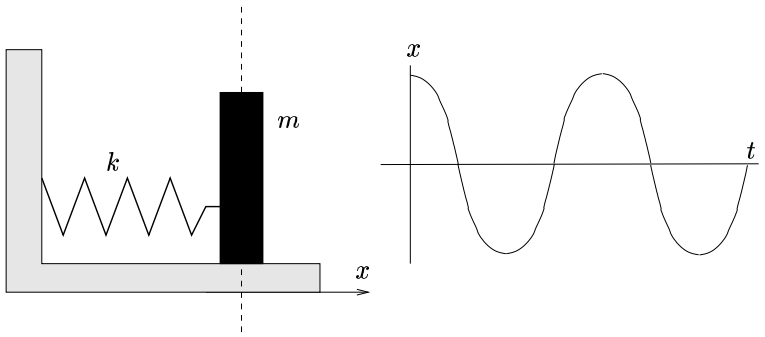


Fig. 2: Mechanical analogy of the synchrotron oscillation

be its position at time t, the equation of motion is

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0 \quad . \tag{14}$$

Starting with an initial displacement x_0 and zero velocity, the solution is

$$x(t) = x_0\,cos\,(\Omega t) \tag{15}$$

with $\Omega = \sqrt{\frac{k}{m}}$. The oscillation lasts forever because there is no damping term (first order derivative) in the equation of motion. The motion can also be represented in the $(x,\, \frac{dx}{dt})$ plane, called the phase plane. Since

$$\frac{dx}{dt} = -x_0\Omega\,sin\,(\Omega t) \quad , \tag{16}$$

we have

$$\left(\frac{x}{x_0}\right)^2 + \left(\frac{\frac{dx}{dt}}{x_0\Omega}\right)^2 = 1 \quad . \tag{17}$$

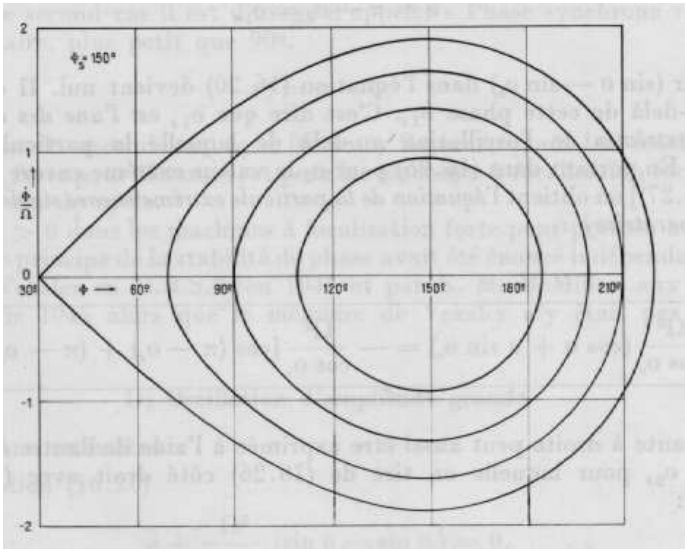


Fig. 3: Phase space trajectory in the $(\phi, \frac{1}{\Omega_s}\frac{d\phi}{dt})$ plane. Case above transition: The particles move clockwise along the trajectories. The stable phase is $150°$. The vertical axis can also be labelled in momentum, energy, frequency or radial position. (Reproduced from Ref. [2]).

The trajectory is an ellipse in phase plane. Similarly the particles that have small phase and momentum deviations with respect to the synchronous particle describe small ellipses around the $(\phi_s, 0)$ point in the $(\phi, \frac{1}{\Omega_s}\frac{d\phi}{dt})$ plane (Fig. 3). As we consider particles with

larger and larger deviations, their trajectory become less elliptic. Recall that we have linearized the equation of motion (8). The largest trajectory shown is called the *separatrix*: its width depends on the stable phase only. In normalized units $\frac{1}{\Omega_s}\frac{d\phi}{dt}$, its height, also depends on the stable phase only. One often uses momentum or energy for the vertical axis. With these units the height of the separatrix is proportional to the square root of the RF voltage [2]. Outside the separatrix the particle trajectories are not closed: These particles are not *captured* by the RF and will be lost as the acceleration proceeds. The area inside the separatrix is called the *RF bucket*. The formula for the synchrotron frequency (13) was derived using the linearized equation of motion. It is exact for small trajectories around the synchronous particle. For larger deviations, the frequency of the oscillation is decreased and finally vanishes on the separatrix [2].

1.4 Transition energy

The slippage factor η is negative at low energy $(\gamma \leq \gamma_t)$ and positive above the transition energy. This change of sign is easily understood: a particle with a momentum in excess (positive $\frac{\delta p}{p}$) will travel faster than the synchronous particle but it will travel on an off-centred orbit since its mass is also larger. Its orbit will thus be longer. At low energy the velocity increase wins over the orbit lengthening and the frequency will increase (negative η). As the energy gets higher, the particle velocity approaches the speed of light and the winning factor is the orbit lengthening (positive η). The equation of motion (11) will be stable only if Ω_s^2 is positive. We thus conclude from Eq. (13) that the sign of $cos\,(\phi_s)$ must change at transition: below transition the particles move counter-clockwise in phase space and $cos\,(\phi_s)$ is positive (ϕ_s will typically be between $0°$ and $30°$). Above transition $cos\,(\phi_s)$ is negative (ϕ_s between $150°$ and $180°$). The phase of the RF must jump at transition as shown in Fig. 1 for the proton cycle in the CERN SPS. Such a critical RF manipulation is possible because the synchrotron frequency goes to zero at transition. The dynamics of the bunch are thus very slow at that critical moment.

1.5 The accelerating bucket

The bunch consists of many particles, each undergoing synchrotron oscillations in phase space. The *longitudinal emittance* of the bunch is the area that it fills in (energy,time) space. It is often measured in eVs. Given the triplet (E_s, V, ϕ_s) we can compute the RF bucket area A in eVs [4]

$$A(E_s, V, \phi_s) = \left(\frac{16\sqrt{e}}{(2\pi)^{\frac{3}{2}}\sqrt{h}}\right) \times \left(\frac{\beta_s\sqrt{E_s}}{f_{rf}\sqrt{|\eta_s|}}\right) \times \left(\sqrt{V}\alpha(\phi_s)\right) \quad . \tag{18}$$

The first factor is constant (if the harmonic number is kept constant). The second factor depends on the energy only. The third factor shows the influence of the RF voltage and the stable phase: $\alpha(\phi)$ is a non-linear function of the stable phase angle. It is equal to one for 0° and 180° (stationary bucket) and drops to 0.3 for 32° or 148° [4]. At injection, the energy spread and bunch length are defined by the injector. We must match the RF bucket to these parameters and this defines the bucket area A_0. From Liouville's theorem we know that the emittance is invariant during acceleration [2]. The bucket area should thus also be kept constant. In practice, however, the RF manipulations will blow up the emittance and we wish to increase the bucket area at some critical points in the cycle (just after transition for example). Once the desired bucket area is defined through the cycle, we can merge Eqs. (18) and (6) to get a system of two equations with two unknowns V and ϕ_s. By solving this we obtain the appropriate values for V (see Fig. 4) and ϕ_s (see Fig. 1) through the acceleration cycle.

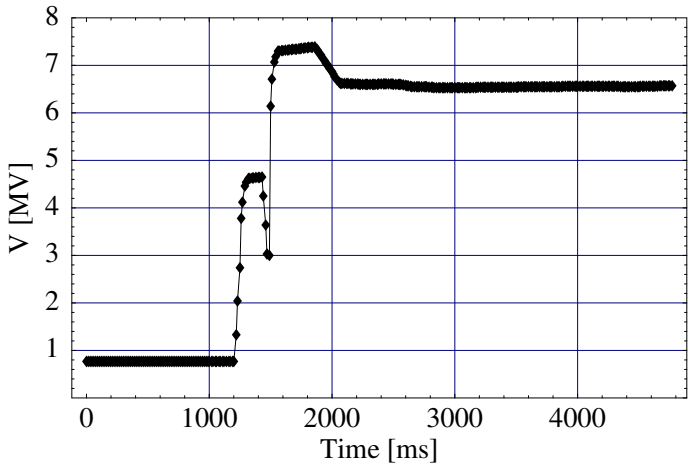


Fig. 4: Variation of the RF voltage during the proton cycle in the CERN SPS (acceleration from 14 GeV to 450 GeV. Transition at 23 GeV). Courtesy of T. Bohl, CERN SL/HRF.

1.6 Bunch transfer function for dipolar motion

In Section 1.3 we examined the motion of a single particle. Let us now consider the bunch as a statistical collection of many particles, each particle oscillating around the synchronous phase. We wish to derive an equation for the motion of the centre of charge of the bunch. Let us freeze the motion at a given instant t and let $f(x)$ be the probability density function for the phase ϕ of the particles. ($f(x)dx$ is the probability that the RF phase be between x and $x + dx$ when the particle crosses the cavity.) The first order moment (centre of charge) $\hat{\phi}$ is

$$\hat{\phi} = \int \phi f(\phi) d\phi \tag{19}$$

where the integration covers the entire bunch. Let $\delta\hat{\phi}$ be $\hat{\phi} - \phi_s$. We have

$$\delta\hat{\phi} = \hat{\phi} - \phi_s = \int \phi f(\phi) d\phi \; - \phi_s \quad . \tag{20}$$

By definition the probability density function integrates to 1

$$\int f(\phi) d\phi = 1 \tag{21}$$

so that Eq. (20) can be rewritten

$$\hat{\delta\phi} = \int (\phi - \phi_s)\, f(\phi) d\phi = \int \delta\phi f(\phi) d\phi \ . \tag{22}$$

Let us now assume a stationary distribution so that $f(x)$ does not depend on t. Taking the second derivative of the above equation we get

$$\frac{d^2\hat{\delta\phi}}{dt^2} = \int \frac{d^2\delta\phi}{dt^2} f(\phi) d\phi \ . \tag{23}$$

Now using the synchrotron equation (11) for each particle in the bunch

$$\frac{d^2\hat{\delta\phi}}{dt^2} = \int \left(-\Omega_s^2\, \delta\phi + \frac{d\delta\omega_{rf}}{dt} \right) f(\phi) d\phi \tag{24}$$

$$\frac{d^2\hat{\delta\phi}}{dt^2} + \Omega_s^2 \int \delta\phi f(\phi) d\phi = \frac{d\delta\omega_{rf}}{dt} \int f(\phi) d\phi \tag{25}$$

$$\frac{d^2\hat{\delta\phi}}{dt^2} + \Omega_s^2\, \hat{\delta\phi} = \frac{d\delta\omega_{rf}}{dt} \ . \tag{26}$$

The above equation is identical to the synchrotron equation describing the motion of each particle. These cannot be individually observed by the instrumentation. On the other hand the motion of the centre of charge of the bunch, also called *dipolar motion*, can easily be monitored.

1.7 The RF frequency

During the acceleration the RF frequency must stay locked to the revolution frequency of the bunch

$$f_{rf} = h f_{rev} = h \frac{v_s}{2\pi R_0} = \frac{hc}{2\pi R_0} \beta_s \tag{27}$$

where the normalized velocity β_s is a function of the momentum p_s

$$\beta_s = \frac{1}{\sqrt{1 + \left(\frac{E_0}{c p_s}\right)^2}} \ . \tag{28}$$

During acceleration the momentum follows the B field according to Eq. (5) and we get

$$f_{rf} = \frac{hc}{2\pi R_0} \frac{B}{\sqrt{B^2 + \left(\frac{1}{c\rho}\frac{E_0}{q}\right)^2}} = f_\infty \sqrt{1 - \frac{1}{\gamma_s^2}} \ , \tag{29}$$

where

$$\gamma_s = \frac{E_s}{E_0} \tag{30}$$

and

$$f_\infty = \frac{hc}{2\pi R_0} \ . \tag{31}$$

The above equation for f_{rf} is called the *frequency programme*. We observe that:

- the RF frequency varies with the energy (or the magnetic field) in a non-linear fashion;
- it can be controlled from a measurement of the magnetic field;
- the frequency swing depends on the range of γ from injection to extraction. For a highly relativistic machine (leptons) where $\gamma \gg 1$, the RF frequency can be kept constant. On the other hand low energy hadron machines need a precise frequency programme. This is also the case for ion acceleration (large $\frac{E_0}{q}$ ratio). Some examples:

  Lepton ($E_0 = 0.511$ MeV) acceleration in the CERN SPS from 3 GeV to 22 GeV at constant frequency 200.395 MHz.

  Proton ($E_0 = 938$ MeV) acceleration in the CERN LHC from 450 GeV (400.789 MHz) to 7 TeV (400.790 MHz).

  Lead ions (208Pb82+) acceleration in the CERN SPS from 5.114 GeV/u (energy per nucleon) at 197.072 MHz to 160 GeV/u (200.393 MHz).

  Original proton acceleration in the CERN PS (1959, h = 20) from 50 MeV (2.9 MHz) to 25 GeV (9.54 MHz) [5];
- the choice of accelerating cavity is dictated by the frequency swing: for a large frequency swing, ferrite cavities are preferred for their important tuning range [6]. For relativistic beams, high Q cavities operated at higher frequencies are the common choice [7].

2 WHY DO WE NEED A LOW-LEVEL SYSTEM?

2.1 The simplest RF system

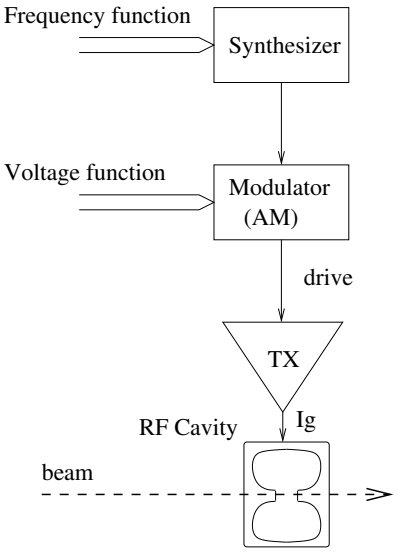


Fig. 5: Simplest RF system. No feedback loop.

The simplest accelerating system for a circular accelerator or collider is shown in Fig. 5. It consists of:

- an accelerating cavity;
- driven by a power amplifier (TX);
- whose drive is adjusted in amplitude by a modulator (to vary the accelerating voltage during the ramp in order to keep the desired bucket area A);
- and whose frequency is controlled by a synthesizer to follow the desired function $f_{rf}(B)$.

2.2 What will go wrong?

In general the performance of the above low-level system will not be good for the following reasons:

- The magnetic field fluctuates, so the RF frequency f_{rf} is not correct. This causes a displacement of the orbit δR given by Eq. (80)
$$\frac{\delta R}{R} = \frac{\gamma^2}{\gamma_t^2 - \gamma^2} \frac{\delta f_{rf}}{f_{rf}} \tag{32}$$
- If the transfer from the injector is of the bunch into bucket type, the parameters of the injected beam will not be perfectly stable: the phase will be slightly wrong and the energy will not be perfectly matched to the magnetic field of the receiving machine (radial position error on the first turn).
- The RF synthesizer injects phase noise that shakes the bunches as is evident from Eq. (26).
- The power amplifiers have ripples at multiples of 50 Hz.
- The gain and phase shift of the power amplifiers may drift.
- The beam current will modify the accelerating voltage (beam loading).
- etc ...

The classic control solution is to use feedback loops measuring the slowly drifting parameters and adjusting the RF settings accordingly.

3 WHAT CAN WE MEASURE?

3.1 The magnetic field

The magnetic field can be measured with a high precision: in the CERN SPS for example it is measured with a precision of 3×10^{-5} T over a range of 2 T.

3.2 The radial position

Three factors limit the precision of single turn measurements of the radial position of the beam:

- The sensitivity to the beam intensity: in the SPS the precision is 0.5 mm for a 1 to 10 range of beam intensity [8]. In the LHC the design precision is 0.1 mm for a 1 to 30 (pilot to nominal) range of beam intensity [9].
- The transverse betatron oscillation: the RF is interested in the *mean* radial position of the beam (that is related to the beam energy). If the beam is not injected on the centred trajectory, it will execute a transverse betatron oscillation that must be filtered out of the part interesting the RF.
- The local orbit distortion: usually we use only one or two radial pick-ups in the RF low-level system. These sample the position at one point in the machine. As a consequence the measurement will differ from the mean radial position in the case of a closed orbit distortion at the pick-up location.

3.3 The beam phase (longitudinal position)

The phase of the beam is defined as *the phase of the Fourier component of the beam current at the RF frequency*. (Note that the beam phase is defined by the beam *current* while the RF phase refers to the *voltage* in the cavity. For a beam at stable phase zero below transition, the beam phase will thus lead the cavity phase by $\pi/2$. Above transition, a beam at stable phase π will lag the cavity phase by $\pi/2$.) If the buckets are not evenly filled around the machine the beam current will have a strong amplitude modulation at the revolution frequency. Its spectrum thus shows side-bands at $f_{rf} \pm n f_{rev}$. These must be filtered out of the beam phase signal to avoid exciting higher order coupled bunch dipole oscillations, ($n \geq 1$), with the phase loop.

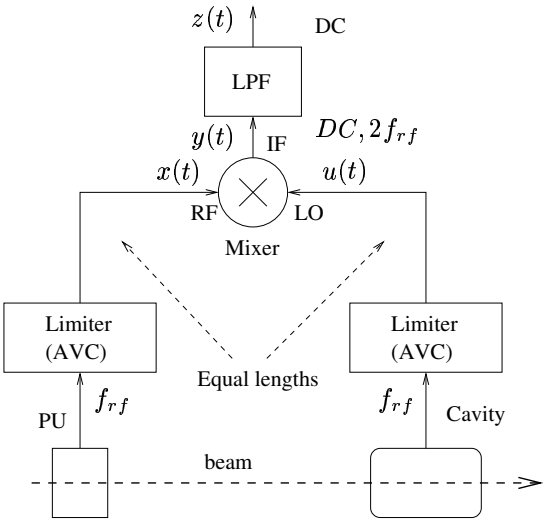


Fig. 6: Direct measurement of the beam/cavity phase. Processing at the RF frequency.

Figure 6 shows a method to measure the phase of the beam with respect to the cavity. The beam signal is generated by a narrow-band pick-up (cavity resonator) centred at the RF frequency. It is followed by a limiter (AVC) that provides a constant RF level at its output so that the phase measurement is insensitive to the beam intensity. The signal from the cavity is processed in an identical chain with an equal delay so that the signals at the input of the mixer stay in phase as the RF frequency varies. Let ϕ_b be the phase of the beam and ϕ_c be the phase of the cavity

$$x(t) = cos(2\pi f_{rf} t + \phi_b) \ , \tag{33}$$

$$u(t) = cos(2\pi f_{rf} t + \phi_c) \ . \tag{34}$$

The mixer multiplies the signals on its RF and LO ports

$$y(t) = cos(2\pi f_{rf} t + \phi_b) cos(2\pi f_{rf} t + \phi_c) = \frac{cos(4\pi f_{rf} t + \phi_b + \phi_c) + cos(\phi_b - \phi_c)}{2} \tag{35}$$

$$z(t) = \frac{cos(\phi_b - \phi_c)}{2} \ . \tag{36}$$

The Low Pass Filter (LPF) must provide a large attenuation at multiples of the revolution frequency. An alternative is to sample its output at the revolution frequency to produce one phase information per turn.

In theory, phase measurements are insensitive to the beam intensity. In practice the phase shift introduced by the limiter in the beam signal path will vary with the level of its input signal. The system of Fig. 6 is used to accelerate the LHC proton beam in the CERN SPS. The frequency swing is only 130 kHz at 200 MHz. Narrow-band limiters have been designed with $\pm 3^\circ$ phase shift for 60 dB input dynamic range in the above frequency range. When the frequency swing is large the system shown in Fig. 7 is preferred. The beam signal from a wide band pick-up (Electrostatic or Wall Current Monitor) is heterodyned to an Intermediate Frequency (IF) and followed by a fixed Band-Pass Filter (BPF) at the IF. For large frequency swings this solution presents the advantage of processing the signal at a fixed intermediate frequency (AGS booster at BNL [10], fixed target beam in the CERN SPS, CERN PS [11]). The LO frequency remains at a fixed frequency offset with respect to the RF during the acceleration: $f_{lo} = f_{rf} - f_{if}$. The BPF at the intermediate frequency selects two bands in the pick-up signal: $f_{lo} + f_{if} = f_{rf}$ and $f_{lo} - f_{if} = f_{rf} - 2f_{if}$. This latter band is called the image. If necessary the noise present in this band can be rejected with the RF BPF. The IF BPF must be narrow enough to reject the sidebands at $f_{rf} \pm n f_{rev}$. (In the CERN SPS we use an IF at 10.7 MHz. The BPF is a 10.7 MHz crystal filter that rejects the 43 kHz side-bands.)

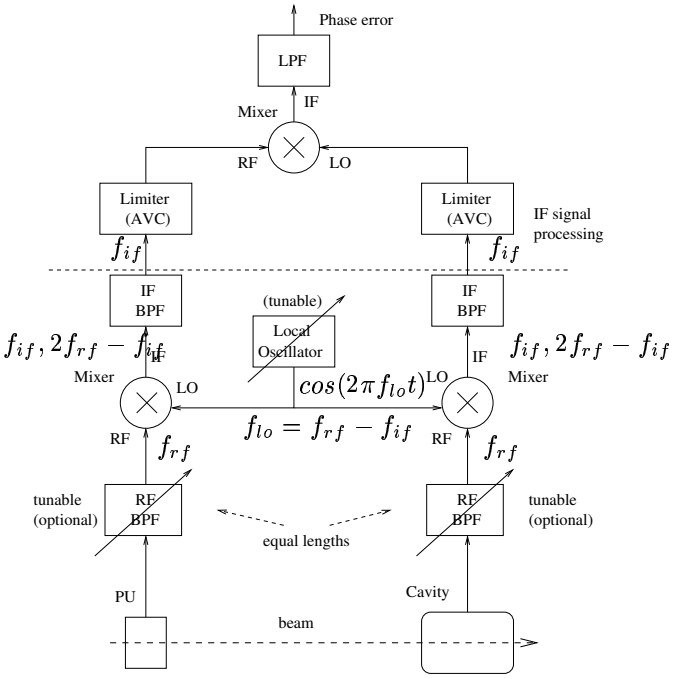


Fig. 7: Heterodyne system for the measurement of the beam/cavity phase. Processing at the IF frequency.

4 WHAT SHOULD BE CONTROLLED?

4.1 Controlling the RF phase seen by the beam: the phase loop

Motivation: The noise in the RF system will excite coherent longitudinal oscillations of the beam. Such an oscillation will also be triggered by a phase or energy error at injection if transfer is of the bunch-into-bucket type. In lepton machines the synchrotron light may provide sufficient natural damping. In hadron machines, however, these oscillations will persist, resulting in emittance blow-up due to filamentation and finally beam loss because the bucket area will be too small. Figure 8 shows a simulation of the injection into the 200 MHz LHC bucket with a small phase and energy error. The spread in synchrotron frequency as a function of the deviation from the synchronous particle causes a *filamentation* of the bunch: the particles with large deviations have smaller synchrotron frequencies and drag behind the core of the bunch. The end result is a severe and often unacceptable blow-up of the emittance.

The cure is easily found if we refer to our mechanical analogy shown in Fig. 2. We must add a friction term (i.e. a force proportional to the velocity as shown in Fig. 9) so that the phase or energy error is quickly damped to zero.

The phase loop is shown in Fig. 10:

The phase of the beam ϕ_b is compared to the phase of the RF in the cavity ϕ_c and the error[3] $\delta\phi_{b,c} = \phi_c - \phi_b + \pi/2 - \phi_s$ is used to correct the frequency of the RF (generated by the Voltage Controlled Oscillator, VCO) via the feedback $\delta\omega_{rf}$. By averaging over the bunches in the beam, Eq. (26) for the synchrotron oscillation of the centre of charge of each bunch transforms into an identical equation for the whole beam

$$\frac{d^2\delta\phi_{b,c}}{dt^2} + \Omega_s^2\,\delta\phi_{b,c} = \frac{d\delta\omega_{rf}}{dt} \quad . \tag{37}$$

We now introduce a feedback term

$$\delta\omega_{rf} = -k_\phi\,\delta\phi_{b,c} \tag{38}$$

[3] The $\pi/2$ term is due to the comparison of the phase of the beam *current* to the phase of the cavity *voltage*. For a beam at stable phase zero below transition, the beam phase will thus lead the cavity phase by $\pi/2$.

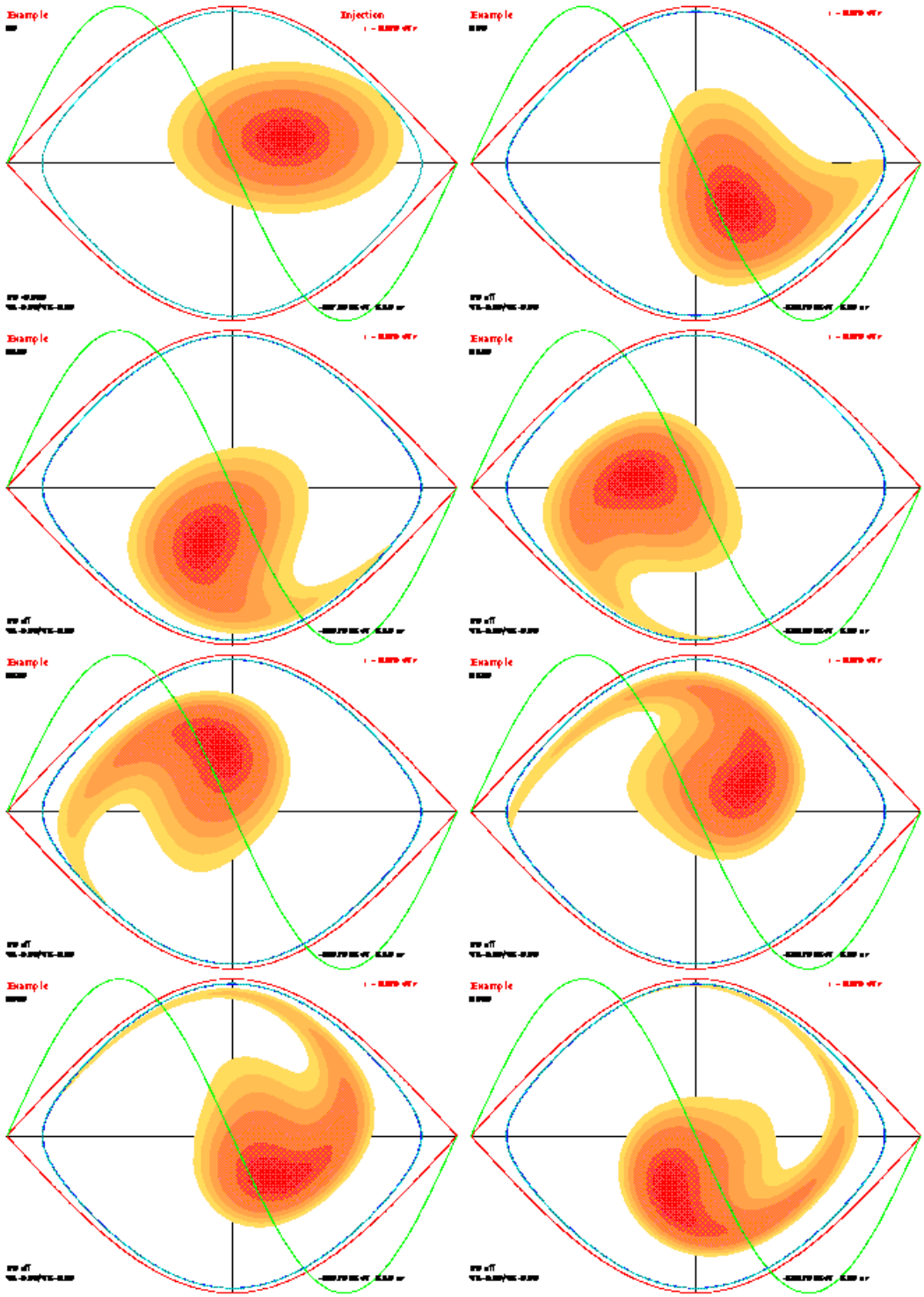


Fig. 8: Phase and energy error at injection into the LHC (450 GeV, 4 MV RF at 200 MHz, $f_{rev} = 11$ kHz). Simulation of the filamentation of the bunch. First turn at top left. Later plots to be read from left to right and from top to bottom. 50 turns between plots. Courtesy of J. Tuckmantel, CERN SL/HRF.

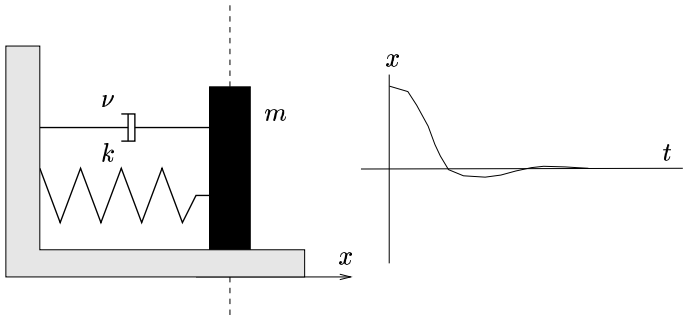


Fig. 9: Damping of our mechanical analogy for the synchrotron oscillation

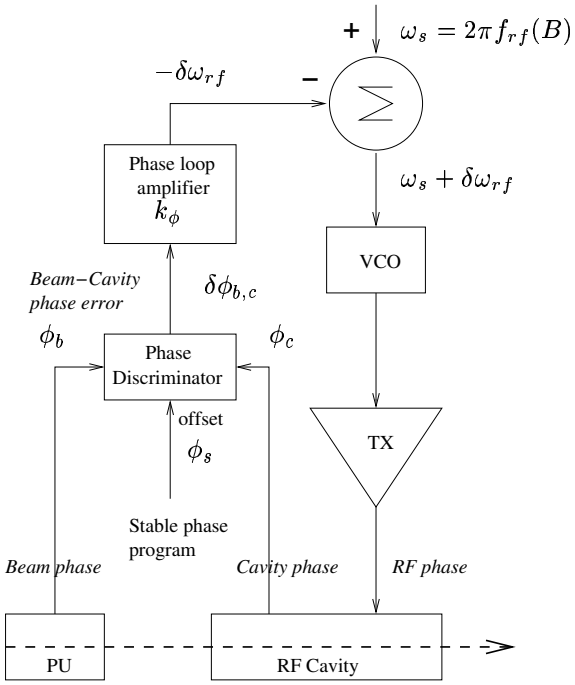


Fig. 10: Phase loop

and this provides the desired damping

$$\frac{d^2\delta\phi_{b,c}}{dt^2} + k_\phi \frac{d\delta\phi_{b,c}}{dt} + \Omega_s^2\,\delta\phi_{b,c} = 0 \ . \tag{39}$$

Remarks

- The stable phase ϕ_s is computed from a measurement of the RF voltage and the rate of change of the magnetic field (stable phase programme (7)) and fed as an offset in the phase discriminator so that the range and linearity of this device are not important. (Note that the systems presented in Section 3.3 generate a signal proportional to the cosine of the phase error).
- A static error in the stable phase programme will introduce a frequency error if the phase loop amplifier is DC-coupled. To avoid this, the phase loop can be AC-coupled. This method is used in the PS accelerator at CERN. It is explained in detail in Ref. [12].
- The phase of the RF must change by 180° at transition. The stable phase programme jumps from ϕ_s^- to $\phi_s^+ = \pi - \phi_s^-$.

- As already mentioned, the components at nf_{rev} in the phase error signal $\delta\phi_{b,c}$ must be carefully filtered out to avoid exciting coupled bunch instabilities.
- The poles of the closed-loop transfer function are

$$p_{\pm} = -\frac{k_\phi}{2} \pm \frac{k_\phi}{2}\sqrt{1 - \left(\frac{2\Omega_s}{k_\phi}\right)^2} \quad . \tag{40}$$

  For small phase loop gain $k_\phi < 2\Omega_s$ we have two complex conjugate poles. The impulse response is a damped sine wave. For $k_\phi = 2\Omega_s$ the two poles merge on the real axis. The system is critically damped and no oscillatory behaviour occurs. For larger gains $k_\phi > 2\Omega_s$ one pole moves to a large negative value on the real axis (i.e. very fast damping) but the other pole moves to the origin. The system becomes unstable at very low frequency. It presents the behaviour of an integrator. *With a phase loop alone one cannot increase k_ϕ much above critical damping*. But this limitation will disappear when we combine the phase loop with another loop.
- With a phase loop the RF noise is less likely to blow up the emittance (see Section 5.1).

4.2 Controlling the radial position: the radial loop

Motivation: In large machines the transverse aperture is usually very small. In the CERN SPS for example ($R = 1100$ m) we accept an average orbit displacement of a few millimetres only ($\frac{\delta R}{R} = 10^{-6}$). An error in the RF frequency will cause a displacement of the orbit given by Eq. (80)

$$\frac{\delta R}{R} = \frac{\gamma^2}{\gamma_t^2 - \gamma^2}\frac{\delta f_{rf}}{f_{rf}} \quad . \tag{41}$$

The required precision for the RF frequency therefore depends on the energy. It is infinite at transition! As a consequence it is difficult to control the beam radial position if the RF frequency is adjusted from a measurement of the B field only (Eq. 29). With a radial loop the required precision can be much relaxed.
Method: Figure 11 shows the radial feedback loop. We measure the radial displacement of the beam and correct the RF frequency accordingly.

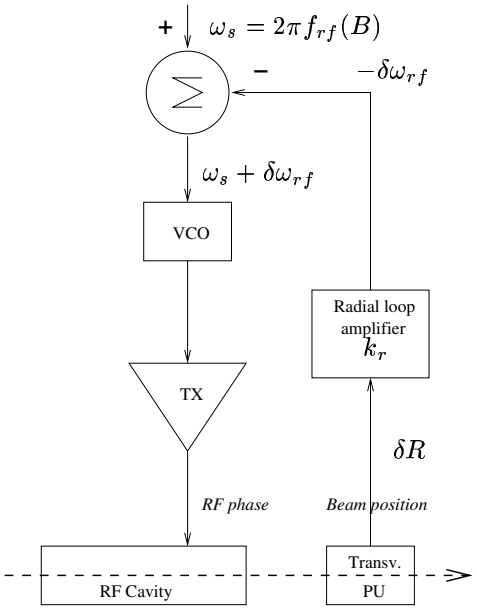


Fig. 11: Radial loop

The radial displacement δR is proportional to the momentum deviation δp, Eq. (81)

$$\delta R = \frac{R_0}{p_s \gamma_t^2} \delta p \tag{42}$$

With the above equation we can rewrite Eq. (12)

$$\frac{d^2}{dt^2} \delta R + \Omega_s^2 \, \delta R = \frac{qV cos\,(\phi_s)}{2\pi p_s \gamma_t^2} \delta \omega_{rf} = a \, \delta \omega_{rf} \tag{43}$$

with

$$a = \frac{qV cos\,(\phi_s)}{2\pi p_s \gamma_t^2} \tag{44}$$

With the feedback loop ($\delta \omega_{rf} = -k_r \, \delta R$) the equation of motion becomes

$$\frac{d^2}{dt^2} \delta R + \Omega_s^2 \, \delta R = -a k_r \, \delta R \tag{45}$$

or

$$\frac{d^2}{dt^2} \delta R + \left(\Omega_s^2 + a k_r \right) \delta R = 0 \ . \tag{46}$$

The radial loop does not provide damping since the above differential equation has no first order derivative term. It only reduces the effect of a frequency error on the radial position.

Consider a small static error ϵ_{rf} in the frequency programme $f_{rf}(B)$. This produces an error $\delta \omega_{rf} = 2\pi \epsilon_{rf}$ in the right-hand side of Eq. (43). Without the radial loop the resulting radial error would be $(2\pi a \epsilon_{rf}) / \Omega_s^2$ wher as the radial loop will reduce it to $(2\pi a \epsilon_{rf}) \, / \, (\Omega_s^2 + a k_r)$. The radial loop gain k_r cannot be made arbitrarily large, however, because the transient response of the loop is indeed changing the beam energy, and our derivation of the longitudinal dynamics assumed that the energy change per turn was small, so the momentum increment of Eq. (3) could be approximated by the continuous derivative of Eq. (4). This condition is called adiabatic evolution. The radial loop will respect this condition only if its time constant is long compared to the synchrotron period $2\pi / \Omega_s$. A classic low-level system for hadron machines will include both a fast phase loop and a slow radial loop as shown in Fig. 12. With these two loops the equation of motion now becomes

$$\frac{d^2 \delta \phi_{b,c}}{dt^2} + k_\phi \frac{d \delta \phi_{b,c}}{dt} + \left(\Omega_s^2 + a k_r \right) \delta \phi_{b,c} = 0 \ . \tag{47}$$

The poles of the closed-loop transfer function are

$$p_\pm = -\frac{k_\phi}{2} \pm \frac{k_\phi}{2} \sqrt{1 - 4 \left(\frac{\Omega_s^2 + a k_r}{k_\phi^2} \right)} \ . \tag{48}$$

If we choose $k_\phi >> \Omega_s$ and $k_\phi^2 >> a k_r$ we get

$$p_- \approx -k_\phi \tag{49}$$

$$p_+ \approx -\frac{\Omega_s^2 + a k_r}{k_\phi} \tag{50}$$

The first pole corresponds to the very fast transient defined by the phase loop gain k_ϕ. Thanks to the radial loop gain k_r the second pole can now be placed on the real axis at the desired distance from the origin. The corresponding time constant must however be long compared to the synchrotron period so that the motion remains adiabatic

$$\frac{\Omega_s^2 + a k_r}{k_\phi} < \Omega_s \ . \tag{51}$$

This places a bound on the gain of the radial loop.

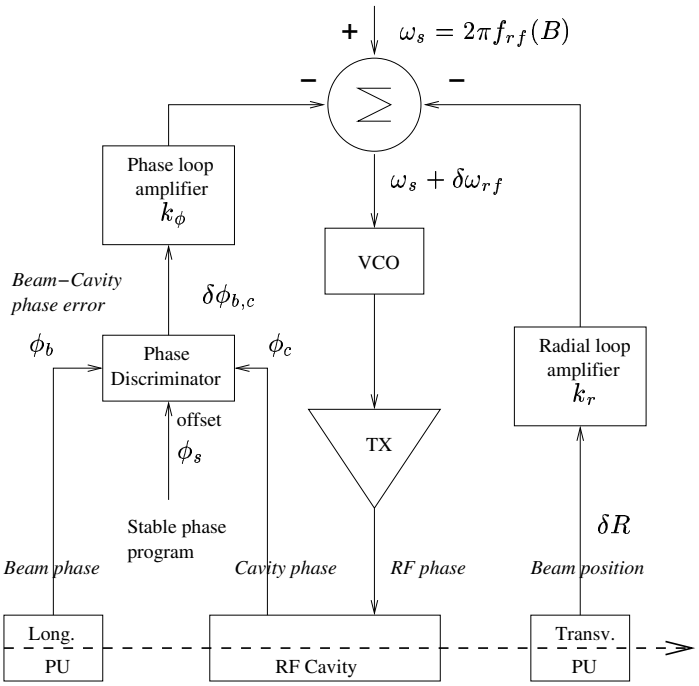


Fig. 12: Classic low-level system for hadron machines: phase loop and radial loop

Remarks

- The phase loop/radial loop tandem has a very good behaviour during transients. At injection, for example, if the beam is injected with a phase and energy error the first turn will be on an off-centred orbit and the beam will see a non-zero RF voltage. The phase loop reacts in a few turns and the RF jumps on the bunch, thereby preventing emittance blow-up. Thereafter the radial loop will slowly modify the beam energy to drive it back to the centre orbit.
- If one neglects the delays, the combination of phase loop/radial loop is unconditionally stable. A comprehensive treatment of the low-level loops in the presence of long delays can be found in Ref. [13]. Other interesting references are [14] (low-level loops with delays), [15] (application to the CERN PS in during the 1970s), and [16] (application to the CERN PS Booster).
- The gain of the radial loop amplifier must be inverted at transition because the sign of the coefficient a then changes (Eq. 43).
- With a radial loop, the noise in the loops changes the radial position and thus the beam momentum. This may be a problem when the beam is transferred to the next machine or sent to the target.
- One problem of the radial loop is that it estimates the *mean* radial position from one or two measurements at discrete points in the machine. As mentioned in Section 3.2, eventual transverse betatron oscillation and local closed orbit distortion at the pick-up location will introduce errors.

4.3 Controlling the beam frequency: the frequency loop

Motivation: The same dynamic behaviour can be achieved if the radial loop is replaced by a frequency loop. The advantage of the frequency loop is the better dynamic range and lower noise achievable with phase measurements than with radial position measurements.

Method: At constant B field the radial displacement δR is proportional to the beam frequency deviation $\delta\omega_b$ Eq. (80)

$$\frac{\delta R}{R} = \frac{\gamma^2}{\gamma_t^2 - \gamma^2}\frac{\delta\omega_b}{\omega_b} \tag{52}$$

We can thus replace the radial loop by a frequency loop as shown in Fig. 13.

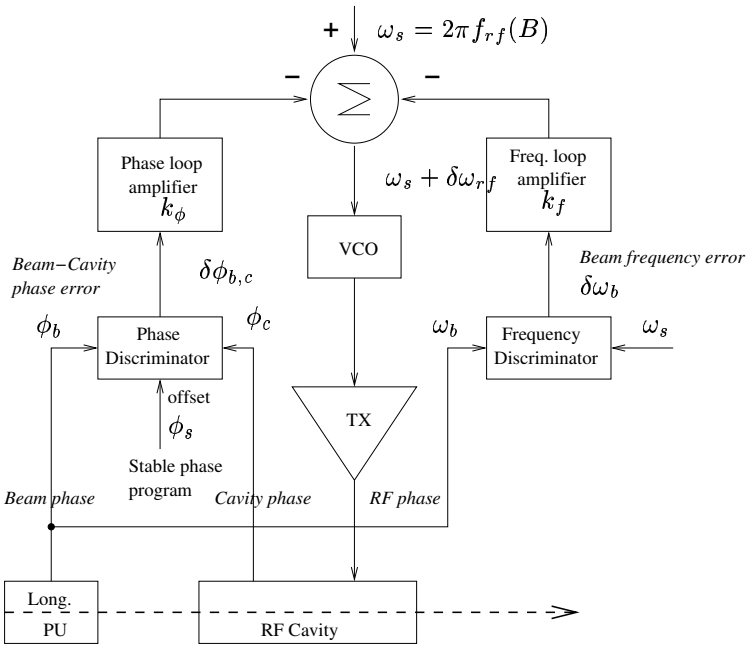


Fig. 13: Alternative low-level system for hadron machines: phase loop and frequency loop

Remark: Although the transient behaviour of the systems Figs. 12 and 13 can be made identical by varying the frequency loop gain as a function of the energy, to follow Eq. (52), there is no control of the actual radial position with the frequency loop. Beam centring is a function of the accuracy of the frequency programme $f_{rf}(B)$ only. The problem is particularly severe at transition where the beam orbit can be displaced without any change in the beam frequency (Eq. 52).[4] A variant of the system was used in the CERN SPS when it was operating as a proton–antiproton collider: the measurement of the beam frequency was replaced by a measurement of the Voltage Controlled Oscillator (VCO) frequency. The beam was injected above transition.

4.4 Controlling the beam phase: the synchronization loop

Motivation: If the transfer is of the bunch-into-bucket type, the beam in the injector must be synchronized with the RF of the receiving machine before transfer.

Method: We compare the phase of the beam with the phase of the external reference and feed the error back into the VCO after proper filtering by the synchronization loop amplifier (Fig. 14). The overall system is described by a third order differential equation. Its analysis is easier if we use Laplace transforms. Taking the transform of both sides of Eq. (37) we derive the beam transfer function $B_\phi(s)$ (dipolar motion)

$$\delta\phi_{b,c}(s) = \frac{s}{s^2 + \Omega_s^2}\delta\omega_{rf}(s) = B_\phi(s)\,\delta\omega_{rf}(s) \tag{53}$$

with

$$B_\phi(s) = \frac{s}{s^2 + \Omega_s^2}\ . \tag{54}$$

In the above equation the same symbols are used for time domain signals and for their Laplace transforms. The argument (s) identifies the transform domain. Equation (9) can also be written for the

[4] With very precise control of frequency and stable phase it is not impossible to accelerate through transition: this has been done in the acceleration of ions in the CERN SPS.

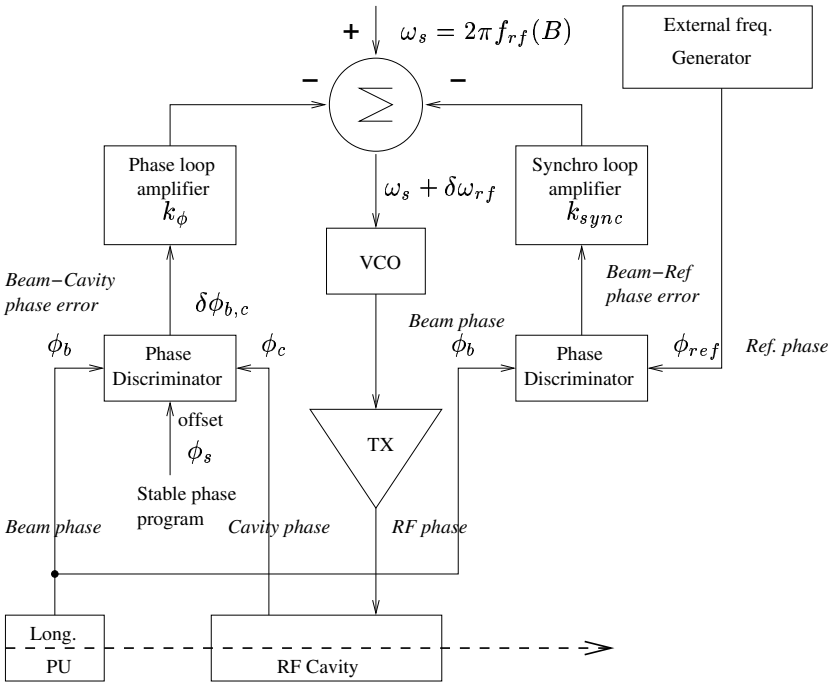


Fig. 14: Common low-level system for synchronizing two machines before transfer: phase loop and synchronization loop

beam frequency ω_b (in radian per second)

$$\frac{d\phi_{b,c}}{dt} = -\delta\omega_b + \delta\omega_{rf} \tag{55}$$

After taking Laplace transforms and rearranging the terms we get the frequency response $B_\omega(s)$ relating the beam frequency (actually its deviation with respect to the synchronous particle) to the RF frequency modulation

$$\delta\omega_b(s) = \delta\omega_{rf} - s\,\delta\phi_{b,c}(s) \tag{56}$$

$$\delta\omega_b(s) = (1 - sB_\phi(s))\,\delta\omega_{rf} = B_\omega(s)\,\delta\omega_{rf} \tag{57}$$

with

$$B_\omega(s) = \frac{\Omega_s^2}{s^2 + \Omega_s^2} \ . \tag{58}$$

We now redraw the phase and synchronization loops of Fig. 14 with the above transfer functions (Fig. 15). The beam frequency deviation $\delta\omega_b$ is integrated to give $\delta\phi_b$, that is, the beam phase minus the linear ramp at the synchronous frequency

$$\delta\phi_b = \phi_b - \omega_s t \ . \tag{59}$$

The phase loop is kept closed. We derive the open loop transfer function from the synchronization loop correction $\delta\omega_{sync}$ to the the phase of the beam $\delta\phi_b$

$$H_{ol}(s) = \frac{\delta\phi_b(s)}{\delta\omega_{sync}(s)} = \frac{1}{1 + k_\phi B_\phi(s)} B_\omega(s) \frac{1}{s} = \frac{\Omega_s^2}{s\,(s^2 + k_\phi\, s + \Omega_s^2)} \ . \tag{60}$$

The phase loop gain k_ϕ is typically much larger than Ω_s so that the open loop can be approximated by two integrators in series, i.e. 180° phase shift. $H_{ol}(j\omega)$ is plotted in Fig. 16 after multiplication by k_ϕ to render it dimensionless.

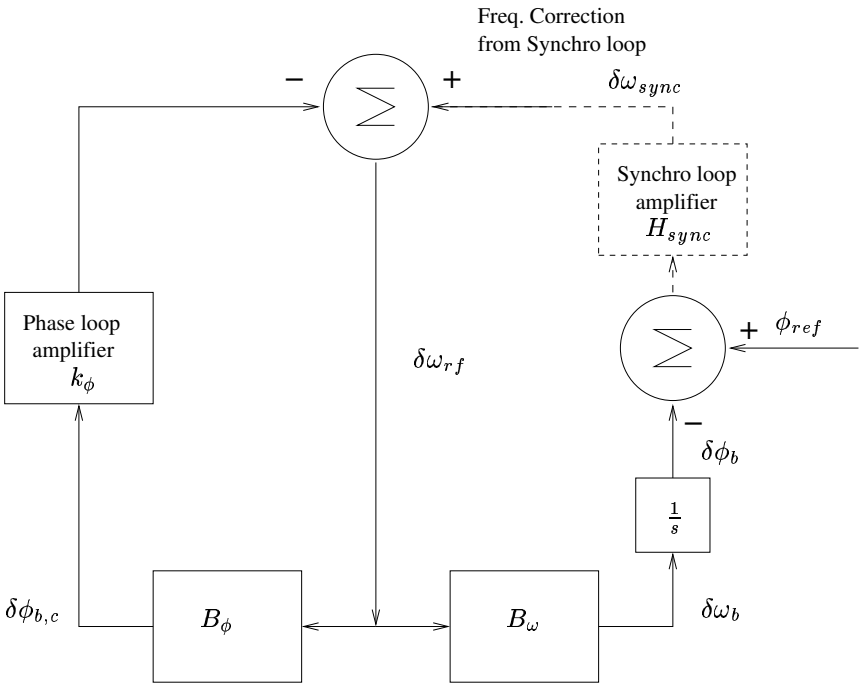


Fig. 15: Phase loop and synchronization loop in the Laplace domain

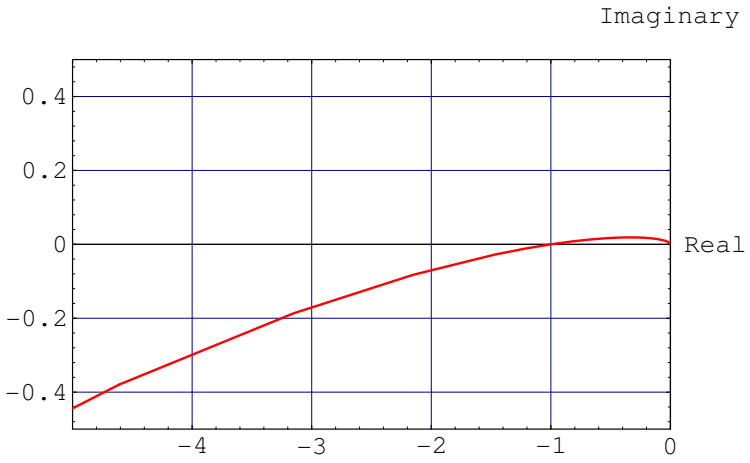


Fig. 16: Nyquist diagram of $k_\phi H_{ol}(j\omega)$ for $k_\phi = 0.6 f_{rev}$, $\Omega_s = 2\pi 200$ (rad/s). Case of the LHC proton beams in the SPS at injection ($f_{rev} = 43$ kHz).

Even with a very small gain, k_{sync}, the closed loop will be on the edge of instability. A classic indicator of stability is the phase margin, defined as the amount by which the phase of the open loop response $k_\phi H_{ol}(j\omega)$ exceeds $-180°$ when the modulus of its gain is one [17]. The phase margin should be minimum $45°$ for reasonable stability. Figure 16 shows that this condition cannot be fulfilled even for very small values of the synchronization loop gain k_{sync}. The solution is to introduce a lead compensation network $H_{sync}(s)$ in series with the open loop [17].

$$H_{sync}(s) = k_{sync}\frac{1+\alpha\tau s}{1+\tau s} \quad . \tag{61}$$

It adds a positive phase shift (dependent on the value of α), thereby increasing the phase margin. In our case the desired phase shift is $45°$, achievable with a parameter α of 7 minimum [18]. We choose $\alpha = 10$. The time constant τ adjusts the frequency ω_m at which the phase shift is maximum

$$\omega_m = \frac{1}{\tau\sqrt{\alpha}} \quad . \tag{62}$$

Using Kuo's method [18][5] we get, for a strong phase loop ($k_\phi \gg \Omega_s$)

$$\tau \approx \frac{1}{\Omega_s}\sqrt{\frac{k_\phi}{k_{sync}}}\frac{1}{\alpha^{3/4}} \quad . \tag{63}$$

The open-loop response now becomes

$$G_{ol}(s) = \frac{\Omega_s^2}{s\left(s^2 + k_\phi\, s + \Omega_s^2\right)}\frac{1+\alpha\tau s}{1+\tau s} \quad . \tag{64}$$

It is plotted in Fig. 17. With $k_{sync} = \frac{1}{4}k_\phi$ and $\alpha = 10$ the phase margin is $55°$.

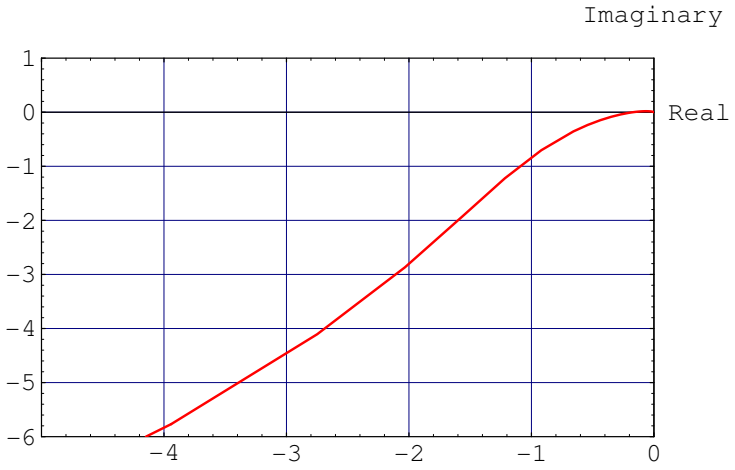


Fig. 17: Nyquist diagram of $k_\phi G_{ol}(j\omega)$ for $k_\phi = 0.6 f_{rev}$, $\Omega_s = 2\pi 200$ (rad/s) with the lead compensation network ($\alpha = 10$, $\tau = \frac{0.36}{\Omega_s}$)

Figure 18 shows the corresponding step response: Beam phase ϕ_b for a step in the reference phase ϕ_{ref}.

Remarks:

- The above design is used in the CERN SPS to accelerate the proton beam for the LHC [19],[20]. The phase loop must be much faster than the synchronization loop: in the above design its time constant[6] is 60 μs while the synchronization loop responds in a few milliseconds (Fig. 18).

[5] We choose τ such that the open loop gain, after compensation, will be 1.0 at the frequency ω_m, where the phase shift is maximum (Eq. 62).

[6] This means that it takes 60 μs to reduce a phase error at injection by a factor $1/e$.

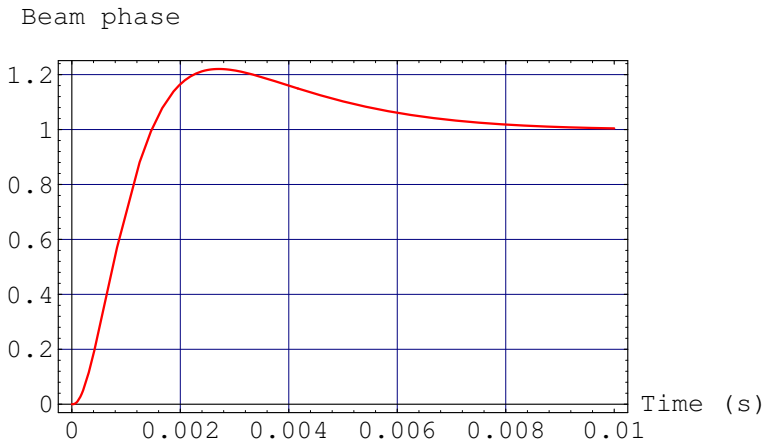


Fig. 18: Synchronization loop step response: beam phase ϕ_b for a step in the reference phase ϕ_{ref}. ($k_\phi = 0.6 f_{rev}, \Omega_s = 2\pi 200$ (rad/s), $k_{sync} = \frac{1}{4} k_\phi$, $\alpha = 10$)

- The use of a synchronization loop is common in injectors for bunch-into-bucket transfer. The external frequency generator is then the RF of the receiving machine. Synchronization scenarios comprise two steps [13]: the frequency of the injector is brought a desired offset $\Delta\omega$ from the frequency of the receiving machine (step 1). The synchronization loop is kept open and the output of the phase discriminator $(\phi_{ref} - \phi_b)$ is then a slowly-beating signal at the frequency $\Delta\omega$. The synchronization loop is closed on a zero crossing of this signal (step 2).
- A synchronization loop can also be used during acceleration if the external frequency generator implements the frequency programme $f_{rf}(B)$ Eq. (29) [19],[20]. The time constant of the lead compensation network is inversely proportional to Ω_s and must be varied during the acceleration. Acceleration with a synchronization loop is thus practical only if the synchrotron frequency does not change too much during the acceleration ramp (i.e. injection well above transition).
- One can also switch onto a synchronization loop before sending the beam to the targets. The advantage of this is that the beam energy (radial position) is precisely defined by the external frequency if the B field is assumed constant and stable. The noise and beam intensity-dependent offsets in the low-level RF will not affect the momentum of the beam delivered to the physicist, this being a weak point of the radial loop.

5 IMPLEMENTATION

5.1 The all-analog beam control

The all-analog beam control it is a straightforward implementation of the loops. The key component is the Voltage Controlled Oscillator (VCO). At RF frequencies, varactor-tuned oscillators are common. An external LC circuit (tank) sets the centre frequency and a variable capacitance diode is used to do the tuning: By varying the DC voltage applied to this diode one varies the VCO output frequency (Fig. 19).

A typical component broadly used at CERN (PS and SPS) is the MC1648 from Motorola. Its successor, the MC12149, can be tuned to oscillate at frequencies up to 1.3 GHz. Observed on a spectrum analyser the output of the VCO is not a pure line but rather a narrow lobe, whose width is caused by the phase noise. An essential figure of merit is the noise spectrum expressed in dBc/Hz as a function of the frequency offset from the carrier (Fig. 20). (One plots the ratio of the Power Spectral Density (PSD) over the power of the RF output on a logarithmic scale.)

Let us now study the effect of this noise on the beam in the presence of a phase loop.

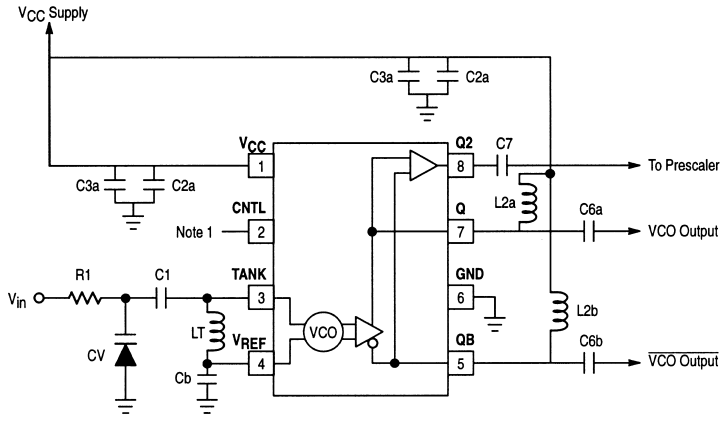


Fig. 19: Varactor-tuned oscillator (Motorola documentation)

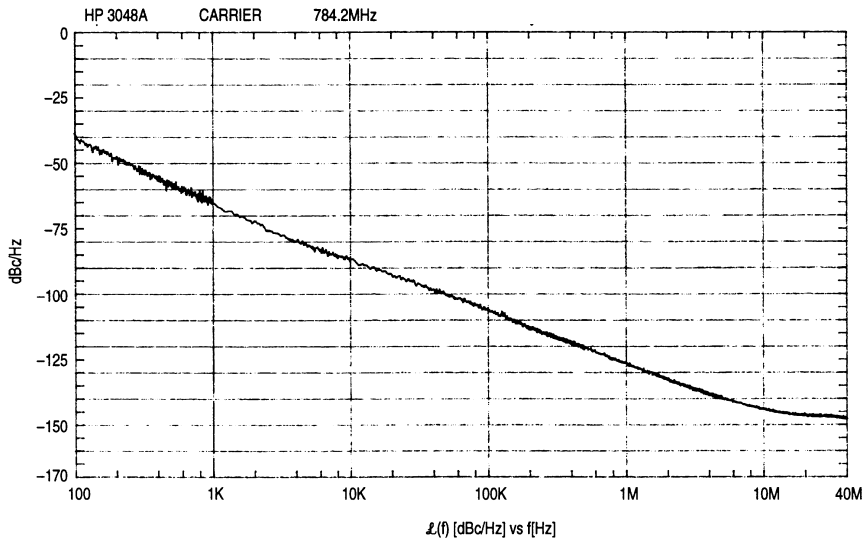


Fig. 20: Typical phase noise power spectral density $S_n(\omega)$ for MC12149 with a centre frequency at 750 MHz (Motorola documentation)

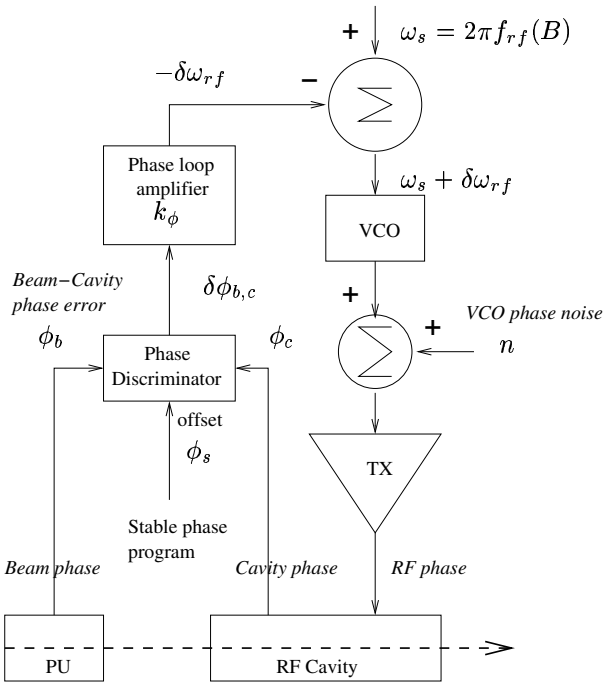


Fig. 21: Phase loop with the VCO phase noise n

The phase noise is introduced as a small random signal n added at the VCO output (Fig. 21). This noise modifies the driving term in Eq. (37). Since the RF frequency is the derivative of the RF phase the noise appears as a second derivative in the right-hand side of the equation

$$\frac{d^2\delta\phi_{b,c}}{dt^2} + \Omega_s^2\delta\phi_{b,c} = \frac{d\left(\delta\omega_{rf} + \frac{dn}{dt}\right)}{dt} = \frac{d\delta\omega_{rf}}{dt} + \frac{d^2n}{dt^2} \ . \tag{65}$$

With the phase loop closed we get

$$\frac{d^2\delta\phi_{b,c}}{dt^2} + k_\phi\frac{d\delta\phi_{b,c}}{dt} + \Omega_s^2\delta\phi_{b,c} = \frac{d^2n}{dt^2} \ . \tag{66}$$

The above equation tells us how the VCO phase noise n transforms into cavity/beam phase noise $\phi_{b,c}$. From Eq. (66) we derive the PSD $S_\phi(\omega)$ of the latter[7]

$$S_\phi(\omega) = \frac{\omega^4}{\left(\omega^2 - \Omega_s^2\right)^2 + \left(k_\phi\,\omega\right)^2}S_n(\omega) \ . \tag{67}$$

The second factor on the right-hand side is the spectrum plotted in Fig. 20. The first factor shows the effect of the phase loop. We shall call it the noise enhancement factor (Fig. 22). It is equal to zero at DC because slow drifts of the RF phase are followed by the beam and the cavity/beam phase remains zero. At very high frequencies the factor equals one. The beam does not react any more and the VCO noise is transmitted directly to the cavity/beam phase. Around the synchrotron frequency the enhancement factor depends on the phase loop gain. With $k_\phi = 0$ (no phase loop) it is infinite: the centre of charge of the bunch will thus see large RF phase errors. This will cause filamentation and will result in emittance blow-up and eventually loss of particle, as previously shown for an injection phase error (Fig. 8). As we increase the loop gain k_ϕ the factor is reduced in an increasing frequency band around the synchrotron frequency (Fig. 22).

[7] The PSD $S_y(\omega)$ of the output y of a linear system with transfer function $H(j\omega)$ is given by $S_y(\omega) = |H(j\omega)|^2 S_x(\omega)$ where $S_x(\omega)$ is the PSD of the input x [21].

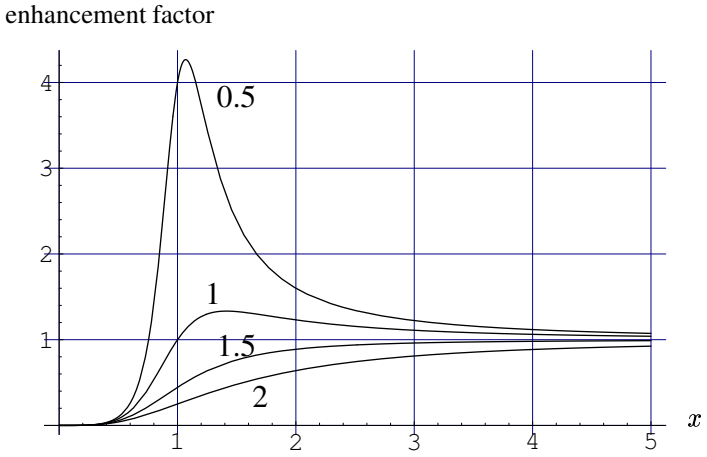


Fig. 22: Enhancement of the VCO noise: $\frac{x^4}{\left(x^2-1\right)^2+\xi^2 x^2}$ as a function of $x = \frac{\omega}{\Omega_s}$ for different values of the normalized phase loop gain $\xi = \frac{k_\phi}{\Omega_s}$ (from top trace to bottom: $\xi = 0.5, 1, 1.5, 2$)

A similar analysis can be done for the noise injected elsewhere in the beam control system: the noise in the frequency programme, for example, is injected in the adder at the top of Fig. 21. The resulting PSD of the cavity/beam phase will be given by a formula similar to Eq. (67) but the numerator will be ω^2 instead of ω^4 because of the integrator characteristic of the VCO. The phase loop will also reduce the damaging effect of this noise.

Nowadays the all-analog implementation is of limited use: the lack of precision in the analog voltage provided by the frequency programme implies poor control of the orbit since the gain of the radial loop cannot be infinite.

5.2 Analog beam control with DDS

To increase the precision of the frequency programme the analog VCO is replaced by a digital frequency synthesizer with a high resolution (i.e. many bits in the frequency word). These are called DDS, for Direct Digital Synthesizer. They present continuity in the phase of the RF output while the frequency is changing during the acceleration ramp. Figures 23 and 24 explain the DDS principle: at each clock pulse the content of the accumulator S is incremented by the digital frequency word F. The accumulator

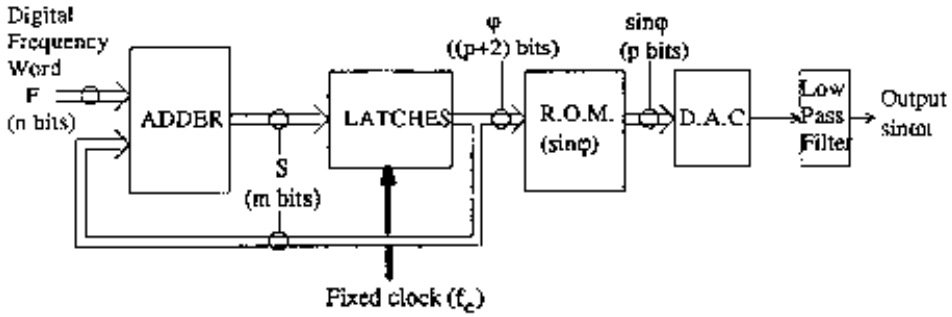


Fig. 23: Elementary block diagram for DDS. Reproduced from Ref. [22].

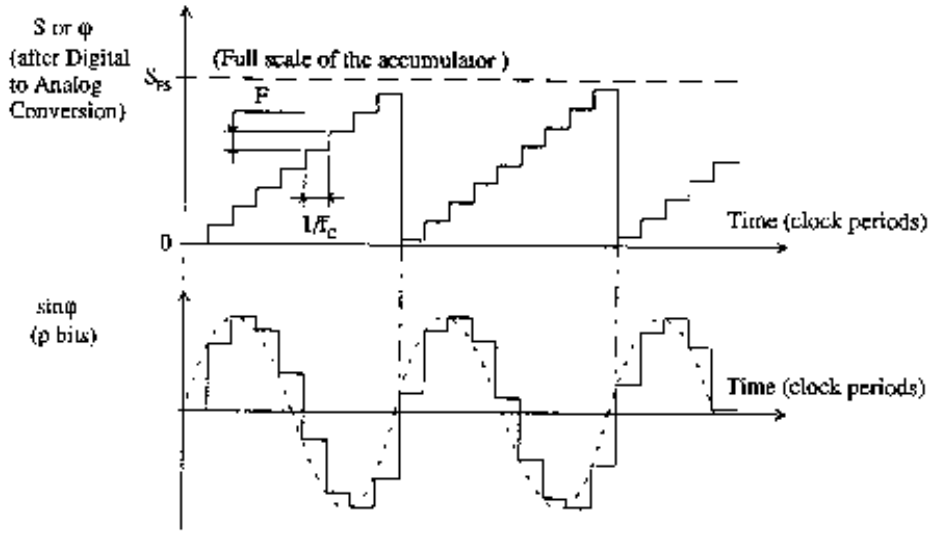


Fig. 24: Signal generation in DDS. Reproduced from Ref. [22].

regularly overflows, creating a quantized sawtooth whose frequency is proportional to F. The sawtooth is transformed into a sine wave by the Read Only Memory (ROM) and finally converted into an analog output. The frequency resolution depends on the number of bits for the control word F and the frequency range depends on the clock frequency f_c. For example, the AD9854, from Analog Devices, can be clocked at 300 MHz and has a 48-bit frequency word, thus providing a 1μHz resolution for a 100 MHz range. The noise at the output of a DDS has two components: the first comes from the division of the phase noise of the clock. It is similar to the noise of an analog VCO but will typically be very small because it is easy to get high spectral purity from a clock at a fixed frequency. The second component comes from the non-linearity of the digital quantization: this produces spurious lines whose amplitude varies in a complex fashion when changing the output frequency. Figure 25 shows the corresponding spectrum. The Spurious Free Dynamic Range (SFDR) is the component's important figure of merit: it measures the minimal guaranteed attenuation of the spurious lines with respect to the desired output signal.

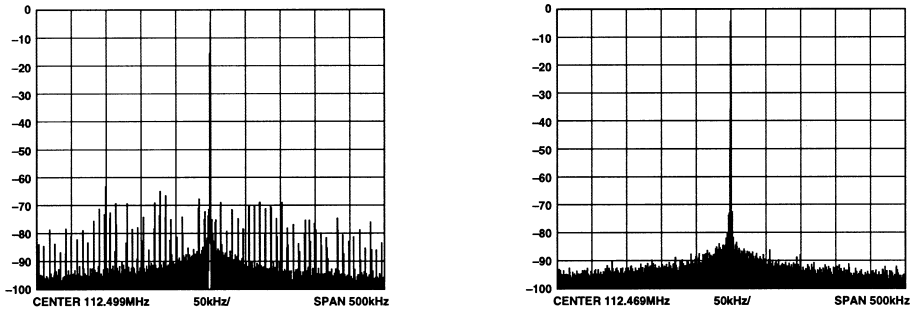


Fig. 25: Typical spectrum of the output of a DDS in a 500 kHz span centred on the carrier (0 dBm). Left: carrier at 112.499 MHz; right: carrier at 112.469 MHz. Notice the severe degradation of the spectrum for a small change in output frequency. The Spurious Free Dynamic Range (SFDR) in this 500 kHz band is here about 65 dB (Analog Devices documentation).

5.3 The digital beam control

In order to take advantage of the digital precision of the DDS, the low frequency part of the low-level loops should also be digital. Only the fast acquisition part remains analog. Figure 26 shows the beam control system used in the CERN PS [11]. The output of the Phase Loop Amplifier (PLA) is digitized by a 10 MS/s (10^7 samples per second) 12-bit ADC and added to the Digital Frequency Program (DFP) output in the Digital Loop Processor (DLP). The digitally generated frequency is sufficiently accurate to centre the beam in the vacuum chamber so there is no need for a radial loop. Correction of the frequency programme is achieved with the Digital Recorder (DR), which reproduces the correction memorized during a reference acceleration where the beam was accurately centred. This correction is added in the Digital Arithmetic Unit (DAU).[8] In this design the PLA is analog. One could also digitize the phase error

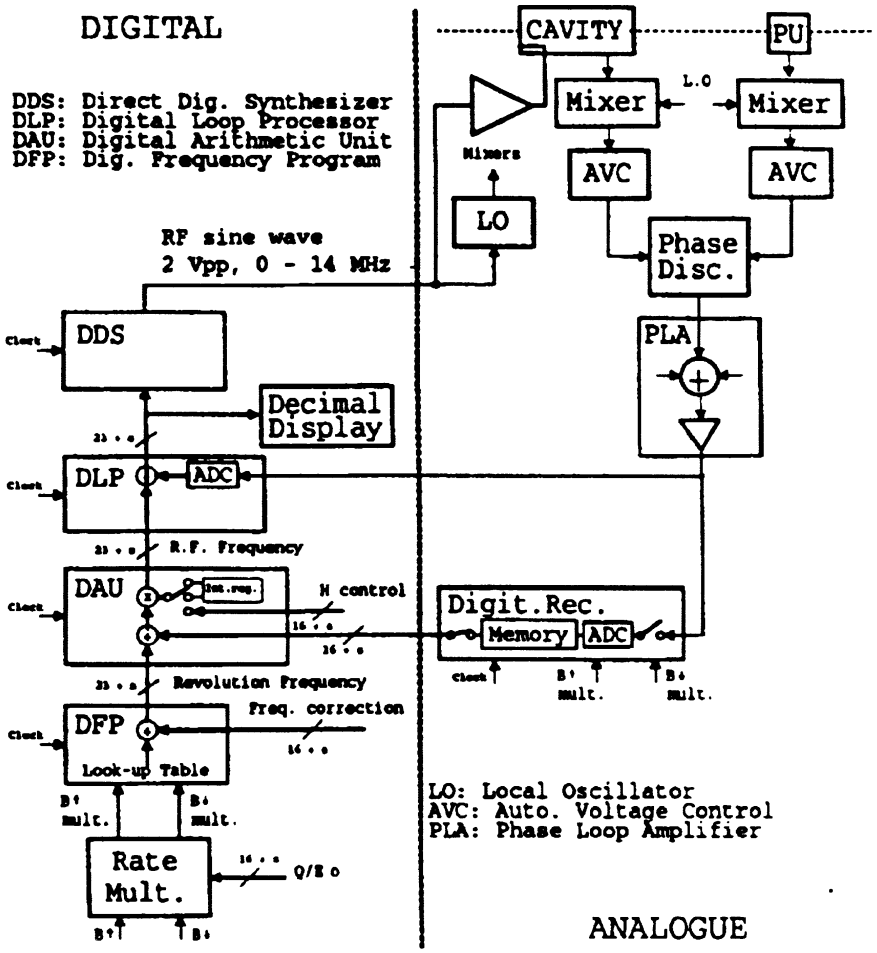


Fig. 26: Digital beam control system used in the CERN PS (reproduced from Ref. [11])

at the output of the phase discriminator and implement the low frequency part of the phase loop in digital [24].

[8] Performances were, however, much more reproducible with the radial loop in service. At present, this is the case with most beams accelerated in the CERN PS [23].

5.4 State variable and DSP implementation

We started our analysis with the description of the beam motion using a second order differential equation (Eq. 11). Alternatively we can keep the two first order differential equations. Let us use the variables $(\delta\phi, \delta R)$[9]. From Eqs. (9) and (80) we get

$$\frac{d}{dt}\delta\phi = -\frac{h\beta_s c}{R_0^2}\left(\frac{\gamma_t^2 - \gamma^2}{\gamma^2}\right)\delta R + \delta\omega_{rf} = b\,\delta R + \delta\omega_{rf} \tag{68}$$

with

$$b = -\frac{h\beta_s c}{R_0^2}\left(\frac{\gamma_t^2 - \gamma^2}{\gamma^2}\right) \quad . \tag{69}$$

and from Eqs. (8) and (81)

$$\frac{d}{dt}\delta R = \frac{qV\cos(\phi_s)}{2\pi p_s \gamma_t^2}\delta\phi = a\,\delta\phi \tag{70}$$

with

$$a = \frac{qV\cos(\phi_s)}{2\pi p_s \gamma_t^2} \quad . \tag{71}$$

The state of the beam is now given by the vector $(\delta\phi, \delta R)$. Its time evolution follows the equations

$$\frac{d}{dt}\begin{pmatrix} \delta\phi \\ \delta R \end{pmatrix} = \begin{pmatrix} 0 & b \\ a & 0 \end{pmatrix}\begin{pmatrix} \delta\phi \\ \delta R \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}\delta\omega_{rf} \tag{72}$$

The coefficients of the square matrix vary slowly with time since a and b vary during the acceleration. Of course $a.b = -\Omega_s^2$. The phase loop and radial loop feedback can now be introduced

$$\delta\omega_{rf} = -\begin{pmatrix} k_\phi & k_r \end{pmatrix}\begin{pmatrix} \delta\phi \\ \delta R \end{pmatrix} \tag{73}$$

and the equations become

$$\frac{d}{dt}\begin{pmatrix} \delta\phi \\ \delta R \end{pmatrix} = \begin{pmatrix} -k_\phi & b - k_r \\ a & 0 \end{pmatrix}\begin{pmatrix} \delta\phi \\ \delta R \end{pmatrix} \quad . \tag{74}$$

The transient behaviour of the system is determined by the eigenvalues λ of the above matrix. The characteristic equation is

$$\lambda^2 + k_\phi\lambda + (ak_r - ab) = 0 \ . \tag{75}$$

The eigenvalues are identical to the poles of the second order differential equation (47). We can choose the eigenvalues and compute the gain coefficients k_ϕ and k_r during the ramp in order to keep the same loop behaviour through the cycle. This method is called pole placement in control theory [17] : one first chooses a *cost function* that measures the error between the desired and actual responses of the system in presence of a perturbation. The optimal positions of the poles minimize this cost function. The use of state variable is somewhat academic for the phase loop/radial loop system. For more complex loops (synchronization loop with lead compensation for example) a set of four first order differential equations is required. Pole placement with the above formalism should then bring a more exact derivation of the optimal values of the gains than the Nyquist analysis presented in Section 4.4. The design of a beam control using state variables was first proposed at BNL for the AGS [25].

To implement pole placement we need a processor that adjusts the feedback gains during acceleration. Nowadays this can easily be done using a Digital Signal Processor (DSP). A DSP is a microprocessor optimized for fast-floating point operations. The ADSP-2106x from Analog Devices and the TMS320C6x from Texas Instruments are used in particle accelerators. Figure 27 shows the hardware

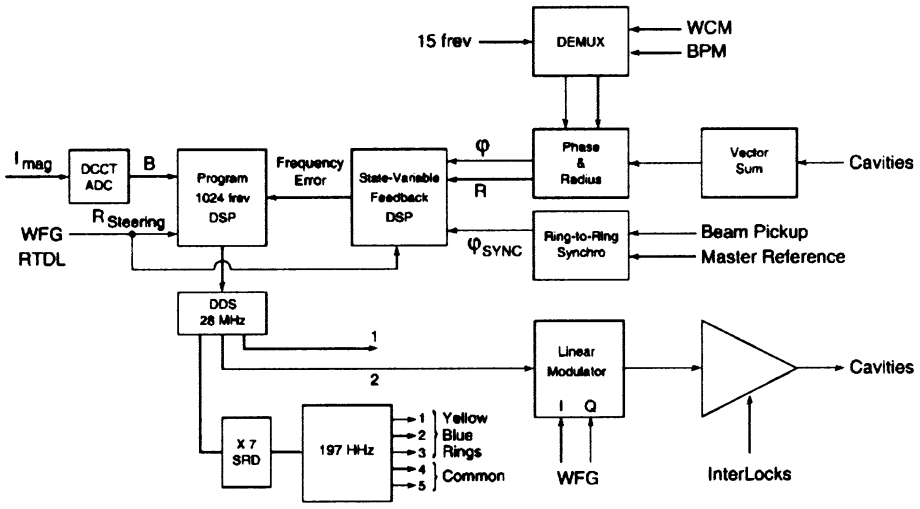


Fig. 27: Beam control system with DSP used at Fermilab for the RHIC (reproduced from Ref. [26])

used at BNL for the Relativistic Heavy Ion Collider (RHIC) [26]. One DSP computes the frequency (frequency programme) from a measurement of the B field (in fact the magnet current). It also receives a correction called the radial steering. The output is a 32-bit word providing 6 mHz resolution at 28 MHz. A second DSP, called the State-Variable Feedback DSP in the figure, implements the feedback loops. The beam phase and radius are measured once per turn (every 12 μs). The feedback gains are recomputed 200 times during the 74 s acceleration ramp. The output is the frequency error added to the DDS control word. The loops implemented by the DSP can be switched on and off at will. This is simply programmed as a C-language if/then statement. The phase loop is normally closed with either the radial loop (during acceleration) or the synchronization loop (before colliding) [26]. A similar system is in use in the Fermilab main ring and Tevatron [27].

6 ANNEXE

6.1 Differential relations

Among the four variables B, p, R, f only two can be chosen independently. The other two are then functions of these. Differential relations describe the variations of any three of these around a valid working point [3],[4] :

$$\frac{\delta p}{p} = \gamma_t^2 \frac{\delta R}{R} + \frac{\delta B}{B} \tag{76}$$

$$\frac{\delta p}{p} = \gamma^2 \frac{\delta f}{f} + \gamma^2 \frac{\delta R}{R} \tag{77}$$

$$\frac{\delta B}{B} = \gamma_t^2 \frac{\delta f}{f} + \frac{\gamma^2 - \gamma_t^2}{\gamma^2} \frac{\delta p}{p} \tag{78}$$

$$\frac{\delta B}{B} = \gamma^2 \frac{\delta f}{f} + \left(\gamma^2 - \gamma_t^2\right) \frac{\delta R}{R} \tag{79}$$

For example, from Eq. (79) we derive that, at B constant, an error in the RF frequency creates a radial position error

$$\frac{\delta R}{R} = \frac{\gamma^2}{\gamma_t^2 - \gamma^2} \frac{\delta f}{f} \quad . \tag{80}$$

[9] To simplify the notations the subscripts b, c are dropped in this section so that $\delta\phi$ now represents $\delta\phi_{b,c}$.

From Eq. (76) we can relate the radial displacement to the momentum deviation at B constant

$$\frac{\delta R}{R} = \frac{1}{\gamma_t^2}\frac{\delta p}{p} \ . \tag{81}$$

The slippage factor η is defined as the opposite of the ratio of the relative frequency deviation to the relative momentum deviation at constant B. From Eq. (78) we get

$$\eta = -\frac{\frac{\delta f}{f}}{\frac{\delta p}{p}} = \frac{1}{\gamma_t^2} - \frac{1}{\gamma^2} \ . \tag{82}$$

6.2 Notation

ϕ_s the stable phase.

ϕ_c the phase of the RF in the cavity.

ϕ the phase of the RF in the cavity when the particle crosses it.

$\delta\phi = \phi - \phi_s$.

$\hat{\phi}$ the phase of the RF in the cavity when the centre-of-charge of the bunch crosses it.

ϕ_b the phase of the beam = the phase-of-the Fourier component of the beam current at the RF frequency.

$\phi_{b,c}$ cavity/beam phase. ($\phi_{b,c} = \phi_c - \phi_b + \pi/2$).

$\delta\phi_{b,c}$ cavity/beam phase minus the stable phase. ($\delta\phi_{b,c} = \phi_c - \phi_b + \pi/2 - \phi_s$).

Ω_s the synchrotron frequency in radian/s (Eq. 13).

ω_{rf} the RF frequency in radian/s. ($\omega_{rf} = \frac{d\phi_c}{dt}$).

ω_s the synchronous frequency in radian/s = the RF frequency that keeps the beam on the centre orbit. It is a function of the B field: $\omega_s = 2\pi f_{rf}(B)$ with $f_{rf}(B)$ given by Eq. (29).

$\delta\omega_{rf}$ a small modulation of the RF frequency around the synchronous frequency. $\delta\omega_{rf} = \omega_{rf} - \omega_s$.

ω_b the beam frequency = $2\pi h$ times the revolution frequency ($\omega_b = \frac{d\phi_b}{dt}$).

$\delta\omega_b$ a small modulation of the beam frequency around the synchronous frequency. $\delta\omega_b = \omega_b - \omega_s = \frac{d\,\delta\phi_b}{dt}$.

$\delta\phi_b$ the deviation of the beam phase from the linear ramp at the synchronous frequency. ($\delta\phi_b = \phi_b - \omega_s t$).

REFERENCES

[1] John J. Livingood, *Principles of Cyclic Particle Accelerators* (D. Van Nostrand, Princeton, NJ, 1961).

[2] H. Bruck, Accélérateurs circulaires de particules, [Circular Particle Accelerators] (Bibliothèque des Sciences et Techniques Nucléaires, Paris, 1966).

[3] L. Rinolfi, Longitudinal beam dynamics, Application to synchrotron, CERN/PS 2000-008 (LP), course given at the Joint Universities Accelerator School (JUAS), 26 April 2000.

[4] C. Bovet, R. Gouiran, I. Gumowski, K.H. Reich, A selection of formulae and data useful for the design of A.G. synchrotrons, CERN/MPS-SI/Int. DL/70/4 (1970).

[5] E. Regenstreif, Le synchrotron à protons du CERN (1ère partie, CERN 58-6 a 1958).

[6] A. Schnase, Cavities with a swing, CAS, Seeheim, 8–16 May 2000, these proceedings.

[7] J. Le Duff, High-frequency non-ferrite cavities, CAS, Seeheim, 8–16 May 2000, these proceedings.

[8] C. Boccard, CERN SL/BI, private communication.

[9] D. Cocq, CERN SL/BI, private communication.

[10] J.M. Brennan, RF beam control for the AGS booster, Brookhaven Natl. Lab., BNL-52438 (1994).

[11] F. Blas, J. Boucheron, B.J. Evans, R. Garoby, G.C. Schneider, J.P. Terrier, J.L. Vallet, Digital beam controls for synchrotrons and storage rings in the CERN PS complex, CERN/PS 94-24 (RF), presented EPAC London, 1994.

[12] R. Garoby, Low level RF and feedback, Proc. Joint US–CERN–Japan International School, Frontiers of accelerator technology, Tsukuba, 1996.

[13] S. Koscielniak, RF systems aspects of longitudinal beam control (in the low current regime), AIP conference proceedings 249 (AIP, 1992) Vol. 1.

[14] W. Schnell, Equivalent circuit analysis of phase-lock beam control systems, CERN 68-27 (1968).

[15] D. Boussard, Une présentation élémentaire du système Beam Control du PS [An elementary presentation of the PS Beam Control System], MPS/SR/Note/73-10 (1973).

[16] G.C. Schneider, RF beam control and stability for newcomers, CERN/PS 90-59 (RF) (1990).

[17] Gene F. Franklin, J. David Powell, Abbas Emami-Naeini, *Feedback Control of Dynamic Systems*, (Addison-Wesley, Reading, MA, 1994).

[18] J.-Ch. Gille, P. Decaulne, M. Pélegrin,*Théorie et calcul des asservissements Linéaires*, [Theory and Calculation of Linear Servo Control] (Dunod, Paris 1987).

[19] P. Baudrenghien, T. Linnecar, D. Stellfeld, U. Wehrle, SPS beams for LHC: RF beam control to minimise rephasing in the SPS, CERN-SL-98-027-RF, presented at EPAC, Stockholm 1998.

[20] P. Baudrenghien, Beam control for protons and ions, CERN-OPEN-99-077 (1999), presented at the 9th LEP-SPS Performance Workshop, Chamonix 1999, CERN-SL-99-007-DI.

[21] A. Papoulis, *Probability, Random Variables and Stochastic Processes* (McGraw-Hill, New York, 1965).

[22] R. Garoby, Low level RF building Blocks, RF Engineering for Particle Accelerator, Cern Accelerator School, Exeter College, 1991.

[23] R. Garoby, CERN PS, private communication.

[24] L.K. Mestha, V. Brouk, R.C. Webber, J. Mangino, T. Uher, A digital beam phase loop for the low energy booster, presented at PAC, Washington, DC, 1993.

[25] E. Onillon, J.M. Brennan, The new BNL AGS phase, radial and synchronisation loops, presented at EPAC, Sitges, 1996.

[26] J.M. Brennan, A. Campbell, J. DeLong, T. Hayes, E. Onillon, J. Rose, K. Vetter, RF beam control system for the Brookhaven relativistic heavy ion collider RHIC, presented at EPAC, Stockholm, 1998.

[27] B.E. Chase, B. Barnes, K. Meisner, Digital low level RF systems for Fermilab main ring and Tevatron, presented at PAC, Vancouver, 1997.

LOW-LEVEL RF SYSTEMS FOR SYNCHROTRONS
Part II: High intensity. Compensation of beam-induced effects

P. Baudrenghien
CERN, Geneva, Switzerland

Abstract

The high-intensity regime is reached when the voltage induced by the beam in the RF cavities is of an amplitude comparable to the desired accelerating voltage. In steady state this *beam loading* can be compensated by providing extra RF power. Transient beam loading occurs at injection or in the presence of a beam intensity that is not uniform around the ring. The transients are periodic at the revolution frequency. Without correction transient beam loading can be very harmful: the stable phase and bucket area will not be equal for all bunches. Strong beam loading often occurs with longitudinal instabilities because the RF cavities are a large contributor to the total ring impedance. The low-level systems that reduce the effect of transient beam loading will also increase the threshold intensity of longitudinal instability caused by the cavity impedance at the fundamental RF frequency. Four classic methods are presented here: feedforward, RF feedback, long delay feedback, and bunch-by-bunch feedback. The first three fight against both transient beam loading and longitudinal instability, if caused by the cavity impedance (at the fundamental). The last cures longitudinal instability (dipole mode) caused by any impedance in the machine but has no effect on beam loading. These techniques have been made possible by the recent advent of fast digital circuitry and an emphasis will be put on implementation.

1 THE PROBLEM OF BEAM-INDUCED VOLTAGE

1.1 Transient beam loading

This Section is a brief presentation of the basics of beam loading. Refer to the literature for more details [1]–[5].

The beam crossing the RF cavity induces an electromagnetic field within it, thereby creating a decelerating voltage V_b acting on the beam in return (Fig. 1). The accelerating voltage seen by the beam is thus the vector sum of the voltage produced by the generator V_g and the beam-induced voltage V_b

$$V_t = V_g + V_b = Z_g I_g + Z_b I_b \tag{1}$$

where I_g is the RF drive and I_b is the beam current.

In the case of a *standing wave cavity*[1], the voltages V_g and V_b are simply the accelerating voltages at the cavity gap created by the generator and by the beam, respectively [6]. Around one resonance ω_0 the standing wave cavity can be modelled as a lumped Resistor–Inductor–Capacitor (RLC) parallel circuit. The two impedances Z_g and Z_b are proportional, with a ratio that is a function of the main coupler transformation ratio. To simplify the notations, we will make them equal

$$Z_b(\omega) = Z_g(\omega) \tag{2}$$

$$Z_g(\omega) = \frac{R}{1 + j2Q\frac{\Delta\omega}{\omega_0}} \tag{3}$$

[1] This applies to ferrite-loaded cavities as well.

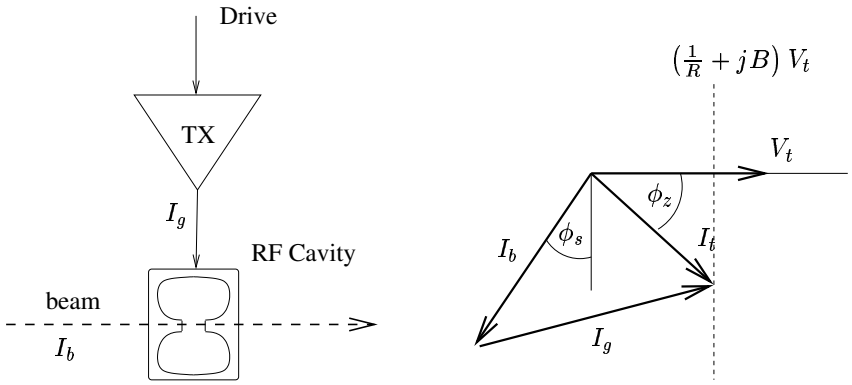


Fig. 1: Left: principle of beam loading. The beam current I_b induces a voltage in the cavity that modifies the total accelerating voltage V_t. Right: vector diagram relevant to stationary beam loading in a standing-wave cavity.

where Q is the quality factor and $\Delta\omega = \omega - \omega_0$.

For a *travelling-wave cavity*, which looks like a matched transmission line for the power generator, $Z_g(\omega)$ is proportional to $\sin\tau/\tau$, with τ being the transit time factor. $Z_g(\omega)$ is purely real. First it decreases for increasing $|\Delta\omega|$; it crosses zero at a frequency offset equal to the inverse filling time of the cavity; and it changes sign thereafter (deceleration) [7]. The impedance $Z_b(\omega)$ is also a function of τ but it is not simply proportional to $Z_g(\omega)$. It has both real and imaginary parts.

If the bunch intensity is uniform around the ring and if we analyse the situation well after injection, we are in the *stationary* situation. In the frequency range of the RF system, the beam current I_b is a single spectral line at the RF frequency f_{rf}. The currents and voltages of Eq. (1) can be represented as vectors in the complex plane (Fig. 1). Consider a standing-wave cavity. Let $Y(\omega)$ be the cavity admittance

$$Y(\omega) = 1/Z_g(\omega) = 1/R + jB(\omega) \tag{4}$$

with

$$B(\omega) = 2\frac{Q}{R}\frac{\Delta\omega}{\omega_0} \ . \tag{5}$$

Making $Z_b = Z_g$, Eq. (1) can be rewritten

$$I_g + I_b = YV_t \ ; \tag{6}$$

I_g must compensate I_b to keep the modulus of the total accelerating voltage V_t at the desired value (desired bucket area). This is done by adjusting the amplitude of the generator drive and by detuning the cavity (angle ϕ_z in Fig. 1). This latter changes the value of the cavity susceptance B, while the conductance $1/R$ remains constant. In the vector diagram of Fig. 1 the vector $I_t = I_g + I_b$ follows the dashed line perpendicular to V_t as the cavity tune is changed. The amplitude $|I_g|$ is minimum when V_t and I_g are in phase. This is the desired working point for the power generator. If the stable phase angle ϕ_s is not zero (acceleration), extra RF power is needed to compensate stationary beam loading. For a standing-wave cavity, the low-level system needed includes a so-called cavity field amplitude loop, which adjusts the amplitude of the generator drive $|I_g|$ to get the desired $|V_t|$, and a slow cavity tuning loop, which maintains V_t and I_g in phase (matched conditions at the cavity input minimizing the generator drive $|I_g|$) [1],[3],[5]. No tuning loop is needed with a *travelling-wave cavity*. It remains a matched load in the presence of beam loading [7]. Stationary beam loading is compensated by the cavity field amplitude loop (identical to the one used with standing-wave cavities).

A *transient* situation is encountered at injection. In the time domain, the beam current suddenly jumps from zero to the ring distribution and we cannot, in the frequency domain, reduce our analysis to

the single spectral component at f_{rf}. Equation (1) is still valid, but the two sides are Fourier transforms with a broad frequency spectrum. Because the RF system is only reacting to a narrow band around f_{rf} we can consider these signals as a band-limited modulation of the carrier frequency f_{rf}. When injecting a bunched beam, the voltage V_b varies from zero to the full beam loading. The low-level system must react in a short time compared with the synchrotron period, to restore the proper bucket area and the correct phase of the total voltage V_t with respect to the beam current I_b. Two loops come into action. The fast beam phase loop, present in most hadron machines [8], tries to restore the stable phase between $I_b(\omega_{rf})$ and $V_t(\omega_{rf})$, while the cavity field amplitude loop adjusts the drive amplitude $|I_g(\omega_{rf})|$ to restore the desired $|V_t(\omega_{rf})|$. For the beam phase loop, V_t is the vector sum of the total voltages in all the cavities. In the case of a standing-wave cavity a third loop (cavity tuning loop) tries to keep the generator current $I_g(\omega_{rf})$ in phase with the total voltage $V_t(\omega_{rf})$ (matched conditions at the cavity input). The argument $\omega_{rf} = 2\pi f_{rf}$ emphasizes that these loops work on the carrier component only, i.e. the component at the exact RF frequency. This works fine for small beam current, that is, when the accelerating voltage is predominantly determined by the generator current. For higher beam currents, a variation of the amplitude of $I_g(\omega_{rf})$ not only results in a variation of the amplitude of $V_t(\omega_{rf})$ but also of its phase. The loops that were independent at low beam currents become coupled, and unstable behaviour of the system results above a certain beam current threshold. Let I_0 be the generator current required to produce the accelerating voltage $|V_t|$ without beam loading and with the cavity tuned to resonance: $I_0 = |V_t|/R$. The beam loading is characterized by the ratio $|I_b|/I_0$, which is called the relative beam loading. It can be shown that the loops become unstable when this ratio approaches two [9]. The result is that the loops do not lock at injection: the beam is not captured.[2]

Transient beam loading also occurs if the ring is not filled uniformly (such as in the presence of a hole reserved for the possible triggering of a beam dump) or if the filling time of the cavities is not very large compared to the revolution period T_{rev}. The value of V_t will vary in amplitude and phase along the batch because of the modulation of the beam current I_b. The bucket area and the phase of the bunch with respect to V_g will also be modulated along the batch. They will not be correct for some bunches in the batch since the beam phase loop and the cavity field amplitude loop only adjust their average values (values at the carrier frequency f_{rf}). Modulation of the bucket area may result in a loss of particles from some bunches in the batch because the bucket area is too small (Fig. 2).

In the case of a collider the modulation of the longitudinal position of the bunches (phase of the bunch with respect to V_g) will displace the collision point for some bunches. In the case of an injector it will reduce capture efficiency when the bunched beam is injected into the receiving machine because some bunches will not see the correct RF phase (if the transfer is of the bunch-into-bucket type). For high-intensity proton machines, the beam loading can be greater than the RF voltage. The nominal LHC beam in the CERN SPS, for example, (1.05×10^{11} protons per bunch, DC current 0.67 A) induces 4.5 MV total in the RF cavities, compared with a matched capture voltage of 0.65 MV (bunch emittance 0.35 eVs, bunch length 4 ns). Fortunately transient beam loading compensation is only needed at a few discrete frequencies: as explained in Section 4.1 (Eq. (69)), the voltage induced by the beam in the cavity consists of a line at the carrier frequency f_{rf} (stationary beam loading) plus sidebands at multiples of the revolution frequency on each side of the carrier (transient beam loading). The strength of the sidebands decreases as we move away from f_{rf} with an envelope that is a function of the ring pattern. Therefore, transient beam loading compensation is only needed at the frequencies

$$f = f_{rf} \pm n f_{rev} \quad . \tag{7}$$

In the above equation the index n goes from 0 to ∞. In practice, however, the spectrum of the beam-induced voltage is limited to the bandwidth of the RF cavities. Ideally the compensation should cover a significantly larger bandwidth to get a good correction of the fast components present when the head of the batch enters the cavity, or when the beam is injected into the machine.

[2] If we avoid the strong transient beam loading at injection, by slow accumulation scheme for example, the loops would still become unstable above the same beam current threshold.

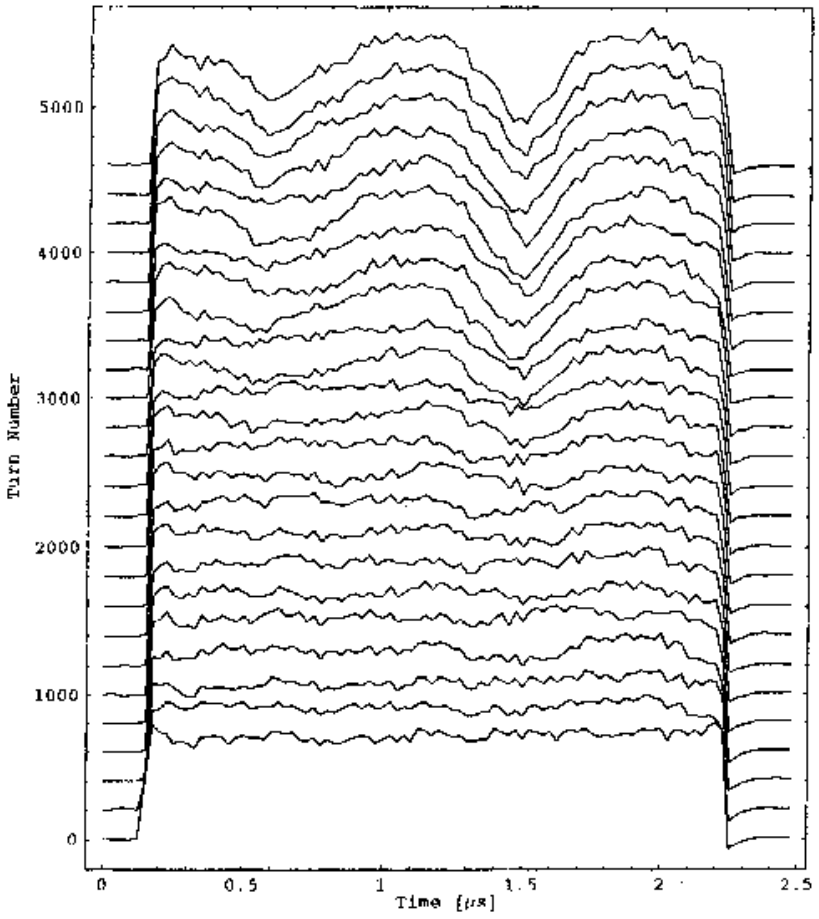


Fig. 2: Beam loss due to transient beam loading in the CERN SPS (proton beam for LHC, 0.5×10^{11} protons per bunch in one batch of 81 bunches). The beam consists of one batch filling 1/11 th ring ($\approx 2\mu$s). The cavity filling time is 800 ns (much shorter than the revolution period 23 μs). Each trace shows the envelope of the bunch intensity along the batch. The bottom trace shows the first turn (26 GeV/c). Traces are separated vertically by 200 turns. The capture voltage is 550 kV. No acceleration. A modulation of the bunch intensity develops along the batch after about 2000 turns.

1.2 Longitudinal instabilities

This Section is a brief qualitative presentation of the longitudinal instabilities caused by cavity impedance at its fundamental resonance. A complete theory can be found in Ref. [10]. At high beam current, the interaction between a bunched beam and an RF cavity can lead to longitudinal instabilities. This process can be broken down into three steps:

- Step 1: when crossing the cavity, the beam induces an electro-magnetic wave called wakefield.
- Step 2: this wakefield modifies the accelerating voltage seen by the beam (phase of V_t when the beam crosses the cavity and bucket height) and acts in return on the current profile along the beam.
- Step 3: this modifies the wakefield created on the next crossing.

If the gain and phase shift of the above natural beam/cavity feedback is unfavourable, instability will grow. The bunches start a longitudinal oscillation at the synchrotron frequency f_s (Fig. 3).

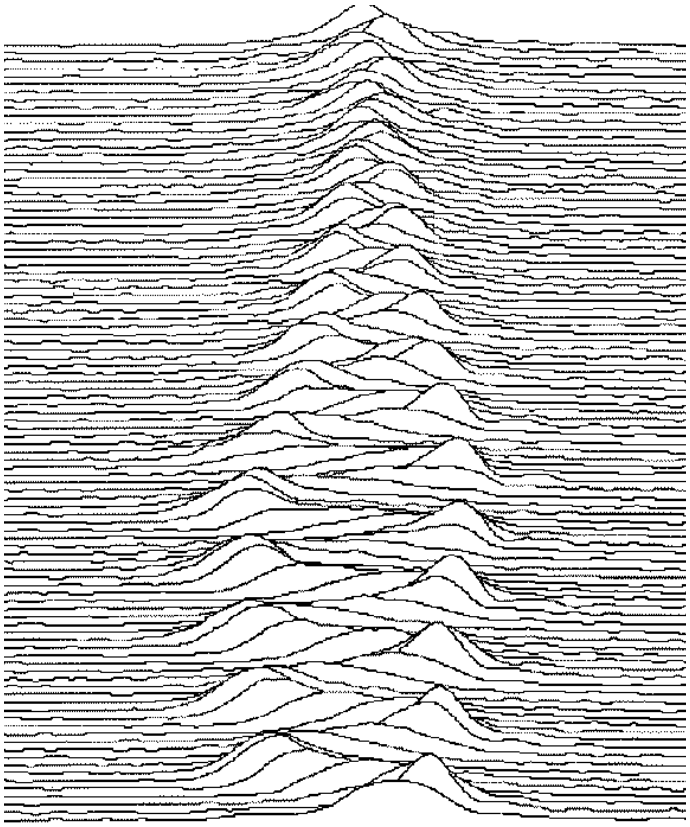


Fig. 3: Mountain range display of a positron bunch in the CERN SPS, from 100 ms after injection (3.8 GeV/c, bottom trace) to 122 ms after injection (4.2 GeV/c, top trace). Time goes from bottom to top. Ten turns (0.23 ms) between traces. Horizontal window = 10 ns. $f_s = 430$ Hz (1 synchrotron period = 100 turns = 10 traces). The amplitude of the dipole oscillation rises to 3 ns maximum. The RF consists of a fixed 0.750 MV at 100 MHz, plus a second harmonic at 200 MHz whose amplitude is zero on the bottom trace, and rises to 0.7 MV on the last trace. The oscillation is here excited by an external cause. Its amplitude (in ns) decreases as the acceleration proceeds thanks to the addition of the higher frequency RF and natural damping (synchrotron radiation). A strong quadrupole oscillation is also present.

In the presence of dipolar motion the bunches create a beam current that is periodic at f_{rev} with a phase modulation at the synchrotron frequency f_s. In Section 4.1 (Eq. (78)) we show that its spectral power is concentrated at the frequencies

$$f = f_{rf} \pm n f_{rev} \pm m f_s \tag{8}$$

with dominant sidebands at $m = 1$. Pure dipole oscillation goes without change in the bunch shape. But the cavity impedance can excite higher order shape oscillations of the bunch. Figure 4 shows a quadrupole oscillation: there is (almost) no motion of the centre of the charge distribution but the second order moment (related to the bunch length) oscillates at twice the synchrotron frequency. Higher order

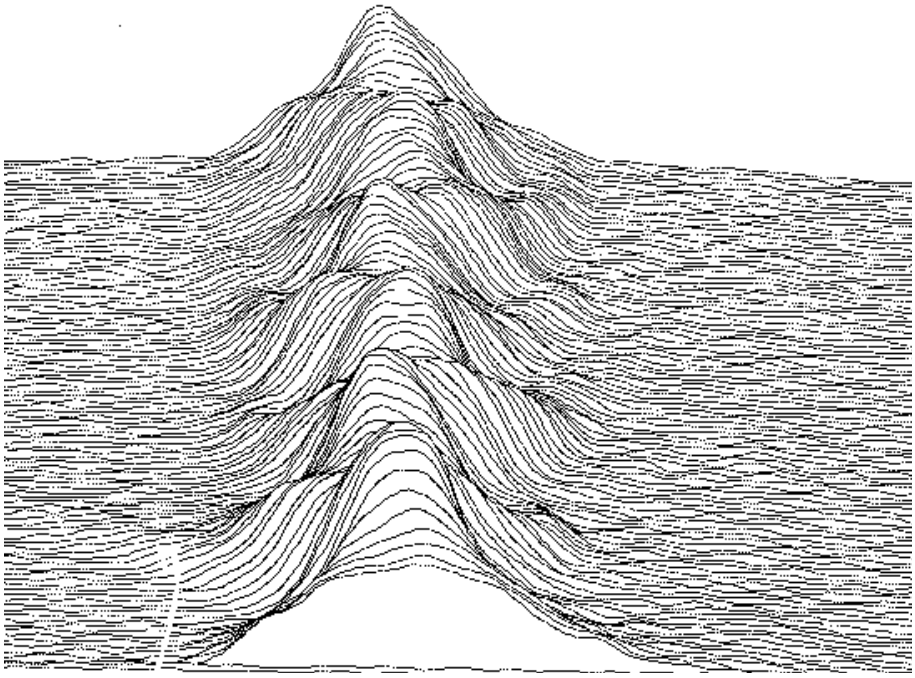


Fig. 4: Quadrupole oscillation (plus a small dipole oscillation) of a proton bunch at injection from the CPS into the CERN SPS (26 GeV/c, 200 MHz RF). The first non-zero trace at the bottom shows the first turn of the injected beam. Horizontal window = 10 ns. Proton beam for the LHC. Note that this oscillation is not due to an instability but to an RF mismatch.

modes can also be excited: sextupole, octopole, etc. [10]. Several modes can occur at the same time. The oscillation shown in Fig. 3 is the superpositions of a dipole and a quadrupole mode. These oscillations will sample the cavity impedance at the frequencies (Eq. (79))

$$f = f_{rf} \pm n f_{rev} \pm m f_s \tag{9}$$

and the low-level system must thus reduce the apparent cavity impedance there.[3] Comparing the above equation to Eq. (7) we notice that it is more difficult to prevent the longitudinal instabilities than to correct the beam loading: the synchrotron frequency f_s is typically much smaller than the revolution frequency f_{rev} and a complete prevention of the instabilities calls for a reduction of the cavity impedance at closely spaced frequencies around each revolution frequency line. Fortunately, higher order modes of oscillation are not as easily excited by cavity impedance (at the fundamental frequency) due to the form factor and are usually damped by the ever-present synchrotron frequency spread, so impedance reduction is hopefully only required on the first few synchrotron sidebands [10].

[3] The fundamental resonance of the RF cavity is most efficient at driving mode m, when the bunch width is comparable to m times $T_{rf}/2$ [10]. The quadrupole mode (or even higher modes) can thus be excited by the cavity fundamental with a full bucket, as is often the case at injection into a hadron machine [11]. In lepton machines the bunches are typically much shorter than the bucket width and the cavity will mainly excite the dipole mode. Higher order modes can always be excited by the undesired high-frequency resonances in the cavity or elsewhere in the beam pipe. The low-level systems presented here have no action on these external causes.

2 CURES

2.1 Feedforward

Method: Figure 5 shows the feedforward technique: the beam current I_b measured by a pick-up is filtered by a Bandpass Filter (BPF) centred at the centre frequency of the cavity response f_0 and fed into the feedforward filter. The feedforward transfer function H_{opt} corrects the generator drive such that the resulting generator current I_g^{comp} gives a cavity voltage V_g equal but opposite to V_b.

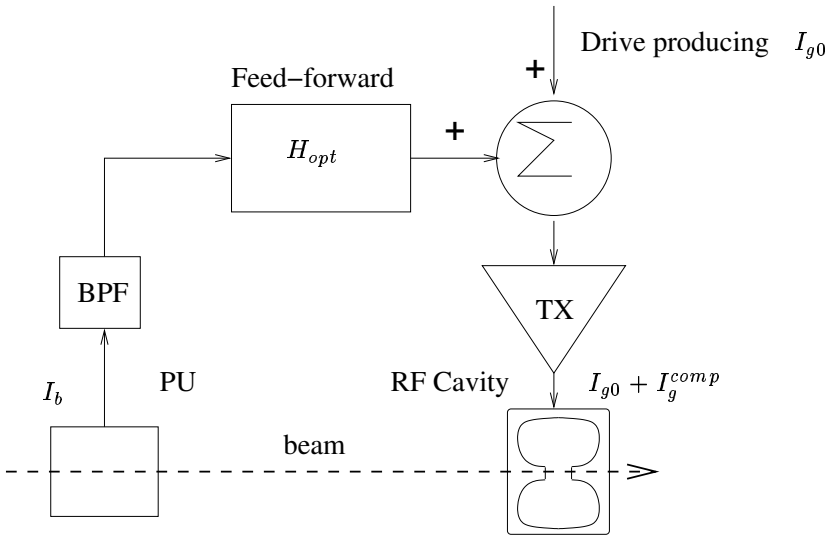


Fig. 5: Block diagram of the feedforward method

From Eq. (1) we get

$$Z_g I_g^{comp} = -Z_b I_b \ . \tag{10}$$

For a standing-wave cavity $Z_g = Z_b$ so H_{opt} is a constant gain. In the case of a travelling-wave cavity perfect compensation is impossible because Z_g is zero at frequencies where Z_b is non-zero [7]. But partial correction is possible with the help of a more complex filtering function H_{opt} [12] (see Section 3.4). In practice the delays in the cables and in the electronics do not permit measurement of the beam current and correction of the cavity voltage in the same turn. An intentional delay of one full turn is thus implemented in the chain.[4] **Advantage:** The feedforward loop is closed through the beam, but from the electronics point of view it is an open-loop system. As such it is not limited by the time constant inherent in a closed-loop system such as the long delay feedback (Section 2.3). At injection the beam current is measured on the first turn and the full compensation can be applied on the second turn. This makes locking of the beam phase loop possible at injection, even with high beam currents.
Limitation: The drawback is that this open-loop system is very sensitive to gain and phase drifts of the RF generator. This limits the long-term performance of the method. It is also difficult to set up for a varying RF frequency (acceleration ramp) because the pick-up to cavity delay must be varied continuously. A phase error of 5° at the RF frequency will degrade a perfect beam loading compensation to only 91% compensation ($2\sin 2.5^\circ = 0.09$).
Example: Figure 6 shows results obtained with the feedforward system installed on the 200 MHz travelling-wave cavities in the CERN SPS [12],[13].

[4] The synchrotron frequency f_s is typically much smaller than the revolution frequency f_{rev}. The current profile along the beam therefore varies very little in one turn.

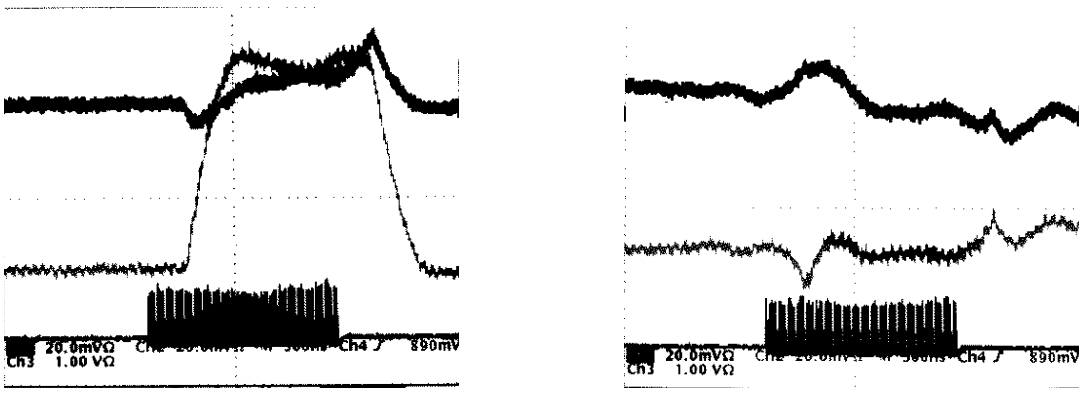


Fig. 6: Transient beam loading voltage measured in the CERN SPS 200 MHz TWC cavities with the LHC beam (1 batch = 81 bunches spaced by 25 ns, 4×10^{12} protons total, 500 ns/div). Left: feedforward OFF. Right: feedforward ON. The bottom trace is the beam current I_b measured with a wide-band pick-up, the upper two traces are the I and Q components (AC coupled) obtained by demodulating the total cavity voltage V_t with a Local Oscillator (LO) at the RF frequency (see Section 3.1 for details on I/Q demodulation). $I^2 + Q^2$ measures the power of the cavity voltage error.

2.2 RF feedback

Method: RF feedback around an amplifier is generally used to reduce the effects of the drifts in gain and phase encountered with power amplifiers. It also reduces distortion by flattening their response: this closed loop makes the overall response substantially independent of the response of the amplifier itself [14]. We can use it here to reduce the influence of the beam-induced voltage by effectively reducing cavity impedance. The principle is shown in Fig. 7. A probe measures the total accelerating voltage in the cavity V_t. It is compared to the desired voltage V_{ref} and the error is used to regulate the drive of the power amplifier. The RF feedback is a closed-loop system. As such it is relatively insensitive to the

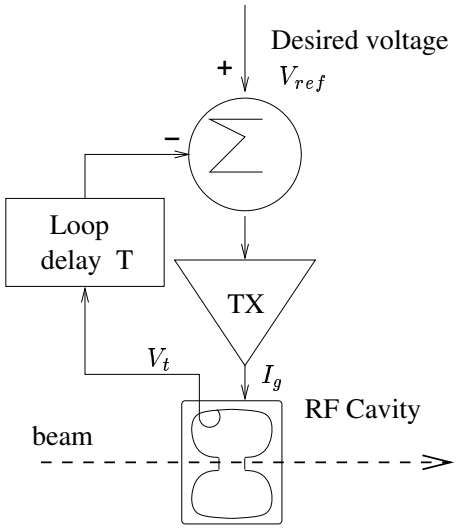


Fig. 7: Block diagram of the RF feedback method

small drifts of the gain and phase of the power generator. It is very easy to implement on a single-cell cavity. RF feedback on a multi-cell cavity is much more complex: a cavity consisting of N identical cells coupled together will show N closely spaced resonances corresponding to its N modes (phase shift of $2\pi n/N$ between cells). Since the coupling between two adjacent cells is slightly different for the

different modes, the phase shift from I_g to V_t will be different. A filter must be placed in the probe signal to adjust the open loop gain and phase independently for each resonance within the bandwidth of the system [15]. Otherwise the undesired resonances will make the loop unstable even though the corresponding field in the cavity may be harmless to the beam.

Limitation: The unavoidable loop delay T clearly appears in Fig. 7 because it limits the achievable impedance reduction. Figure 8 shows a block diagram of the RF feedback in the Laplace domain: G is the gain of the return path, A is the amplifier gain and e^{-Ts} is the delay operator.

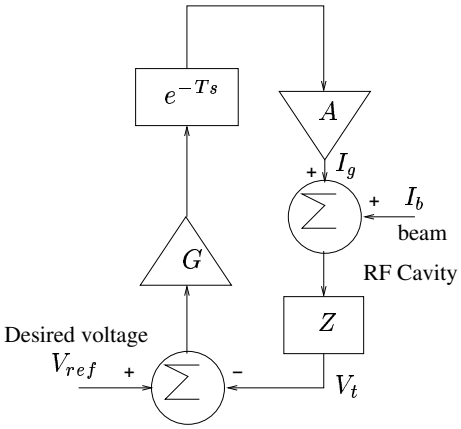


Fig. 8: Block diagram of the RF feedback in the Laplace domain

The beam current I_b is an added perturbation at the input of the standing-wave cavity of impedance $Z_g(\omega) = Z_b(\omega) = Z(\omega)$. The overall delay in the loop is T. It is the sum of the cable delay and the delay in the electronics (inversely proportional to the bandwidth of the amplifier).

Near its resonant frequency ω_0 a standing-wave cavity can be represented as an RLC circuit

$$Z(\omega) = \frac{R}{1 + j2Q\frac{\Delta\omega}{\omega_0}} \tag{11}$$

where Q is the quality factor and $\Delta\omega = \omega - \omega_0$. With the feedback loop closed, the beam loading voltage is

$$V_b(\omega) = \frac{Z(\omega)}{1 + GAe^{-jT\Delta\omega}Z(\omega)} I_b(\omega) \ . \tag{12}$$

In the above we have assumed that the phase of the return path has been adjusted so that the phase shift of the total loop is zero degrees at the frequency ω_0. Therefore a large open loop gain GAR leads to a good impedance reduction. The stability of the loop imposes a limit: outside the ω_0/Q band the cavity is purely reactive

$$Z(\omega) \approx \frac{R}{j2Q\frac{\Delta\omega}{\omega_0}} \ . \tag{13}$$

The phase shift is thus $-\frac{\pi}{2}$. A classic indicator of stability for a feedback loop is the phase margin, defined as the amount by which the phase of the open loop response exceeds $-\pi$ when the modulus of its gain is one [16]. To keep a phase margin of $\frac{\pi}{4}$, the open loop gain must have decreased to 1 when the delay T has added an extra $-\frac{\pi}{4}$ phase shift, that is at $\Delta\omega = \frac{\pi}{4T}$

$$GA|Z(\frac{\pi}{4T})| \leq 1 \tag{14}$$

$$GAR \leq \frac{\pi}{2}\frac{Q}{\omega_0}\frac{1}{T} = G_{max}AR \ . \tag{15}$$

Let $Z_{fbk}(\omega)$ be the apparent cavity impedance with the feedback loop. From Eq. (12) we get

$$Z_{fbk}(\omega) = V_b(\omega)/I_b(\omega) = \frac{Z(\omega)}{1 + GAe^{-jT\Delta\omega}Z(\omega)} \ . \tag{16}$$

Using the dimensionless variable $x = \Delta\omega T$ and the value of $G_{max}A$ given by Eq. (15) we can rewrite $Z_{fbk}(\omega)$ as

$$Z_{fbk}(\omega) = \frac{\frac{1}{GA}}{\frac{1}{GAR} + e^{-jx} + j\frac{4}{\pi}\frac{G_{max}A}{GA}x} \ . \tag{17}$$

At resonance, we have

$$Z_{fbk}(\omega_0) = \frac{R}{1 + GAR} \ . \tag{18}$$

A reduction of the apparent cavity impedance is thus possible only if GAR is larger than 1. Assuming that $GAR >> 1$, the first term in the denominator of Eq. (17) can be neglected and Z_{fbk} can be plotted as a function of $x = \Delta\omega T$ for different values of G (Fig. 9). For $G = G_{max}$ ($k = 1$), the frequency response presents a 3 dB overshoot on the edges of the passband. This is avoided by reducing the gain to $0.7G_{max}$ ($k = 0.7$). This flattens the frequency response in the passband, but the achieved value of the impedance at resonance ($x = 0$) is 3 dB larger. The figure also shows the effect of increasing the feedback gain above G_{max} ($k = 1.3$): the strong overshoot in the frequency response indicates that we are approaching the instability limit.

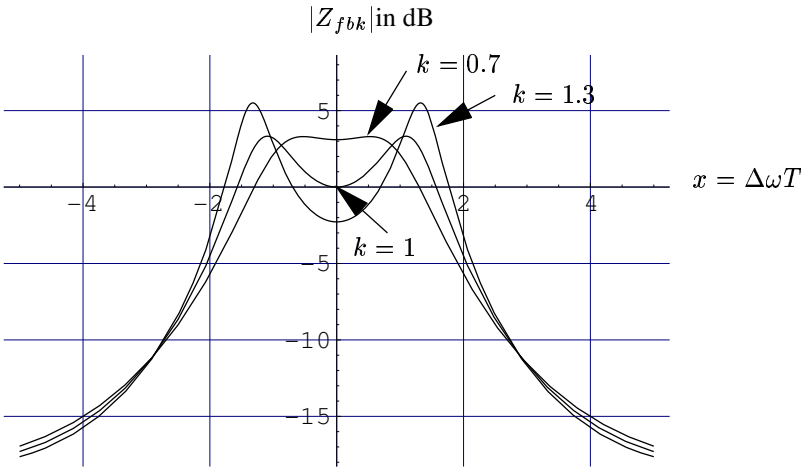


Fig. 9: Modulus of the apparent cavity impedance for three values of the normalized feedback gain $k = G/G_{max}$ as a function of $x = \Delta\omega T$

The minimum achievable value for the apparent cavity impedance at resonance ($\Delta\omega = 0$) is [5]

$$R_{min} = \frac{R}{1 + G_{max}AR} \approx \frac{2}{\pi}\frac{R}{Q}\omega_0 T \ . \tag{19}$$

The ultimate performance depends only on the cavity geometry R/Q and the loop delay T. The lesson is that the delay must be kept short, i.e. a broadband amplifier located as close to the cavity as possible. The achievable bandwidth is a function of the loop delay only. For the flat frequency response ($G = 0.7G_{max}$) we get

$$\Delta\omega_{-3dB} \approx \frac{1.3}{T} \ . \tag{20}$$

[5] If the flat frequency response ($G = 0.7G_{max}$) is preferred, the impedance will be 3 dB higher.

Examples: The present RF system of the storage ring at the Synchrotron Radiation Research Center (SRRC) in Taiwan consists of two independent and identical chains [17]: a 60 kW klystron is connected to a warm pill-box cavity via a circulator (Doris cavity, 500 MHz centre frequency, $Q_L = 14500$). It has been decided to replace this system by a single superconducting cavity [18]. As a first step in this upgrade project, an RF feedback has been tested on the warm cavities. Figure 10 shows the apparent impedance Z_{fbk} with a feedback gain $GAR = 11.2$ (or 21 dB). This measurement is obtained by sweeping the frequency of the klystron drive and recording the amplitude of the cavity voltage: it is the transfer function from V_{ref} to V_t in Fig. 8 and it is proportional to $|Z_{fbk}|$, with a proportionality factor equal to GAe^{-Ts}. The most important parameter is the loop delay T, here equal to 450 ns, including the 150 ns group delay of the klystron. From Eq. (15) we derive the maximal loop gain $G_{max}AR = \frac{\pi}{2}\frac{Q}{\omega_0}\frac{1}{T} = 16.111$. The flat response of Fig. 9 should thus be obtained with a 30% lower gain, that is $GAR = 11.28$, in good agreement with the experiment ($GAR = 11.2$). Eq. (20) predicts a bandwidth of 920 kHz (two sided), that is a Q equal to 543. This again is in good agreement with the measured Q of 562.29. On this system, the RF feedback reduces the apparent cavity impedance by 11.2.

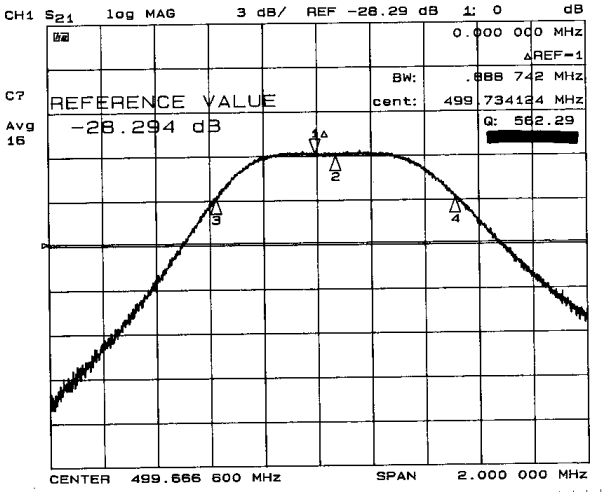


Fig. 10: Apparent cavity impedance of the SRRC RF system with an open-loop gain $GAR = 11.2$

Another example is the RF feedback installed in the CERN PS on the 40 MHz cavities [19]. A tube amplifier is connected to the warm re-entrant cavity ($Q_L = 10000$) via a short (60 cm) coaxial line. The total loop delay, including the amplifier group delay, is only 220 ns. The maximal loop gain is derived from Eq. (15) and we get $G_{max}AR = \frac{\pi}{2}\frac{Q}{\omega_0}\frac{1}{T} = 284$ or 49 dB. The system is operated with a gain of 43 dB, i.e. half the maximal value. This feedback reduces the cavity impedance at the resonance by a factor of 140.

2.3 Long delay feedback

Method: In the previous section it was concluded that the amplifier was best located close to the cavity for a good reduction of beam-induced effects with RF feedback. The amplifier is, however, often placed outside the accelerator tunnel in order to ease maintenance and reduce down-time in case of failure. In that case, impedance reduction is still possible if the cavity is not narrow band. It is first noted that beam loading needs only to be compensated at the frequencies (Eq. (69))

$$f = f_{rf} \pm n f_{rev} \tag{21}$$

while the prevention of instability calls for a reduction of the cavity impedance at the frequencies (Eq. (79))

$$f = f_{rf} \pm n f_{rev} \pm m f_s \ . \tag{22}$$

The synchrotron frequency f_s is typically much smaller than the revolution frequency f_{rev}. We thus conclude that our feedback needs gain around the revolution frequency lines only, and in a bandwidth sufficient to include the first synchrotron sidebands. The long delay in the loop will not affect the phase at these frequencies *if it is an exact multiple of the revolution period* T_{rev}. Figure 11 shows the block diagram of the long delay feedback [20]: the cable and electronics delay T is extended to one full turn in the feedback loop. The phase of the correction is thus $0°$ at multiples of f_{rev}. It is wrong by $180°$ at the centre, between two revolution frequency lines. The open-loop gain must have dropped to a value smaller than 1 in order to maintain the stability of the loop there. The frequency response of the filter (including the delay of one turn) is shown in Fig. 12. It is similar to a comb with high gain plus zero phase shift on the revolution frequency lines, and low gain plus $180°$ phase shift in between.

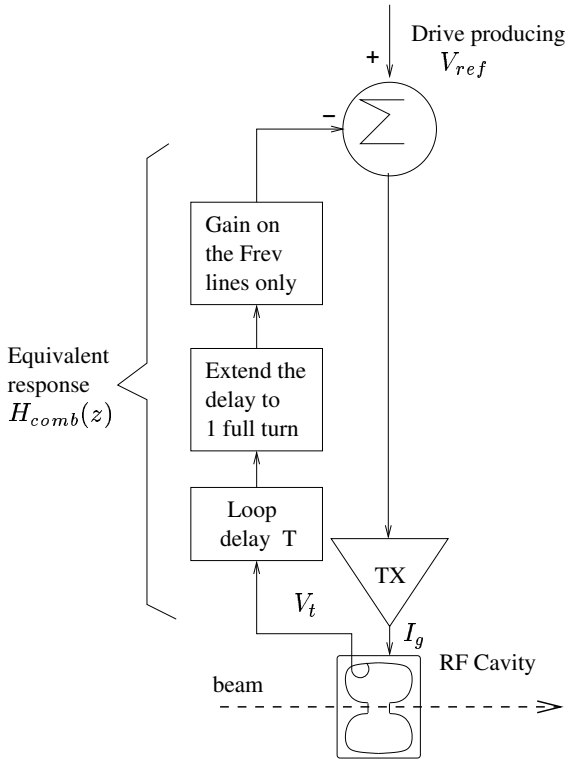


Fig. 11: Block diagram of the long delay feedback method

The comb filter is easily implemented with digital technology.[6] The clock f_c, used to sample the input signal, is obtained by division of the RF frequency f_{rf} so that

$$f_c = M f_{rev} \ . \tag{23}$$

Inside the digital filter a delay of M clock periods therefore implements an exact revolution period, i.e. a delay of one turn. And this remains true with the RF frequency varying during acceleration. The transfer function of the filter[7], including the one-turn delay, is

$$H_{comb}(z) = G \frac{1-a}{1-az^{-M}} z^{-M} \tag{24}$$

[6] Section 4.2 presents the basics of digital filters.
[7] Transfer functions and z-transforms are explained in Section 4.2.3.

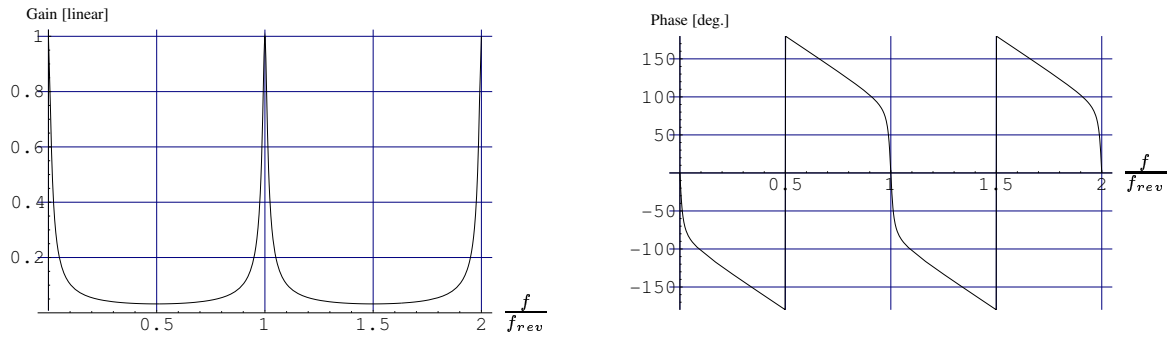


Fig. 12: Frequency response of the revolution frequency comb $H_{comb}(z)$ ($a = 15/16$). Although the phase slips by π halfway between two successive revolution frequency lines, the gain has decreased enough to maintain stability.

Limitation: The parameter a fixes the bandwidth of the filter around each revolution frequency line

$$\Delta f_{-3db} \approx \frac{1}{2\pi M}(1-a)f_c = \frac{1}{2\pi}(1-a)f_{rev} \tag{25}$$

where the approximation holds for a value of a close to 1 (small passband). It also fixes the reduction of the transfer function at half distance between passbands (where $z^M = -1$)

$$H_{min} = G\frac{1-a}{1+a} \approx G(1-a)/2 \ . \tag{26}$$

To maintain reasonable stability, the open-loop gain must be significantly below 1 when the phase is π [16]. It is usual to impose a gain margin of 10 dB ($\approx 1/3$ linear). We get

$$G(1-a)/2 \leq 1/3 \ . \tag{27}$$

The impedance reduction in the passbands is equal to G. Equation (27) shows that a large reduction is possible with a value of the parameter a close to 1. However, Eq. (25) shows that the passband then vanishes. This may not be a problem for the compensation of the transient beam loading (except at injection, see below) but a bandwidth covering at least the first synchrotron sidebands is necessary if the long delay feedback is to act against longitudinal instabilities. The optimal value of a therefore depends on the synchrotron tune $Q_s = f_s/f_{rev}$. The smaller the tune, the more efficient the long delay feedback can be.

Another solution is to use a dedicated system to compensate the beam loading with a transfer function $H_{comb}(z)$ given by Eq. (24) and a parameter a very close to 1. A second system with a different transfer function having gain only on the first synchrotron sidebands (double peaked comb filter) will fight against the dipole mode longitudinal instabilities [21]. An implementation of this double-peaked comb filter is presented in Section 3.3. The transient response at injection may also limit the value of a even though we are only concerned by beam loading. The time needed at injection to compensate beam loading is inversely proportional to the bandwidth Δf_{-3db}. If this time is too long compared to the synchrotron frequency, the mismatch between the bunch emittance and the beam loaded bucket will last sufficiently to create loss of particles. It may even prevent the locking of the beam phase loop. This is unlike a feedforward compensation, which corrects the transient beam loading on the second turn (Section 2.1).

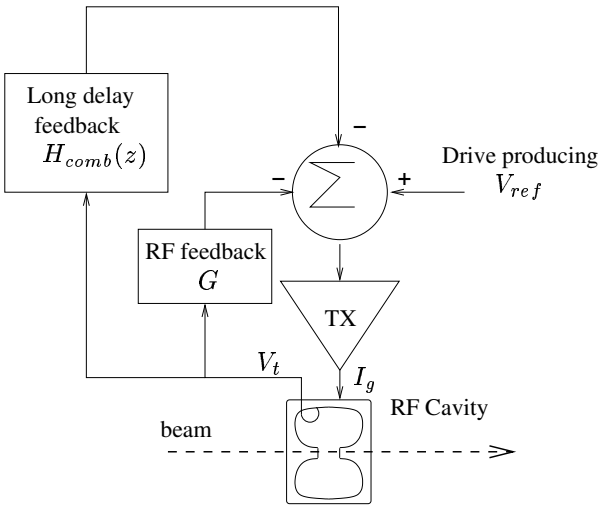


Fig. 13: Long delay feedback and RF feedback on a high Q cavity

Remark: In this section we have not included the cavity impedance Z_g in the open-loop response. If the long delay feedback is used alone on a narrow band cavity, the open-loop response will be modified by the cavity impedance Z_g and the gain and phase shift will not be equal on all frequency lines. Good and stable overall performances will not be easily achieved. In the case of a high Q cavity the long delay feedback is best used in conjunction with an RF feedback (Section 2.2) that effectively transforms the narrowband impedance of the cavity into a broadband response. The long delay feedback is then further reducing the impedance on the revolution frequency lines, in the band where the RF feedback has flattened the cavity response. The whole system is shown in Fig. 13. The drive producing V_{ref} will be generated by the slow cavity amplitude loop and the beam phase loop of the low-level system, which, respectively, try to keep the modulus and phase of V_t at the desired values. They only act on the line at the exact RF frequency f_{rf}. In order to avoid the interference of the long delay feedback with these loops the passband, centred on f_{rf}, is best cancelled in the comb response. An elegant way of rejecting this band is presented in Section 3.1. This combination of RF feedback and long delay feedback is used on the 352 MHz superconducting cavities in the CERN SPS. In the CERN PS it is used on the ferrite cavities [22]. Travelling-wave cavities are different [7]: their bandwidth is large enough for a long delay feedback alone, without RF feedback. As seen from the generator they present a matched load even in the presence of beam loading, so no tuning is needed. The two impedances Z_g and Z_b are, however, different and vary in a complex fashion with frequency. When designing the long delay feedback it is therefore necessary to consider the Z_g impedance. The 200 MHz travelling-wave cavities of the CERN SPS are equipped with a long delay feedback whose response includes a Post Filter $H_{pf}(z)$ in series with the classic comb $H_{comb}(z)$. The Post Filter compensates the effects of the cavity response Z_g [12],[13].

2.4 Bunch-by-bunch feedback

Method: The last method does not involve the RF cavity, neither as a probe nor acting on the beam. It will have no effect on beam loading: its goal is to prevent dipole mode instabilities. It is broadly used in high-intensity lepton machines (synchrotron light sources). The principle is shown in Fig. 14. The longitudinal oscillation of each bunch (dipole mode, shown in Fig. 3) is measured independently. The corresponding signals are processed in parallel in order to generate a dedicated longitudinal kick on each bunch. The kicks are adjusted so that they reduce the amplitude of the oscillations [23].

Let $\hat{\phi}_k$ be the phase of the RF when the centre of charge of the bunch of index k crosses the pick-up. We saw in the first part of these two lectures that, in the presence of a modulation of the RF

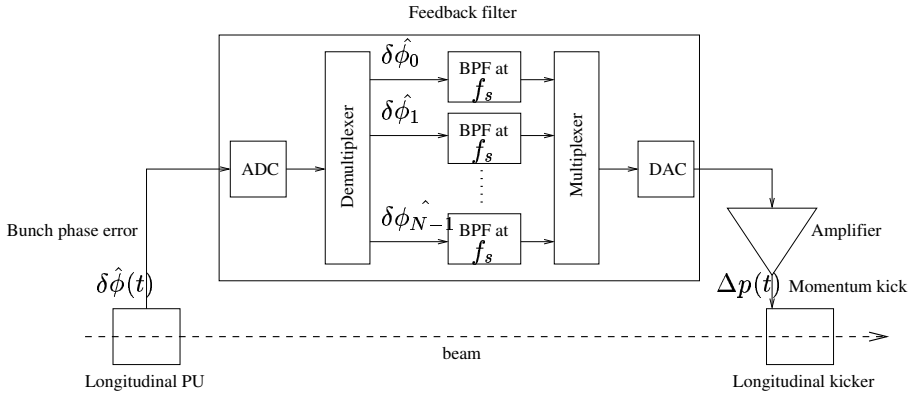


Fig. 14: Block diagram of the bunch-by-bunch feedback method (N bunches)

frequency $\delta\omega_{rf}$, the longitudinal motion of the centre of charge of each bunch obeys the equation

$$\frac{d^2\delta\hat{\phi}_k}{dt^2} + \Omega_s^2\,\delta\hat{\phi}_k = \frac{d\delta\omega_{rf}}{dt} \ , \tag{28}$$

where $\delta\hat{\phi}_k = \hat{\phi}_k - \phi_s$. ϕ_s is the stable phase and $\Omega_s = 2\pi f_s$ [8]. A similar equation can be written if the driving term is a small momentum kick Δp

$$\frac{d^2\delta\hat{\phi}_k}{dt^2} + \Omega_s^2\,\delta\hat{\phi}_k = \frac{\eta h\,(\beta c)^2}{2\pi R_0^2 p}\Delta p \ , \tag{29}$$

where η is the slippage factor related to the energy

$$\eta = \frac{1}{\gamma_t^2} - \frac{1}{\gamma^2} \tag{30}$$

with $\gamma = E/E_0$ (E_0 being the rest energy), γ_t is the value of γ at the transition energy, h is the harmonic number, β is the normalized velocity ($\beta = v/c$), c being the speed of light, $2\pi R_0$ is the machine circumference, and p is the average momentum of the bunch. To obtain damping, we must introduce a term proportional to the first derivative of $\delta\hat{\phi}_k$. This is achieved if we make the momentum kick Δp_k proportional to the *derivative* of the phase error $\delta\hat{\phi}_k$

$$\Delta p_k = a\frac{d\delta\hat{\phi}_k}{dt} \ . \tag{31}$$

The equation of motion then becomes

$$\frac{d^2\delta\hat{\phi}_k}{dt^2} + \Omega_s^2\,\delta\hat{\phi}_k = -\alpha_f\frac{d\delta\hat{\phi}_k}{dt} \tag{32}$$

or

$$\frac{d^2\delta\hat{\phi}_k}{dt^2} + \alpha_f\frac{d\delta\hat{\phi}_k}{dt} + \Omega_s^2\,\delta\hat{\phi}_k = 0 \tag{33}$$

with

$$\alpha_f = -a\frac{\eta h\,(\beta c)^2}{2\pi R_0^2 p} \ . \tag{34}$$

Taking as initial conditions for bunch k a non-zero phase error and a zero frequency error, the feedback would make the phase error go to zero with a time constant $\tau_f = 2/\alpha_f$ (if $\alpha_f \leq 2\Omega_s$) in the

absence of the instability mechanism. The dipole mode instability can be modelled as an additional term, proportional to the first derivative of $\delta\hat{\phi}_k$, with a negative factor

$$\frac{d^2\delta\hat{\phi}_k}{dt^2} + (\alpha_f - \alpha_i)\frac{d\delta\hat{\phi}_k}{dt} + \Omega_s^2\,\delta\hat{\phi}_k = 0 \quad . \tag{35}$$

If the damping rate $\alpha_f/2$ is larger than the instability growth rate $\alpha_i/2$ the feedback will keep the centre of charge of the bunches stable: no dipole mode instability will occur.
Advantage: Bunch-by-bunch feedback fights against *all* sources of dipole mode instabilities no matter where they are located in the machine.
Limitation: Beam loading is not compensated. There is no effect on the instabilities of modes higher than one either (quadrupole mode, sextupole mode, etc.).
Remark: If the feedback filter has the characteristic of a pure differentiator, the high-frequency noise in the pick-up signal will severely degrade the performance (emittance blow-up and reduction of life-time). Since the bunches oscillate at the synchrotron frequency, we can use a bandpass centred at the synchrotron frequency f_s and implement a phase shift of $\pi/2$. For the synchrotron oscillation the band-pass will act as a differentiator and the noise present outside the band will be rejected. Rejection of the DC component is also desirable. We actually measure $\hat{\phi}_k$ but must provide a kick proportional to the derivative of $\delta\hat{\phi}_k = \hat{\phi}_k - \phi_s$. The stable phase offset ϕ_s appears as a DC component in the signal $\hat{\phi}_k$ and the filter will automatically remove it. Other offsets in the acquisition electronics will also have no consequence for the kick. A large bandwidth is required for the acquisition of the phase of each bunch and for the generation of the kicks. Assuming N evenly spaced bunches, the minimum bandwidth is half the bunch frequency $Nf_{rev}/2$ (250 MHz for PEPII and for the Japanese b-Factory KEKB [24]). At first sight the requirement for processing power seems discouragingly high. The synchrotron frequency is, however, much smaller than the revolution frequency (small Q_s). For a given bunch, the momentum kick need not be recomputed at each turn. We can thus reduce the sampling rate (downsampling), i.e. acquire the phase of a bunch only every D turns, compute the corresponding kick, and apply the same kick on the bunch for the coming D turns (interpolation). It may be sufficient to process the signal ten times per synchrotron period instead of every turn. This would result in a large saving if Q_s is small. The downsampling/interpolation method is treated in detail in Section 3.6.

3 IMPLEMENTATION

3.1 Heterodyning

Motivation: The feedforward and long delay feedback imply complex signal processing in a relatively narrow band around the fundamental RF frequency f_{rf}. The bandwidth is rarely more than a few tens of MHz. It is limited by the bandwidth of the RF power chain. On the other hand, the RF frequency can be as high as 500 MHz. When using a digital filter, the processing bandwidth is typically limited to one third of the sampling frequency (see Section 4.2.1). If the RF frequency is low, the cavity signal can be sampled directly. This is the case in the long delay feedback of the CERN PS. The RF frequency remains below 10 MHz. The cavity voltage is sampled at $80f_{rev}$ (from 33.2 MHz at lower energy to 38.4 MHz at higher energy) [22]. Treatment is thus possible up to 13 MHz. The RF frequencies used in the CERN SPS are 200 MHz (travelling-wave cavities) and 352 MHz (superconducting standing-wave cavities). They are both equipped with a long delay feedback whose processing bandwidth covers about 13 MHz on each side of f_{rf}. The sampling frequency should thus be around 1 GHz (for the 352 MHz cavities) if the cavity signal was processed directly, while a more reasonable sampling rate of 40 MHz will be sufficient for the heterodyne system.
Method: The input signal $x(t)$ at the RF frequency is either the cavity voltage V_t (the long delay feedback of Fig. 11) or the beam current I_b (feedforward of Fig. 5). It is fed into the RF input of the I/Q demodulator (Fig. 15).

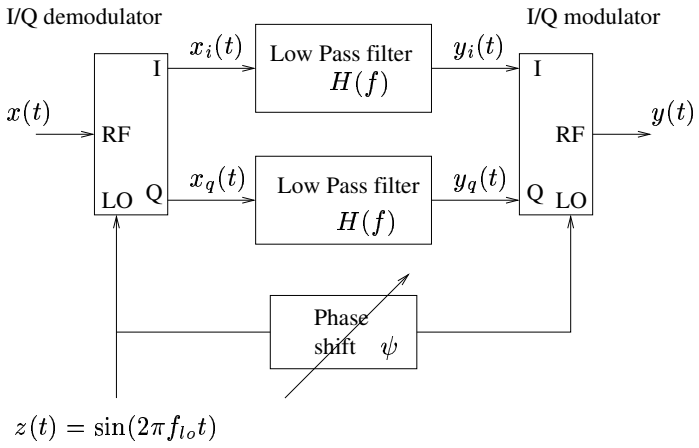


Fig. 15: Heterodyning: bandpass filtering at the RF frequency using two identical low pass filters

A pure sine wave $z(t)$ at frequency f_{lo} is fed into the LO input. The demodulator decomposes the signal $x(t)$ into two components that are in phase and in quadrature with $z(t)$, respectively

$$x(t) = x_i(t)\sin(2\pi f_{lo}t) + x_q(t)\cos(2\pi f_{lo}t) \tag{36}$$

with

$$x_i(t) = x(t)\sin(2\pi f_{lo}t) \tag{37}$$

$$x_q(t) = x(t)\cos(2\pi f_{lo}t) \ . \tag{38}$$

Let $X(f)$ be the Fourier transform of $x(t)$; the Fourier transforms of $x_i(t)$ and $x_q(t)$ are

$$X_i(f) = \frac{X(f - f_{lo}) - X(f + f_{lo})}{2j} \tag{39}$$

$$X_q(f) = \frac{X(f - f_{lo}) + X(f + f_{lo})}{2} \ . \tag{40}$$

The signals $x_i(t)$ and $x_q(t)$ are fed into two identical Low Pass Filters (LPF) whose outputs are $y_i(t)$ and $y_q(t)$

$$Y_i(f) = H(f)X_i(f) \tag{41}$$

$$Y_q(f) = H(f)X_q(f) \ . \tag{42}$$

The I/Q modulator receives on its LO input the sine wave at f_{lo} shifted in phase by ψ. It produces an RF output signal $y(t)$

$$y(t) = y_i(t)\sin(2\pi f_{lo}t + \psi) + y_q(t)\cos(2\pi f_{lo}t + \psi) \ . \tag{43}$$

The Fourier transform of $y(t)$ is

$$Y(f) = \frac{Y_i(f - f_{lo})e^{j\psi} - Y_i(f + f_{lo})e^{-j\psi}}{2j} + \frac{Y_q(f - f_{lo})e^{j\psi} + Y_q(f + f_{lo})e^{-j\psi}}{2} \ . \tag{44}$$

Now using Eqs. (39)–(42) in (44) we get, after simplifications

$$Y(f) = \frac{H(f - f_{lo})e^{j\psi} + H(f + f_{lo})e^{-j\psi}}{2}X(f) \ . \tag{45}$$

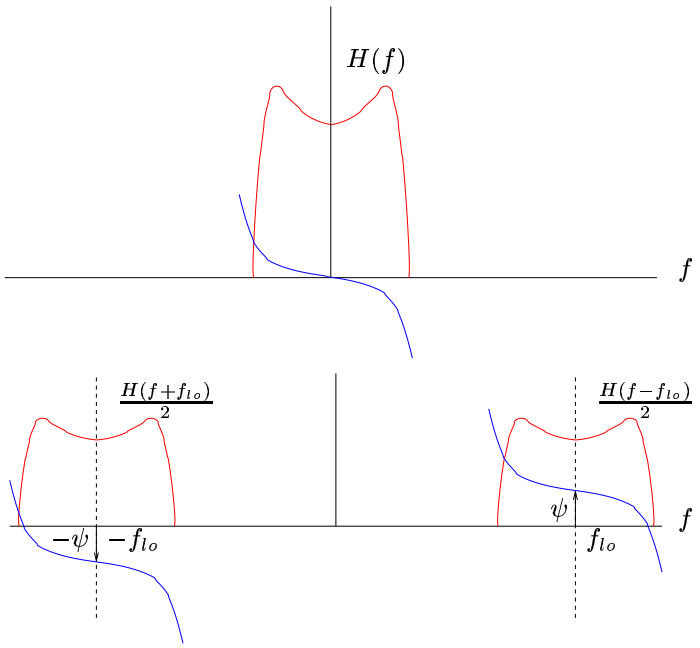


Fig. 16: Top: low pass model. Bottom: BPF implemented by the heterodyne system.

Figure 16 illustrates the above equation: the response at the top is the modulus and phase of the frequency response of the LPF $H(f)$. The response at the bottom is the filtering described by Eq. (45). The response $H(f - f_{lo})$ is identical to $H(f)$, but shifted by f_{lo}, while $H(f + f_{lo})$ is identical to $H(f)$ shifted by $-f_{lo}$. The result is a BPF centred at f_{lo} with an amplitude response, in the band, identical to the low pass model while the phase response is shifted by a constant value ψ. In a heterodyne implementation of the long delay feedback, $x(t)$ is the total cavity voltage V_t. The signal at the LO input of the modulator and demodulator is the RF frequency ($f_{lo} = f_{rf}$). The signals $x_i(t)$ and $x_q(t)$ are AC-coupled to the digital LPF $H_{comb}(z)$ (z-transform given by Eq. (24)). The AC coupling introduces a zero in the overall filtering at the exact frequency f_{rf} so that the long delay feedback does not interfere with the other low-level loops (cavity field amplitude loop, beam phase loop, and the possible cavity tuning loop). The clock frequency for the digital filters is a multiple of the revolution frequency. It is obtained by dividing the RF frequency. The leakage of the LO at the output of the modulator is not important. It introduces a small error in the cavity voltage at f_{rf}, but this is corrected by the other low-level loops. The phase shifter between the LO references fed into the demodulator and modulator is an easy way to finely adjust the phase of the output at the RF frequency. This implementation is used in the CERN SPS for the long delay feedback on the 200 MHz travelling-wave cavities. The digital filters are clocked at 40 MHz. The processing bandwidth is 13 MHz on each side of the RF frequency [12]. In a heterodyne implementation of the feedforward, the input $x(t)$ is the beam current after bandpass filtering to isolate the interesting band around the cavity centre frequency f_0. The signal at the LO inputs is a sine wave at f_0. The digital filters are DC coupled and implement the delay of one turn. Their clock frequency f_c must be a multiple of the revolution frequency so that the overall delay remains one full turn when the RF varies. In the case of a travelling-wave cavity the response H_{opt} is more complex (see Section 3.4). The leakage of the LO at the output of the modulator must be minimized because it is not at the RF frequency f_{rf} and will thus not be corrected by the low-level loops. The CERN SPS 200 MHz cavities are equipped with a feedforward working in tandem with a long delay feedback for the acceleration of the proton beam for the LHC. The RF frequency ramps from 200.264 MHz to 200.395 MHz. The cavity centre frequency (f_{lo}) is constant at 200.222 MHz. The digital filters are clocked at $1/10$ the RF frequency (20 MHz) [12].

In the range of frequencies used in particle accelerators, I/Q modulators and demodulators are readily available, for example:

- MIQ family from Mini-Circuits: RF/LO 9 MHz to 1.9 GHz, I/Q DC to 5 MHz, 40 dB sideband rejection.
- IM/ID family from Pulsar: RF/LO 10 MHz to 1.9 GHz, I/Q DC to 50 MHz, 30 dB sideband rejection.
- QM/SM family from Synergy: RF/LO 20 MHz to 1.9 GHz, I/Q DC to 50 MHz, 30 dB sideband rejection.

The heterodyne system of Fig. 15 implements an RF bandpass filter using two identical digital low pass filters. An alternative is shown in Fig. 17: instead of mixing the carrier frequency down to DC, we mix it down to an intermediate frequency (IF). In order to implement the IF bandpass filter $H_{if}(f)$ with digital technology, the IF output of the mixer must be sampled.

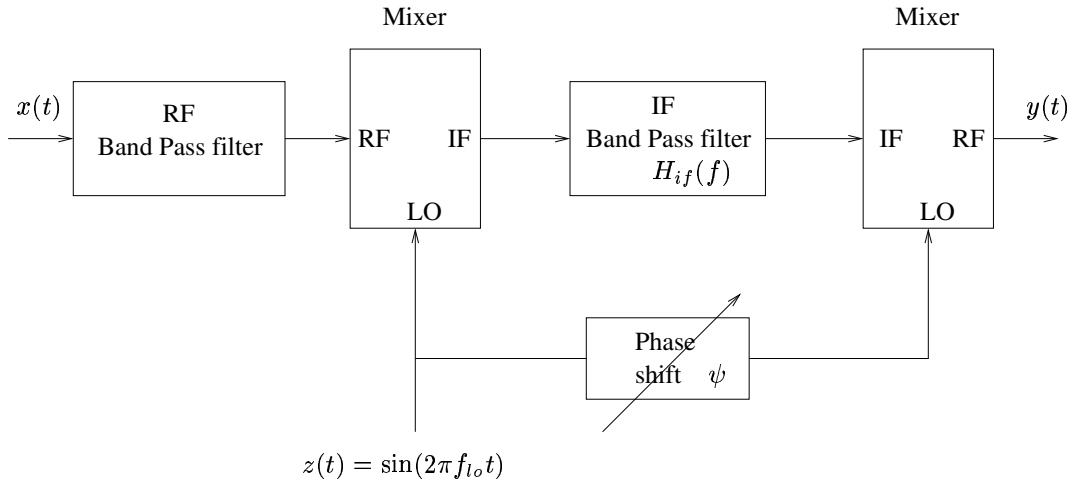


Fig. 17: Heterodyning: Bandpass filtering at the RF frequency using a BPF at the IF frequency

Let us take, as an example, the long delay feedback presented above. The desired bandwidth is 13 MHz on each side of the RF frequency at 200 MHz. The use of an IF frequency at 13 MHz, that is $f_{lo} = 187$ MHz, is not recommended because the noise in the band from 161 MHz to 187 MHz will create, at the mixer output, image signals corrupting the interesting signal band (from DC to 26 MHz). The RF bandpass filter centred at 200 MHz first selects the desired signal band with minimal distortion (from 187 MHz to 213 MHz), and ideally rejects the rest of the spectrum so no image signals are present at the mixer output. In practice, accepting 1 dB attenuation at the extreme of the signal band ($f_{rf} \pm$ 13 MHz), we will get 20 dB attenuation at $f_{rf} \pm 25$ MHz using an RF bandpass filter available from industry (of 4 to 5 sections).[8] The image at the mixer output will thus be rejected by 20 dB if we use an intermediate frequency at 25 MHz ($f_{lo} = 175$ MHz). At the input of the digital filter the signal will cover the band from 12 MHz to 38 MHz, implying a minimum sampling frequency of around 100 MHz, compared to the 40 MHz sufficient in the system of Fig. 15. This example shows that the system using the I/Q demodulator and modulator is less demanding on digital bandwidth. But two identical digital filters are needed while one is sufficient in the system of Fig. 17.

[8] Theoretical attenuation curves give, respectively, 22 dB and 29 dB attenuation at ± 25 MHz for a five section Butterworth and a four section 0.3 dB passband ripple Chebishev, with 1 dB attenuation at ± 13 MHz. In practice, given these passband specifications, companies designing custom RF filters (RLC Electronics for example) will propose a $4-5$ section filter providing 20 dB attenuation at ± 25 MHz.

In the following sections, we study heterodyne implementations of the feedforward and long delay feedback systems. One could also implement an RF feedback using an I/Q demodulator, followed by two digital low pass filters and an I/Q modulator, as shown in Fig. 15. The I and Q components of the desired accelerating voltage should be subtracted from the I and Q output of the demodulator (cavity voltage) so that the loop gives precise control of the accelerating voltage [25]. This architecture is used to linearize power amplifiers in mobile communication systems, where it is called Cartesian feedback [26]. For its application to an accelerator cavity, and if the main concern is beam loading, one must make sure that the digital processing does not contribute significantly to loop delay, thereby limiting the achievable impedance reduction. A variant of the I/Q feedback is proposed for the RF cavities of the synchrotron light source Soleil [27].

3.2 Analog-to-digital conversion and back

The performance (speed and resolution) of the Analog to Digital and Digital to Analog Converters (ADC and DAC) available on the market are sufficient for our applications. State-of-the-art examples are available from Analog Devices: 12-bit ADC at 105 Msamples/s (AD9432), 14-bit DAC at 125 Msamples/s (AD9754). Note that, with a 12-bit ADC covering the analog range of 1 V peak-to-peak, one Least Significant Bit (LSB) corresponds to 250 μV, i.e. -65 dBm. The noise in the RF signal plus eventual glitches on the power supply lines and noise because of poor grounding must stay below that level to take full advantage of the digital resolution. An efficient ground plane is essential to prevent the pollution of the sensitive analog signals by the noise due to the fast commutations of the digital circuitry. The designers of the ADC and DAC provide valuable information (see for example Ref. [28]).

3.3 Implementation using discrete numerical operators

Principle: The filtering realized by a digital filter can be represented as a difference equation relating the output sequence y_n to the input sequence x_n. In our applications n will be the time index: $x_n = x(nT_c)$, where T_c is the period of the clock ($T_c = 1/f_c$). For example, the output y_n of the comb filter H_{comb} (z-transform given in Eq. (24)) is related to the input x_n by the difference equation

$$y_n = a\, y_{n-M} + G(1-a)x_{n-M} \quad , \tag{46}$$

where the sequence x_{n-M} is obtained by delaying the sequence x_n by M samples. (Section 4.2.4 shows how to derive a transfer function from the difference equations.) A linear, constant-coefficients, difference equation can be implemented with only three basic operations: addition, delay and multiplication by a constant. The digital filter can thus be realized using three discrete components:

- Addition/Subtraction using 16-bit Arithmetic/Logical Units (ALU) such as IDT7381 (Integrated Device Technology) or L4C381 (Logic Devices). Operation in 15 ns.
- Delay of any depth using First-In First-Out memories (FIFO) such as CY7C42X5 (Cypress) or IDT722X5 (Integrated Device Technology). An 18-bit FIFO of variable depth up to 4 k words. Operation in 10 ns.
- Multiplication using 16×16-bit multipliers such as IDT7216 (Integrated Device Technology). Operation in 16 ns.

Example: As an example we present the double-peaked comb filter installed on the 352 MHz superconducting cavities of the CERN SPS. Four cavities providing a total accelerating voltage of 32 MV were installed to accelerate leptons to 22 GeV for LEP. They are equipped with a strong RF feedback (Section 2.2) that flattens the cavity response in a band of 1 MHz around 352 MHz [15]. The tetrode amplifiers are located next to the cavities and the overall loop delay is 500 ns. The SPS is also accelerating a high-intensity proton beam and the impedance of these cavities triggers longitudinal instability for this beam. Beam loading is not a problem. The ring is almost completely filled with bunches spaced by 5 ns, so that the beam current I_b has very little power around 352 MHz. An additional long delay

feedback was designed to further reduce the cavity impedance at the synchrotron sidebands of the revolution frequency lines. The two systems work together as shown in Fig. 13. The long delay feedback uses the heterodyne method shown in Fig. 17. The intermediate frequency (IF) is 4 MHz. The LO is thus at a 4 MHz offset from the centre frequency 352 MHz. The RF bandpass filter selects a band of ± 1.2 MHz around 352 MHz (1 dB bandwidth) and gives 13 dB attenuation at ± 4 MHz. Note that the interesting band is rather small (2.4 MHz). The digital filter is clocked at 20 MHz. It is a variant of the double-peaked comb filter proposed in Ref. [21]. The synchrotron tune is so small ($f_s \leq 1$ kHz, $f_{rev} \approx 43$ kHz) that the peaks on the sidebands can be realized by placing zeros and double poles on the revolution frequency lines. The z-transform is[9]

$$H_{sbd}(z) = \frac{1 - z^{-M}}{(1 - az^{-M})(1 - az^{-M})} z^{-M} \tag{47}$$

with $M = 462$, $f_c = 20$ MHz. The parameter a can be varied. As it gets closer to 1, the achievable impedance reduction increases but the peaks move closer to the revolution frequency lines. We use $a = 31/32$ or $a = 63/64$ providing an impedance reduction of 20 dB and 26 dB respectively on the synchrotron sidebands with the classic 10 dB gain margin. Figure 18 shows the modulus of the frequency response.

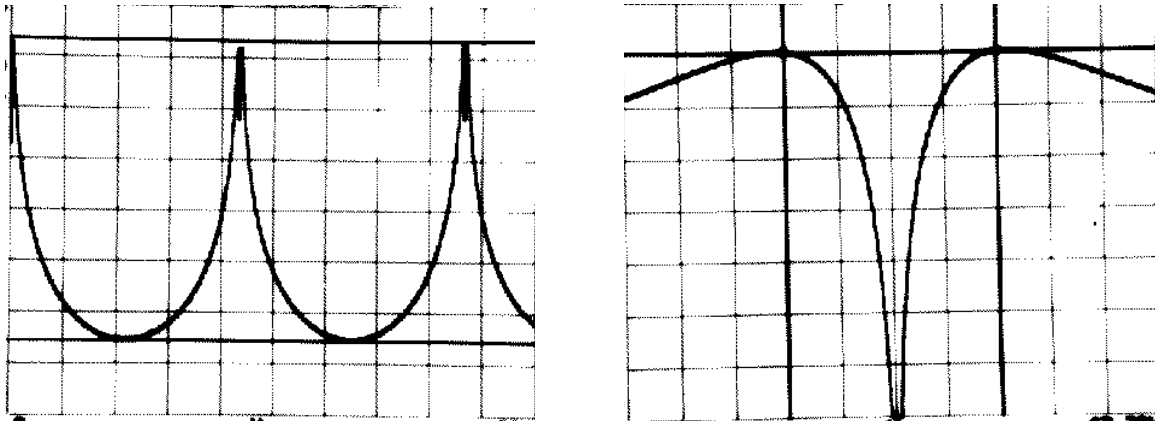


Fig. 18: Frequency response of the double-peaked comb filter, showing peaks on the synchrotron sidebands of the revolution frequency lines. Left: 100 kHz span, 5 dB/div. Right: enlargement (1 kHz span, 3 dB/div) around one f_{rev} line, showing the two peaks on the sidebands ($a = 31/32$, $f_{rev} = 43$ kHz).

The corresponding impedance reduction is shown in Fig. 19. The smooth trace shows the cavity impedance with the RF feedback. The comb shows the additional reduction achieved around the revolution frequency lines with the double-peaked long delay feedback.

In the time domain, the filtering realized by $H_{sbd}(z)$ can be implemented as the following set of difference equations (see also Section 4.2.4). The term x_n is the input and y_n the output:

$$v_n = a\, v_{n-M} + x_n \tag{48}$$

$$w_n = a\, w_{n-M} + v_n - v_{n-M} \tag{49}$$

$$y_n = w_{n-M} \ . \tag{50}$$

Figure 20 shows a direct implementation of the above equations. The delays are implemented using FIFOs. The multiplications by $31/32$ are replaced by subtraction of a shifted version of the operand

[9] This z-transform is analysed in Section 4.2.3.

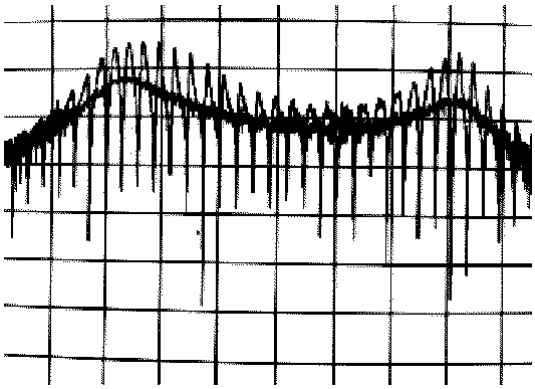


Fig. 19: Apparent impedance of the CERN SPS 352 MHz superconducting cavity. The smooth trace shows the impedance with the RF feedback alone. The second trace shows the additional reduction around the revolution frequency lines with the double-peaked long delay feedback: $a = 31/32$, 150 kHz/div, 3 dB/div. The resolution of the measurement is not sufficient to show the full 20 dB reduction or to separate the sidebands.

from itself. (In two's complement arithmetic a scaling by 32 is easily implemented by shifting the binary word five positions towards the LSB, with extension of the most significant bit.) Additions and subtractions are realised in the ALUs.

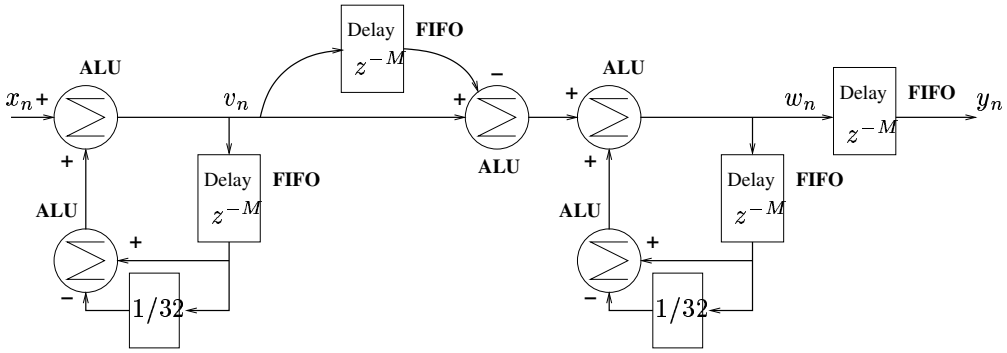


Fig. 20: Implementation of the double-peaked comb filter using ALUs and FIFOs

The advantage of such an implementation using discrete numerical operators is its speed. A new output signal is available on each clock pulse. A 50 MHz clock rate is easily achievable with 16-bit words. It is a straightforward implementation for digital filters containing loops, i.e. with poles in their z-transform (also called Infinite Impulse Response filters or IIR).
The main limitation is related to the use of fixed-point arithmetic:

- Rounding noise: after the shift by five positions (multiplication by $1/32$) in the above loops, the binary word must be truncated (or rounded) to remain a 16-bit word. This rounding creates an error (called rounding noise) that will propagate to the output of the filter.
- Rounding of filter coefficients: the coefficients of the filter must also be quantized. In the above design we chose $a = 31/32$, which can be implemented exactly with 16 bits but, in general, the coefficients must be rounded to fit the binary word and this will modify the filter response.
- Scaling: the input must be scaled properly to guarantee that the outputs of the ALUs do not overflow.

- Choice of structure: given a z-transform, many sets of difference equations can realize it. But the choice of the realization is important because some structures are much more sensitive to the negative effects of fixed-point arithmetic than others. In general, if the filter has a high order (many poles and zeros), a cascade of low-order sections is preferred (bi-quad sections having a maximum of two zeros and two poles) instead of direct form realization.

This type of implementation does not have much future. There is no new development from chip designers regarding discrete ALUs and multipliers. New FIFOs are, however, periodically being introduced.

3.4 Implementation with video products

Principle: Recently (end 1998) a new family of products was introduced for digital image processing in High Definition TV (HDTV) applications. Their speed and resolution make them ideal for our applications. The LF3320 (Horizontal Digital Image Filter) from Logic Devices implements a 32-tap Finite Impulse Response filter (FIR, see Section 4.2.2) at data rate of 83 MHz with a resolution of 12 bits for data and coefficients $h_0, h_1, ..., h_{L-1}$

$$y_n = h_0 x_n + h_1 x_{n-1} + h_2 x_{n-2} + \cdots + h_{L-1} x_{n-L+1} \quad . \tag{51}$$

The advantages of such an implementation are

- Accuracy: if the input data x_n and the coefficients $h_0, h_1, ..., h_{L-1}$ are 12-bit words in the above equation, the result y_n must be extended to 29 bits to avoid overflow and rounding (for $L = 32$). Internally, the chip keeps all these bits, and by programming the output scaling one chooses which 16 bits come out.
- Flexibility: by programming (with an on-board ROM for example or external host), one changes the filter coefficients, the output scaling and the output limiting. This facility to bound the output is very attractive in set-ups where the filter output is the drive to a power amplifier. An overflow in two's complement arithmetic results in the output of the DAC dropping from the maximum positive value to the minimum negative value. This fast transient is likely to trip the power amplifier.
- It is cascadable for larger filters (more than 32 coefficients).
- It supports decimation up to 16:1 with a correspondingly increased number of filter coefficients.
- It is targeted at a promising new market and new developments can be expected.

The disadvantages are

- It can implement FIR filters only. These have only zeros in their transfer function (no pole).
- The range is limited by the use of fixed-point arithmetic (12 bits). It suffers from the need to quantize the coefficients. Rounding noise is not a big problem. It is here limited to one LSB at the output since the FIR keeps all bits in the computation.

Example: The performance of the feedforward on the CERN SPS 200 MHz travelling-wave cavities was presented in Section 2.1 (Fig. 6). The implementation uses the heterodyne method shown in Fig. 15. Each of the two identical digital filters is realized using two FIRs (LF3320) clocked at $f_c = 20$ MHz. The optimal transfer function H_{opt} is complex for a travelling-wave cavity because Z_g is not equal to Z_b. We have [29]

$$Z_g(\Delta\omega) = l\sqrt{\frac{Z_0 r_2}{2}} \left(\frac{\sin \tau/2}{\tau/2}\right) \tag{52}$$

where the phase slip $\tau(\Delta\omega)$ is

$$\tau(\Delta\omega) = \frac{l}{v_g}\left(1 - \frac{v_g}{v}\right)\Delta\omega \tag{53}$$

with $\Delta\omega = \omega - \omega_0$. The term ω_0 is the angular frequency of the cavity fundamental resonance, l is the interaction length of the cavity, v is the particle velocity, v_g is the group velocity in the cavity, r_2 is the series impedance (Ω/m^2), and Z_0 is the characteristic impedance of the RF chain (50 Ω). Notice that Z_g is purely real but its sign changes periodically as a function of frequency. The impedance Z_b is

$$Z_b(\Delta\omega) = -\frac{l^2 r_2}{8}\left[\left(\frac{\sin\tau/2}{\tau/2}\right)^2 - 2j\left(\frac{\tau-\sin\tau}{\tau^2}\right)\right] \tag{54}$$

It has both a real and an imaginary part. The naive solution $H_{opt} = -Z_b/Z_g$ will not work for a travelling-wave cavity because Z_g vanishes at frequencies where the imaginary part of Z_b is non-zero. Let us decompose H_{opt} into a real and an imaginary part:

$$H_{opt} = H_{opt}^{re} + jH_{opt}^{im} \ . \tag{55}$$

Since Z_g is real valued, our design goal is

$$H_{opt}^{re} Z_g + jH_{opt}^{im} Z_g \approx -Re[Z_b] - jIm[Z_b] \ . \tag{56}$$

By inspection of Z_g, (Eq. (52)), and Z_b, (Eq. (54)), it follows that the real parts on the two sides can be made identical:

$$H_{opt}^{re} = \frac{l}{4}\sqrt{\frac{r_2}{2Z_0}}\left(\frac{\sin\tau/2}{\tau/2}\right) \tag{57}$$

and the compensation of the resistive part of beam loading is exact. The impulse response of H_{opt}^{re} is rectangular (inverse Fourier transform of H_{opt}^{re}), lasting for a time equal to $\frac{l}{v_g}\left(1-\frac{v_g}{v}\right)$. In the case of the SPS cavities this corresponds to 12 samples at 20 MHz. This is implemented with the first FIR filter (12 coefficients are equal to 1, the other coefficients being 0). The value jH_{opt}^{im} is implemented with a second FIR filter. Its impulse response h_n^{im} is limited to 31 samples ($-15 \le n \le 15$) and must be odd-symmetric ($h_{-n}^{im} = -h_n^{im}$) so that its frequency response is purely imaginary. Perfect compensation is not possible, however, because Z_g is zero at frequencies where the imaginary part of Z_b is non-zero. We must therefore choose a criterion for computing the optimal coefficients $h_1^{im}, h_2^{im}, ..., h_{15}^{im}$. The details are presented in Ref. [12]. The outputs of the two FIR filters are finally added together using an ALU (IDT7381 or L4C381).

3.5 Implementation with Programmable Gate Array (PGA)

The PGA (or Programmable Logic Device, PLD) is a high-density gate array (up to 200 000 gates on one chip) that the user can configure to implement the desired function. Modern devices also include a memory and some specialized functions on the chip (embedded programmable logic arrays).

The advantages of the PGA are:

- It helps ensure very compact hardware by minimizing the external connections between chips.
- It can be re-configured on the board.
- Powerful development and debugging tools are available. It is possible to simulate both the logic function and the delays to evaluate the possible overall processing rate.
- PGAs are widely used and we can therefore expect new developments in the field.

The drawback is that the resulting design will implement fixed-point arithmetic only.

The bunch-by-bunch feedback at the Japanese b-Factory KEKB is an example of an implementation with PGA [24].

Figure 21 shows a block diagram of the feedback filter. The longitudinal position of each bunch is measured with a wide band phase detection system (not shown here) capable of distinguishing individual bunches spaced by less than 2 ns (maximum bunch frequency 509 MHz) [31]. The result is an analog

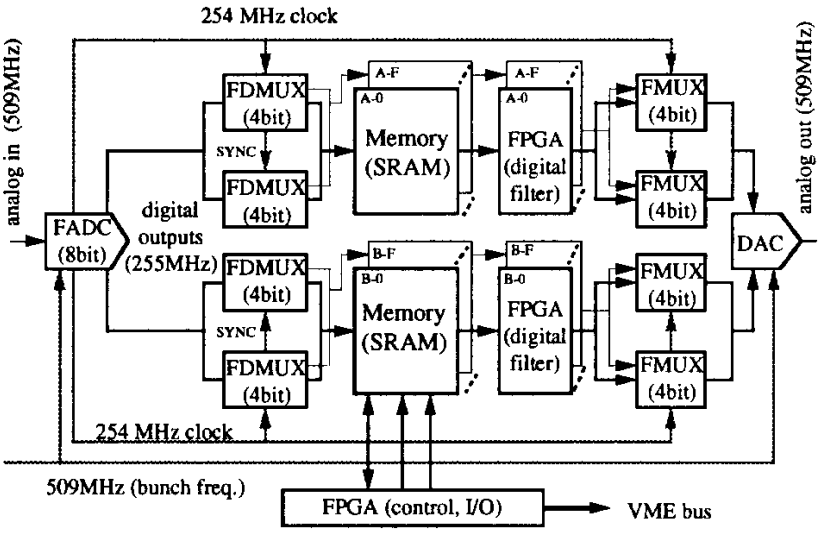


Fig. 21: Digital filter of the KEKB bunch-by-bunch feedback (reproduced from Ref. [30]). The Field Programmable Gate Array (FPGA) implements the 2-tap filtering of Eq. (58).

signal whose average value, in each 2 ns window, is a measurement of the phase $\hat{\phi}_k$ of the corresponding bunch. This signal is fed, on the left, into the digital filter shown on the figure. The ADC output is first demultiplexed in 2 channels, each being further demultiplexed in 16 channels. The processing rate is thus finally reduced to 16 MHz. The Field Programmable Gate Array (FPGA) QL16 $\times$ 24B made by Quick Logic [32] implements a subtraction between its two 8-bit operands. These are actually the same time sequence x_n, but one is delayed by L samples with respect to the other, thereby implementing a two-tap FIR filter with fixed coefficients 1 and -1. The output of the subtractor y_n is related to its input x_n by the difference equation

$$y_n = x_n - x_{n-L} \ . \tag{58}$$

The delay L is adjusted to get the required 90° phase shift at the synchrotron frequency. With these coefficients (1 and -1) the FIR filter also eliminates the DC component (stable phase).

The outputs of the 32 channels are finally multiplexed and converted into analog (the output signal is shown on the right of the figure). This analog signal is fed into the IF port of a mixer, receiving twice the bunch frequency on its LO (not shown here). The RF output drives the amplifiers feeding the wide band kicker. The design of the digital filter dates from the mid-1990s. The FPGA is only used to implement the subtraction between two 8-bit words, which are properly aligned in time by the memories (Static Random Access Memories, marked SRAM on Fig. 21). These two functions could be integrated in a single Embedded PGA chip nowadays, which would make the implementation more attractive. Since the speed of the components has improved, using the latest technology would also reduce the number of channels. As an example of the power of modern PGAs, Altera proposes software called MegaCore to implement FIR filters using its FLEX10KE family. With this tool, it quotes the realization of a 19-tap FIR filter (8-bit input) at 101 MHz throughput rate [33]. In the CERN SPS we have used this technology to upgrade the electronics of the transverse damper for the LHC beam: a digital filter for closed-orbit rejection[10] has been designed with a single chip FLEX 10K100E from Altera. It works with 12-bit inputs and has a throughput of 80 MHz [30].

[10] The minimal digital filter of a *transverse* bunch-by-bunch feedback produces notches at the multiple of the revolution frequency and introduces a delay of one full turn [34]: $H(z) = (1 - z^{-M})z^{-M/2}$. The required 90° phase shift can be obtained from the betatron phase advance between pick-up and kicker.

3.6 Implementation with digital signal processors

Digital Signal Processors (DSPs) are microprocessors that specialize in performing repetitive mathematical operations. Most DSPs are based on the Harvard architecture. The core processor is connected to two separate memories by two separate buses so that two memory accesses can be made in one cycle. For example a new instruction can be fetched from one memory while data is fetched from the other. In the implementation of an FIR (Eq. (81)), one item of data x_{n-i} can be fetched from one memory and one coefficient h_i fetched from the other memory (if the instruction is already in a third smaller memory called the cache). All DSPs incorporate a hardware multiplier/accumulator unit, which, combined with the Harvard architecture, makes it possible to perform a complete floating-point multiply-accumulate operation in a single clock cycle, including the fetch of the two operands. Older DSPs implement only fixed-point arithmetic. The ADSP-2106x from Analog Devices and the TMS320C6x from Texas Instruments are in use in particle accelerators. In the first part of these two lectures [8] the beam control system of the RHIC at Brookhaven National Laboratory was presented. It uses two DSPs (TMS320C40) clocked at 50 MHz [35]. Some of the systems presented here can also be implemented with DSPs.

The advantages of the DSPs are:

- Modern DSPs implement floating-point operations in the 32-bit IEEE standard. It was shown above that the limited dynamic range of fixed-point arithmetic results in the need for careful scaling of inputs and intermediate signal levels to avoid overflow. Here, the need for such scaling is essentially eliminated by using floating-point arithmetic. Rounding noise is still introduced at each operation but the signal-to-noise ratio at the output of the filter is significantly better for floating point arithmetic as compared with fixed-point arithmetic.[11] The filter coefficients still need to be quantized but this is done at constant relative precision.
- The industry provides complete DSP systems. For example a printed circuit board containing several DSPs linked together, external memories, input/output ports and a VME bus for connection to an external host. Very complete software tools are also provided, such as extensive signal processing libraries to implement FIR and IIR filters, Fast Fourier Transform (FFT). The prototype can be tested in an emulator. Debugging tools are also available. The algorithm implemented by the DSP can be re-programmed at will.
- The filtering realized by the DSP can be very complex. For example one could imagine adjusting the coefficients of the bunch-by-bunch feedback BPF continuously (Fig. 14) so that it remains centred on the synchrotron frequency f_s during the acceleration. The low-level system developed at RHIC and presented in the first part of these two lectures [8] is a good example of the sophistication made possible by DSPs: the feedback gains (radial loop, beam phase loop and synchronization loop) are varied during the acceleration to keep the poles of the closed-loop response at the optimal location while the synchrotron frequency varies. Notice, however, that these gains are adjusted at a very slow rate: only 200 times during the 74 s long acceleration ramp [35].
- Control and diagnostic functions can easily be integrated. Additional software code can be added to change the filtering function or to read interesting signals (see the example of ALS shown below).
- The market for DSPs is growing fast. We can expect continuing improvements in their performance over coming years.

The drawback is that the DSPs are still very slow for RF applications. They may have a very fast clock frequency, they still have only one multiplier/accumulator unit and will perform only one multiplication at a time.[12] It therefore takes many clock cycles to produce a single output sample when

[11] This is obvious if we compare the 32-bit floating point IEEE standard (23-bit in the mantissa) to the classic 8- or 16-bit fixed point format. But even with a fair comparison taking the same number of bits in the mantissa of the floating format as in the fixed-point format one concludes that the output signal-to-noise ratio is better with the floating point implementation [36].

[12] The Single-Instruction-Multiple-Data (SIMD) DSPs from Analog Devices have two arithmetic units executing the same instruction. But performance increases only if two independent filtering channels have to be implemented.

the DSP must implement a filtering operation. For example, suppose that we wish to use a DSP to implement a L taps FIR filter for the feedforward presented in Section 3.4 (feedforward on a travelling-wave cavity). Generating the code using the optimized assembly-language library of the ADSP-21000 family, it takes $5 + L$ cycles to compute a single output value y_n [37]. Thirty-six cycles are needed to generate one output of the 31-tap FIR. If the DSP is clocked at 100 MHz,[13] we get an output signal every 360 ns, compared with the 12 ns needed by the video filter LF3320 clocked at 83 MHz. In this application, the sampling rate of the DSP implementation would be 2.8 MHz and the useful bandwidth less than 1 MHz (one-third the sampling rate). In our RF applications DSPs are only a good candidate if the processing rate can be much reduced by downsampling and interpolation. If parallel processing is possible, one could also compensate the low rate by using several DSPs in parallel, as shown in the following example.

The bunch-by-bunch feedback system of the ALS (synchrotron light source) at LBNL is implemented using DSPs (Fig. 22) [23].

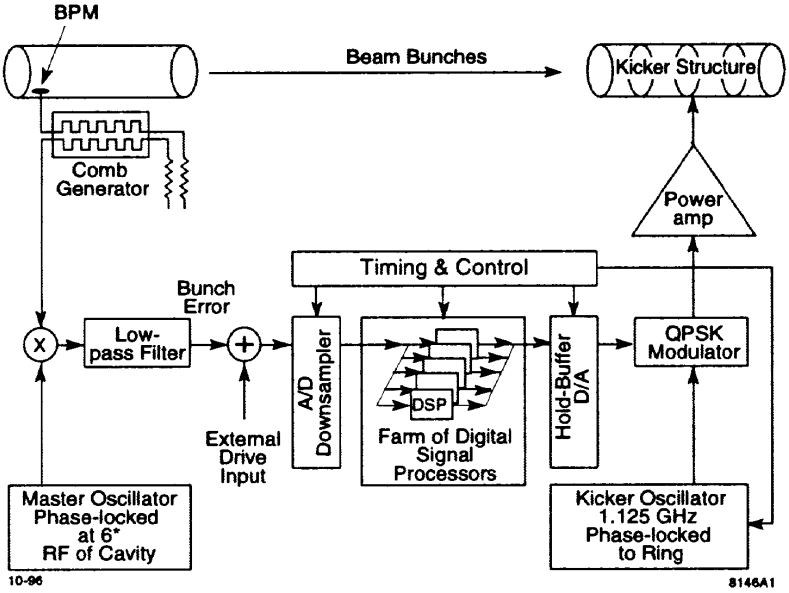


Fig. 22: Block diagram of the longitudinal bunch-by-bunch feedback used at ALS (reproduced from Ref. [23])

The machine parameters are 328 bunches, 2 ns spacing, revolution frequency $f_{rev} = 1.5$ MHz, and synchrotron frequency $f_s = 11$ kHz. The acquisition part is similar to the one implemented at KEKB and we shall not detail it. The bunch phase signal $\hat{\phi}$ is fed into the ADC. Recall that, for each bunch k, the bunch-by-bunch feedback must implement a separate BPF processing the average bunch phase $\hat{\phi}_k$. Since the processing of each bunch is independent of the others, we can treat them in parallel, using a bank of DSPs, each branch implementing one BPF with an L-tap FIR. The throughput rate could be reduced to a comfortable 1.5 MHz (f_{rev}) with 328 DSPs. Realization becomes possible once we notice that the synchrotron tune is small ($Q_s = f_s/f_{rev} = 7.3 \times 10^{-3}$). The phase of each bunch does not change much between two turns. Let us therefore reduce the processing rate by a factor D: for each bunch, we treat its signal every D turns only. This process is called downsampling. A reasonable value for the downsampling ratio D is a fraction of $1/Q_s$. For N bunches the total number of multiply–accumulate

[13] Texas Instruments advertises a new 600 MHz DSP including four multiplier/accumulator units [38] (TMS320C64x). We have not evaluated it yet.

cycles per second (MAC/s) using an L-tap FIR is[14]

$$\frac{NLf_{rev}}{D} \quad . \tag{59}$$

In the case of the ALS, taking 10 samples per synchrotron period ($D = 1/(10Q_s) \approx 14$) and with a 20-tap FIR, the processing power required is 7.03×10^8 MAC/s. Considering a modern DSP clocked at 100 MHz, realizing one multiply–accumulate per cycle, we see that the above bunch-by-bunch feedback can be implemented using only eight such DSPs, each one processing the signal from 41 bunches and producing one output per bunch every 14 turns. The kick value for each bunch will thus be kept constant over 14 turns (Hold–Buffer DAC).[15] This derivation is, however, overly optimistic: the ALS design actually uses 40 DSPs in parallel. In addition the DSPs provide a powerful diagnostic tool. The phase signal from each bunch can be observed individually. The growth rate of each mode of instability can be measured by switching the feedback off, letting the oscillation develop, and switching the feedback back on. This allows measurements in the small-signal regime [23]. Acquisition of these signals is easily achieved through the existing interface between the DSP and its host computer.

4 ANNEXE

4.1 Spectrum of the beam-induced voltage

First consider a uniform ring distribution with all buckets filled and suppose that the bunch length is infinitely short. The beam current measured in a pick-up then consists of a series of infinitely narrow pulses spaced in time by the RF period.

$$i_\delta(t) = I_0 T_{rf} \sum_{l=-\infty}^{+\infty} \delta(t - lT_{rf}) \tag{60}$$

where I_0 is the DC component of the beam current. Using the identity

$$\sum_{l=-\infty}^{+\infty} \delta(t - lT) = \frac{1}{T} \sum_{m=-\infty}^{+\infty} e^{j2\pi mt/T} \tag{61}$$

the beam current can be rewritten

$$i_\delta(t) = I_0 \sum_{m=-\infty}^{+\infty} e^{j2\pi m f_{rf} t} \quad . \tag{62}$$

The corresponding spectrum is an infinite set of discrete lines at the RF frequency and its harmonics: $f = \pm m f_{rf}$. In practice the bunches are not infinitely short. Let $\lambda(t)$ be the normalized longitudinal charge density, the beam current is

$$i_\lambda(t) = I_0 T_{rf} \sum_{l=-\infty}^{+\infty} \lambda(t - lT_{rf}) = i_\delta(t) * \lambda(t) \quad . \tag{63}$$

The current i_λ is the convolution of the two time domain signals $i_\delta(t)$ and $\lambda(t)$. The spectrum of the result is thus obtained by multiplying the spectrum of i_δ by the Fourier transform of the bunch charge density λ. This spectrum still consists of a set of discrete lines at $f = \pm m.f_{rf}$ but the amplitude of the spectral lines in the frequency domain will decrease with an envelope equal to the Fourier transform of the charge density $\lambda(t)$. The width of the envelope is inversely proportional to the bunch length. The

[14] We have neglected the small overhead of 5 cycles in the total number of cycles $5 + L$.

[15] A proper downsampling/interpolation filter (also called multirate filter) is somewhat more complex. See Section 4.2.5 for more details.

bandwidth of the accelerating system around f_{rf} being much narrower than the RF frequency we can neglect all the harmonics. The non-zero bunch length then only introduces a scaling factor in the spectral component at f_{rf}. For our analysis the beam current can thus be simplified to

$$i_{uniform}(t) \propto \sin(2\pi f_{rf} t + \phi) \ . \tag{64}$$

that is, a single line at the RF frequency. This remains valid if the bunch spacing is a multiple of the RF period (uniform ring but not all buckets filled), assuming that the bunch frequency (inverse of the bunch spacing) is much larger than the bandwidth of the RF system so only the spectral component at f_{rf} need to be considered.

We now consider a non-uniform ring distribution. This is the case if the ring pattern consists of one or several batch(es) of bunches and one (or more) hole(s) without bunches. As seen in a pick-up (or in a cavity), the beam current is then modulated by a function $a(t)$ representing the batch envelope

$$i_{batch}(t) \propto a(t)\sin(2\pi f_{rf} t + \phi) \ . \tag{65}$$

The same bunch pattern periodically crosses the pick-up. The modulation function $a(t)$ is thus periodic in time with a period equal to the revolution period T_{rev}. We can expand it as a Fourier series and the beam current becomes

$$\begin{aligned} i_{batch}(t) \quad \propto \quad & [a_0 + a_1\cos(2\pi f_{rev} t) + a_2\cos(4\pi f_{rev} t) + a_3\cos(6\pi f_{rev} t) + \cdots \\ & + b_1\sin(2\pi f_{rev} t) + b_2\sin(4\pi f_{rev} t) + b_3\sin(6\pi f_{rev} t) + \cdots]\sin(2\pi f_{rf} t + \phi) \end{aligned} \tag{66}$$

where $a_0, a_1, a_2, \ldots, b_1, b_2, \ldots$ are the coefficients of the Fourier series. The amplitude of the coefficients will typically decrease with increasing index in a manner that is a function of the shape of the batch. Consider the classic situation where only a fraction α of the ring is evenly filled. By choosing the time origin so that $a(t)$ is even-symmetric, we get

$$a_n = \frac{\sin\pi\alpha n}{\pi\alpha n} \tag{67}$$

$$b_n = 0 \ . \tag{68}$$

Figure 23 shows the spectrum of the beam current $i_b(t)$ in that case. It consists of the carrier frequency f_{rf} plus a set of sidebands at $f_{rf} \pm n.f_{rev}$. The voltage V_b induced by the beam in the cavity will thus consist of a discrete set of lines at the frequencies

$$f = f_{rf} \pm n f_{rev} \ . \tag{69}$$

We now return to the uniform ring distribution and consider a beam undergoing longitudinal dipole oscillations, as shown in Fig. 3. Consider a single infinitely narrow bunch undergoing a pure dipole oscillation, the measured current is

$$i_{bunch,dipole}(t) = I_0 T_{rev} \sum_{l=-\infty}^{+\infty} \delta(t - lT_{rev} - \hat{\tau}\sin(2\pi f_s t + \psi)) \ , \tag{70}$$

where $\hat{\tau}$ is the amplitude of the synchrotron oscillation (in time) and f_s is the synchrotron frequency. Using the identity from Eq. (61) we rewrite the bunch current

$$i_{bunch,dipole}(t) = I_0 \sum_{n=-\infty}^{+\infty} e^{j2\pi n f_{rev}[t - \hat{\tau}\sin(2\pi f_s t + \psi)]} \ . \tag{71}$$

Now using the identity

$$e^{-jx\sin\phi} = \sum_{m=-\infty}^{+\infty} (-1)^m J_m(x) e^{jm\phi} \tag{72}$$

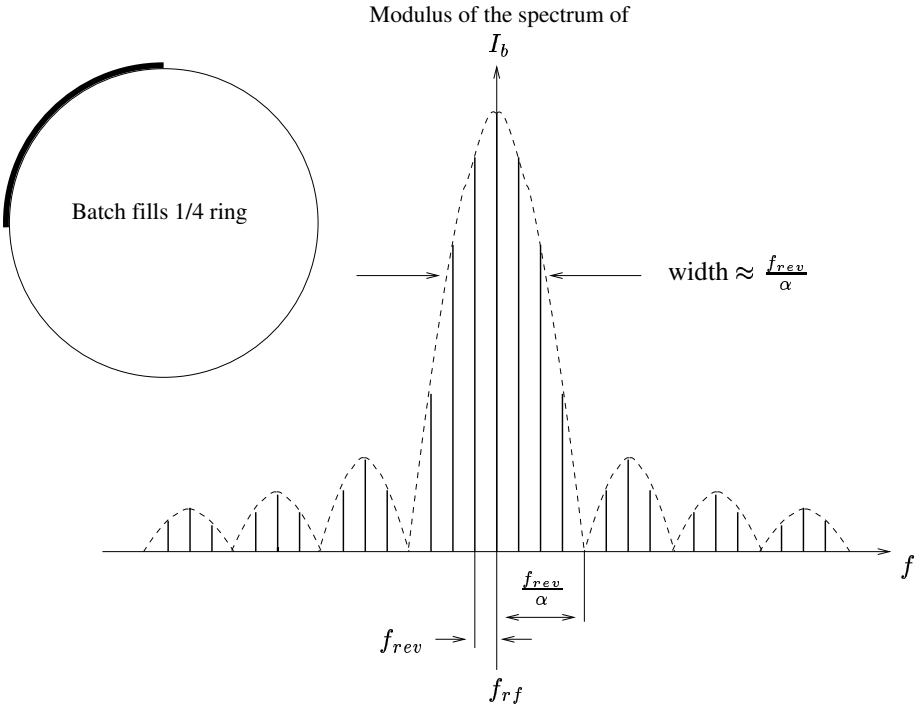


Fig. 23: Frequencies where beam loading must be compensated: spectrum of the beam current I_b in the case where only a fraction α of the ring is filled with bunches ($\alpha = 1/4$)

we get

$$i_{bunch,dipole}(t) = I_0 \sum_{n=-\infty}^{+\infty} e^{j2\pi n f_{rev} t} \sum_{m=-\infty}^{+\infty} (-1)^m J_m(2\pi n f_{rev}\hat{\tau}) e^{j[2\pi m f_s t + m\psi]} \tag{73}$$

$$i_{bunch,dipole}(t) = I_0 \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} (-1)^m J_m(2\pi n f_{rev}\hat{\tau}) e^{j[2\pi (n f_{rev} + m f_s) t + m\psi]} \tag{74}$$

Around each harmonic of the revolution frequency (at nf_{rev}) there is an infinite number of synchrotron sidebands (at $nf_{rev} + mf_s$). The spectral amplitude of the m th sideband of the n th revolution line is given by $J_m(2\pi n f_{rev}\hat{\tau})$ (Bessel function of order m). The spectral lines falling in the RF cavity impedance (near the fundamental) have index $n \approx h$ (where h is the harmonic number). The amplitudes of their sidebands are proportional to

$$J_m(2\pi n f_{rev}\hat{\tau}) \approx J_m(2\pi h f_{rev}\hat{\tau}) = J_m(2\pi \frac{\hat{\tau}}{T_{rf}}) \ . \tag{75}$$

When the instability starts growing, the amplitude of the longitudinal oscillation is much smaller than the RF period. The dominant sidebands are thus the first ones (at $nf_{rev} \pm f_s$) since, for a small value of their argument δ, the Bessel function of higher order are close to zero ($J_m(\delta) \approx (\delta/2)^m / m!$). Taking, for example, an oscillation of $\pm 10^\circ$ in the RF bucket ($\tau/T_{rf} = 10/360$), the strength of the sidebands relative to the revolution frequency line will be 0.09 for $m = 1$, 0.004 for $m = 2$, 0.0001 for $m = 3$, etc. Figure 3 shows an oscillation of larger amplitude: ± 1.5 ns in the 100 MHz bucket ($\tau/T_{rf} = 1.5/10$), and the sidebands of large index are not negligible. Relative to the revolution frequency line we get an amplitude of 0.53 for $m = 1$, 0.13 for $m = 2$, 0.02 for $m = 3$, etc. We show only one bunch in Fig. 3. The other bunches execute similar dipole oscillations at the same frequency f_s but each bunch k has its own phase ψ_k. For N evenly spaced bunches of equal intensity, each bunch executing a dipolar

longitudinal oscillation of the same amplitude but with different phases, the beam current is

$$i_{beam,dipole}(t) = I_0 T_{rev}/N \sum_{k=0}^{N-1} \sum_{l=-\infty}^{+\infty} \delta(t - lT_{rev} - \frac{k}{N}T_{rev} - \hat{\tau}\sin(2\pi f_s t + \psi_k)) \ . \quad (76)$$

Following the derivation used for a single bunch, we can rewrite the beam current as

$$i_{beam,dipole}(t) = I_0 \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} (-1)^m J_m(2\pi n f_{rev}\hat{\tau}) e^{j[2\pi(nf_{rev}+mf_s)t]} \sum_{k=0}^{N-1} e^{j[m\psi_k - 2\pi kn/N]} \ . \quad (77)$$

The spectrum is similar to the one created by a single bunch. The only difference is the last factor, which combines the phases ψ_k of the dipolar oscillations of the individual bunches to enhance or reduce the spectral lines. We conclude that the spectrum of the voltage V_b induced in the cavity by a beam undergoing a small longitudinal dipole oscillation contains power only at the frequencies

$$f = f_{rf} \pm nf_{rev} \pm mf_s \quad (78)$$

with the amplitude of the sidebands decreasing quickly with increasing index m.

The quadrupole oscillation shown in Fig. 4 does not modulate the phase of the beam current but only its amplitude. Since the frequency of the quadrupole mode is twice the synchrotron frequency we can expect that this mode of oscillation will create dominant sidebands at the frequencies $f = f_{rf} \pm nf_{rev} \pm 2f_s$. Generalizing to the higher order modes, we conclude that the voltage V_b induced in the cavity (around its fundamental resonance) by a beam undergoing a longitudinal oscillation has power only at the discrete frequencies

$$f = f_{rf} \pm nf_{rev} \pm mf_s \ . \quad (79)$$

4.2 Digital filters

4.2.1 Nyquist rate

Nyquist theorem states that a continuous-time signal can be sampled without loss of information if the sampling rate is at least twice the highest frequency present in the signal spectrum. In a mixed signal set-up, one first filters the analog signal with an analog anti-aliasing filter (low pass filter). This filter has a stop band starting at a frequency lower or equal to half the sampling frequency. In order not to distort the signal band, some frequency margin must be left for the transition band (between passband and stop band), so that the maximum interesting frequency in the continuous-time signal does not exceed one-third of the sampling frequency.

4.2.2 Finite Impulse Response and Infinite Impulse Response filters

A digital filter is an operator that maps an input sequence x_n to an output sequence y_n. If the filter is linear and invariant in time, the output y_n is related to the input x_n via the convolution

$$y_n = h_n * x_n \ , \quad (80)$$

h_n is the filter impulse response, defined as the output when the filter is excited by a Dirac pulse δ_n at its input ($\delta_n = 0$ for $n \neq 0$, and $\delta_0 = 1$). If the impulse response is of finite duration the filter is called a Finite Impulse Response (FIR) filter. Otherwise it is called an Infinite Impulse Response (IIR) filter. An FIR filter, of impulse duration L samples, realizes the following filtering operation

$$y_n = h_0 x_n + h_1 x_{n-1} + h_2 x_{n-2} + \cdots + h_{L-1} x_{n-L+1} \quad (81)$$

It is easily implemented with a tap delay line of length L, fed at its input by the signal x_n, and with an adder that computes the weighted sum of the tap outputs, the weights being the values of the impulse

response. IIR filters have feedback loops relating the value of the output at time n with the past values of the output. For example, the filter implementing the recursion

$$y_n = a\, y_{n-1} + x_n \tag{82}$$

has the impulse response $h_n = a^n$ for $n \geq 0$ and $h_n = 0$ for $n < 0$. The impulse response lasts forever.

4.2.3 z-transform

The Laplace transform is a very powerful tool for analysing continuous-time linear time-invariant systems. It provides answers to important questions of stability and frequency response. The equivalent for the analysis of discrete-time linear time-invariant systems is the z-transform, defined as

$$X(z) = \sum_{n=-\infty}^{+\infty} x_n z^{-n} \ . \tag{83}$$

The transfer function of the filter is the z-transform of its impulse response

$$H(z) = \sum_{n=-\infty}^{+\infty} h_n z^{-n} \ . \tag{84}$$

Because the z-transform of the convolution of two sequences is the product of the z-transforms of the two sequences, it follows from Eq. (80) that the z-transform $Y(z)$ of the filter output is the product of the z-transform of its input and the transfer function $H(z)$,

$$Y(z) = H(z)\, X(z) \ . \tag{85}$$

The variable z is the equivalent of the variable s for the Laplace transform. It is complex valued. The poles (zeros) of the transfer function are defined as the values of z for which the denominator (numerator) of $H(z)$ equals zero. Let us consider the double-peaked comb filter of Eq. (47). We have M zeros at

$$z_k = e^{j2\pi \frac{k}{M}} \tag{86}$$

for $k = 0, 1, 2, \cdots, M-1$. And M poles at

$$z_p = a^{\frac{1}{M}} e^{j2\pi \frac{p}{M}} \tag{87}$$

for $p = 0, 1, 2, \cdots, M-1$. Each pole has a multiplicity of two: it appears twice in the denominator. Recall that, for continuous-time filters, the frequency response is obtained by evaluating the Laplace transform on the imaginary axis. This gives the Fourier transform. Similarly, the frequency response of the digital filter is obtained from the evaluation of its transfer function on the *unit circle*. Let $\Omega = \frac{f}{f_c}$ be the normalized frequency of the input sine wave. The filter output will be a sine wave at the same frequency, but its amplitude and phase are given by the modulus and phase of $H(e^{j2\pi\Omega})$. Let us consider the double-peaked comb filter again. Because the zeros are located on the unit circle (Eq. (86)), the filter will show zeros of transmission at the exact multiples of the revolution frequency as shown in Fig. 18. The parameter a is very close to one, so that the pairs of poles (Eq. (87)) will be located very close to the corresponding zero but slightly inside the unit circle. This produces the desired pair of resonances on the synchrotron sidebands (Fig. 18).

4.2.4 From difference equations to transfer function

From a set of difference equations, we can easily get the filter transfer function by applying the z-transform to both sides of the equations. Two properties of the z-transform must be recalled:

Linearity: the z-transform of a sum of two sequences x_n and y_n is the sum of the z-transform of each sequence, $X(z) + Y(z)$.

Delay property: the z-transform of a sequence delayed by one sample is z^{-1} times the z-transform of the original sequence.

Using these properties, we derive from Eqs. (48),(49) and (50)

$$V(z) = az^{-M} V(z) + X(z) \tag{88}$$

$$W(z) = az^{-M} W(z) + \left(1 - z^{-M}\right) V(z) \tag{89}$$

$$Y(z) = z^{-M} W(z) \ , \tag{90}$$

where $X(z)$, $Y(z)$, $V(z)$ and $W(z)$ are the z-transforms of the sequences x_n, y_n, v_n and w_n respectively. We can now eliminate $V(z)$ and $W(z)$ from the above three equations and we get

$$Y(z) = \frac{1 - z^{-M}}{(1 - az^{-M})(1 - az^{-M})} z^{-M} X(z) \ . \tag{91}$$

This gives the transfer function $H_{sbd}(z)$ of Eq. (47). Notice, however, that several different sets of difference equations will lead to the same transfer function and they therefore realize the same filter. For example H_{sbd} is also realized by the following difference equation

$$y_n = 2a\, y_{n-M} - a^2\, y_{n-2M} + x_{n-M} - x_{n-2M} \ . \tag{92}$$

However, some realizations will be much more sensitive to quantization effects and the cascade form of Fig. 20 is often preferred.

4.2.5 Multirate filters

Throwing away $D - 1$ data out of every D data as proposed in Section 3.6 will unavoidably worsen the signal-to-noise ratio. Also, simply holding the output constant for D turns will introduce distortion. It is, however, possible to reduce the necessary processing without loss of precision [39]. The input sequence x_n is first processed by a so-called decimation filter (impulse response h_n^d). It is a low pass filter with a passband extending to the maximal synchrotron frequency $f_{s,max}$ (so that it covers the band of the BPF) and a stopband starting at $f_{rev}/D - f_{s,max}$. This puts a first limit on the downsampling ratio $D < f_{rev}/2f_s = 1/2Q_s$. All input data is used as input to the decimation filter but its output is computed every D turns only (downsampling). This sequence is fed into the original BPF, called the kernel filter in multirate filtering, and now operated at the reduced rate f_{rev}/D (rejection of the DC component and of the noise outside the synchrotron frequency band, $\pi/2$ phase shift at the synchrotron frequency). The saving in processing time in the BPF only is proportional to D^2 because the number of coefficients required to implement a given passband characteristic (in Hz) scales linearly with $1/D$ if the transition bandwidth is kept constant (from passband to stopband in Hz). We here assume that all filters are implemented as FIR. The original sampling rate is recovered at the BPF output using an interpolation filter h_n^u (upsampling) that produces an output, at every turn, from its input sequence at the rate f_{rev}/D. The interpolation filter must reject the image spectra created by the upsampling process. This requirement is satisfied if we choose the same filter for interpolation and decimation ($h_n^u = h_n^d$). Notice that, out of every D samples at the input of h_n^u (rate f_{rev}), $D - 1$ samples are equal to zero. The unnecessary multiplications by zero are not performed, thereby saving on the overall processing time. With this multirate scheme there is no degradation in performance compared to the implementation of the BPF at the rate f_{rev}, but a significant saving in processing time is possible if $f_{rev} >> 2f_{s,max}$. While the processing time required by the BPF decreases as $1/D^2$, the time spent computing the outputs of the decimation and interpolation filters increases with D and the overall processing time is therefore minimal for some optimal value of D.

REFERENCES

[1] A. Gamp, Servo control of RF cavities under beam loading, these proceedings.

[2] D. Boussard, Beam loading, Proc. CERN Accelerator School, Oxford, 1985, CERN 97-03 (1997).

[3] D. Boussard, Control of cavities with high beam loading, *IEEE Trans. Nuc. Sci.* **NS-32** (1985) No. 5.

[4] M.M. Karliner, Beam–cavity interaction, beam loading, in *Frontiers of Accelerator Technology*, Proc. Joint US–CERN–Japan International School, Tsukuba, 1996 (World Scientific, Singapore, 1999).

[5] R. Garoby, Beam loading in RF cavities, *Beam Intensity Limitations*, Joint US–CERN School, Hilton Head Island, South Carolina, 1990 (Lecture Notes in Physics No. 400, Springer, Berlin).

[6] J. Le Duff, High frequencies non-ferrite cavities, these proceedings.

[7] D. Boussard, Travelling-Wave Structures, in *Frontiers of Accelerator Technology*, Proc. Joint US–CERN–Japan International School, Tsukuba, 1996 (World Scientific, Singapore, 1999).

[8] P. Baudrenghien, Low-level RF systems for synchrotrons, Part I: Low intensity case, CAS course on RF Engineering, Seeheim, May 8-16 2000, these proceedings. (Also published as SL-Note-2001-008 HRF, 2001).

[9] F. Pedersen, Beam loading effects in the CERN PS booster, *IEEE Trans. Nucl. Sci.* **NS-22** (1975) No. 3.

[10] F.J. Sacherer, A longitudinal stability criterion for bunched beams, *IEEE Trans. Nucl. Sci.* **NS-20** (1973) No. 3.

[11] D. Boussard, G. Dôme, T.P.R. Linnecar, Acceleration in the CERN SPS, present status and future developments, *IEEE Trans. Nucl. Sci.* **NS-26** (1979) No. 3.

[12] P. Baudrenghien, Reducing the impedance of the travelling wave cavities: feed-forward and one turn delay feed-back, Proc. Workshop on LEP-SPS Performance, Chamonix, 2000, CERN-SL-2000-007 DI.

[13] P. Baudrenghien, Control of strong beam loading: results with beam, Proc. Workshop on LEP-SPS Performance, Chamonix, 2001, CERN-SL-2001-003 DI.

[14] F.E. Terman, *Radio Engineering* (McGraw-Hill, London, 1951).

[15] D. Boussard, H.P. Kindermann, V. Rossi, RF feedback applied to a multicell superconducting cavity, Proc. EPAC, Rome, Italy, 1988.

[16] Gene F. Franklin, J. David Powell, Abbas Emami-Naeini, *Feedback Control of Dynamic Systems* (Addison-Wesley, Reading, MA, 1994).

[17] R.C. Sah, J.R. Chen, C.C. Kuo, and G.H. Luo, Status report on the Synchrotron Radiation Research Center, Proc. APAC, Tsukuba, 1998.

[18] G.H. Luo, L.H. Chang, C.C. Kuo, M.C. Lin, R. Sah, T.T. Yang, Ch. Wang, The superconducting RF cavity and 500 mA beam current: upgrade project at Taiwan Light Source, Proc. EPAC, Vienna, 2000.

[19] R. Garoby, D. Grier, E. Jensen, A. Mitra, R.L. Poirier, The PS 40 MHz bunching cavity, Proc. PAC, Vancouver, 1997, CERN/PS/97-03.

[20] D. Boussard, G. Lambert, Reduction of the apparent impedance of wide band accelerating cavities by RF feedback, *IEEE Trans. Nucl. Sci.* **NS-30** (1983) No. 4.

[21] F. Pedersen, RF cavity feedback, CERN PS 92-59 RF (1992).

[22] F. Blas, R. Garoby, Design and operational results of a 'one-turn-delay feedback' for beam loading compensation of the CERN PS ferrite cavities, Proc. PAC, San Francisco, CA, 1991.

[23] J.D. Fox, Bunch feedback systems and signal processing, Proc. Joint US–CERN–Japan–Russia School on Particle Accelerators, Montreux, 1998.

[24] E. Kikutani, M. Tobiyama, S. Kurokawa, Development of a high-speed digital signal-process board for the KEKB bunch feedback systems, Proc. EPAC, Sitges, 1996.

[25] R. Garoby, Low level RF and feedback, in *Frontiers of Accelerator Technology*, Proc. Joint US–CERN–Japan International School, Tsukuba, 1996 (World Scientific, Singapore, 1999).

[26] P.B. Kenington, *High Linearity RF Amplifier Design* (Artech House, 2000).

[27] A. Mosnier, F. Orsini, B. Phung, Analysis of the heavily beam-loaded SOLEIL RF system, Proc. EPAC, Stockholm, 1998.

[28] *Mixed-Signal and DSP Design Techniques* (Analog Devices, 2003).

[29] G. Dôme, The SPS accelerating system, travelling wave drift-tube structure for the CERN SPS, CERN-SPS/ARF/77-11 (1977).

[30] E. Kikutani, J. Flanagan, M. Tobiyama, Limitations of multibunch feedback systems and extrapolation, Proc. EPAC, Vienna, 2000.

[31] M. Tobiyama, E. Kikutani, T. Obina, Y. Minagawa, T. Kasuga, Initial test of a bunch feedback system with a two-tap FIR filter board, Proc. 7th Beam Instrumentation Workshop (BIW96), Argonne National Laboratory, IL, 1996 (KEK Preprint 96-22, 1996).

[32] Y. Minagawa, E. Kikutani, S. Kurokawa, M. Tobiyama, Study of a transverse bunch-by-bunch feedback system based on the two tap FIR filter, *Nucl. Instrum. Methods Phys. Res.* **416** (1998).

[33] FIR Compiler MegaCore Function, version 1.12, ALTERA, http://www.altera.com .

[34] V. Rossi, CERN SL/HRF, private communication.

[35] J.M. Brennan, A. Campbell, J. DeLong, T. Hayes, E. Onillon, J. Rose, K. Vetter, RF beam control system for the Brookhaven Relativistic Heavy Ion Collider RHIC, presented at EPAC, Stockholm, 1998.

[36] *ADSP-21000 Family, Applications Handbook Vol. 1* (Analog Devices, 1995).

[37] A. Oppenheim, R. Schafer, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[38] Texas Instruments, Technology Innovations Bulletin **8** March 2001, http://www.ti.com/sc/techinnovations8 .

[39] N.J. Fliege, *Multirate Digital Signal Processing* (Wiley, Chichester, 1994).

HIGH-POWER RF TRANSMISSION

R.K. Cooper *
Los Alamos National Laboratory, New Mexico, USA

R.G. Carter
Lancaster University, Lancaster, U.K

Abstract
We deal primarily with the practice, as opposed to the theory, of high-power RF transmission systems. The emphasis is on commercially available power transmission components, with an explanation of their functioning.

1 INTRODUCTION

This paper discusses the means of transmission of RF power from the source to the accelerating structure or other load. The discussion cannot be comprehensive due to limitations of space, but the intention is to give a discussion in sufficient depth that the reader will have a feeling for what is really involved in transmitting high powers of RF. At a minimum the reader should feel comfortable picking up a manufacturer's catalogue and understanding what it is that all those devices do.

2 CHOOSING THE TRANSMISSION LINE

The first choice that has to be made in putting together a transmission system is that of selecting what kind of transmission line to use. There are really only two choices: coaxial line and waveguide. For the waveguide one can choose between rectangular, elliptical, or circular waveguides. Only the rectangular waveguide is discussed here, since it is by far the most common choice and has the greatest availability of circuit devices. Figure 1 shows the nomenclature used to described coaxial lines and rectangular waveguides.

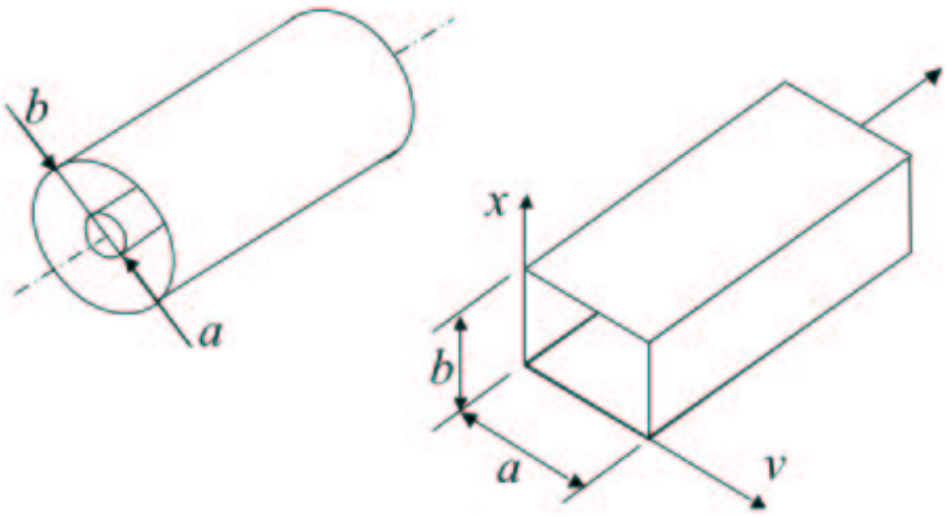


Fig. 1: Coaxial line and rectangular waveguide

* *Work supported by the US Department of Energy, Office of High Energy and Nuclear Physics.*

2.1 Coaxial line

The coaxial transmission line supports a TEM mode which has no cut-off frequency, that is, coax can be used down to d.c. This mode, in which the electric field is radial and the magnetic field azimuthal, has phase velocity and characteristic impedance given by

$$\nu_{\rm p} = c/\sqrt{\varepsilon_{\rm r}} \ , \tag{1}$$

$$Z_0 = 60 \ \ln(b/a) \ , \tag{2}$$

where c is the velocity of light in free space and $\varepsilon_{\rm r}$ is the relative permittivity of the material filling the space between the conductors. For high powers it is usual to use air-spaced lines to minimize the losses, but these require the use of dielectric spacers to support the inner conductor, as shown in Fig. 2.

Fig. 2: Typical high-power coaxial components

Coaxial transmission lines for high-power transmission are commonly available in 50 Ω and 75 Ω characteristic impedances, the former representing a compromise between breakdown field strength and power handling capacity, and the latter being selected for minimum attenuation. The ratio b/a is fixed by the characteristic impedance of the line at 2.3 for the 50 Ω line and 3.49 for the 75 Ω line. Propagation on a coaxial line is as $\exp j(\omega t - \gamma z)$ where $\gamma = \alpha + j\,\beta$. The loss parameter is given by

$$\alpha_{\rm coax} = \frac{1}{2\zeta_0 \log(b/a)} \left(\frac{R_{\rm a}}{a} + \frac{R_{\rm b}}{b} \right) \ , \tag{3}$$

where $R_{\rm a}$ and $R_{\rm b}$ are the surface resistances of the outer and inner conductors, respectively, and ζ_0 is the wave impedance of free space (= 377 Ω). The large coaxial transmission lines used for high-power transmission may use copper for the inner conductor and aluminium for the outer conductor to save

weight. The use of aluminium as the outer conductor increases the transmission losses by about 10%. The average power carried on a coaxial line is related to the peak electric field by

$$\bar{P}_{\text{coax}} = \frac{E_0^2 a^2}{\zeta_0} \pi \log\left(\frac{b}{a}\right) \quad . \tag{4}$$

The derivations of Eqs. (2), (3) and (4) are given in Appendix A. It is customary to assume that the peak electric field is the breakdown electric field strength in dry air at standard pressure and temperature, 3 MV/m. It must be remembered that this figure provides no margin of safety and that allowances must be made for changes with altitude, humidity and for the presence of dust particles in the air.

Higher-order modes (TE and TM modes) can propagate in coax at higher frequencies, and one wants to avoid these modes because mode conversion from TEM to TE or TM modes represents a source of power loss. The cut off wavenumber for the mode with the lowest cut off frequency, the TE_{11} mode, is approximately given by

$$k_c^2 (b+a)^2 \approx 4 \left[1 + \frac{1}{3}\left(\frac{b-a}{b+a}\right)^2\right] \quad . \tag{5}$$

As a numerical example, let us consider a 500 MHz power system using a rigid, air-filled, 75 Ω aluminum 14 inch outer diameter outer conductor transmission line. The radii of the outer and inner conductors are then a = 51 mm and b = 178 mm. Neglecting higher order waveguide mode effects and $\sigma_{\text{Cu}} = 5.8 \times 10^7$ mhos/m and $\sigma_{\text{Al}} = 3.5 \times 10^7$ mhos/m, so that the surface resistances are $R_a = (\sigma_{\text{Cu}}\delta_{\text{Cu}})^{-1}$ = 5.8×10^{-3} Ω per square and $R_b = (\sigma_{\text{Al}}\delta_{\text{Al}})^{-1} = 7.5 \times 10^{-3}$ Ω per square, the attenuation constant is

$$\alpha_{\text{coax}} = \frac{1}{2 \cdot 377 \cdot 1.25}\left(\frac{5.8 \times 10^{-3}}{0.051} + \frac{7.5 \times 10^{-3}}{0.178}\right) = 1.66 \times 10^{-4}\,\text{m}^{-1} \tag{6}$$

The power handling capacity is (with no safety factor or allowance for variation of breakdown field strength due to elevation, say)

$$\bar{P}_{\text{coax,max}} = \frac{(3 \times 10^6)^2 \pi (0.051)^2}{377} 1.25 = 244\ \text{MW} \tag{7}$$

The cut-off wavenumber is obtained from Eq. (5) as

$$k_{\text{c}}^2 \approx \frac{4\left[1 + \frac{1}{3}\left(\frac{17.8-5.10}{17.8+5.10}\right)^2\right]}{(17.8+5.10)^2} = 8.41 \times 10^{-3}\,\text{cm}^{-2} \quad . \tag{8}$$

Whence

$$k_{\text{c}}^2 = 0.092\,\text{cm}^{-1} \tag{9}$$

and

$$f_{\text{c}} = \frac{c}{2\pi} k_{\text{c}} = 439\ \text{MHz} \quad . \tag{10}$$

Thus we would not want to use this particular coaxial line much above 400 MHz, for example. Raising the TE_{11} cut-off frequency to 550 MHz for the same characteristic impedance would require a = 41 mm and b = 142 mm, yielding α = 2.1×10^{-4} m^{-1} and a maximum average power of 155 MW.

2.2 Rectangular waveguide

Standard rectangular waveguides have aspect ratios close to 2:1, but reduced-height waveguides are sometimes used for special purposes. The propagation constant in rectangular waveguide is given by

$$\beta^2 = \beta_0^2 - k_{\rm c}^2 \ , \tag{11}$$

where the cut-off wavenumber $k_{\rm c} = \pi/a$ and the cut-off frequency $f_{\rm c} = c/2a$ for the lowest mode of propagation. In this mode, the $\rm TE_{10}$ (or $\rm H_{10}$) mode, the electric field is normal to the broad wall and the magnetic field is parallel to the broad wall. Assuming that $b = a/2$ the attenuation constant is found to be

$$\alpha_{\rm wg} = \frac{2\,R_s\left[1 + (f_{\rm c}/f)^2\right]}{\alpha\zeta_0\left[1 - (f_{\rm c}/f)^2\right]^{0.5}} \tag{12}$$

and the maximum power is

$$\bar{P}_{\rm wg} = \frac{E_0^2 a^2}{\zeta_0}\, ab\,\sqrt{1 - (f_{\rm c}/f)^2} \ . \tag{13}$$

The derivations of Eqs. (12) and (13) are given in Appendix B. The power-handling capacity of a transmission line decreases with increasing altitude [1] because the breakdown field strength is a decreasing function of pressure. This is sometimes counteracted by pressurizing the waveguide. Thus, for example, the 1300 MHz RF power system of the Los Alamos (elevation 7000 feet = 2134 m) free-electron laser linac is pressurized to 10^{-12} psig with $\rm SF_6$. Such pressurization is the exception rather than the rule. For one thing, pressurization may cause the waveguide to deform, thus altering the propagation constant, which might then become a function of atmospheric pressure. (To counter this effect the waveguide manufacturers produce heavy-wall high-pressure waveguides.) It is, however, fairly common practice to pressurize the waveguide system with approximately 1 psig or so of dry nitrogen, to serve as a monitor for the tightness of the system, under the philosophy that if there are no air leaks there will be no RF leaks. This has the added advantage of ensuring that the accelerator structure is filled with dry nitrogen, rather than atmospheric air, if a high-power window should fail. That significantly reduces the time needed to recommission the accelerator following the failure.

Higher order waveguide modes are also a consideration in using rectangular waveguides. For the conventional choice of a $2:1$ aspect ratio in the transverse dimensions, the first higher order mode is the $\rm TE_{20}$ mode ($\rm H_{20}$ mode), for which the cut-off frequency is just twice the cut-off frequency of the dominant $\rm TE_{10}$ mode ($\rm H_{10}$ mode). Common practice is to use a rectangular waveguide with a $\pm 20\%$ bandwidth about a centre frequency which is 1.5 times the waveguide cut-off frequency. Roughly put, one operates in a band from approximately 1.25 f_c to 1.90 f_c. Table 1 shows the standard waveguides for high-power transmission in the frequency range commonly used for particle accelerators.

In the grey area around 200 to 400 MHz, where one might want to choose either coaxial transmission line for its more compact size, or a waveguide for its lower attenuation, one must bear in mind both attenuation and power-handling capacity.

For the high-power transmission system discussed previously but using a WR1800 waveguide made of aluminium, with a = 18 inches = 0.457 m, the attenuation constant is

$$\alpha_{\rm wg} = \frac{2 \cdot 7.5 \times 10^{-3}\left[1 + (327.9/500)^2\right]}{0.457 \cdot 377\ \left[1 - (327.9/500)^2\right]^{0.5}} = 1.65 \times 10^{-4}\,{\rm m}^{-1} \tag{14}$$

Table 1: Standard waveguide characteristics

Waveguide designation	**Inside dimensions (inches)**	**TE_{10} mode operating Range (MHz)**	**Cut-off frequency (MHz)**	**Cut-off wavelength (cm)**
WR2300	23.0×11.5	320–490	256	116.84
WR2100	21.0×10.5	350–530	281	106.68
WR1800	18.0×9.0	410–625	328	91.44
WR1500	15.0×7.5	490–750	393	76.20
WR1150	11.5×5.75	640–960	513	58.42
WR975	9.75×4.875	750–1120	605	49.53
WR770	7.7×3.85	960–1450	766	39.12
WR650	6.5×3.25	1200–1700	908	33.02

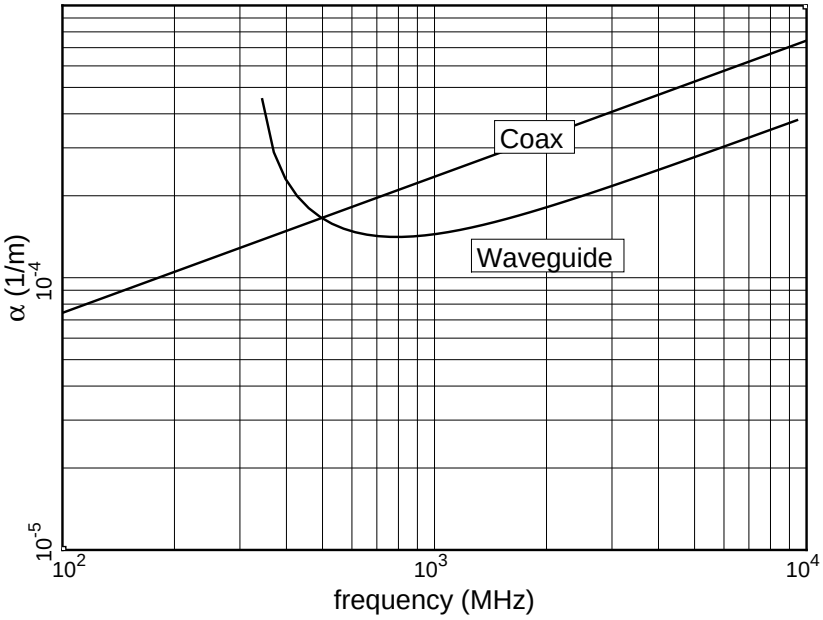


Fig. 3: Attenuation constants for a 14 inch coaxial line and for WR1800 waveguide

and the maximum power is

$$\bar{P}_{\text{wg,max}} = \frac{(3 \times 10^6)^2 \, 0.457 \cdot 0.229}{4.377} \sqrt{1 - (327.9/500)^2} = 471 \text{ MW} \ . \tag{15}$$

Thus at 500 MHz choosing a waveguide over a coax line is fairly clear cut. In order to avoid higher order waveguide mode losses, the diameter of the coax must be reduced, leading to higher attenuation and lower power-handling capacity as shown above. For short distances of transmission, however, the higher losses of coaxial transmission lines may be acceptable. Figure 3 shows attenuation versus frequency for a 14-inch coaxial line and a WR1800 waveguide. The figure is intended to show the frequency dependence of both coaxial line and waveguide and should not be interpreted as implying a choice between these two transmission lines. The attenuation constant α of a coaxial line increases with increasing frequency as the square root of the frequency, because of the surface resistance term, which contains a skin-depth factor. The attenuation constant for a waveguide has a more complicated frequency behaviour, as given

in Eq. (12) and shown in Fig. 3. As a function of frequency, the attenuation constant has a minimum at 2.471 f_c [2], but this frequency is not used for transmission, because several higher order modes can also propagate at the same time.

Figure 4 shows the attenuation constant α versus frequency for standard waveguides throughout their recommended frequency bands. The attenuation increases with frequency because of the decrease in the skin depth. Figure 5 shows the power-handling capacity of standard waveguides assuming a breakdown field strength of 3 MV/m.

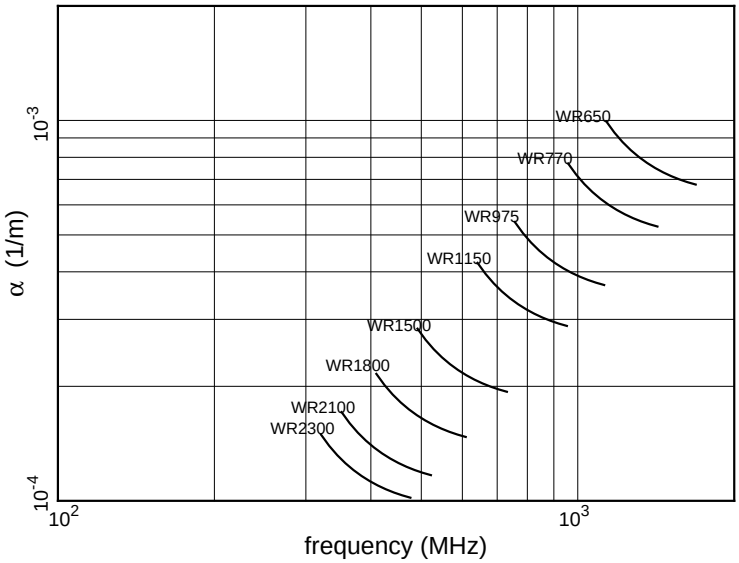


Fig. 4: Attenuation constant vs frequency for standard waveguides

Figure 6 shows a 425 MHz system set-up for evaluation of its performance. At the left is seen a klystron with its output to waveguide. Up from the klystron is a 90° mitred H-plane bend. Just before the Y-shaped circulator is a dual directional coupler. In the horizontal direction after the circulator is a 90° mitred E-plane bend. Following the bend is a broad-wall directional coupler. The operation and characteristics of these components, and more, are discussed in the remainder of this paper.

3 TWO-PORT DEVICES

3.1 Bends

Bends in waveguide systems typically leave the plane of the top (broad) waveguide wall undisturbed, or they leave the plane of the side (narrow) wall unchanged. Bends of the former type are known as H-plane bends, while the latter type are known as E-plane bends. In practice, both E-plane and H-plane bends commonly come in two forms, swept and mitred.

A swept bend forms a gradual change of direction. Figure 7 shows schematically a swept H-plane bend and a mitred H-plane bend. Swept bends tend to be rather long; a bend of two guide wavelengths in length is common and is good engineering practice. The reasoning behind this last statement is based on the observation that if two transmission lines of characteristic impedance Z_0 are connected by a transmission line of characteristic impedance Z_1, then there will be no reflection arising from the change of impedance if the length of the intermediate line is an integral number of half guide wavelengths long. That is, the impedance seen looking into the first junction between Z_0 and Z_1 is, assuming the second Z_0 line is properly terminated and the length of the intermediate line is L,

$$Z_{\text{in}} = \frac{Z_0 \cos \beta L + j Z_1 \sin \beta L}{j_0 (Z_0/Z_1) \sin \beta L + \cos \beta L} \quad , \tag{16}$$

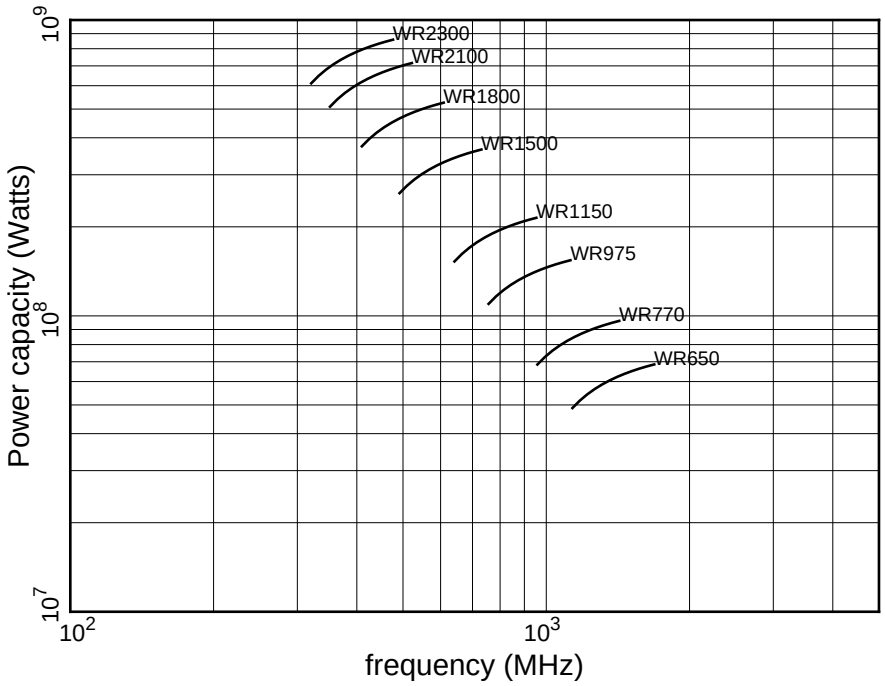


Fig. 5: Waveguide power handling capacity at a beakdown voltage of 3 MV/m

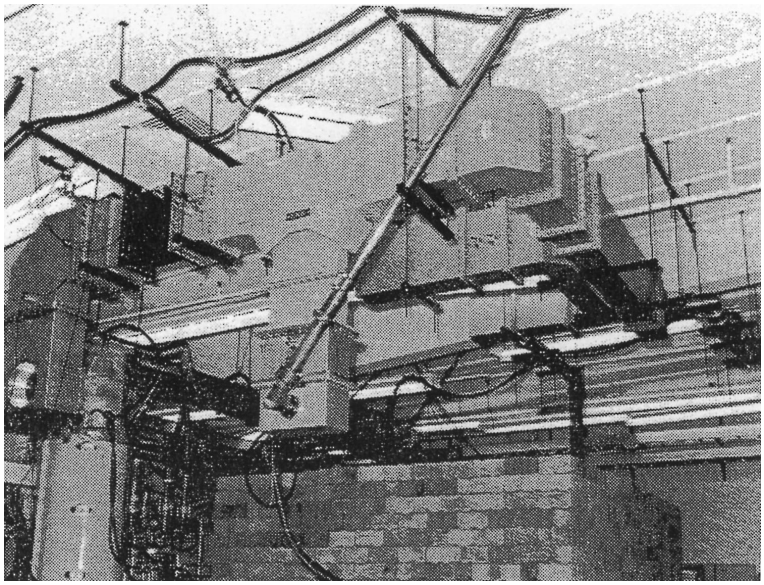


Fig. 6: A 425 MHz system set up for evaluation

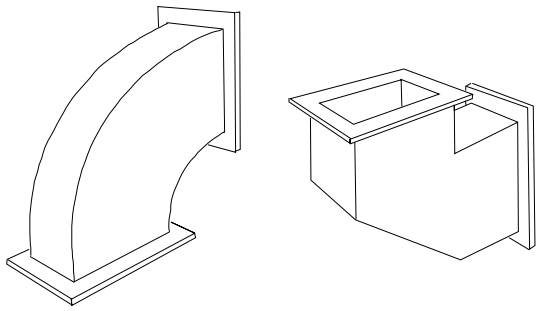


Fig. 7: Outline of a swept H-plane bend (left) and a mitred H-plane bend (right)

which will have the value Z_0 whenever βL is a multiple of π.

A mitred bend, on the other hand, can be quite compact, slightly longer than one electrical wavelength. A 90° H-plane bend is essentially a right-angled junction with a reflecting plane replacing the outer corner. The location of this plane is discussed in Ref. [1]. Other mitred bends have both corners replaced by 45° planes.

Commercially available swept H-plane bends have a maximum Voltage Standing Wave Ratio (VSWR, the ratio of the sum of the incident and reflected wave amplitudes to the difference of the incident and reflected amplitudes) of 1.03 over the 40% waveguide band, while the mitred bends have a maximum VSWR of 1.02, but only over a 10% band.

Essentially the same remarks can be made about the swept E-plane bends and the mitred E-plane bends as were made about the H-plane bends, including performance characteristics.

3.2 Twists and tapers

It is often the case that the plane of polarization of the electric field must be rotated by 90°. One can choose to make a gradual change (twist) of the waveguide, or one can make the change of plane in a series of discontinuous rotations. The problem with gradual changes in general is that they take up length in the transmission system. The step twist, on the other hand, accomplishes the rotation in a distance as short as three-quarters of a guide wavelength. Figure 8 shows a 90° step twist. The basic idea of the

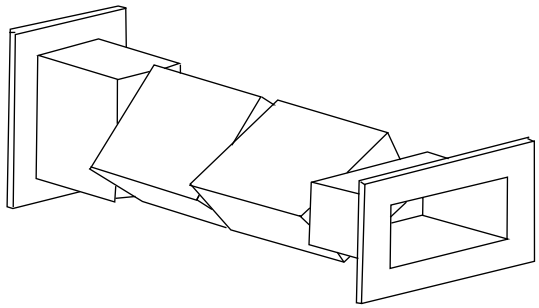


Fig. 8: Outline drawing of a 90° step twist

step twist is that the discontinuities in the orientation of the component waveguide sections should occur at one-quarter guide wavelength separation, so that reflections from the discontinuities are allowed to cancel. If we denote by θ the rotation angle between the sides of two sections of waveguide, then the reflection coefficient arising from the discontinuity will be proportional [3] to θ^2 for small enough θ. If we want to accomplish a 90° twist with, say, three rotations of angles θ_1, θ_2, and θ_1, then we would want

the reflection coefficients to be in the ratio 1:2:1 for pairwise cancellation of the reflected waves. Thus we want to have:

$$\begin{array}{ccccccc} 1 & : & 2 & : & 1 & & \\ \theta_1^2 & : & \theta_2^2 & : & \theta_1^2 & & \\ 26.4^\circ & + & 37.2^\circ & + & 26.4^\circ & = 90^\circ & . \end{array}$$

In practice these angles need to be adjusted slightly for optimum performance. One expects to obtain a VSWR resulting from reflection from a step twist of the order of 1.05 or less over a bandwidth of at least 10%. Step twists with five rotations are also available for broader-band performance and higher power-handling capacity. Other angles of twist such as 30°, 45°, and 60° are also available. The power-handling capacity of the step twist is about half that of the waveguide to which it is mated, and therefore may represent the weakest link in the transmission chain.

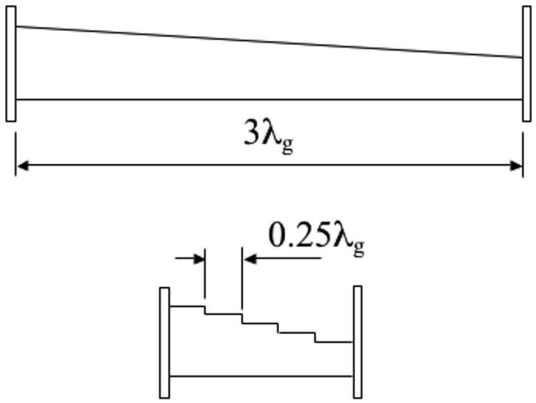


Fig. 9: Outline drawings of gradual and stepped height tapers

Where it is necessary to make a transition from standard waveguide to reduced-height waveguide we may use either a gradual or a step taper, as shown in Fig. 9. The gradual taper must have a length of at least three guide wavelengths to achieve an acceptable VSWR. The step taper, employing three or four quarter-wavelength steps, is much shorter and can be made with good VSWR if it is only to operate over a narrow band. Figure 10 shows a waveguide system with swept and mitred bends and a step transformer to reduced-height waveguide.

3.3 Windows

Since the high-power waveguide system is normally at atmospheric pressure, it is necessary to provide RF windows between the waveguide and the accelerating structure. There are, of course, similar windows on the outputs of high-power klystrons, which are an integral part of those tubes. The RF windows are often the most critical part of a high-power RF system, since their failure will cause the high vacuum system to go ‘down-to-air’, leading to the need for a lengthy period of reconditioning before the system is fully operational again.

Figure 11 shows the arrangement of a number of common types of RF window. The window is typically made from high purity alumina ceramic which is brazed into the transmission line. The simplest are those in coaxial line [Fig. 11(a) and circular waveguide Fig. 11(b)]. These have the advantage that the thermal stresses produced in the window by dielectric heating are symmetrical and that the field pattern in the ceramic matches that in the empty transmission line, avoiding mode conversion. To achieve a good match the window must either be thin, half a wavelength thick, or provided with additional matching

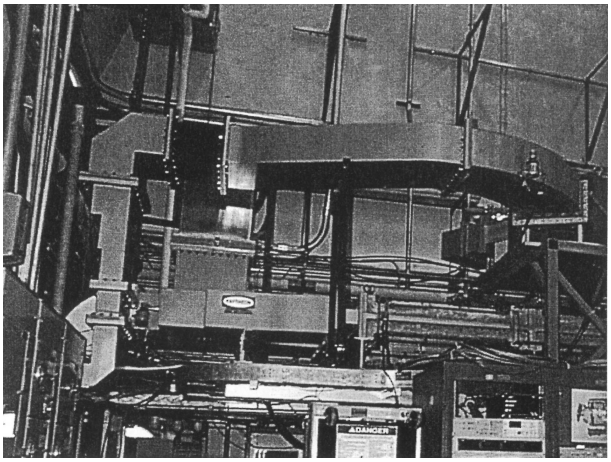


Fig. 10: Arrangement of a typical high-power waveguide system

elements. The lighthouse window [Fig. 11(c)] employs a cylindrical ceramic around the junction between a coaxial line and a waveguide. This arrangement is particularly suitable when the coupler that feeds RF power into the accelerator is in coaxial line. The creation of a window in a rectangular waveguide by fixing a rectangular piece of ceramic into the waveguide is not suitable for use at high power levels because of the stress concentrations that occur at the corners of the ceramic. A possible alternative is to use an iris window, as shown in Fig. 11(d). The iris is designed to be resonant at the centre frequency of the band of frequencies required. Additional matching elements must be used if broadband operation is required. Finally, the pill-box window [Fig. 11(e)], is essentially a circular waveguide window placed between a pair of transitions from circular to rectangular waveguide. Because the ceramic is larger than that of an iris window, the power-handling capacity is higher. With careful design a pill-box window can give a good VSWR over a band of frequencies.

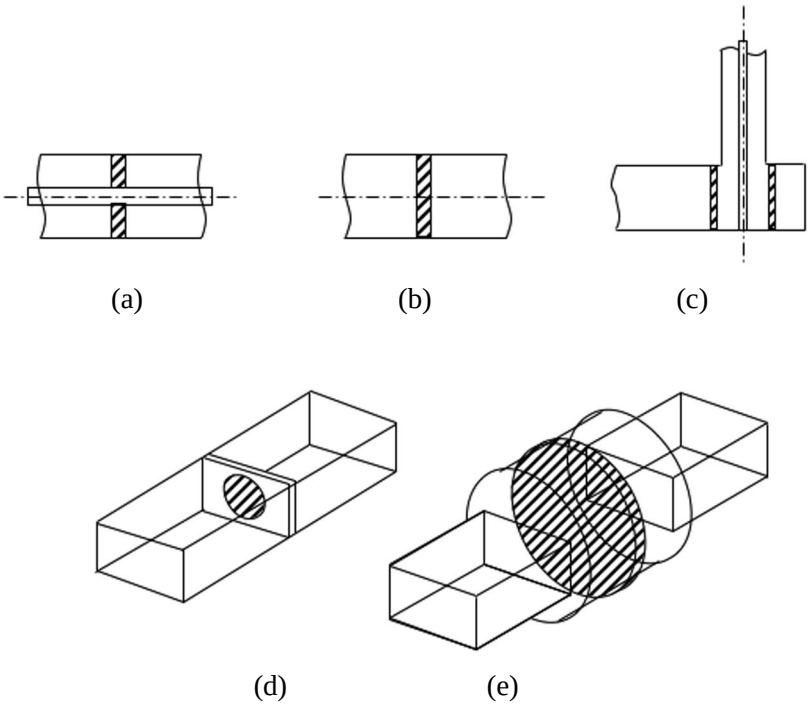


Fig. 11: Various types of RF windows

Because the consequences of window failure can be catastrophic, it is usual to take precautions against such an event. Common precautions are provision of air-blast cooling on the air side of the window, monitoring of the window temperature, and the use of double windows in series.

3.4 Mode transformers

It is sometimes necessary to couple power from one type of transmission line to another with a low VSWR. In order to achieve this, the basic principle is to try to ensure that the electric and magnetic field patterns on the two sides of the junction are roughly similar. Since the field patterns are not normally an exact match, some of the RF power is scattered into higher order modes. Provided these are cut off, no power can propagate in them and it therefore remains stored as reactive energy in the region close to the junction. The mismatch caused by this reactance can be tuned out by the addition of matching elements (posts or irises) to give a good overall VSWR. Rather different problems arise when the higher order modes can propagate because significant power can then be scattered into undesired modes. This should not normally happen at the fundamental frequency, but it can cause serious problems at harmonic frequencies. A high-power RF system always contains some harmonic power from the output of the RF source or from power induced in the accelerating structure by beam bunches. Harmonic power can cause standing waves in the waveguide and trigger waveguide breakdown.

Figure 12 shows, as an example, how power can be coupled from the TE_{10} rectangular waveguide mode to the TM_{10} circular waveguide mode (or *vice versa*). When the rectangular waveguide is terminated in a short circuit the resulting standing wave has magnetic field lines, which are approximately circular, as shown on the left of Fig. 12. These are a good match to the magnetic field in the circular waveguide. In order to match the electric field lines it is necessary to add a post to the junction, as shown on the right of Fig. 12. The dimensions of the post are adjusted to match the junction and additional matching elements can be added in either, or both, of the waveguides if necessary.

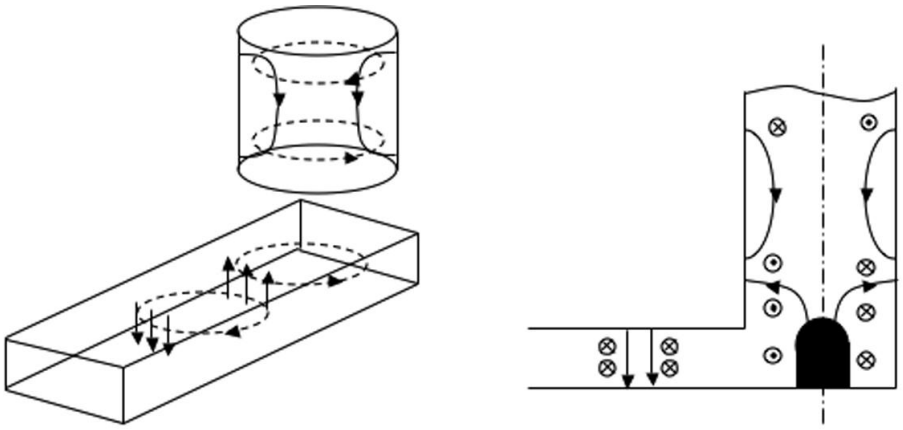


Fig. 12: Arrangement of a transition from rectangular to circular waveguide

The most common types of mode transformer are those for coupling power from a coaxial line to a rectangular waveguide. There are a number of ways of achieving this but the only ones that are really suitable for use at high power levels are the door-knob transition and the tee bar transition illustrated in Fig. 13. Note that in this case part of the problem arises from the mismatch of impedance between the waveguide (typically 200 Ω) and the coaxial line (50 Ω or 75 Ω).

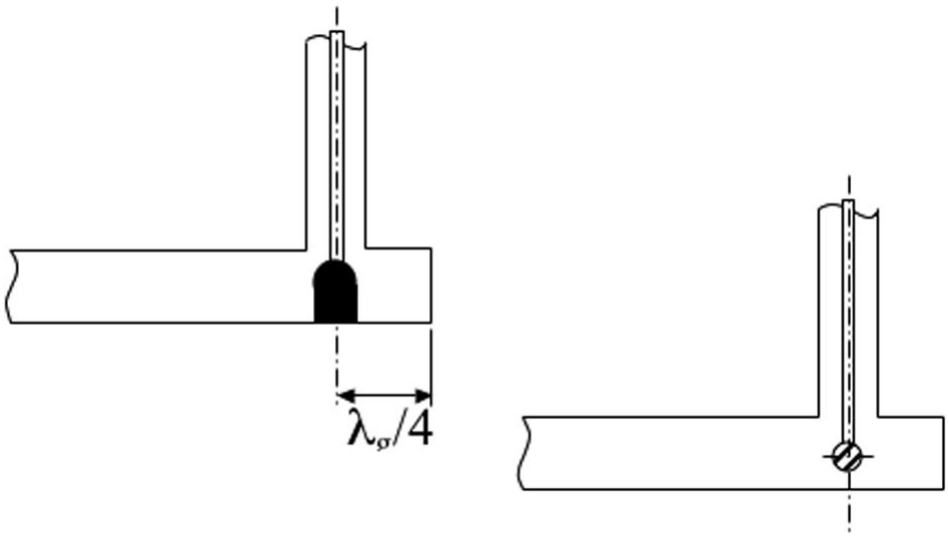


Fig. 13: Arrangements of high-power transitions from coaxial line to waveguide

4 THREE AND FOUR-PORT DEVICES

4.1 Directional couplers

In waveguide transmission systems it is often useful and necessary to couple energy from one waveguide system into another system. Such coupling might be to provide power to the second system, or to provide a measure of the power flow in the first system. Consider, first, two rectangular waveguides side by side with a common side wall, with a hole connecting the two waveguides, as shown in Fig. 14. The action of the hole can be described by considering it as a dipole radiator that is excited by the wave S_0 in the primary waveguide. The dipole radiates waves symmetrically into both waveguides as shown, with the condition that the sum of the powers emerging from the four ports is exactly equal to that in the incident wave (for a theoretical lossless device). Since the action of the hole is to couple some of the incident power to ports 3 and 4 and some power re-emerges from port 1, it follows that the wave S_{1+} must combine destructively with S_0 so that the power emerging at port 2 is reduced by the presence of the hole. By using two or more coupling holes it is possible to make devices, known as directional couplers, which couple power selectively to particular ports.

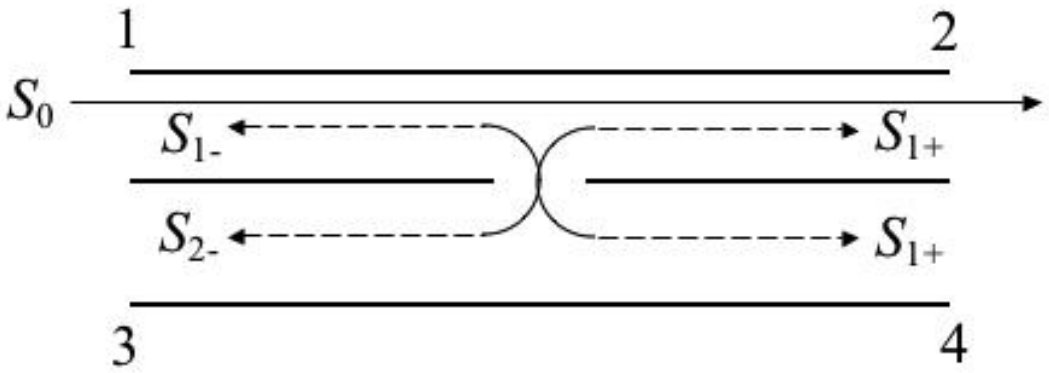


Fig. 14: Waveguides coupled by a hole in their common wall

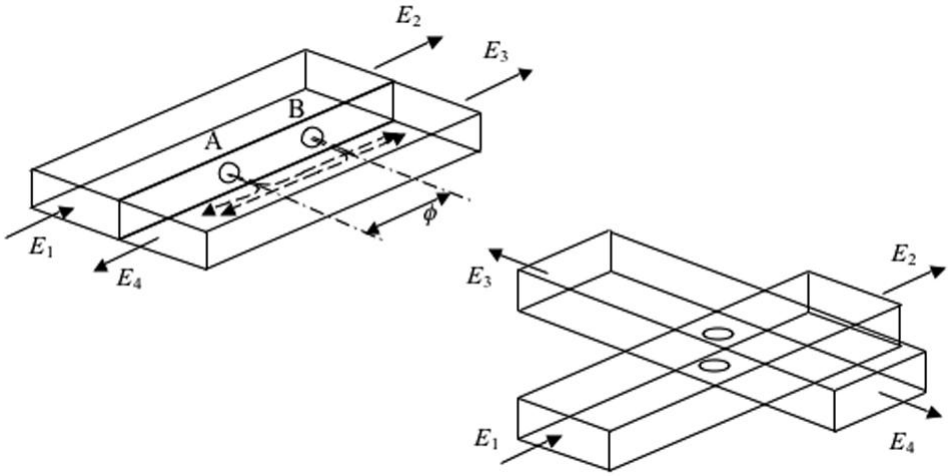


Fig. 15: Directional couplers

The principle of two common types of directional coupler is illustrated in Fig. 15. The left-hand diagram shows two waveguides coupled by a pair of identical holes in their common, narrow wall. If we treat port 1 as the input port then hole B will be excited with a phase shift ϕ relative to hole A, where ϕ is the electrical length of the guide between the holes. Thus the forward waves radiated by the two holes into the second guide are always in phase with each other and add together, and the amplitude of E_3 is non-zero. If we choose $\phi = \pi/2$ then the backward wave in the second guide radiated by hole B is in antiphase with that radiated by hole A at the plane of A and $E_4 = 0$. Thus the power input at port 1 has been divided between ports 2 and 3 in a proportion determined by the sizes of the holes. Since the device is symmetrical, the same argument can be applied to power input at any of the ports. It is important to note that, as described, this coupler works perfectly at only one frequency. The coupling can be increased and made broadband by the use of arrays of holes. For details on the theory of design of multihole couplers see Ref. [4]. The S matrix of a directional coupler has already been given in these proceedings by Caspers [5]. The coupler described is known as a 'side-wall coupler'. It is also possible to make couplers with holes in the common broad wall. Couplers for use at low power levels often have three ports, since one port is terminated in an internal matched load. This is not normally suitable for high-power use, where a four-port coupler with an external high-power load is required. The properties of a directional coupler excited at port 1 are described by

$$\text{Coupling} = 20 \log_{10} \left| \frac{E_1}{E_3} \right| \tag{17}$$

and

$$\text{Directivity} = 20 \log_{10} \left| \frac{E_3}{E_4} \right| \ . \tag{18}$$

The coupling measures the transfer of power to the second guide. For purposes of monitoring and control this is normally quite small and the coupled power is typically 30 dB or 40 dB below that in the main guide. A special case is the power divider, with 3 dB coupling, which splits the power equally between ports 2 and 3. This is examined further in Sections 4.2 and 4.3. The directivity is a measure of how close the coupler is to perfection. It is usually at least 30 dB and exceeds 40 dB for couplers designed

for measurement systems. Typical characteristics of a commercially available coupler over a bandwidth of 10% are VSWR, 1.05; directivity, 30 dB; coupling variation ±0.5 dB; coupling factor range, 6 dB to 50 dB (a fixed value of, for example, 20 dB for any individual coupler).

The diagram on the right of Fig. 15 shows the arrangement of a cross-guide coupler. These couplers do not have good broadband characteristics because they normally have only two coupling holes. The principle of operation is similar to that already described. Cross-guide couplers have the advantage of being very compact. They are particularly useful in situations where accuracy is not critical, such as in the monitoring of reflected power to detect waveguide arcs. See Ref. [6] for details of various directional coupler schemes.

In the larger waveguide sizes another form of directional coupler is available. This type of coupler, described by Klein [6], consists of a circuit loop penetrating into the waveguide top wall. The loop will have currents induced in it both by the electric field of the waveguide mode and by the magnetic field of the mode. By proper orientation and termination of the loop, the electric and magnetic currents can be made to be equal in magnitude. Then a wave in the forward direction, say, will produce a current in the loop, which is the constructive sum of the electric and magnetic currents, while a backward-travelling wave produces no current in the loop because the electric and magnetic currents cancel each other. See Ref. [7] for a circuit analysis of the operation of these loop-type couplers. These couplers are commercially available packaged singly (for power monitoring) and doubly (for reflectometer use). Typical characteristics of such couplers over a 10% bandwidth are coupling factor range, 40–75 dB (a fixed value for any particular coupler); VSWR, less than 1.05; directivity, greater than 25 dB; insertion loss, less than 0.1 dB.

4.2 The Riblet coupler

We discuss next the short-slot hybrid coupler. To quote from Harvey [7]: “The hybrid junction is a four terminal-pair device which ideally has the property that power supplied to a given terminal is divided, usually equally, between two of the three remaining terminal-pairs and nothing is coupled to the fourth terminal-pair.” The Riblet short-slot hybrid coupler [8] is one of these devices and forms the basis of a number of useful waveguide devices, such as the phase shifter and the variable attenuator. Note also that the magic tee, described below, is also a hybrid junction according to this definition. The Riblet coupler is depicted in Fig. 16. Two waveguides with a common side wall have a section of that wall removed.

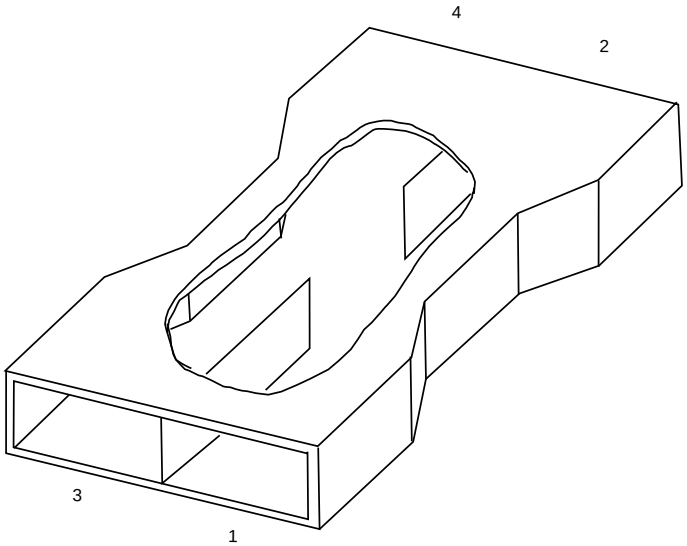


Fig. 16: Outline drawing of a Riblet coupler

Table 2

Location	**Port 1**	**Port 2**	**Port 3**	**Port 4**
Entering	E_0	0	0	0
Leaving	0	$E_0/\sqrt{2}$	0	$j\,E_0/\sqrt{2}$
Entering	0	$\rho e^{j\phi}\,E_0/\sqrt{2}$	0	$j\,\rho e^{j\phi}\,E_0/\sqrt{2}$
Leaving	0	0	$j\,\rho e^{j\phi}\,E_0$	0

In that portion of the coupler with the side wall removed there is a double-width waveguide in which both the TE_{20} mode and the TE_{10} can propagate. (The width is sometimes narrowed, as shown, to keep the H_{30} mode from propagating.) By choosing the length of the removed wall section properly, an equal division of the power incident in the first guide can result. The forward-going wave in the coupled guide has in addition a phase shift of 90°, whereas there is no backward-going wave in the coupled guide. The scattering matrix for the Riblet coupler is given by

$$S = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 & j \\ 1 & 0 & j & 0 \\ 0 & j & 0 & 1 \\ j & 0 & 1 & 0 \end{pmatrix} \quad . \tag{19}$$

As an illustration of a simple application of the Riblet coupler and to introduce a form of analysis to be used later, consider using the coupler to split power equally from a source to two identical loads. The table below shows how the power is first split from port 1 equally into ports 2 and 4, how the reflected powers from the identical loads (reflection coefficient ρ, overall phase shift from coupler to load and back ϕ) re-enter the coupler at ports 2 and 4, and how the reflected powers recombine in the coupler to exit at port 3. Thus no reflected power comes out from port 1, so that the source is isolated from the load reflections.

For purposes of typography and explanation, let us consider instead of the customary S matrix definition,

$$E_{\text{scattered}} = S\,E_{\text{incident}} \quad , \tag{20}$$

where the E's are column vectors with entries representing the amplitudes of the fields at the respective ports, the transposed equation

$$\tilde{E}_{\text{scattered}} = \tilde{E}_{\text{incident}}\,\tilde{S} \quad , \tag{21}$$

where the tilde represents the transpose of the indicated quantity. Thus we have an equation relating row vectors instead of the more commonly used column vectors. If we have a wave incident only on port 1, the incident vector is given by $(E_0\ 0\ 0\ 0)$, and the scattered vector is given by

$$\begin{aligned} S &= (E_0, 0, 0, 0)\,\frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 & j \\ 1 & 0 & j & 0 \\ 0 & j & 0 & 1 \\ j & 0 & 1 & 0 \end{pmatrix} \\ &= (0, E_0/\sqrt{2}, 0, j, E_0/\sqrt{2}) \quad . \end{aligned} \tag{22}$$

We see that each row of Table 2 above labelled 'leaving' can be obtained from the row above it labelled 'entering', simply by multiplying the entering row into the transpose of the scattering matrix.

4.3 The magic tee

H-plane and E-plane tees have already been discussed by Caspers [5] who has also given the S matrix characterization of these junctions. In high-power systems these junctions are used mainly to provide opportunities for matching loads to sources, and will not be discussed further here.

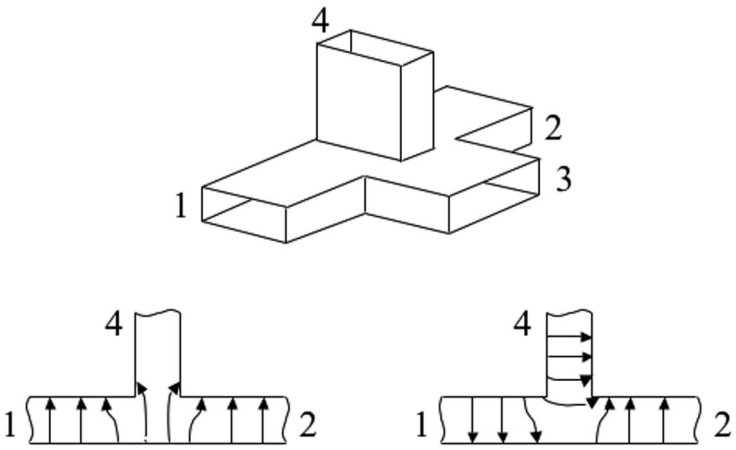


Fig. 17: Arrangement and principle of operation of a magic tee

Figure 17(top) shows the arrangement of a 'magic tee' which comprises an H arm, an E arm, and two co-linear arms. The properties of a magic tee can be illustrated simply by considering the effect of exciting ports 1 and 2 in phase and in anti-phase. When they are excited in phase (Fig. 17, left) the resulting signal is coupled to the TE10 mode of the H arm (port 3) but not to that of the E arm (port 4). When the excitation is in antiphase (Fig. 17, right) the coupling is to port 4 but not to port 3. The magic tee is therefore another example of a hybrid junction as defined in Section 4.2. Since it is a passive device, its properties must be symmetrical so that power entering at port 3 excites waves at ports 1 and 2 in phase with each other and power entering at port 4 excites outputs at ports 1 and 2 in antiphase. The scattering matrix of a magic tee is

$$S = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 1 \\ 1 & -1 & 0 & 0 \end{pmatrix} \quad . \tag{23}$$

As an application of power splitting and combining, consider the output waveguide system of the SLAC 5045 klystron. The output power of 67 MW (peak power) exceeds the capacity of a single ceramic window, so a magic tee is used to split the power equally between the co-linear arms, the power is passed equally between two windows, and then recombined in a second magic tee.

In a load-isolation application, one may input power in the H arm, split it equally between the colinear arms, and thereby send equal power to equal loads. If one of the loads is located an electrical distance 90° further from the magic tee than the other load, then any reflection common to the two loads will combine to go up the E arm.

Commercially available magic tees for high-power application have the following typical characteristics over a 10% bandwidth: balance between the colinear arms, ± 0.1 dB; insertion loss, less than 0.1 dB; isolation of the E and H arms, 30 dB minimum; VSWR, less than 1.10. Folded magic tees are also commercially available (Fig. 18); these tees have slightly poorer characteristics than the standard magic tee, but have the advantage of a more compact structure. These tees are commonly used, for instance, in circulators (see Section 6).

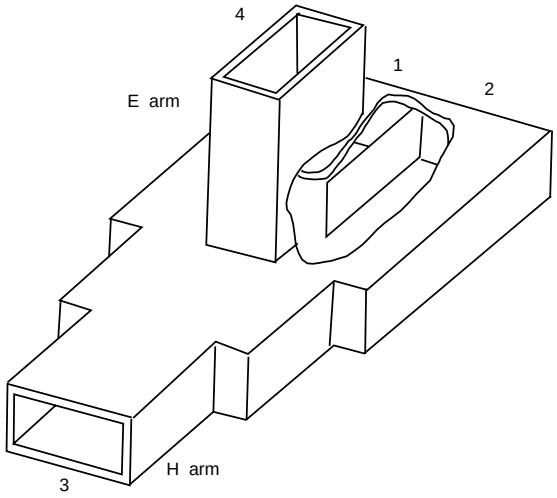


Fig. 18: A folded magic tee

5 PHASE SHIFTERS AND VARIABLE ATTENUATORS

At low power levels it is possible to make phase shifters and attenuators by placing dielectric or lossy vanes within the waveguide. At high power levels this is not possible because the vanes are incapable of handling the power required. An alternative approach that is satisfactory for use in high-power systems is to make use of the properties of hybrid junctions.

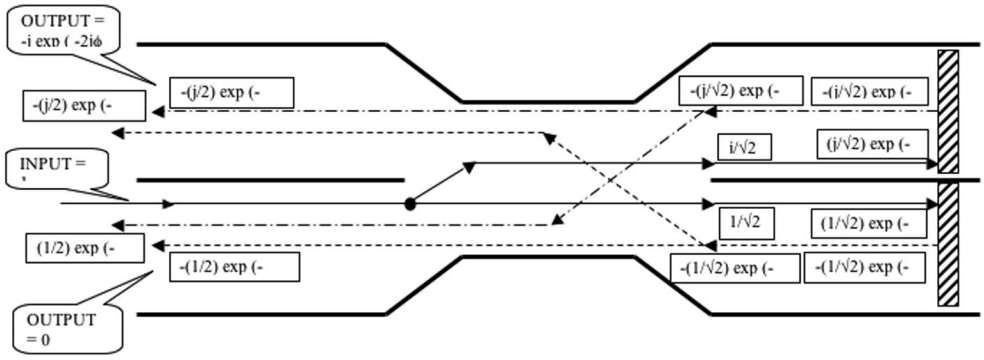


Fig. 19: A Riblet coupler with short circuit terminations

If the 'loads' of the Riblet coupler load isolation system are in fact shorts, then ρ = –1 and the result is that the action of the coupler with two identical shorts is represented by the diagram in Fig. 19. The information in the diagram can be summarized by writing

$$\begin{aligned} E_{\text{incident,port 1}} &= E_0 \\ E_{\text{outgoing,port 3}} &= -j\,E_0\,E^{j\phi} \quad . \end{aligned} \tag{24}$$

Thus the phase of the outgoing wave depends upon the distance between the short circuits and the coupler and can, therefore, be adjusted by moving the short circuits. A word must be said about the phase of the outgoing wave. The scattering matrix given for the Riblet coupler makes an assumption about the reference planes at which the S parameters are measured. These reference planes may not (and most

Table 3

Location	**Port 1**	**Port 2**	**Port 3**	**Port 4**
Entering	E_0	0	0	0
Leaving	0	$E_0/\sqrt{2}$	0	$j\,E_0/\sqrt{2}$
Entering	0	$-j\,e^{j\phi}\,e^{j\Delta\phi}\,E_0/\sqrt{2}$	0	$-j\,e^{j\phi}\,e^{j\Delta\phi}\,E_0/\sqrt{2}$
Leaving	$-j\,e^{j\phi}\,E_0\,\sin\Delta\phi$	0	$-j\,e^{j\phi}\,E_0\,\cos\Delta\phi$	0

likely do not) correspond to the actual bounding planes of the physical coupler. Thus there are some phase shifts in waves moving through the coupler that have not been properly taken into account, so that the ϕ was to account for the electrical phase to the 'load' and back must be modified to take into account a fixed phase shift due to electrical lengths in the coupler. In the following analyses what matters is in fact the relative change in phases, so that an overall constant phase need not concern us, and we continue to disregard common constant phases.

What is done in practice to make a practical phase shifter is to make up movable shorts in parallel waveguides attached to the short-slot hybrid coupler, to locate the shorts the same distance from the coupler, and to gang them mechanically so that they move together. The ganged combination is then moved in and out to accomplish a phase shift linear with distance moved. In high-power systems this movement is commonly accomplished by a motorized drive mechanism. Commercially available phase shifters of the type described here have typical performance characteristics over the waveguide band of VSWR, less than 1.05; insertion loss, less than 0.1 dB; total phase shift available, more than 190°.

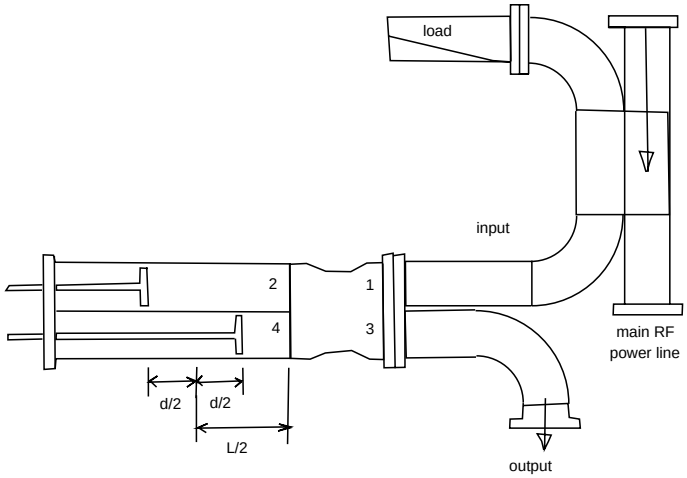


Fig. 20: Attenuator and phase shifter with reflection to the source

Two novel arrangements of Riblet couplers to produce a phase shifter/attenuator combination have been described by Ohsawa *et al.* [9] of KEK. If one can tolerate reflection back to the source, the first arrangement is very simple, as shown in Fig. 20. The arrangement consists of the two sliding shorts as in the phase shifter, but in this arrangement the location of one short is further from the phase shifter position by a distance $d/2$, say, while the other short is closer than the phase shifter position by the same amount. Thus the phase of the wave reflected from the more distant short and arriving back at the coupler is given by $\phi + \Delta\phi$, where $\Delta\phi = d/\lambda_g$, while the phase of the reflection from the nearer short is given by $\phi - \Delta\phi$. An analysis similar to that given for the power distribution system above shows the result of the differential shift $\Delta\phi$ (Table 3).

Notice, as a check, that if $\Delta\phi = 0$ we obtain the result for the straight phase shifter. Thus

Table 4: Action of the attenuator and phase shifter with load isolation

Top coupler	**Port 1**	**Port 2**	**Port 3**	**Port 4**
Entering	E_0	0	0	0
Leaving	0	$E_0/\sqrt{2}$	0	$j\,E_0/\sqrt{2}$
Bottom coupler	**Port 1**	**Port 2**	**Port 3**	**Port 4**
Entering	$j\,e^{j\phi}\,e^{-j\Delta\phi}\,E_0/\sqrt{2}$	0	$j\,e^{j\phi}\,e^{j\Delta\phi}\,E_0/\sqrt{2}$	0
Leaving	0	$j\,e^{j\phi}\,E_0\,\sin\Delta\phi$	0	$j\,e^{j\phi}\,E_0\,\cos\Delta\phi$

the differential distance d between the two sliding shorts controls the amplitude of the output (port 3), multiplying the input amplitude by $\cos\Delta\phi$, while at the same time generating a reflection directed back to the input, with amplitude $\sin\Delta\phi$ times the input amplitude. This configuration thus gives independent control of amplitude and phase; the mean position of the two shorts controls the overall phase shift and the difference in positions controls the amplitude. In the KEK application this attenuator/phase shifter was used for a prebuncher, which did not need particularly high power, and the input power was derived from the main RF power distribution system via a 20 dB directional coupler. Because of the directional coupler, at most 1/10,000th of the input power could be reflected back to the klystron.

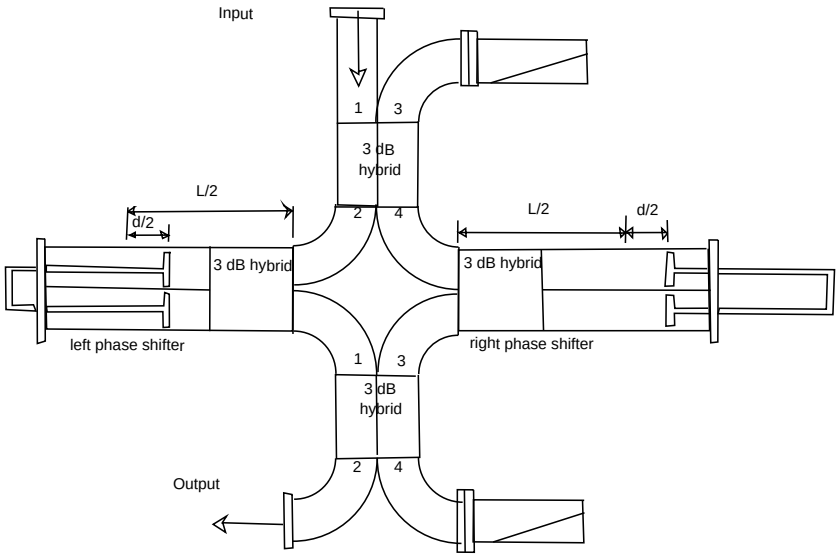


Fig. 21: Attenuator and phase shifter with load isolation

The second configuration of couplers to produce independent phase and amplitude controls reported by the staff of KEK involves the use of four Riblet couplers and completely isolates the power source from the load(s). Figure 21 shows the configuration; the top coupler splits the power equally, the left coupler is configured as a phase shifter with phase shift $\phi - \Delta\phi$, the right coupler is a phase shifter with phase shift $\phi + \Delta\phi$, and the bottom coupler recombines the power. Table 4 describes the action of the device. As before, phase shifts that are common to all signal paths are ignored. Thus once again we have independent controls on amplitude and phase; the phase shift common to both phase shifters controls the overall phase shift, while the difference in phase between the two phase shifters controls the amplitude of the output (either port 2 or port 4 of the bottom coupler).

6 NON-RECIPROCAL DEVICES

The need to isolate sources from the reflections caused by mismatched loads historically prompted the development of non-reciprocal devices at microwave frequencies. The post-war availability of ferrites, with their low electrical conductivity and high permeability, as discussed by Caspers [5], made such devices possible.

The basic property of ferrites used in non-reciprocal microwave devices is the ferrimagnetic resonance, in which the spinning electrons of the ferrite interact strongly with circularly polarized RF magnetic fields at a frequency approximately given by

$$f_{\mathrm{L}} = \frac{e}{2\pi m_{\mathrm{e}}} B \ , \tag{25}$$

where m_{e} is the mass of the electron and B is the strength of the magnetic field. The factor in front of B is approximately 2.8 MHz/G in practical units. In a waveguide one finds circularly polarized RF H fields at a particular distance from the side wall. Referring to Appendix B, where expressions for the RF magnetic fields of the H_{10} waveguide mode are given, we see that if we require that the magnitude of H_z equals that of H_x, this requirement can be met if the condition

$$\sin \frac{\pi x}{a} = \frac{\pi}{\beta a} \cos \frac{\pi x}{a} \tag{26}$$

is met. This condition becomes

$$\tan \frac{\pi x}{a} = \frac{\pi}{\beta a} = \frac{1}{\sqrt{(f/f_{\mathrm{c}})^2 - 1}} \ . \tag{27}$$

At the centre of the waveguide band, where $f = 1.5\ f_{\mathrm{c}}$, this condition yields $x = 0.232a$, i.e. the position at which circularly polarized RF magnetic fields are found is nearly half-way between the side wall and the centre of the waveguide. Thus if a piece of magnetized ferrite with biasing B field chosen for resonance at the waveguide operating frequency is placed in this location, there will be a strong interaction between the RF wave and the spins of the ferrite, for *waves travelling in one direction*. For waves travelling in the opposite direction, the sense of the circular polarization is opposite, and only a weak interaction results. Isolators use just this principle of operation, namely a slab of ferrite with a biasing magnetic field placed in the position of circular polarization, such that waves travelling in one direction are passed with minimal interaction, while those travelling in the opposite direction are strongly absorbed due to the resonant interaction. Figure 22(a) shows the arrangement of an isolator of this kind. The device passes power in one direction with very little loss but with a loss of 20 dB or more in the opposite direction. In moing to high-power operation, it becomes necessary to remove the heat from the ferrite, and designs to do this place the ferrite in maximum contact with the metallic walls of the waveguide, as shown in Fig. 22(b).

Although isolators are used to protect klystrons from reflections (the Los Alamos FEL linac has an isolator at the output of each klystron), it has become more common in high-power systems to use a circulator, since the reflected power is sent to a load rather than dissipated in the ferrite. Circulators come in four-port and three-port varieties. The four-port varieties use Riblet couplers or magic tees to divide the power, and then waveguide sections with low-loss non-reciprocal phase shifts totalling 180° carry the power to a second magic tee or Riblet coupler. The low-loss operation of the non-reciprocal phase shifters results from operating some distance from the resonance of the ferrite spins, where one still has significant interaction with the resonance but minimal loss.

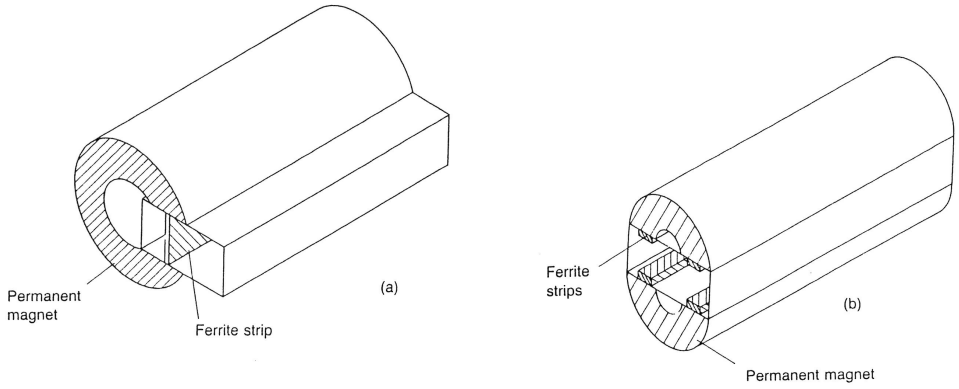


Fig. 22: Outline drawings of ferrite isolators

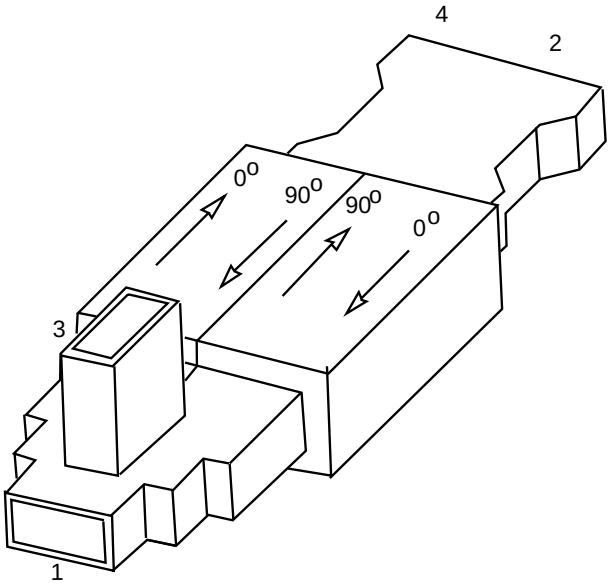


Fig. 23: A four-port (phase shift) circulator

Figure 23 shows a circulator based on a Riblet coupler and a folded magic tee. A commercial isolator of this design made for operation at 2856 MHz has the following characteristics: 10 MHz bandwidth, peak power rating 6 MW, average power rating 6 kW, 0.3 dB maximum insertion loss, 25 dB minimum isolation, VSWR less than 1.10. This water-cooled unit weighs 49 pounds (22.25 kg). A commercially available four-port circulator using two Riblet couplers and operating at 500 MHz has the following characteristics: 10 MHz bandwidth, 0.15 dB insertion loss, 20 dB minimum isolation, 4 MW CW maximum forward power, 600 kW CW maximum reflected power, VSWR less than 1.10. This water-cooled circulator weighs 6000 kg.

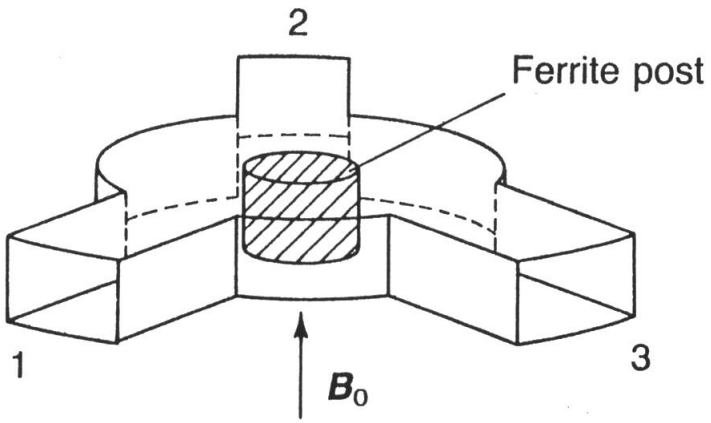


Fig. 24: Arrangement of a ferrite junction circulator

Three-port circulators are typically made with three waveguides symmetrically placed about a circular ferrite disk or stack of disks, as shown in Fig. 24. If one imagines the disks as constituting a resonator that supports the E_{110} circular cavity mode, then the ferrite disks when biased will split the two-fold degeneracy of this mode into modes having positive and negative circular polarization. The RF frequency of operation is then chosen to be midway between the resonant frequency of the positive and negative circularly polarized modes. This driving of the resonant modes off resonance produces a standing wave that couples the input port to only one of the two other ports. A commercially available three-port circulator designed for use at 352.2 MHz has the following characteristics: 10 MHz bandwidth, 0.15 dB maximum insertion loss, 20 dB minimum isolation, 1.2 MW CW maximum forward power, 600 kW CW maximum reflected power, VSWR less than 1.2. This water-cooled circulator weighs 1800 kg. Figure 25 shows part of the high-power waveguide of the CERN LEP system with a three-port circulator in the foreground. Note that the connections to the circulator are made in reduced-height waveguide.

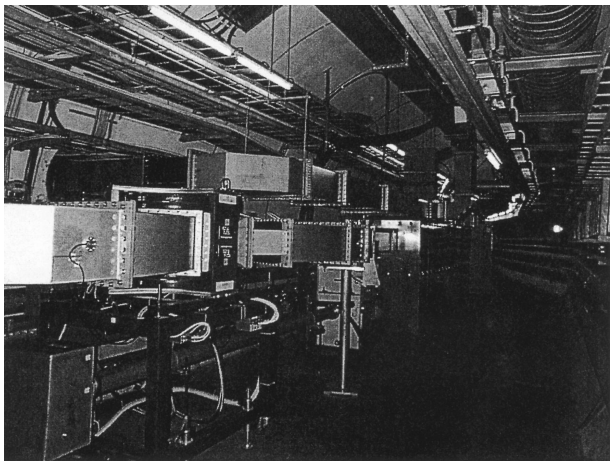


Fig. 25: View of a high-power waveguide system with a ferrite junction circulator in the foreground.

ACKNOWLEDGEMENTS

We would like to thank J.N. Weaver of SLAC for his enthusiastic help in gathering material for this article. Brian Taylor of the Lawrence Berkeley Laboratory was also most helpful. We would also like to thank George Spalek of LANL for his many discussions on RF practice. Finally we would like to thank Jean Browman for her enthusiastic attention to the subject and assistance with the preparation of materials for this work. Colleagues in the RF areas of the Accelerator Technology Division of LANL and the SRS department of the CLRC Daresbury Laboratory have been most helpful in providing material.

APPENDIX A: TEM MODE IN COAXIAL LINE

For a coaxial line with inner radius a and outer conductor radius b, the electric and magnetic fields of the dominant TEM mode are given by

$$E_{\rm r}(r,z,t) = E_0\,\frac{a}{r}\,\cos(\omega t - \beta z)\ \ , \tag{A.1}$$

$$H_\phi(r,z,t) = -\frac{E_0}{\zeta_0}\,\frac{a}{r}\,\cos(\omega t - \beta z)\ \ , \tag{A.2}$$

where ζ_0 is the characteristic impedance of free space. Corresponding to these field quantities one can calculate the integrated quantities of voltage and current (with a somewhat cavalier treatment of algebraic sign):

$$V_r(z,t) = \int_a^b E_r(r,z,t)\,{\rm d}r = E_0\,a\,\log\left(\frac{b}{a}\right)\,\cos(\omega t - \beta z) \tag{A.3}$$

and

$$I(z,t) = \int_0^{2\pi} H_\phi(a,z,t)\,{\rm d}\phi = 2\pi\,a\,\frac{E_0}{\zeta_0}\,\cos(\omega t - \beta z)\ \ , \tag{A.4}$$

where the voltage and the current on the inner conductor are positive for positive E_0.

Taking the ratio of V to I yields the characteristic impedance of the coaxial line

$$Z_{\rm c} = \frac{\zeta_0\,\log(b/a)}{2\pi}\ \ . \tag{A.5}$$

The Poynting vector for this mode is

$$\vec{S}(r,z,t) = \vec{E}\times\vec{H} = \frac{E_0^2}{\zeta_0}\,\left(\frac{a}{r}\right)^2\,\cos^2(\omega t - \beta z)\,\hat{z}\ \ . \tag{A.6}$$

Integrating over the cross sectional area of the coaxial line gives the power flowing at each longitudinal value as a function of time:

$$\begin{aligned} P(z,t) &= \int_a^b r{\rm d}r\int_0^{2\pi}{\rm d}\phi\,S_z = \\ &= \frac{E_0^2\,a^2}{\zeta_0}\,2\pi\,\log\left(\frac{b}{a}\right)\,\cos^2(\omega t - \beta z) \\ &= V(z,t)\,I(z,t)\ \ . \end{aligned} \tag{A.7}$$

so that the relationship between the maximum electric field strength E_0 and the average power transmitted in the coaxial line is

$$\bar{P} = \frac{E_0^2\, a^2}{\zeta_0}\, \pi\, \log\left(\frac{b}{a}\right) \quad . \tag{A.8}$$

If the inner conductor is made of a material with conductivity σ_a, and δ_a is the skin depth, then R_a is the surface resistance $(\sigma_a \delta_a)^{-1}$ of the inner conductor, so the power dissipated in this conductor in a length $\lambda_g = 2\pi/b = \lambda_0 = c/f$ is

$$\begin{aligned} P &= R_{\rm a} \int_0^{\lambda_{\rm g}} 2\pi a {\rm d}z H_\phi^2 \\ &= R_a\, 2\pi\, a \frac{E_0^2}{\zeta_0^2} \int_0^{\lambda_g} \cos^2(\omega t - \beta z)\, {\rm d}z \\ &= \frac{\pi a \lambda_{\rm g} R_{\rm a} E_0^2}{\zeta_0^2} \end{aligned} \tag{A.9}$$

which is independent of time. Thus the power dissipated per unit length in the inner conductor is

$$\left.\frac{{\rm d}P}{{\rm d}z}\right|_{\rm inner} \approx \frac{P}{\lambda_{\rm g}} = \frac{\pi a R_{\rm a} E_0^2}{\zeta_0} \propto \omega^{1/2} \tag{A.10}$$

The outer conductor may be made of a material different from that of the inner conductor. For example, the inner conductor is most often made of copper, while the outer conductor may be made of aluminum to save weight (at a slight increase in attenuation per unit length). If the conductivity of the outer conductor is σ_b, the power dissipated per unit length turns out to be

$$\begin{aligned} \left.\frac{dP}{dz}\right|_{\rm inner} \approx \frac{P}{\lambda_{\rm g}} &= \pi\, b \frac{a^2}{b^2}\, R_{\rm b}\, \frac{E_0^2}{\zeta_0^2} \\ &= \frac{R_{\rm b}}{R_{\rm a}} \left(\frac{a}{b}\right) \left.\frac{dP}{dz}\right|_{\rm inner} \quad . \end{aligned} \tag{A.11}$$

To calculate the attenuation constant for this TEM mode we must divide the power dissipated per unit length by the power being transmitted:

$$\alpha = \frac{1}{2\bar{P}} \frac{{\rm d}P}{{\rm d}z} = \frac{1}{2\zeta_0 \log(a/b)} \left[\frac{R_{\rm a}}{a} + \frac{R_{\rm b}}{b}\right] \quad . \tag{A.12}$$

This attenuation constant agrees with that given by Marcuvitz [10].

APPENDIX B: DOMINANT MODE IN A RECTANGULAR WAVEGUIDE

The fields of the H_{10} (TE_{10}) mode in a rectangular waveguide [width a and height b (usually $b = a/2$)] are given by

$$E_{\rm y}(x,y,z,t) = E_0\, \sin\frac{\pi x}{a}\, \cos(\omega t - \beta z) \tag{B.1}$$

$$H_{\rm x}(x,y,z,t) = -\frac{E_0}{Z_{\rm H}}\, \sin\frac{\pi x}{a}\, \cos(\omega t - \beta z) \tag{B.2}$$

$$H_{\rm z}(x,y,z,t) = -\frac{\pi/a}{\beta}\, \frac{E_0}{Z_{\rm H}}\, \cos\frac{\pi x}{a}\, \sin(\omega t - \beta z) \quad , \tag{B.3}$$

where $\beta^2 = (k^2 - (\pi/a)^2)$, $k = \omega/c$, and

$$Z_{\rm H} \equiv \frac{\omega\mu_0}{\beta} = \frac{k\zeta_0}{\beta} = \zeta_0/\sqrt{1-(f_{\rm c}/f)^2} \ , \tag{B.4}$$

where $f_{\rm c}$ is the waveguide cut-off frequency. The Poynting vector for this mode is

$$\vec{S}(x,y,z,t) = \vec{E} \times \vec{H} = \frac{E_0^2}{Z_{\rm H}} \sin^2\frac{\pi x}{a} \cos^2(\omega t - \beta z)\,\hat{z} \tag{B.5}$$

When integrated over the cross section of the guide the power transmitted as a function of z and t is:

$$P(z,t) = \int_0^b {\rm d}y \int_0^a {\rm d}x\, S_z = \frac{E_0^2}{2Z_{\rm H}}\, ab\, \cos^2(\omega t - \beta z) \ , \tag{B.6}$$

which, when averaged over time, gives the relationship between average power and maximum electric field strength,

$$\bar{P} = \frac{E_0^2}{4Z_{\rm H}}\, ab \ . \tag{B.7}$$

The power dissipated on one side wall in a guide wavelength, $\lambda_{\rm g}$, is

$$\begin{aligned} P_{\rm side} &= R_{\rm s} \int_0^b {\rm d}y \int_0^{\lambda_g} {\rm d}z\, H_{\rm z}^2(a,y,z,t) \\ &= R_{\rm s}\, b\, \frac{E_0^2}{Z_{\rm H}^2} \left(\frac{\pi/a}{\beta}\right)^2 \int_0^{\lambda_g} \sin^2(\omega t - \beta z){\rm d}z \\ &= \frac{\lambda_{\rm g} b}{2} R_{\rm s} \frac{E_0^2}{Z_H^2} \left(\frac{\pi/a}{\beta}\right)^2 = \frac{\lambda_{\rm g} b}{2} R_{\rm s} \frac{E_0^2}{\zeta_0^2} \left(\frac{\pi/a}{k}\right)^2 \\ &= \frac{\lambda_{\rm g} b}{2} R_{\rm s} \frac{E_0^2}{\zeta_0^2} \left(\frac{\lambda}{\lambda_{\rm c}}\right)^2 \ , \end{aligned} \tag{B.8}$$

which is not a function of time. In this last equation λ_c is the cut-off wavelength of the guide. Note that this loss component, arising from transverse wall currents, decreases with increasing frequency as $\omega^{-3/2}$. The power dissipated in a length $\lambda_{\rm g}$ in the top wall is

$$\begin{aligned} P_{\rm top} &= R_{\rm s} \int_0^a {\rm d}x \int_0^{\lambda_g} {\rm d}z \left(H_{\rm z}^2(a,y,z,t) + H_{\rm x}^2(x,b,z,t)\right) \\ &= R_{\rm s} \frac{E_0^2}{Z_{\rm H}^2} \int\int \Big[\left(\frac{\pi/a}{\beta}\right)^2 \cos^2\left(\frac{\pi x}{a}\right) \sin^2(\omega t - \beta z) + \sin^2\left(\frac{\pi x}{a}\right) \cos^2(\omega t - \beta z)\Big] {\rm d}x\, {\rm d}z \\ &= R_{\rm s} \frac{E_0^2}{Z_{\rm H}^2} \frac{a\lambda_g}{4} \Big[\left(\frac{\pi/a}{\beta}\right)^2 + 1\Big] = R_{\rm s} \frac{E_0^2}{Z_{\rm H}^2} \frac{a\lambda_{\rm g}}{4} \left(\frac{k^2}{\beta^2}\right) \\ &= \frac{a\lambda_{\rm g}}{4} R_{\rm s} \frac{E_0^2}{\zeta_0^2} \ . \end{aligned} \tag{B.9}$$

The power dissipated in a length $\lambda_{\rm g}$ includes that dissipated in the top and bottom walls and the two side walls,

$$P = 2 \times \lambda_{\rm g}\, R_{\rm s} \frac{E_0^2}{\zeta_0^2} \Big[\frac{a}{4} + b \left(\frac{\lambda}{\lambda_c}\right)^2\Big] \ . \tag{B.10}$$

Thus the power dissipated per unit length is

$$\frac{{\rm d}P}{{\rm d}z} \approx \frac{P}{\lambda_{\rm g}} = R_{\rm s} \frac{E_0^2}{\zeta_0^2} \Big[\frac{a}{4} + b \left(\frac{\lambda}{\lambda_c}\right)^2\Big] \ . \tag{B.11}$$

For the usual case in practice, $b = a/2$, so that the power dissipated per unit length is

$$\frac{\mathrm{d}P}{\mathrm{d}z} = R_\mathrm{s}\,\frac{a}{2}\,\frac{E_0^2}{\zeta_0^2}\left[1 + \left(\frac{\lambda}{\lambda_\mathrm{c}}\right)^2\right] \ . \tag{B.12}$$

The surface resistance R_s, being inversely proportional to the skin depth, increases as the square root of the frequency, while the quantity inside the brackets decreases with increasing frequency. Note that the contribution to the loss per unit length attributable to the transverse wall currents (arising from H_z) decreases with increasing frequency as $\omega^{-3/2}$, while that contribution from the longitudinal currents (arising from H_x) increases with increasing frequency as $\omega^{1/2}$ (for $\omega \ll \omega_c$). This observation suggests that a mode inducing only transverse wall currents would have wall losses that decrease with increasing frequency. Such a mode is the H_{01} (TE_{01}) mode in circular waveguide. This mode has been the subject of considerable attention for many years for its potential for low-loss transmission. The fact that it is a higher order mode means that any imperfection in the waveguide will result in energy being converted from the desired H_{01} mode into other propagating modes, thus providing an energy loss mechanism other than Joule heating.

Finally, the attenuation constant is calculated by taking the ratio of the power dissipated per unit length to the power being transmitted (for the case $b = a/2$),

$$\alpha = \frac{1}{2\bar{P}}\frac{\mathrm{d}P}{\mathrm{d}z} = \frac{2R_\mathrm{s}\left[1 + (f_\mathrm{c}/f)^2\right]}{a\zeta_0\sqrt{1-(f_\mathrm{c}/f)^2}} \tag{B.13}$$

where the relation $Z_\mathrm{H} = \frac{\zeta_0}{\sqrt{1-(f_\mathrm{c}/f)^2}}$ has been used. This expression for the attenuation constant agrees with that given in Marcuvitz [10].

REFERENCES

[1] T. Moreno, *Microwave Transmission Design Data* (Dover Publications, New York, 1958) (reprint).

[2] F.E. Borgnis and C.H. Papas, *Electromagnetic Waveguides and Resonators, Handbuch der Physik, Vol. XVI, Electric Fields and Waves* Berlin, Springer-Verlag, 1958).

[3] H.A. Wheeler and H. Schwiebert, Step-twist waveguide components, *Trans. I.R.E.*, MTT-3, (1955) 44.

[4] David M. Pozar, *Microwave Engineering* (Addison-Wesley, Reading, MA, 1990), Chapter 8, section 4.

[5] F. Caspers, Basic Concepts II, these proceedings.

[6] Horst Klein, Basic Concepts I, these proceedings.

[7] A.F. Harvey, *Microwave Engineering* (Academic Press, New York, 1963).

[8] H.J. Riblet, The short-slot hybrid junction, *Proc. I.R.E.* **40** (1952) 180.

[9] S. Ohsawa *et al.*, High-Power Hybrid Attenuator and Phase-Shifter Systems, Proc. Linear Accelerator Conf., Albuquerque, New Mexico, 1990, Los Alamos document LA-12004-C.

[10] N. Marcuvitz, *Waveguide Handbook*, M.I.T. Radiation Laboratory Series vol. 10 (McGraw-Hill, New York, 1951), also IEE Electromagnetic Waves Series, 1986.

CAVITIES WITH A SWING

A. Schnase
Jülich Research Centre, Jülich, Germany

Abstract
A description is given of the basic configurations used in variable frequency accelerating cavities and drift tubes that use ferrite. The relevant ferrite parameters that determine cavity performance are discussed. Equivalent circuits for the load on the RF power source and the ferrite biasing source are given. New aspects are cavities filled with amorphous alloys that show broadband behaviour without tuning loops. This text is a rewritten and updated version of 'Ferrite dominated cavities' by I.S.K. Gardner [1].

1. INTRODUCTION

Ferrite cavities are the devices that transfer energy to the beam in proton and ion synchrotrons. The particle velocity starts from the non-relativistic region and may reach the relativistic region. Depending on the circumference of the machine and the harmonic used, the frequency swing is within the approximate range of 10 kHz to 10 MHz. A simplified structure of an RF system is shown in Fig. 1. An oscillator generates a frequency derived from the radial beam position and actual dipole field. The signal is amplified and drives a cavity, where the beam sees the acceleration voltage.

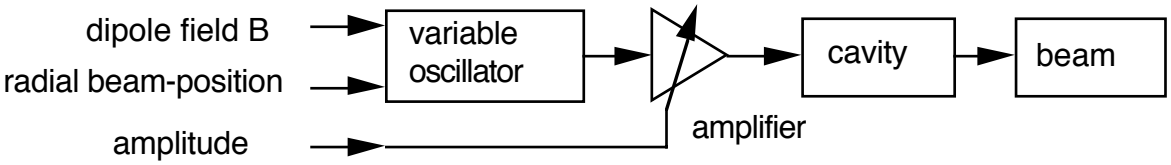


Fig. 1: Simplified structure of an RF system

To create the acceleration voltage along the beam pipe, a gap is inserted (see Fig. 2(a)). This device will radiate, so screening is necessary (see Fig. 2(b)). The outer shielding is a short circuit for the acceleration voltage, so it is inductively isolated from the gap by the use of ferrites (see Fig. 2(c)). Then the inductance, L, of the ferrite combined with the capacitance, C, of the gap forms a resonant circuit, whose frequency has to be tuned to the velocity of the circulating beam.

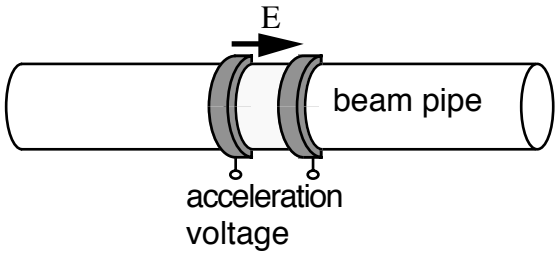


Fig. 2(a): Insertion of an isolating gap into the beam pipe to create a connection for the acceleration voltage

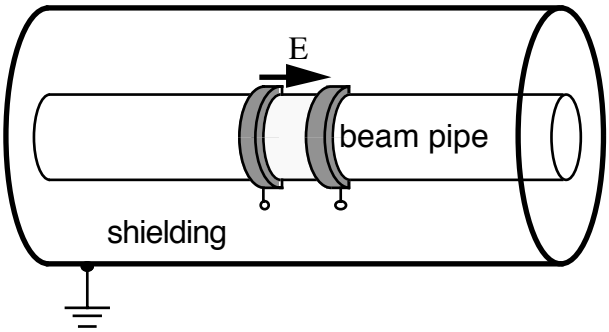


Fig. 2(b): Screening against RF radiation

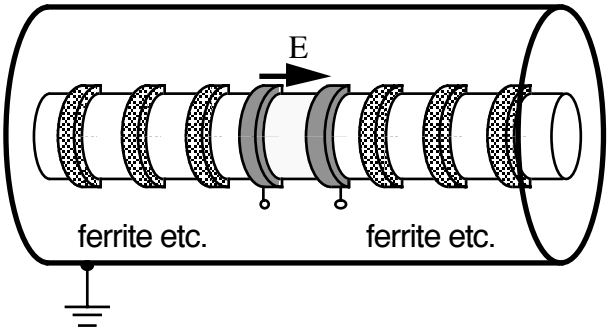


Fig. 2(c): Inductive isolation of the acceleration gap from the beam pipe and outer shielding

An incomplete list of ferrite-filled accelerating structures used in accelerators is given in Table 1. The ferrites for these systems are supplied by a variety of manufacturers. The frequency swing required varies widely. A large frequency swing is required for heavy-ion synchrotrons and a rapid frequency swing is needed for rapid-cycling synchrotrons. Large accelerators have long straight sections with room for many cavities. Other accelerators have little room, few straight sections, or even have stacked rings with small vertical separations, like the PS Booster at CERN. These differences lead to various solutions to the cavity or drift tube design. However, the designs can in general be described by five basic configurations. Finally, designs using materials such as VitroVac, FineMet, and VitroPerm are presented—these achieve broadband behaviour and reduced size compared with ferrite-filled cavities. A comparison between ferrite and VitroVac can be found in Table 2.

2. FERRITE

2.1 Relevant parameters

Naturally occurring ferrous ferrite has the formula Fe_3O_4, often written as $Fe^{3+}(Fe^{2+}Fe^{3+})O^4$ to show the distribution of the metal ions in the crystal structure [2]. A range of ferrites can be produced by replacing the Fe^{2+} ion by any other divalent metal, as long as the ion diameter is comparable to that of the Fe^{2+} ion. Typical divalent metals fulfilling this requirement are nickel, manganese, magnesium, cobalt, copper, zinc, and cadmium.

Ferrite materials with a wide range of magnetic, electrical and mechanical properties can be produced by mixing these different ferrites [2, 3]. The main properties that must be considered for use in RF cavities are magnetic permeability, dielectric constant, resistivity, thermal conductivity, and the effects of temperature and electromagnetic fields on these properties.

Table 1: Parameters of some synchrotrons that use ferrite-tuned cavities

Synchro-tron	No. of cavs.	No. of gaps per cavity	Tuning range (MHz)	Accel-erating time (s)	Max. df/dt (MHz/s)	Gap capacity (pF)	Ind. range (μH)	Type of ferrite	B_{max} in Ferrite (mT)	Bias current range (A)	Tuning system bandwidth (kHz)
ISIS	6	2	1.3–3.1	0.01	325	2200	6.8–1.3	Philips 4M2	10	200–2300	6
CERN PS	11	2	2.8–9.6	0.7				Philips 4L2		3100	
CERN PSB before 1998	1 / ring	1	3–8.4	0.45		80		Philips 4L2		60–800	15
CERN PSB_C02		1	0.6–1.8					Philips 4A11	12		
CERN PSB_C04			1.2–3.9					Philips 4L2	9.4		
CERN PSB_C16			6.0–17.0					Philips 4M2	3.2		
CERN LEAR	2	1	0.38–3.5	0.10		500–3000		Philips 8C12 / Toshiba PE17			
DESY-III	1	2	3.27–10.33	3.6						160–2000	
SACLAY MIMAS	2	1	0.15–2.5	0.2	14			TDK C4 SY7		0–400	
SACLAY SATURNE	2	1	1.7–8.3	0.5				Philips 8C12		-10–200	1–10
COSY	1	1	0.45–1.6	1–3	1.6	1000		Philips 8C12		-10–100	1–10
CELSIUS	1		0.4–2 1–5							1500	
KEK PS	4	2	6–8	0.8	14.5	100	7–4	Toshiba M4B23 μ ~100	7	80–400	3
KEK BOOSTER	2	2	2.2–6	0.025	265	650	8–1	Toshiba M4A23 μ ~150	10	250–2200	1
FNL BOOSTER	18		30.3–52.8	0.033	3000			Stack-pole and Toshiba		50–2250	
BROOKHAVEN AGS	10	4	2.52–4.46	0.6							
BROOK-HAVEN BOOSTER	2	4	2.4–4.2	0.062		395	115–37	Philips 4M2		145–900	
GSI–SIS	2	1	0.85–5.5					Philips 8C12			

The hysteresis loop followed by a ferromagnetic material when subjected to a large field variation is shown in Fig. 3. Added to this is the effect of a smaller, rapidly varying RF magnetic field H_{ac}. As can be seen, the incremental permeability $\mu_r = B_{ac}/(\mu_0 H_{ac})$ varies when the operating point on the hysteresis loop is moved. Thus an inductor made with a ferrite core will have a different value of inductance depending on the magnetic biasing field, and also on the magnitude of the applied H_{ac}. The variation of μ that can be achieved in a particular type of ferrite determines the inductance change that can be obtained. The variation of the incremental permeability with applied field for several grades of ferrite is shown in Fig. 4 [3].

Table 2: Comparison between a ferrite- and a VitroVac-filled cavity

	(h = 1)-cavity	**VitroVac-cavity**
Material	ferrite 8C12, Philips	VitroVac 6025F
Number of toroids	46	24
Frequency range	0.44–2 MHz	0.2– 8 MHz
Tuning current	-20–70 A	0–10 A
Duration of a phase jump	10–50 periods	2 periods
Impedance	ca. 1 kΩ	ca. 300 Ω
RF power	up to 50 kW	10–50 kW
Max. RF amplitude	7100 V peak	2450 V peak
Signal shape	Sinusoidal	Sinusoidal + harmonics

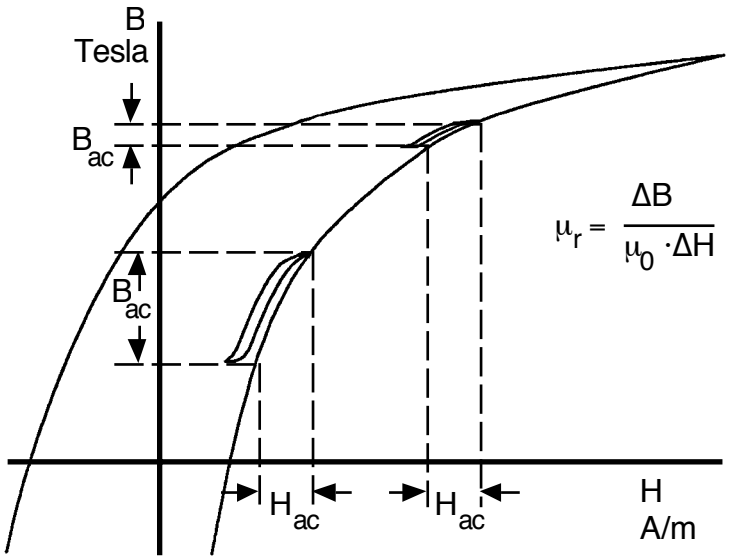


Fig. 3: Ferrite B-H loop with added AC field H_{ac} at two different working points

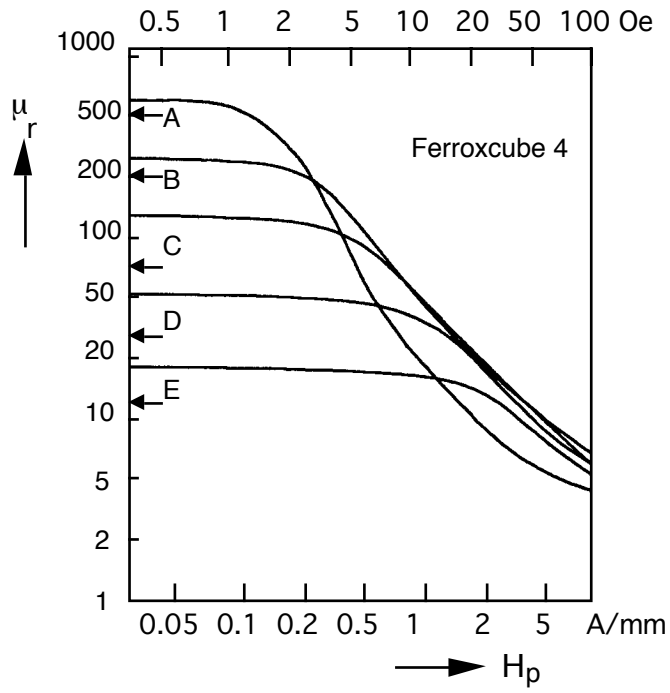


Fig. 4: Plot of μ_r against magnetic bias field H_p for different grades of Philips NiZn ferrite (see also Fig. 6)

2.2 Ferrite inductors

Inductors with ferrite cores offer an increased inductance value compared with an air-cored version, and will also have increased losses due to hysteresis and eddy current loss in the core. The loss can be represented by assigning a complex value to the magnetic permeability:

$$\mu = \mu' - j\mu'' \,. \tag{1}$$

The real part μ' measures the increase in inductance. The imaginary part μ'' is a measure of the core loss. Thus the impedance of the inductor is

$$Z = j\omega\mu L_0 \tag{2}$$

$$= j\omega L_0 \,(\mu' - j\mu'')$$

$$= j\omega\mu' L_0 + \mu''\omega L_0$$

$$= j\omega L + R \,,$$

where L_0 is the value of the air-cored inductor.

$$L = \mu' L_0 \tag{3}$$

$$R = \mu''\omega L_0 \,. \tag{4}$$

The quality of this series circuit is defined as

$$Q = \omega L / R = \omega\mu' L_0 / \omega\mu'' L_0 \tag{5}$$

$$Q = \mu' / \mu'' \,.$$

If this ferrite-loaded inductance is connected across a lossless capacitor, as in Fig. 5, then the impedance of the circuit at resonance can be written as

$$Z = Q\omega L = Q\omega\mu' L_0 \tag{6}$$

$$Z = (\mu' Qf)\, 2\pi L_0 \, .$$

The factor $(\mu' Qf)$ can be regarded as a figure of merit for ferrite materials in a given set-up. Plots of μ' and μ'' against frequency are shown in Fig. 6 for different grades of Philips Ferroxcube 4 [3]. The specifications of the material 8C12 are shown in Fig. 7 [4].

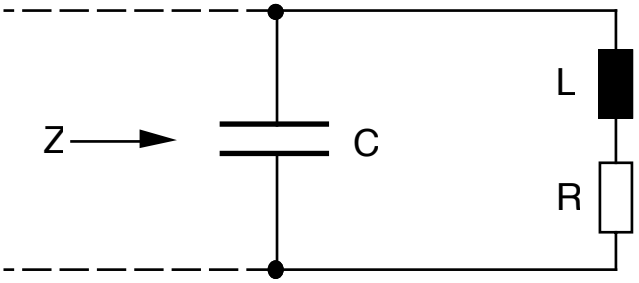


Fig. 5: Parallel resonant circuit with ferrite-cored inductor

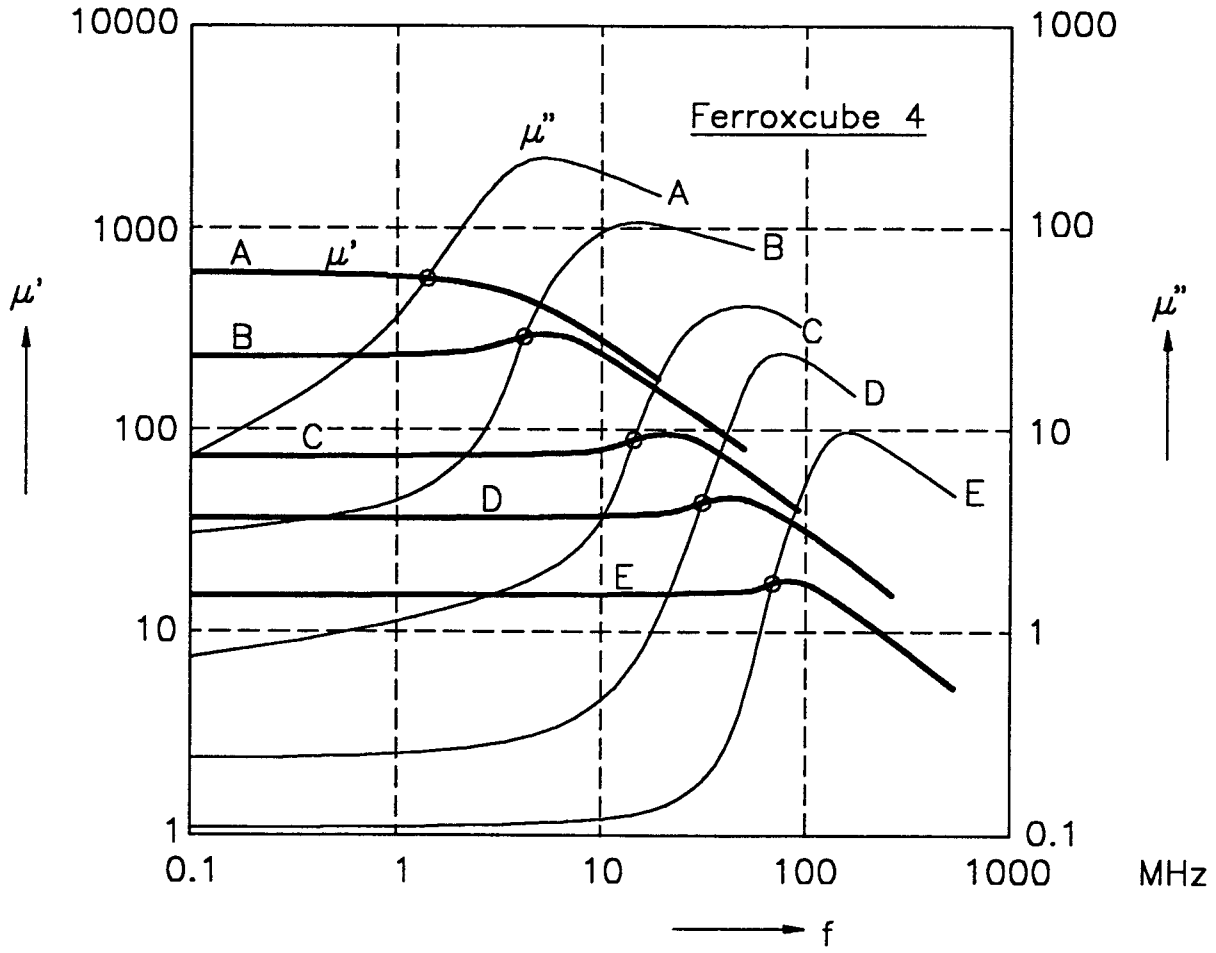
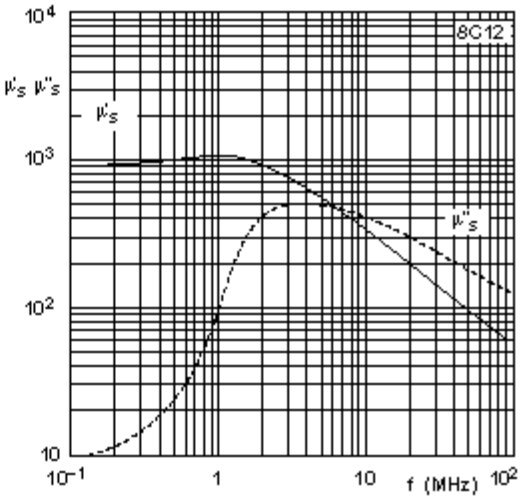


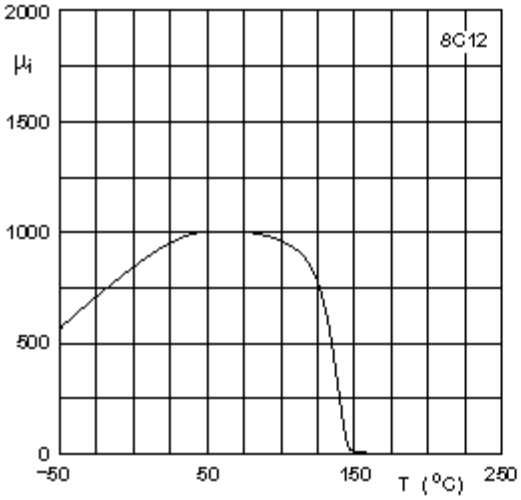
Fig. 6: Plots of μ' and μ'' against frequency for different grades of Philips ferrite

8C12 SPECIFICATIONS

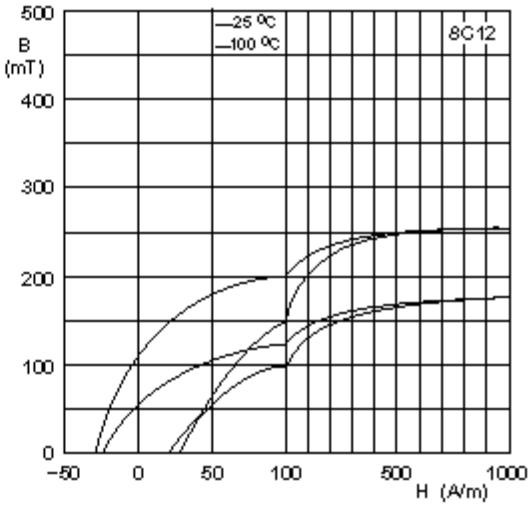
SYMBOL	CONDITIONS	VALUE	UNIT
μ_i	25 °C; ≤10 kHz; 0.1 mT	900 ±20%	
B	25 °C; 10 kHz; 250 A/m 100 °C; 10 kHz; 250 A/m	≈230 ≈150	mT
ρ	DC; 25 °C	$\approx 10^5$	Ωm
T_C		≥125	°C
density		≈5100	kg/m³



Complex permeability as a function of frequency.



Initial permeability as a function of temperature.



Typical B-H loops.

Fig. 7: Specifications of Philips ferrite 8C12

2.3 Dielectric constant of ferrite

The dielectric loss in ferrite can be treated in the same manner as the magnetic loss by assigning a complex value to the permittivity ε, such that

$$\varepsilon = \varepsilon' - j\varepsilon''. \tag{7}$$

This results in a resistance in parallel with a ferrite filled capacitor:

$$R = 1/\omega\varepsilon'' C_0 \tag{8}$$

where C_0 is the value of the air-filled capacitor. The Q-value obtained when the ferrite-filled capacitor is resonated with a lossless coil is then

$$Q = \varepsilon' / j\varepsilon''. \tag{9}$$

However, with NiZn ferrites, ε' remains constant at a value of about 10 and does not vary with electric or magnetic fields [5]. At frequencies of up to 30 MHz the value of ε'' has not been found to be significant [5]. This is particularly the case with resonators that use metallic cooling plates between the ferrite toroids, when there is very little electric field penetration of the ferrite.

3. COAXIAL RESONATORS

3.1 Characteristic impedance and resonance conditions

Most of the variable frequency RF systems found in accelerators use shorted, ferrite-loaded, coaxial transmission lines as inductors to resonate with the accelerating gap or drift tube capacitance. In the case of the coaxial cavity the beam passes through the hollow central coaxial conductor and is accelerated by the electric field across the end capacitive gap. With the drift tube, the beam is accelerated by the electric fields across the gaps at each end. The drift tube forms a capacitor that is connected to the end of the resonator. The energy gain is dependent on the phase change in these electric fields during the transit time of the beam through the drift tube. In both cases the circuit is tuned to the required accelerating frequency by changing the biasing magnetic field in the ferrite of the resonator. In Fig. 8, C_g represents the gap or drift tube capacitance.

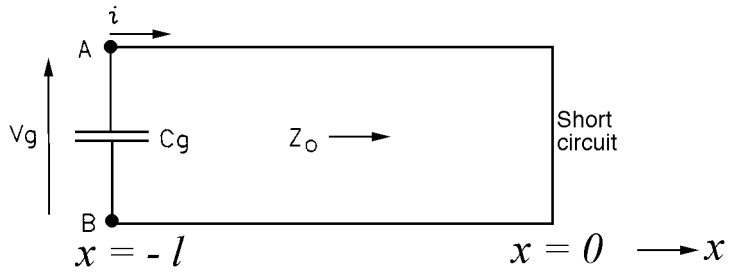


Fig. 8: Basic coaxial resonator circuit as short-circuited transmission line

The reactive part of the impedance of such a resonator at the points AB is

$$Z = j \cdot Z_0 \tan(\omega l / v) \tag{10}$$

where l is the length of the line, v the velocity of the signal in the line, and Z_0 the characteristic impedance of the transmission line.

With $\omega = 2\pi f$ and the wavelength $\lambda = (v / f)$ the relation

$$(\omega l / v) < \pi / 2 \tag{11}$$

is equivalent to

$$(2\pi l / \lambda) < \pi / 2$$

$$\Leftrightarrow 1 / \lambda < 1 / 4 \,. \tag{12}$$

Then Z is inductive

$$Z = j\omega L = j \cdot Z_0 \tan(\omega l / v) \tag{13}$$

with

$$L = (Z_0 / \omega) \tan(\omega l / v) \,. \tag{14}$$

The resonance condition

$$\omega^2 = \frac{1}{LC_g} \tag{15}$$

with the gap capacitor C_g gives the length of the line to reach a desired frequency

$$l = \frac{v}{\omega}\arctan\left(\frac{L\omega}{Z_0}\right) = \frac{v}{\omega}\arctan\left(\frac{1}{Z_0\,\omega C_g}\right) . \tag{16}$$

The inverse function to obtain the frequency has to be solved numerically.

A typical resonator structure is shown in Fig. 9, where the ferrite is in the shape of toroids and each is separated from the next by an air-cooled gap or a water-cooled metal plate. A gap between the inner tube and the cooled metal discs prevents voltage breakdown.

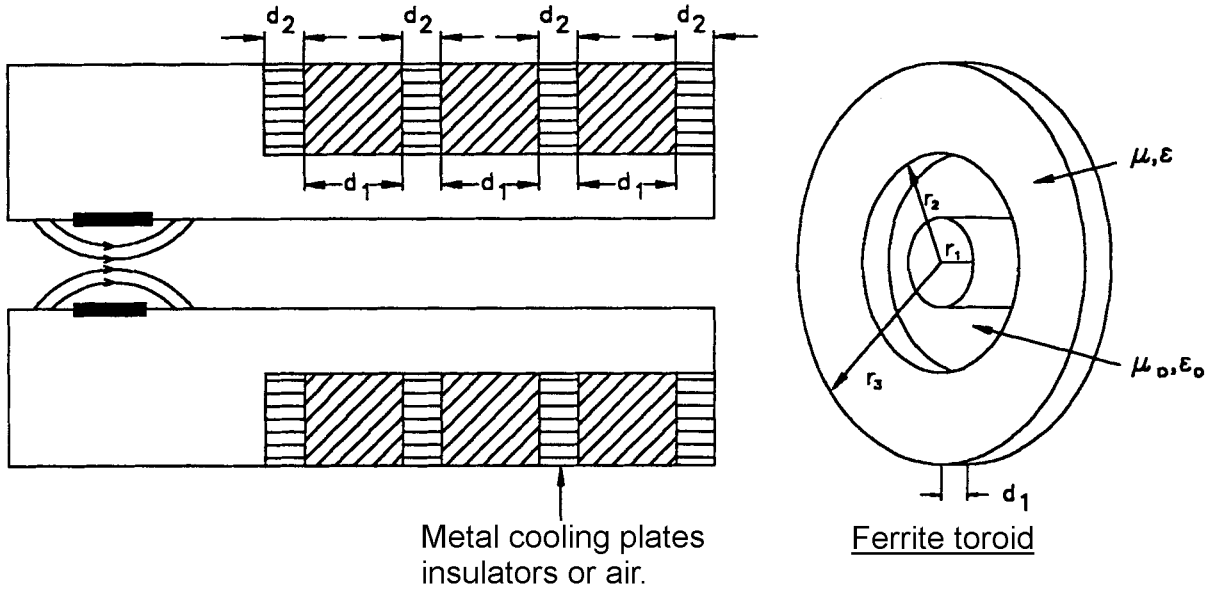


Fig. 9: Typical resonator structure

The capacitance of this line (with the effective ε_e) in farads/metre is

$$C_t = \frac{2\pi\,\varepsilon_e\,\varepsilon_0}{\ln\frac{r_3}{r_1}} \tag{17}$$

and the inductance (with the effective μ_e) in henrys/metre is

$$L_t = \frac{1}{2\pi}\;\mu_e\,\mu_0 \ln\frac{r_3}{r_1} . \tag{18}$$

The effective permittivity ε_e is:

- case 1 for an air-filled gap

$$\varepsilon_e = \frac{\varepsilon}{k + \varepsilon\,(1-k)}\;\frac{d_1}{d_1 + d_2} \tag{19}$$

- case 2 for a resonator with cooled metallic plates

$$\varepsilon_e = \frac{1}{1-k} \tag{20}$$

The effective permeability μ_e is in both cases

$$\mu_e = \left(1 + k\ (\mu' - 1)\right)\ \frac{d_1}{d_1 + d_2} \tag{21}$$

with (derivation in Appendix A)

$$k = \ln\frac{r_3}{r_2}\ \Big/\ \ln\frac{r_3}{r_1}\ . \tag{22}$$

Example: material grade 8C12 used in COSY (h = 1) ferrite cavity:

$r_1 = 0.1$ m, $r_2 = 0.125$ m, $r_3 = 0.25$ m, gives k = 0.76, $\varepsilon_e = 4.1$

$d_1 = 25$ mm, $d_2 = 6$mm, $\mu' = 700$, gives $\mu_e = 427$ (e.g. 60%).

The wave velocity (with c_0 = velocity of light) in such a line is

$$v = \frac{1}{\sqrt{L_t\, C_t}} = \frac{c_0}{\sqrt{\mu_e\, \varepsilon_e}}\ . \tag{23}$$

The characteristic impedance of this loaded line is

$$Z_0 = \sqrt{\frac{L_t}{C_t}} = \frac{1}{2\pi}\ \ln\frac{r_3}{r_1}\sqrt{\frac{\mu_e\, \mu_0}{\varepsilon_e\, \varepsilon_0}}\ = 60\ \Omega\ \ln\frac{r_3}{r_1}\ \sqrt{\frac{\mu_e}{\varepsilon_e}}\ . \tag{24}$$

If the voltage at the capacitor is given as a complex vector $V_g(t) = V_g\, e^{j\omega t}$, the voltage *V(t, x)* and current *i(t, x)* are given by

$$V(t,x) = -V_g\ \ e^{j\omega t}\ \frac{\sin\dfrac{\omega\, x}{v}}{\sin\dfrac{\omega\, l}{v}}\ , \tag{25}$$

$$i(t,x) = jV_g\ \ e^{j\omega t}\ \frac{\cos\dfrac{\omega\, x}{v}}{Z_0\ \sin\dfrac{\omega\, l}{v}} \tag{26}$$

(derivation in Appendix B).

The current is at a maximum at the short circuit where $x = 0$ and the ratio of the current at $(x = -l)$ to the current at $(x = 0)$ is

$$\frac{i(x = -l)}{i(x = 0)}\ = \frac{\cos\dfrac{\omega\ l}{v}}{1} = \cos\frac{\omega\ l}{v}\ . \tag{27}$$

This ratio determines the difference in the RF magnetic fields in the toroids at both ends of the resonator. Resonators are usually designed to keep this difference within about 10%, indicating a value of $(\omega l / v) <$ arc cos (0.9) = 26°.

3.2 RF magnetic induction in the ferrite

In Fig. 10 the magnetic field H at radius r and the induction B due to the current i are

$$H(r) = \frac{i}{2\pi r} \qquad B(r) = \mu\mu_0 \frac{i}{2\pi r} . \tag{28}$$

The maximum induction B_{rfmax} is at the radius $r = r_2$ and longitudinal at the short circuit where i is a maximum in the resonator (I is the current at $x = -l$):

$$B_{rf\ \max} = \mu\mu_0 \frac{I}{2\pi r \cos\dfrac{\omega l}{v}} . \tag{29}$$

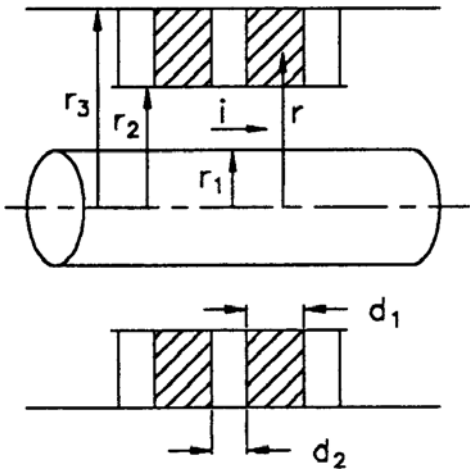


Fig. 10: Magnetic induction in the ferrite at radius r

In NiZn ferrites, where the saturation field is 0.2–0.3 T, the maximum value of B_{rf} is usually kept below 0.01 T. The Q-value of the ferrite will decrease as B_{rf} is increased. A typical graph of $(\mu' Qf)$ against B_{rf} is shown in Fig. 11. [6]

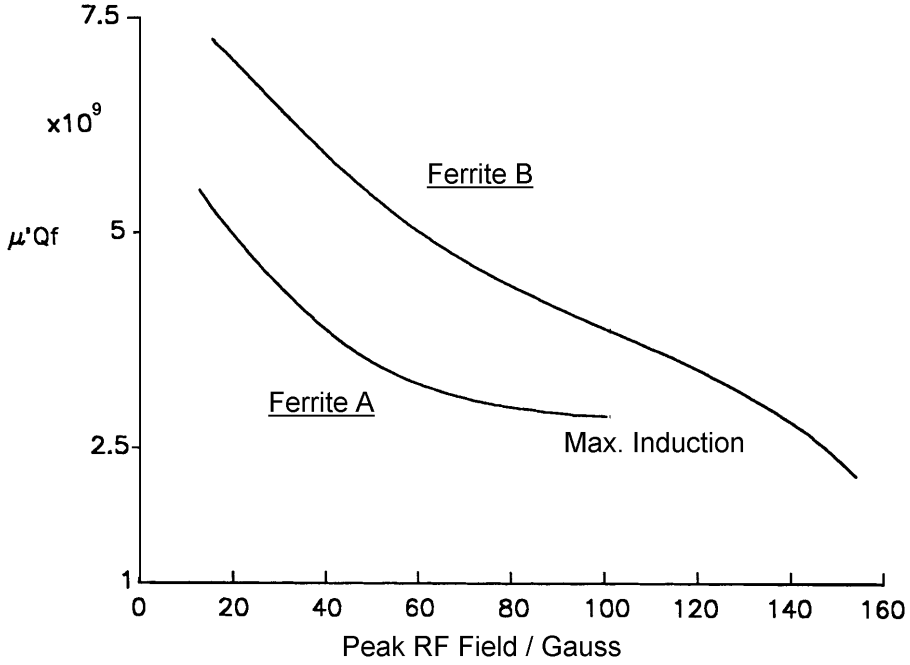


Fig. 11: Plot of $(\mu' Qf)$ against peak RF field at 2.0 MHz

The required resonator length (derivation in Appendix C) for a given cap voltage V_g is estimated from

$$l = \frac{V_g \ \mu}{\mu_e \, r_2 \ln \frac{r_3}{r_1} \ \omega \ B_{rf\,\max}} \ . \tag{30}$$

3.3 Power dissipation in the ferrite

The equivalent circuit of the resonator in Fig. 12 describes losses with $R = Q\omega L$.

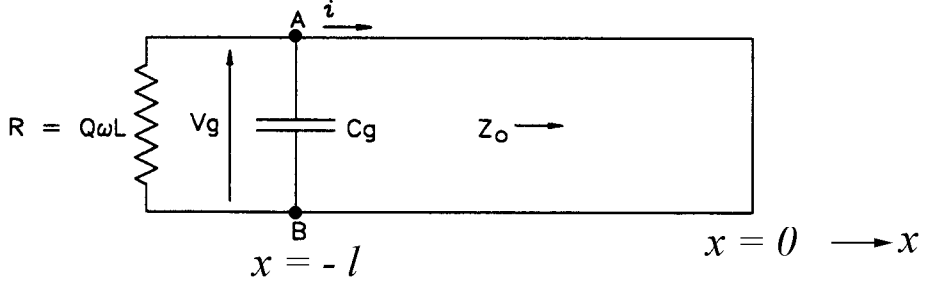


Fig. 12: Diagram of transmission line with C_g and resistor R.

The power dissipation in the resonator is

$$P = \frac{V_g^2}{2R} = \frac{V_g^2}{2Q\omega L} \ , \tag{31}$$

where P is the total power in the resonator, and V_g the peak RF voltage (amplitude) on the end of the resonator. The mean power per unit length in the resonator is

$$P_{mean} = \frac{V_g^2}{2lQ\omega L} \ \text{[W/m]} \ , \tag{32}$$

where Q will vary with voltage and frequency. The power dissipation along the resonator, $P(x)$, depends on the square of the current, i^2.

$$P(x) = P_{\max} \cos^2 \frac{\omega x}{v} \ \text{[W/m]} \tag{33}$$

$$P_{mean} = \frac{P_{\max}}{2} \left(1 + \frac{\sin \frac{2\omega l}{v}}{\frac{2\omega l}{v}} \right) \ \text{[W/m]} \ . \tag{34}$$

The maximum power density in a ferrite toroid and the temperature rise T is given by

$$P_{d\,\max} = P_{\max} \frac{d_1 + d_2}{\pi \left(r_3^2 - r_2^2 \right) d_1} \ \text{[W/m}^3\text{]} \tag{35}$$

$$T = \frac{P_d}{k} \ \frac{d_1^2}{4} \ \text{[degrees]} \ , \tag{36}$$

where k is the thermal conductivity of the ferrite in (W/mK), and P_d the power density in the ferrite in (W/m^3).

With thermal conductivities of k = (3.5 W / mK) and a toroid thickness of 25 mm the power density is limited to 100–300 W/l for most NiZn ferrites. The usable power density depends on ferrite properties and mechanical design. A rise in the ferrite temperature increases the μ-value and decreases the Q-value [5]. Finally, the μ-value decreases rapidly as the Curie temperature of the ferrite is reached. This can lead to thermal runaway in the ferrite.

3.4 High loss effect

High loss effect is exhibited as a time-dependent phenomenon, and it occurs when the stored energy in ferrite is raised above a threshold level of (3 ± 1) x 10^{-7} J/cm^3 [7]. At this power density the Q of the resonator drops after a period of time and the voltage across the resonator becomes unstable.

4. THE EFFECT OF MAGNETIC BIAS FIELDS ON FERRITE

4.1 The comparison of parallel and perpendicular magnetic bias fields

The application of a magnetic field to ferrite leads to a change in the incremental permeability, and this field can be applied either parallel or perpendicular to the RF field. As most of the systems in use today use parallel bias this will be discussed first. Plots of the incremental permeability, μ, as a function of the magnetic biasing field were shown for five grades of ferrite in Fig. 4. In general the ferrites with high values of initial permeability will change the permeability at lower values of applied bias field. The frequency range over which the system can be tuned is given by the change in μ' value.

$$\omega^2 = \frac{1}{LC_g} \qquad \omega \tan\left(\frac{\omega l}{v}\right) = \frac{1}{Z_0\ C_g} \tag{37}$$

Relations to the permeability μ':

$$v \propto \frac{1}{\sqrt{\mu'}} \qquad \omega \propto \frac{1}{\sqrt{\mu'}} \qquad Z_0 \propto \sqrt{\mu'}\ . \tag{38}$$

So the ratio of maximum to minimum frequency is given by $\sqrt{\frac{\mu_{\max}}{\mu_{\min}}}$, (39)

where $\mu_{\max}$ is the maximum value of μ' at the minimum bias field, and $\mu_{\min}$ is the minimum value of μ' at the maximum bias field.

It has been demonstrated [5, 8, 9] that the use of a bias field perpendicular to the RF field significantly reduces losses in the ferrite. It has also been found that the change in μ for a given value of perpendicular bias field is much reduced compared with that for a parallel bias field. This lower loss is explained by the fact that the ferrite is operated closer to magnetic saturation. There is great interest in this type of biasing for resonators working at the higher frequency end of the spectrum (around 50–80 MHz) [8].

4.2 Dynamic loss

Dynamic loss in a resonator is observed as a decrease in Q of the resonator subject to changing bias field as compared with that of a resonator with a fixed bias field [7]. This loss is not associated with the bandwidth of the resonator but is very dependent on the ferrite mixture used. It is therefore important to measure ferrite performance at the rate of change of bias field at which it will be used.

5. COMMON DESIGNS AND THEIR EQUIVALENT CIRCUITS

5.1 General considerations

Accelerating structures can be divided into two main types, depending on whether the system uses drift tubes plus coaxial resonators or coaxial cavities. A third type, which needs to tune over only a small frequency range, uses a coaxial cavity with a loosely coupled external resonator. Most other variations are introduced by the necessity of providing a variable bias field to change the ferrite μ' value. A further division can be made between those that apply the bias field in parallel with the RF field and those that apply it perpendicularly to the RF field.

A typical synchrotron RF system is shown in Fig. 13. The cavity is kept in tune with the frequency required for acceleration by the servo control of the cavity bias field current. When in tune, the cavity presents a purely resistive load to the tetrode. To achieve this, the servo keeps the cavity voltage in antiphase to the grid voltage of the tetrode. The bias field current may swing from a few amps to a few thousand amps while covering the frequency range of the RF [10].

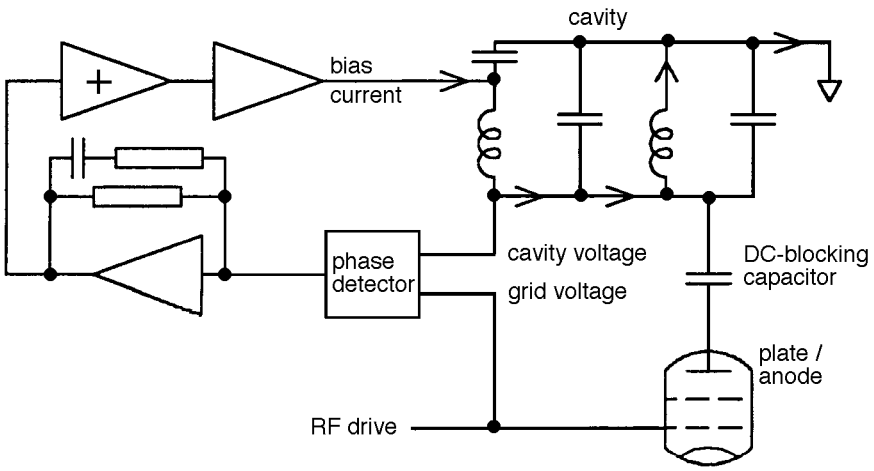


Fig. 13: Servo control of cavity tuning

5.2 Type 1 cavity

Figure 14 shows the main elements of an ISIS cavity. It is very similar to the design of a CERN PS cavity. The water-cooled plates, which provide the ferrite cooling, can be seen separating the 35 ferrite rings in each resonator. The centre tube of the co-ax is made from two coaxial tubes. The outer tube is made from copper to provide low impedance for the bias field current and the RF. The inner tube is made from nickel-plated mild steel to provide the vacuum pipe and to shield the beam aperture from any magnetic field produced on the beam axis by the bias field. The two accelerating gaps are ceramic, metallized at the ends and welded to the mild steel.

The simplest equivalent circuit is shown in Fig. 15. The total capacitance at a gap is represented by C_g, and includes the cap capacitance, the added capacitance, and the stray capacitance. The two gaps are connected in parallel by two 70 Ω coaxial lines that carry both the bias field current and the RF voltage. The RF by-pass capacitor, C_b, forms a low impedance for the circulating RF current in one of the resonators but allows the bias field current to be fed around both resonators, which are in series with the bias current. The bias field is parallel to the RF field but in the opposite direction in each resonator. With a gap voltage of 14 kV, peak C_b must be about 10,000 C_g to limit the RF voltage appearing across the bias field supply to less than a few volts.

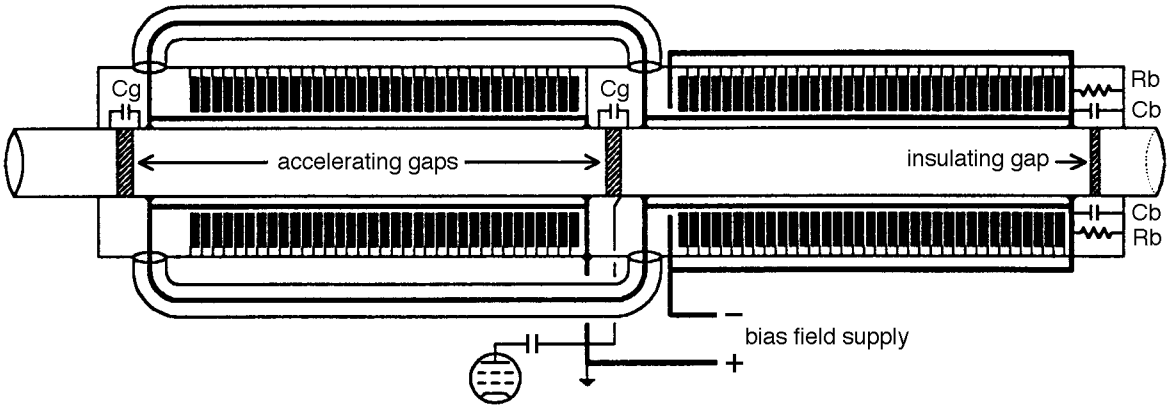


Fig. 14: Type 1 cavity (ISIS)

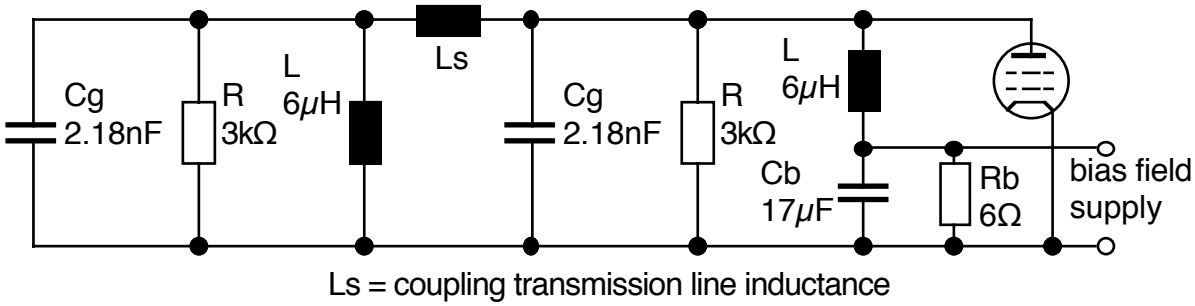


Fig. 15: Type 1 cavity (ISIS) equivalent circuit

The load presented to the RF amplifier is that of the two cavities in parallel. The bandwidth of this circuit is just $f_0/(2Q)$ and the minimum value for this on ISIS is 6.5 kHz in the swept condition. The capacitor C_b and the two resonators also form a tuned LC load for the bias field system. The resonance frequency is

$$f = \frac{1}{2\pi\sqrt{2LC_b}} \quad . \tag{40}$$

In ISIS this frequency has a minimum value of 10.4 kHz. Thus any control loop on the bias supply involving phase modulation of the cavity voltage will incorporate two bandwidth limiting time constants. In ISIS the Q of the resonance produced by C_b is reduced to near unity by a 6 Ω resistor in parallel with C_b. The time constant associated with C_b will decrease as L reduces. The time constant associated with the cavity Q will vary with frequency and gap voltage, and will also depend on whether the cavity is sweeping or not. If stable operation in all conditions is required then the worst-case bandwidths must be catered for. These limiting bandwidths should be taken into consideration in the design of the RF system. An approximation of the phase error in the cavity tuning caused by bandwidth limitations is given by the response to a ramped frequency input.

Phase error

$$\Phi_e = K\left[t + \tau\left(e^{-t/\tau} - 1\right)\right] \tag{41}$$

and Φ_e steady state will be

$$\Phi_e = K\tau \quad , \tag{42}$$

where $1/(2\pi\tau)$ is the tuning system bandwidth, and Φ_e the cavity voltage phase error.

The constant K describes the relation to the acceleration ramp

$$K = \frac{df}{dt}\ \frac{\pi Q}{2 f_0}\ . \tag{43}$$

There (df/dt) is the maximum rate of chance of RF frequency, and f_0 the frequency at which (df/dt) is a maximum.

5.3 Type 2 cavity

Figure 16 shows the main elements of an FNL Booster Synchrotron RF drift tube accelerating system. Each resonator consists of two resonators joined in parallel at the high-voltage end. The bias field is parallel to the RF field and is in the same direction through each resonator half. The RF currents are in opposite directions in each half. The RF voltage induced in a bias winding is zero. There are three resonators in parallel around the drift tube (only one is shown here). The three parallel resonators reduce the minimum inductance value at high bias field to a lower value than could be achieved with a single resonator. The series inductance is minimized between the resonators and the drift tube, and the power tetrode anode and the drift tube. The resonators each have 10 turns for the bias field windings.

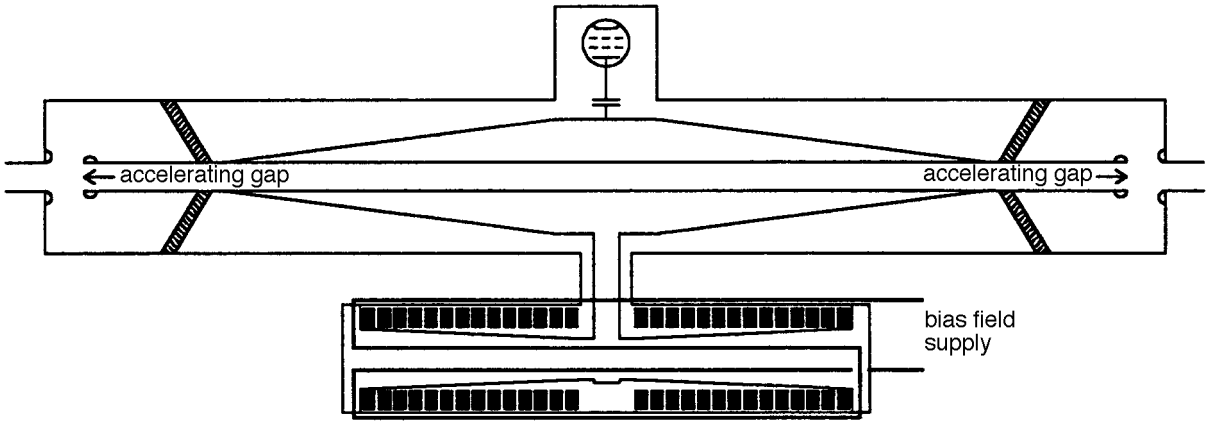
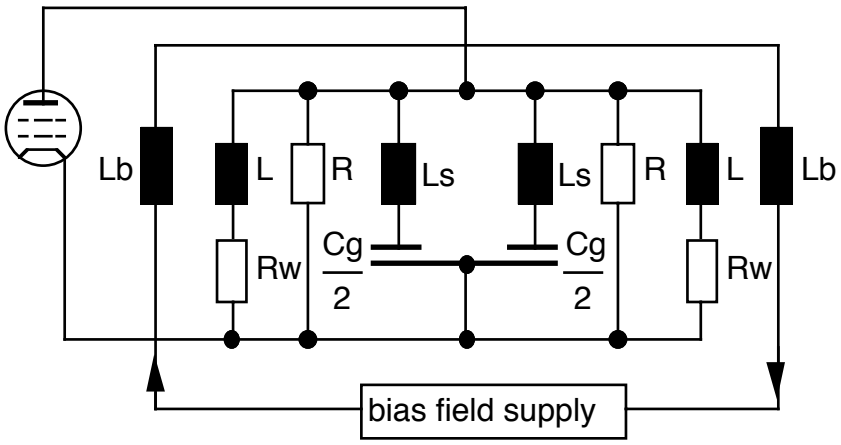


Fig. 16: FNL Booster drift tube system

The equivalent circuit in Fig. 17 is similar to that of the Type 1 system, if a single-turn bias field winding is used. No RF bypass capacitor is used. The inner conductor of the resonator is resistive and the turn formed by the inner and outer resonator conductors is driven by the bias field supply. For multiple-turn bias field windings the induced current is proportionally reduced. The bandwidth limit set by the cavity load for the tuning system is determined only by $f_0/(2Q)$.



C_g = drift tube Capacity

R_w = wall resistance inside co-ax in resonator

R = ferrite loss in resonator

L_b = bias winding inductance

L = inductance of resonator

L_s = series inductance to drift tube end

Fig. 17: Drift tube system equivalent circuit

5.4 Type 3 cavity (PS Booster, COSY, SATURNE)

In this type of cavity (Fig. 18), which is typified by the former CERN PS Booster cavity, a virtual RF ground exists in the middle of the accelerating gap and the voltages on the resonators are in anti-phase. The bias field is parallel to the RF field but in opposite directions in each half of the cavity. The RF currents are in the same direction in each half of the cavity. The RF amplifier can have a single-ended or differential output, as there is strong coupling between the two cavities via the bias field winding. The bias field windings are arranged as a figure of eight to produce balanced RF voltages on each winding and no RF voltage on the ferrite bias source. The equivalent circuit is shown in Fig. 19 with the circuit as seen by the bias supply in Fig. 20. More information about the status of the CERN Booster cavities operating at different harmonics can be found in Ref. [11]. A view of the installation of the ferrite cavity of COSY is given in Fig. 21. Figure 22 shows the arrangement of ferrites, tuning loop, and cooling discs near the acceleration gap.

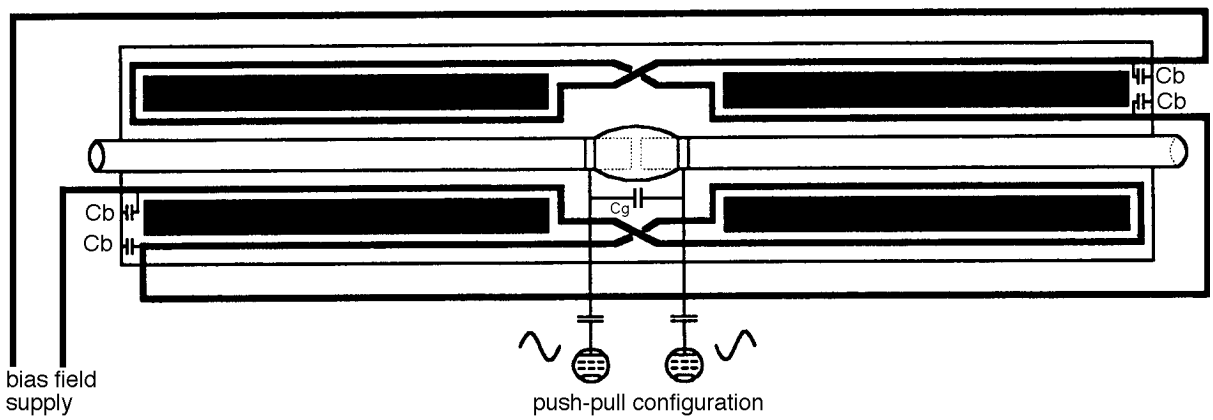


Fig. 18: CERN PS Booster/COSY/SATURNE cavity

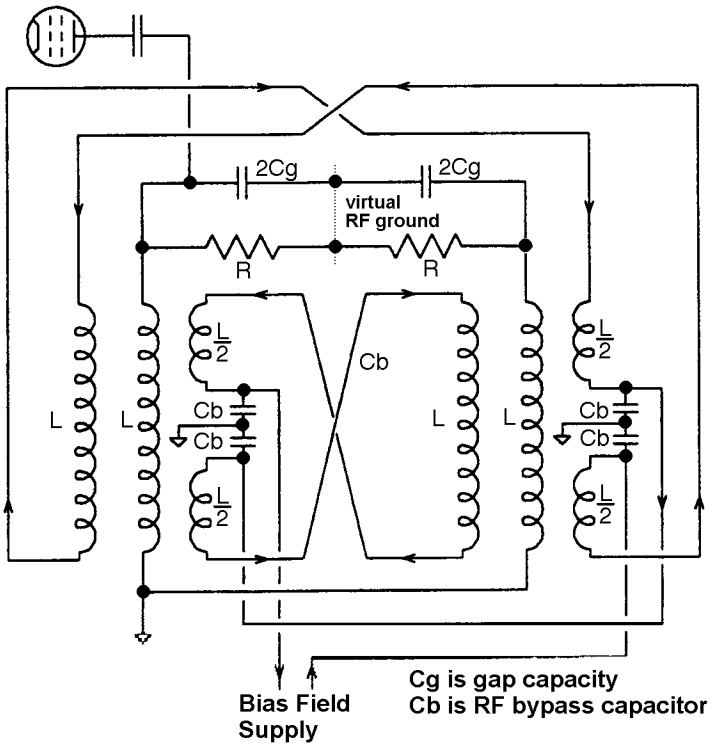


Fig. 19: CERN PS Booster cavity equivalent circuit

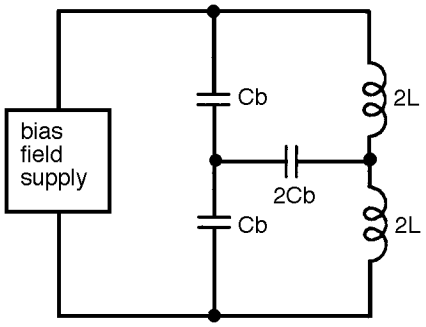


Fig. 20: CERN PS Booster equivalent load for bias field supply

The bandwidth limits set by the cavity on the bias field tuning system are as for Type 1, $f_0 / (2Q)$ and

$$f = \frac{1}{2\pi\sqrt{2LC_b}} \qquad K = \frac{df}{dt}\,\frac{\pi}{2}\,\frac{Q}{f_0} \qquad \Phi_e = K\tau \ . \tag{44}$$

Example for a COSY cavity, shown in Figs. 21 and 22:

(df / dt) = 1.6 kHz / ms at f_0 = 655 kHz,

Q = 10, τ = 1ms gives **K = 0.038 / ms, Φ_e = 2.2°**

As RF voltage is carried by the bias field, winding care must be taken to avoid voltage breakdown from the bias conductors and to avoid RF resonances involving the bias conductors. This becomes increasingly difficult with higher numbers of turns.

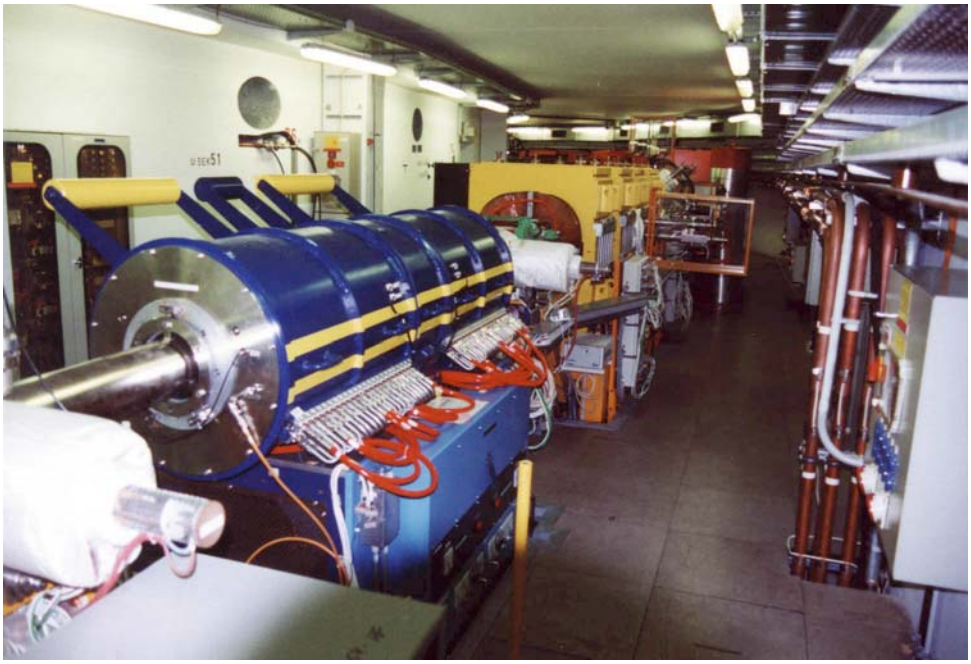


Fig. 21: (h = 1) ferrite cavity of COSY

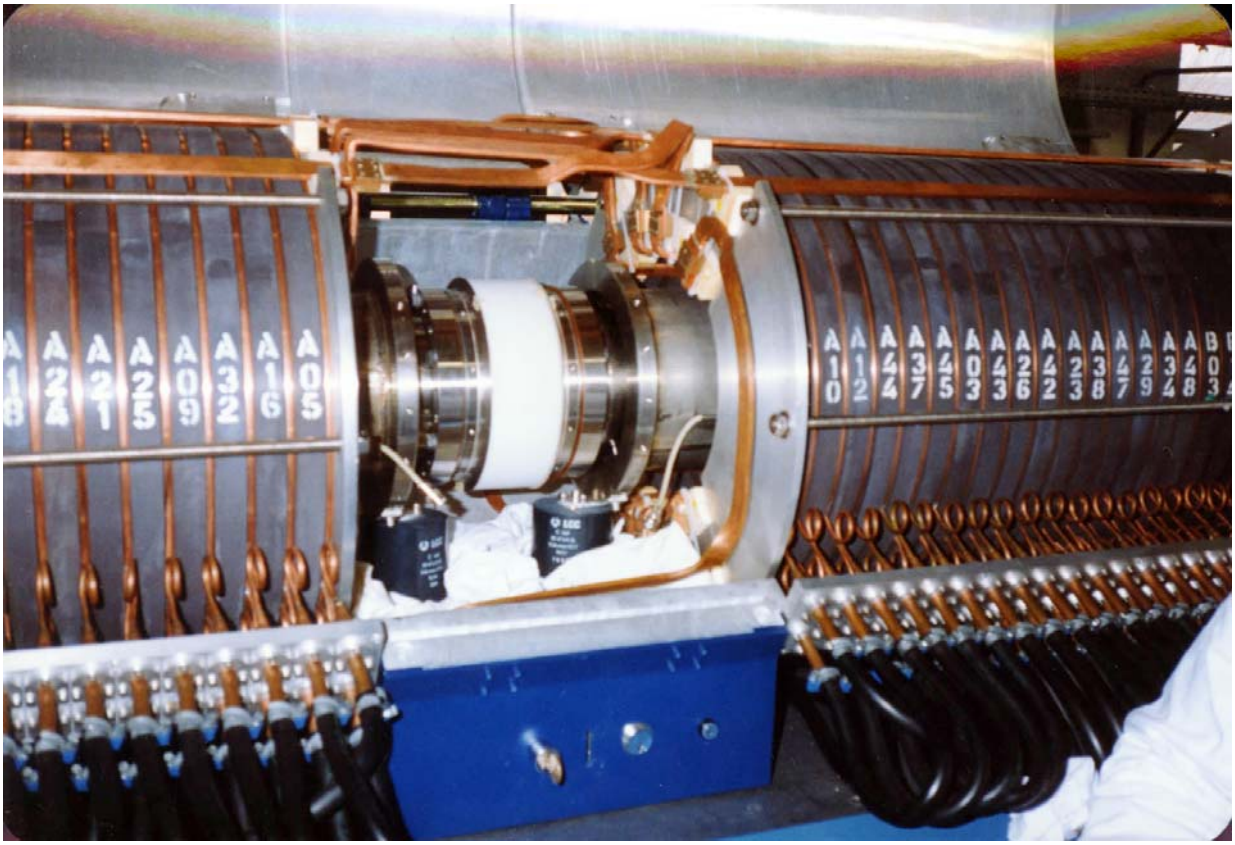


Fig. 22: Arrangement of gap, 8C12 ferrites, and cooling plates in COSY (h = 1) cavity

5.5 Type 4 cavity

This type of cavity (CERN LEAR-type), shown in Fig. 23, is asymmetric and uses a single gap. With this arrangement of bias winding, the induced RF voltage on the winding is constrained to be zero at the bias field supply feed point by splitting the winding at the mid-voltage point on the cavity. The two windings are then crossed in a figure-of-eight mode as in Type 3. In the equivalent circuit (shown in Fig. 24), the parallel bias produces a field that changes direction at the crossover point of the bias winding.

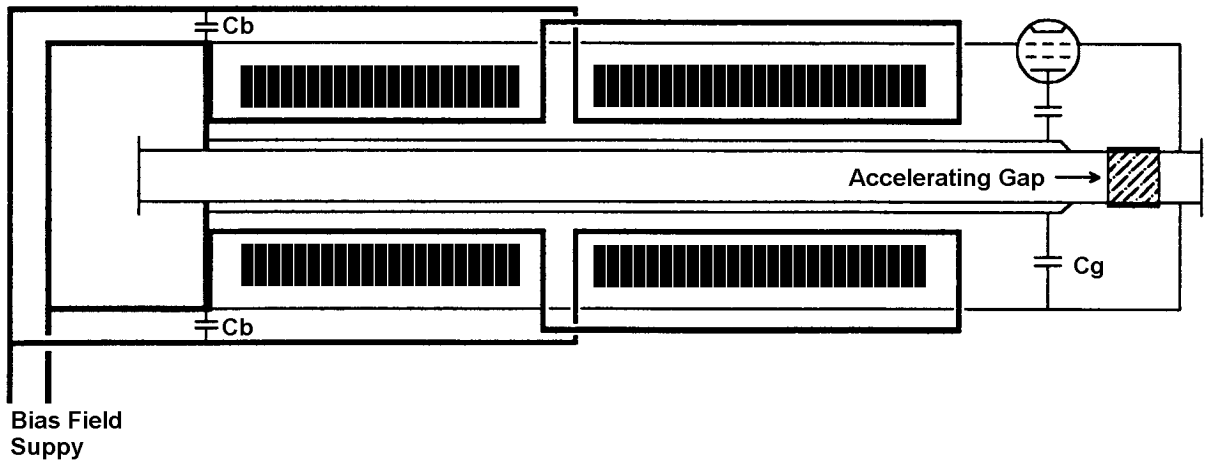


Fig. 23: CERN LEAR-type cavity

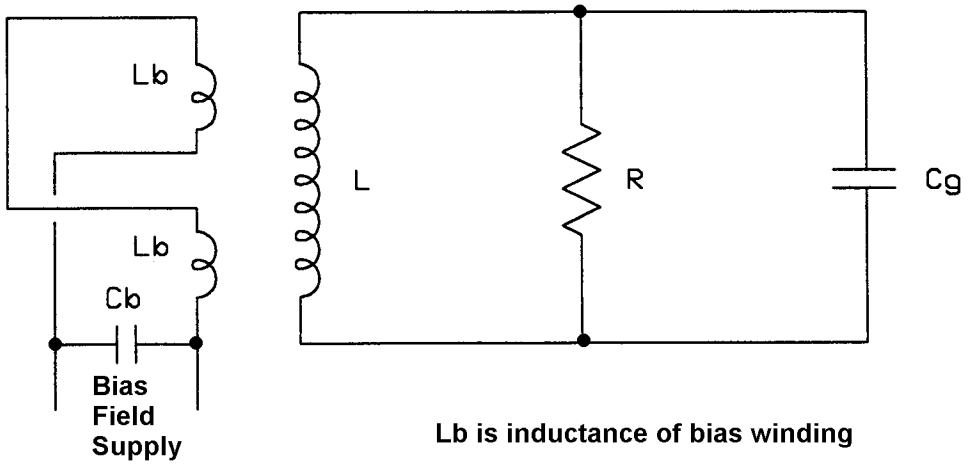


Fig. 24: CERN LEAR-type equivalent circuit

5.6 Type 5 cavity

In this cavity another solution for the single-gap cavity is obtained by using a perpendicular bias field on the ferrite. This is the Los Alamos/TRIUMF type of cavity (Fig. 25). Perpendicular bias and the use of microwave ferrite resulted in a compact cavity with high Q in the 46–61 MHz frequency range. The high Q allows a higher gap voltage and reduces the number of cavities required in the accelerator. The power tube is capacitively coupled to the cavity, whereas the coupling in Types 1 to 4 is a combination of inductive and capacitive coupling.

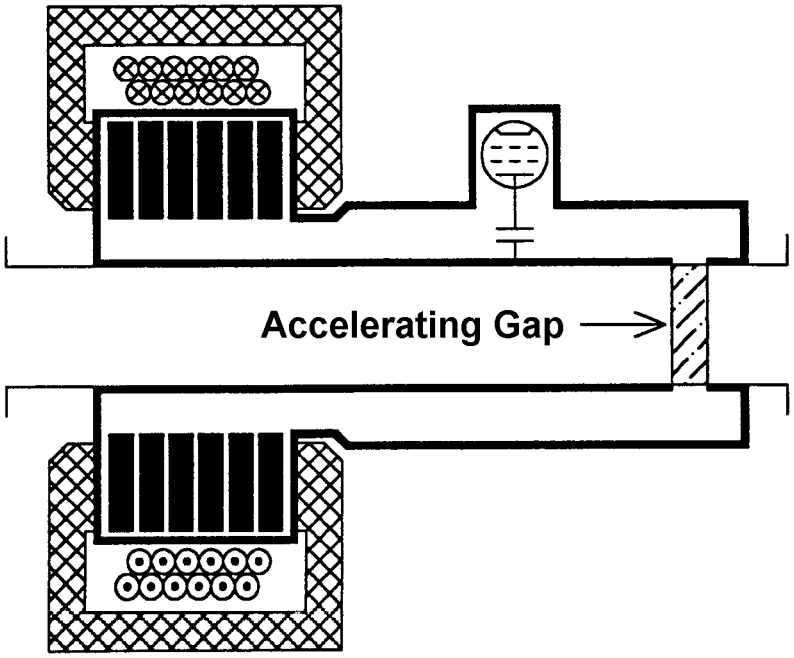


Fig. 25: Los Alamos/TRIUMF cavity

The ferrite is located at the short circuit end of the cavity and the bias field is applied by a solenoidal winding. The cavity outer conductor and short circuit require cuts to reduce the eddy current heating caused when the bias field is changed. The eddy current load is supplied by the bias field supply. This arrangement precludes the use of metallic cooling plates between the ferrite toroids. Furthermore, because a perpendicular bias field is used with low values of μ, then a high value of bias field must be supplied, and this results in some stray field on the beam axis. Figure 26 shows the equivalent circuit. Another possible way of creating a perpendicular field involves the use of a quadrupolar field. An example of this [9] is shown in Fig. 27, a cavity built at the Indiana University Cyclotron Facility. The principle of such a cavity was first demonstrated in 1989 at MPI, Heidelberg.

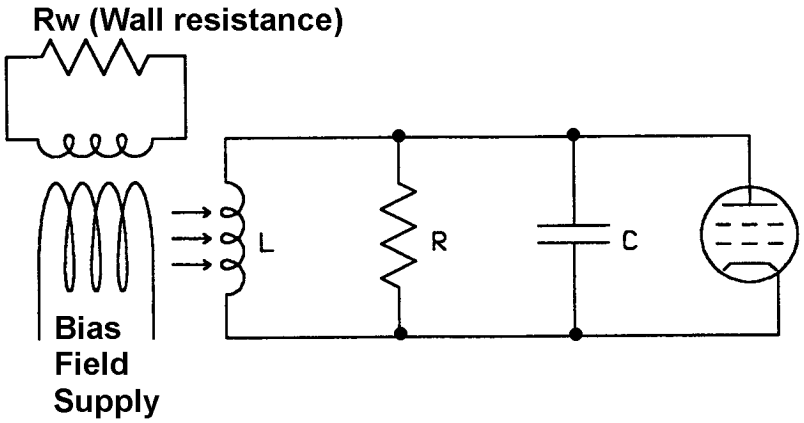


Fig. 26: Los Alamos/TRIUMF cavity equivalent circuit

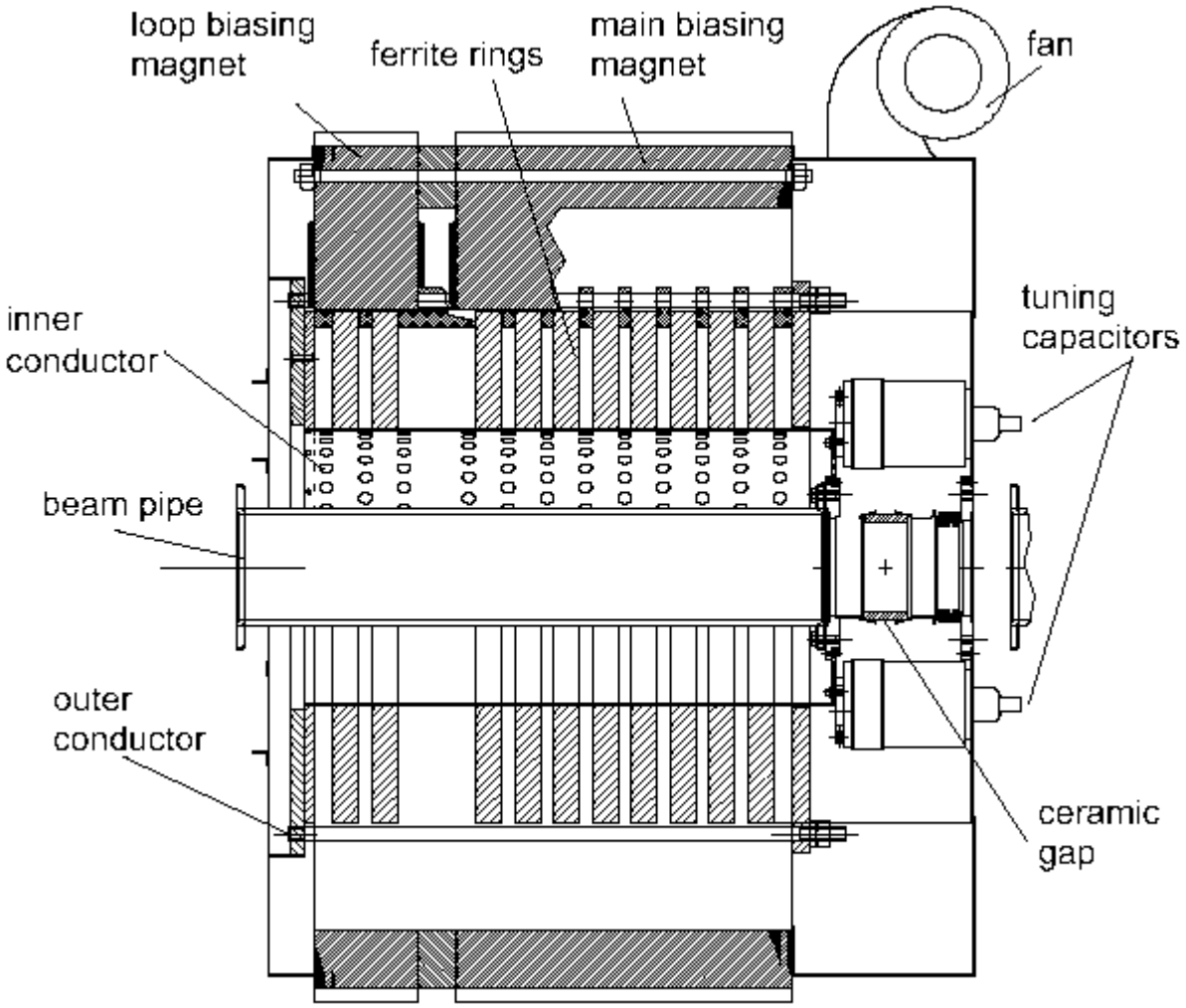


Fig. 27: RF cavity of CIS, Bloomington, Indiana, tuned with a quadrupole field

6. HIGH-ORDER MODES

Unwanted high-order modes in ferrite cavities are only briefly mentioned here. The first resonant mode above the fundamental mode will occur at such a frequency that the resonator appears to be 180° long. As resonators are, in general, designed to be less than 26° long in order to maintain a low variation in B_{RF} in the ferrite, then the next mode is likely to be more than five times the frequency of the fundamental. Additional resonant modes are likely to be caused by multi-turn bias field windings and coaxial coupling arms.

7. BROADBAND CAVITIES

Traditionally, radio-frequency accelerating structures in accelerator rings are made from one or more cavities. Consequently, these structures can transmit only narrow-band frequency signals to the particle beam. While such narrow-band structures require dynamic tuning during the acceleration cycle of a synchrotron, they limit noise transfer to the particle beam to a narrow frequency regime. Moreover, the generation and amplification of harmonic RF signals is a standard procedure, and resonant cavities offer considerable efficiency with respect to power utilization. These are some of the reasons why resonant cavity structures have been used throughout and are used (with very few exceptions) as accelerating structures in accelerators.

Increasing demands on specific accelerating voltage waveforms have led to the installation of more complex accelerating systems, whether they use a set of cavities to run at harmonics [11, 12], or a broadband electrostatic structure [13].

Such specific demands on accelerating voltage waveforms of more than one frequency component, or even strictly non-harmonic waveforms, such as the saw-tooth type, may stem from the need to tailor particle bunch shapes or dynamics, e.g. for the passage of points of instability, or the synchronized transfer to another accelerator [12–14].

In addition to the accelerating structure and the associated power amplifiers, the low-level signal generation and dynamic control has to comply with these specific voltage waveform demands. Because of the broadband feature, signal generators must have unprecedented purity or lack of spurious frequency emissions at high frequency variability.

At CRYRING, a small ring for many different ion species, experiments were performed with a double-gap drift-tube structure without material [15]. A simplified layout is shown in Fig. 28. After the particles cross the first gap, they fly through a drift tube and then cross the second gap. The particles gain an amount of energy determined by the voltage on the central drift tube and the change of the signal during the flight through the drift tube. This structure is broadband (10 kHz to 1.5 MHz) without tuning, so one can apply a waveform that is a repeated sequence of parabolas (Fig. 29), and the particles will experience an energy gain, as if they had seen a saw-tooth signal at a single gap.

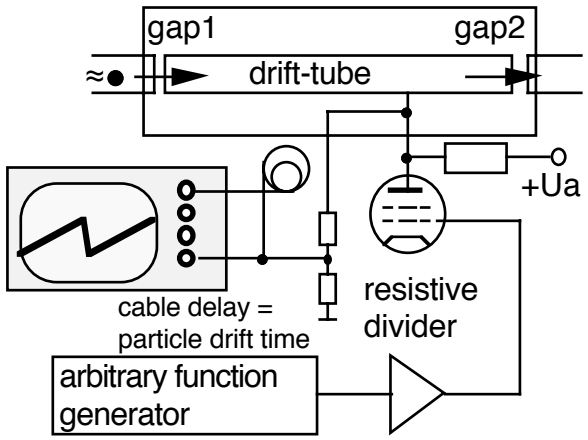


Fig. 28: Drift tube structure at CRYRING

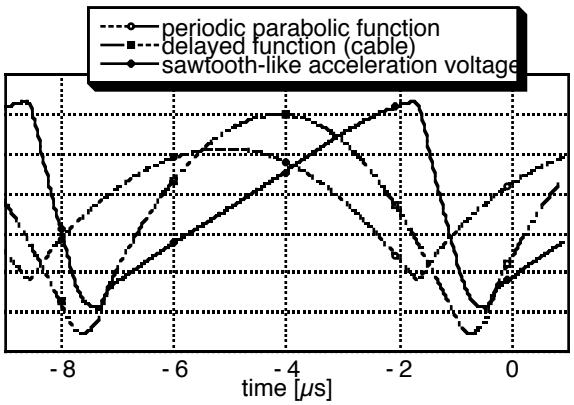


Fig. 29: Application of a parabolic waveform

Another approach to broadband operation is to apply a load resistor in parallel to the gap of a ferrite cavity [16]. This reduces the external quality of the resonator, so one therefore has to balance power efficiency with broadband behaviour.

7.1 VitroVac cavity

Instead of ferrite, one can use amorphous magnetic materials that are wound like magnetic tapes to obtain a toroid shape. A cavity with VitroVac 6025F was constructed at SATURNE for MIMAS as proof-of-principle [17]. A cut-away drawing of this cavity is shown in Fig. 30 and the system layout is described in Fig. 31. The hysteresis curve of the material 6025F is displayed in Fig. 32.

This cavity still needs tuning, but the current is lower compared to a ferrite-filled cavity. The impedance as a function of the tuning current is displayed in Fig. 33. Such a cavity allows almost arbitrary waveforms if enough RF power is available for the higher harmonics, which have to be adjusted in amplitude and phase, is available. Examples are shown in Figs. 34(a) and 34(b).

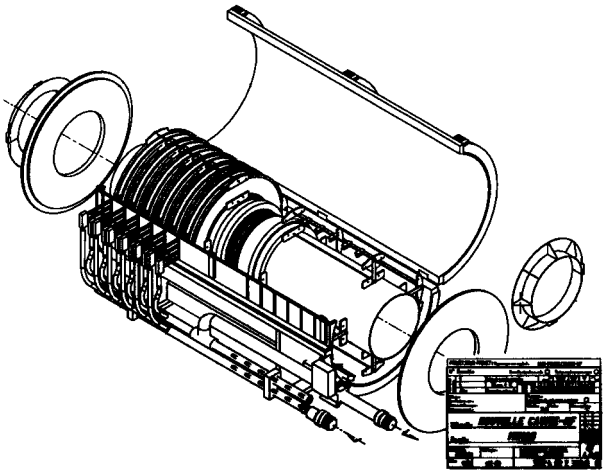


Fig. 30: Cut-away drawing of a VitroVac cavity, showing inner and outer conductor, toroids, and cooling and tuning bars

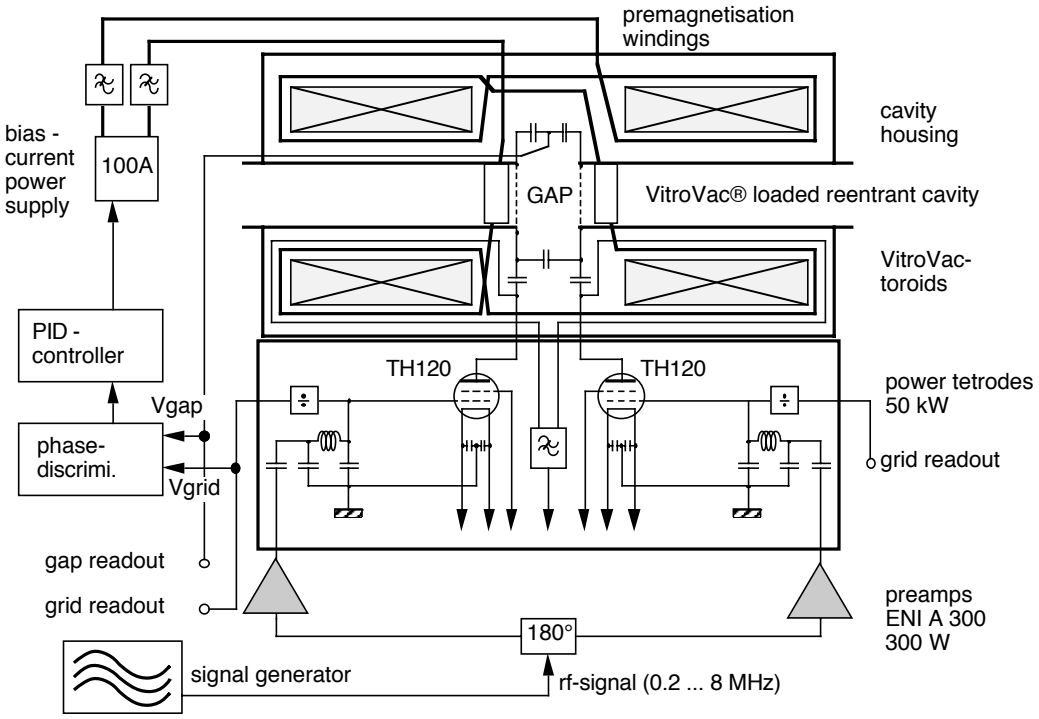


Fig. 31: Layout of the RF system of the VitroVac cavity for operation from 0.2 to 8 MHz

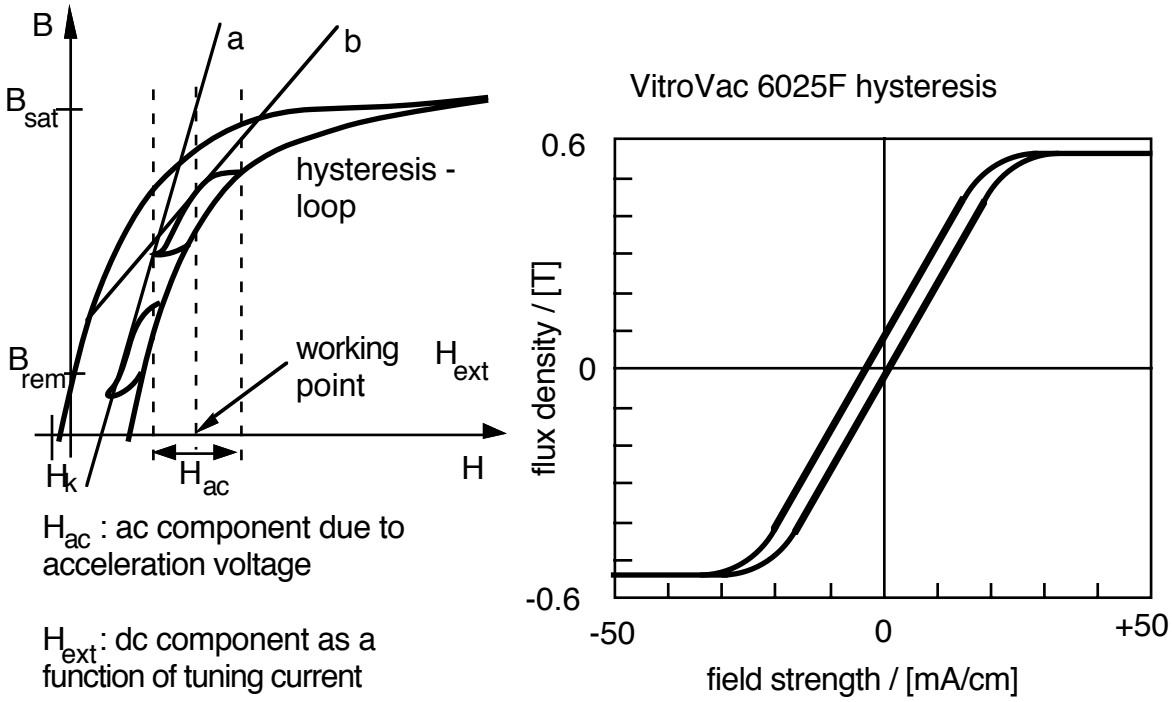


Fig. 32: Hysteresis curve for VitroVac 6025F

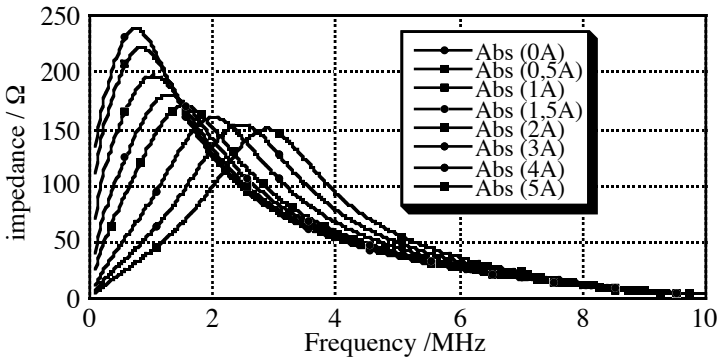


Fig. 33: Absolute value of impedance as a function of frequency and tuning current

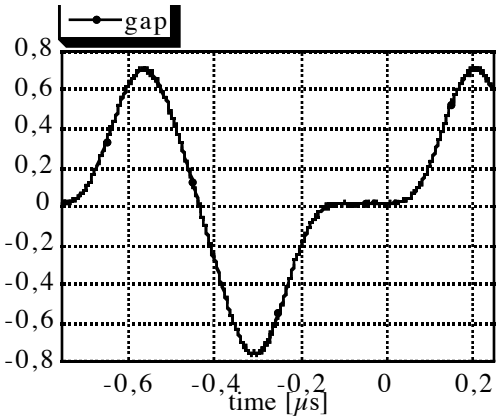


Fig. 34(a): 'Lawn-chair voltage', 0.2 kV; 200 ns per div.

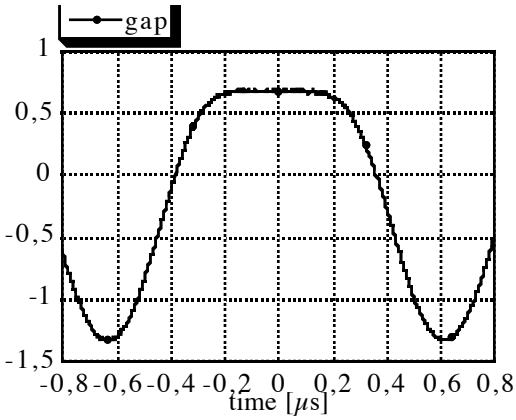


Fig. 32(b): 'Flattened-top voltage', 0.5 kV; 200 ns per div.

The quality of the VitroVac cavity is much lower than the quality of a ferrite cavity. This is shown in Fig. 35, where a 180° phase jump in 20 ns is applied to the cavity. It takes less than three periods to reach a steady state. The same phase jump applied to the COSY ferrite cavity (Fig. 36) shows a much stronger ringing due to the higher quality. Table 2, above, compares some parameters of the VitroVac 6025F-filled cavity with a ferrite 8C12-filled cavity.

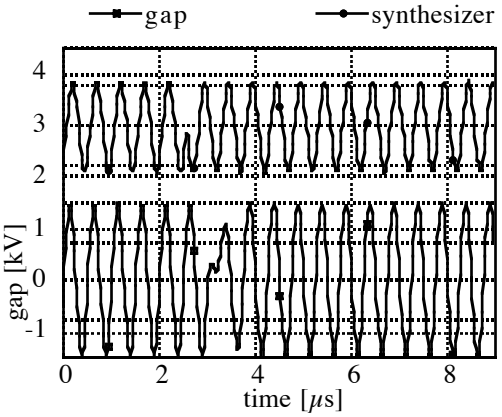


Fig. 35: A rapid 180˚phase jump applied to the VitroVac cavity. Upper trace: synthesizer; lower trace: gap-signal.

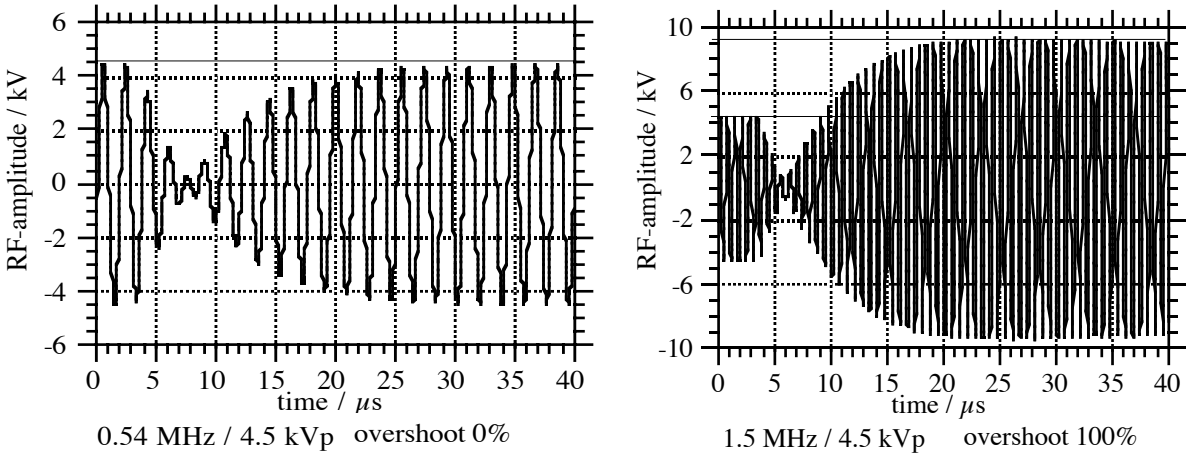


Fig. 36: A rapid 180˚ phase jump applied to an 8C12 ferrite-filled cavity at different frequencies

7.2 Magnetic alloy-filled cavities

In studies for the Japanese Hadron Facility (JHF) [18, 19], NiZn cores were compared with soft magnetic cores made of FineMet. This amorphous material has a very high Curie temperature of about 600°C and allows operation in excess of 100°C. Cores with 67 cm outer diameter and 2.5 cm thickness present an impedance of about 83 Ω. This results in 10 kV/m at 3.4 MHz with 24 cores driven by a 30 kW amplifier.

The broadband behaviour of the impedance of cavities filled with FineMet was clarified in a comparison between cavities for the Barrier Bucket Experiment at AGS, Brookhaven [20, 21]. In the case of ferrite cavities [21], there were eight cells with six cores per cell (one cell reaches 10 kV). In total 80 kV were generated at 3.7 MHz with a 600 kW TH 558 single tetrode amplifier. With the magnetic alloy (MA)-filled cavity [20], a total voltage of 40 kV was obtained with a 60 kW push–pull amplifier.

A cavity filled with high-permeability magnetic alloy was proposed for a high-intensity proton synchrotron [22]. A prototype of only 40 cm in length reached a voltage of 20 kV, which is a gradient of

50 kV/m. Only six cores were used. These were placed directly in a water bath without cooling discs. The prototype and the measured impedance of a half-cavity are plotted in Figs. 37(a) and 37(b). Figure 38 shows the ($\mu'Qf$) product for SY-2 ferrite compared to FineMet FT3 cores. The MA cores exhibit constant shunt impedance up to 2 kG.

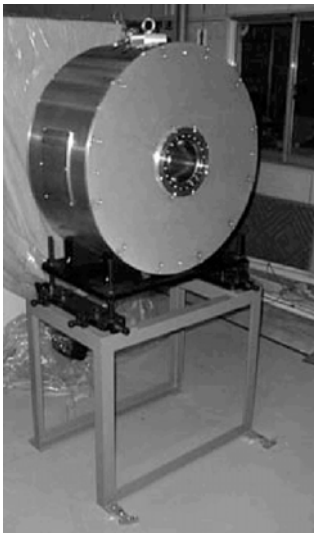


Fig. 37(a): High gradient prototype cavity [23]

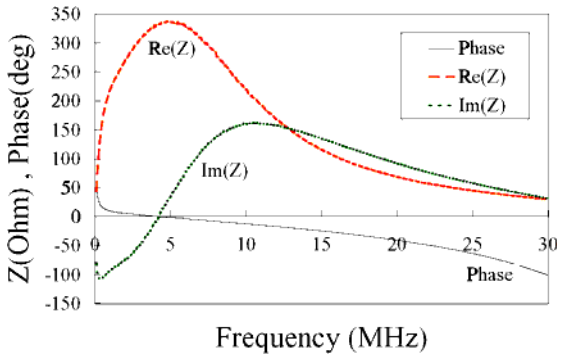


Fig. 37(b): Impedance curve of a half-cavity [22]

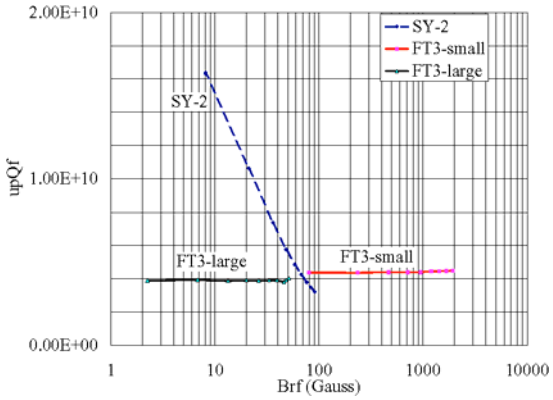


Fig. 38: Dependence of ($\mu'Qf$) values on the RF magnetic field strength (B_{rf}) for Ni Zn ferrite (SY-2) and magnetic alloy (FineMet FT3) cores. FT3-small and -large are cores of 570 mm O.D. and 70 mm O.D. respectively [23].

According to Ref. [23], the characteristics of MA cores can be summarized as follows:

- The ($\mu'Qf$) product of MA cores does not depend on the flux density. Among them, FineMet cores have a large saturation flux density and can be used at the high RF flux density of 2 kG.
- The intrinsic Q-value of the MA core is 0.6–1. It is suitable for a wide frequency sweep without tuning system, barrier bucket, and multiharmonic RF acceleration.
- The permeability of the MA core is about 10 times higher than that of ferrite. Although the Q-value is low, the ($\mu'Qf$) product is high enough to be used for an RF cavity.
- The MA core has a very high Curie temperature of 570°C. The characteristics of the core remain constant at temperatures above 100°C.
- The Q-value of the core can be increased to more than 10 by using the cut core technique. The R/Q-value of the cavity is a variable without changing the shunt impedance.

– The core consists of thin MA tape coated against electrical breakdown by a thin silica insulator. It is possible to manufacture large cores of about 100 cm diameter.

The operation of a High Gradient Cavity (HGC) at HIMAC is described in Ref. [24]. With 60 kW of RF power, a maximum voltage of 4 kV is obtained with a cavity only 40 cm in length. The frequency range is 1–8 MHz; the maximum impedance is 400 Ω between 2–3 MHz.

Another MA-loaded cavity proved the feasibility of barrier-bucket voltages [25]. Because of the high beam current (8 x 10^{12} to 3 x 10^{13}) protons and the broadband behaviour of the cavity, a feedforward system is necessary. The feedforward system has to compensate for the effect whereby the cavity 'talks' to the beam at several harmonics [26]. Some examples of non-sinusoidal waveforms applied to MA-loaded cavities are given in Ref. [27].

7.3 VitroPerm cavity

At COSY, we want to replace the 2.1 m long ferrite cavity (using Philips 8C12) with a much shorter broadband cavity, to gain space for insertions that preserve polarization during acceleration [28]. We have therefore developed the cavity shown in Fig. 39, filled with VitroPerm [29], and which is only about 80 cm long and allows the use of non-sinusoidal waveforms, mentioned above.

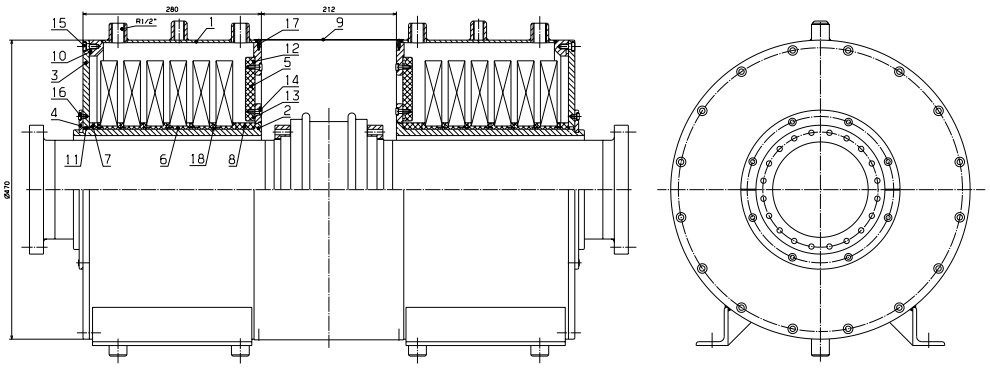


Fig. 39: Layout of a broadband cavity filled with VitroPerm

The parameters of different materials and different cooling media were measured with a small test cavity, shown in Fig. 40 [30]. The measured impedance of a group of three cores as a function of medium is plotted in Fig. 41.

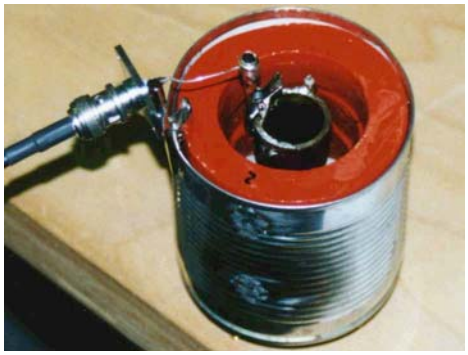


Fig. 40: Small test cavity filled with three cores

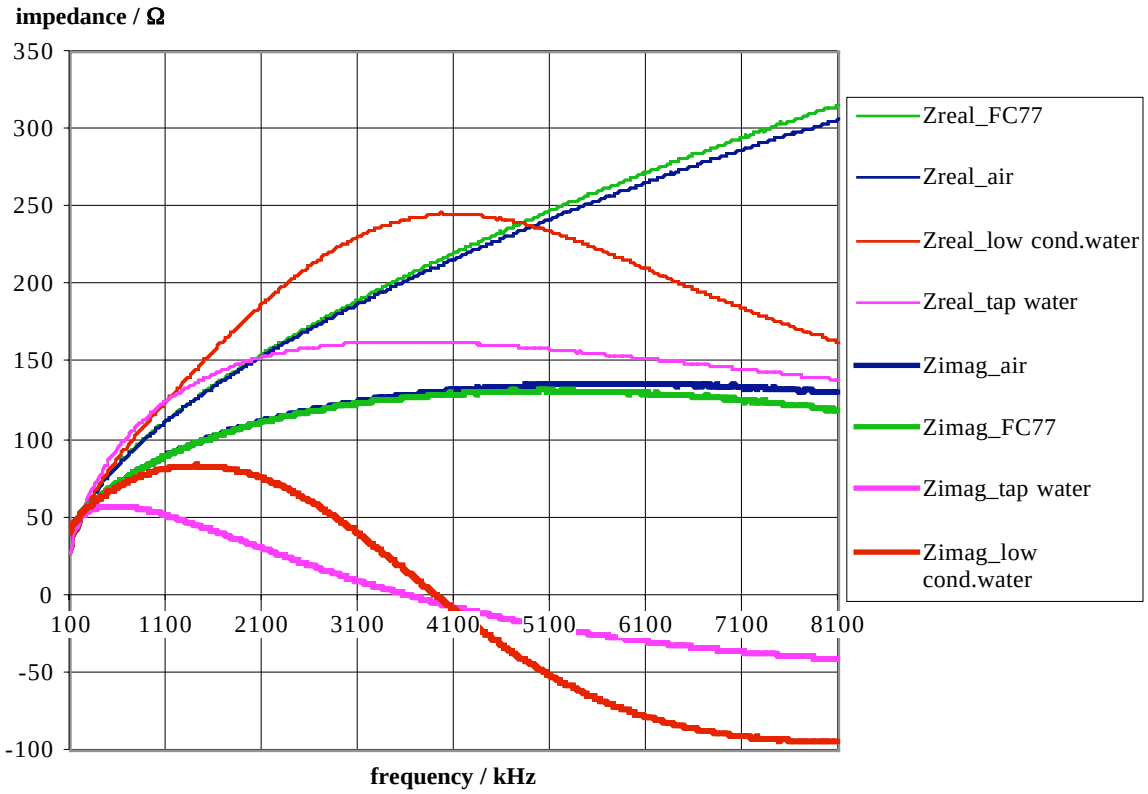


Fig. 41: Impedance of three VitroPerm sample cores as a function of cooling medium

The cooling medium Fluorinert FC_77 [31] with its dielectric constant ε_r = 1.86 leaves the impedance almost untouched compared to air. When using water as a coolant, it is important to use low-conductivity water instead of tap water in order to obtain a high impedance value. The layout of the combination of the VitroPerm-loaded cavity and the tube amplifier is sketched in Fig. 42. The cavity is connected to the amplifier with 200 Ω coaxial transmission lines, shown in Fig. 43. For lower voltages in the kV range, it is sufficient to drive the cavity with a total of four 500 W transistor amplifiers (two per side), connected to (1:4) impedance transformers.

The impedance of the installed air-filled cavity is plotted in Fig. 44. The impedance of the cavity filled with low conductivity water is displayed in Fig. 45. The measurement values for the left and the right half of the cavity are not equal because the value depends on the distribution of the toroids between the left and right sides. Between 500 kHz and 1.6 MHz (the operatimg range of COSY) the total impedance of the cavity is over 500 Ω. An example of the voltages that can be generated with such a cavity, and the reaction of a proton beam at 300 MeV/*c*, are shown in Fig. 46. As expected, this kind of cavity allows modifications of the longitudinal phase space [32].

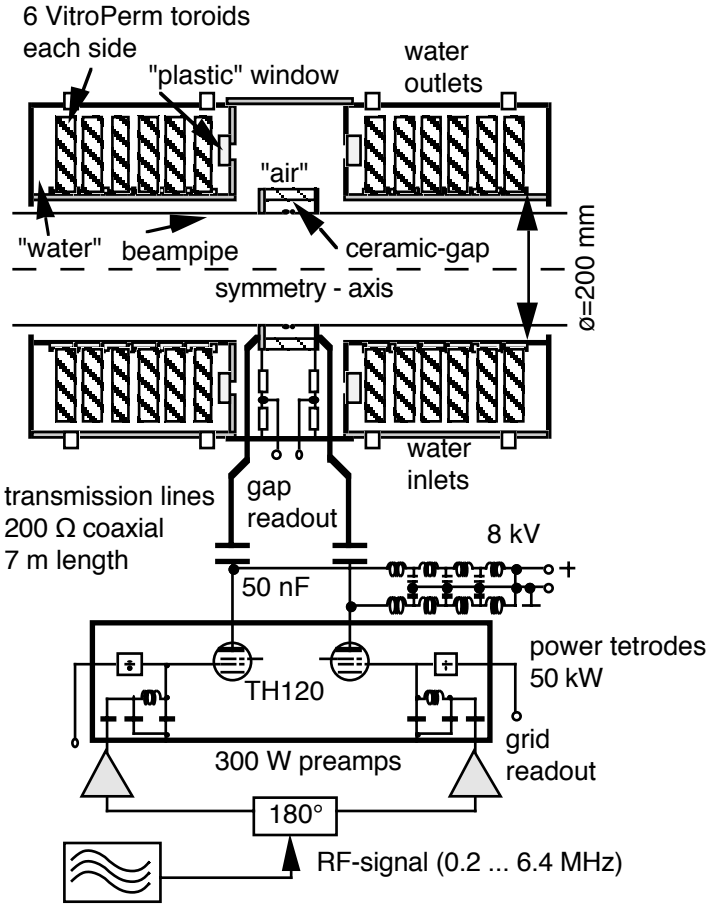


Fig. 42: Combination of VitroPerm cavity and tube amplifier

Fig. 43: Power connection with transmission lines

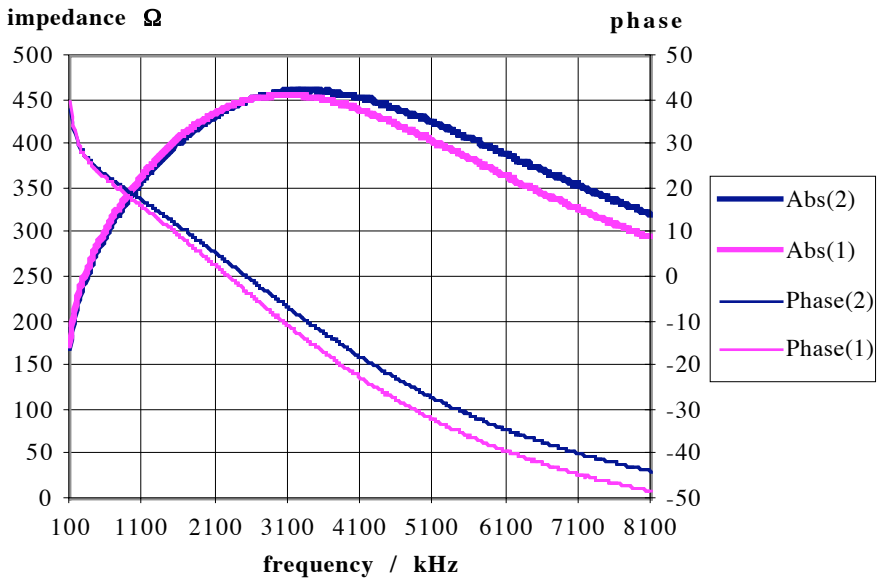


Fig. 44: Impedance of air-filled VitroPerm cavity

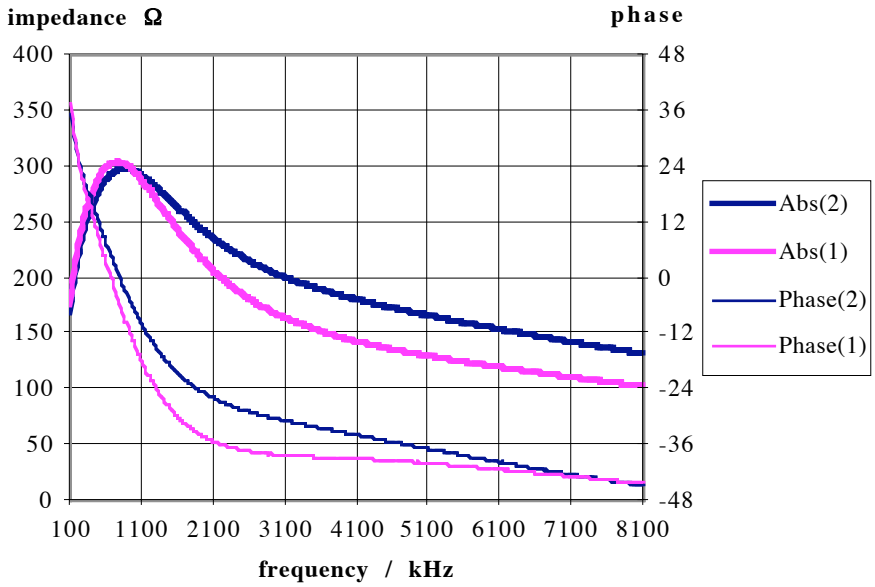


Fig. 45: Impedance of VitroPerm cavity cooled with low-conductivity water

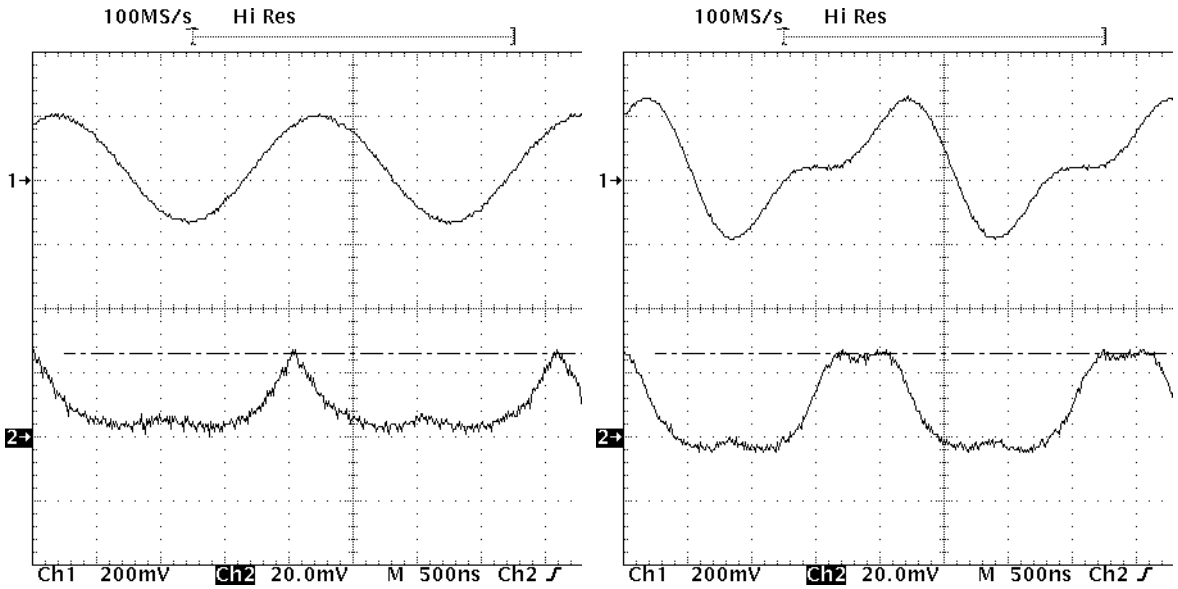


Fig. 46: RF voltage (upper trace) and longitudinal beam signal (lower trace) at COSY injection (500 kHz) with harmonic (h = 1) only (left) and combined harmonics (h = 1) and (h = 2) (right).

REFERENCES

[1] I.S.K. Gardner, Ferrite dominated cavities, 'CERN Accelerator School: RF engineering for particle accelerators', Exeter College, Oxford, 1991, CERN 92–03, CERN, Geneva, 1992.

[2] J. Roberts, *High Frequency Applications of Ferrites* (English Universities Press, 1960).

[3] F.G. Brockman, H. van der Heide and M.W. Louwerse, 'Ferroxcube for proton synchrotrons', Philips Tech. Rev. **30** (1969).

[4] 8C12 Material Grade Specification, Philips Components, 8C12.pdf, December 1999.

[5] P.P. Lombardini, R.R. Schwartz, R.J. Doviak, 'Evaluation of ferrite materials for possible application in the Princeton Pennsylvania 3 BeV proton synchrotron', Moore School Report, No 58-05 (1957).

[6] R.G. Bendall, R.A. Church, 'Ferrite measurements for SNS accelerating cavities', Rutherford Laboratory Report RL-79-04 (1979).

[7] J.E. Griffin, G. Nicholls, 'A review of some dynamic loss properties of NiZn accelerator RF system ferrite', *IEEE Trans. Nucl. Sci.* **NS-26** (1979).

[8] W.R. Smythe, 'Reducing ferrite power loss by bias field rotation', *IEEE Trans. Nucl. Sci.* **NS-30** (1983).

[9] A. Pei, S. Anderson, D. Jenner, X. Kang, S.Y. Lee and D. McCammon, Proc. PAC 97, Vancouver, 1997, p. 2968.

[10] P. Barratt *et al.,* Proc. EPAC-90, Nice, 1990, p. 949.

[11] A. Krusche, M. Paoluzzi, Proc. EPAC 98, Stockholm, 1998, p. 1782.

[12] J.M. Baillod, L. Magnani, G. Nassibian, F. Pedersen, W. Weissflog, 'A second harmonic (6–16 MHz) RF system with feedback-reduced gap impedance for accelerating flat-topped bunches in the CERN PS Booster', Particle Accelerator Conference on Accelerator Engineering and Technology, Santa Fe, 1983 (CERN Report CERN/PS/BR 83–17).

[13] K. Abrahamsson, G. Andler, and C.B. Bigham, A drift tube accelerating structure for CRYRING, *Nucl. Instr. Meth.* **B31** (1988) 475.

[14] J.A. MacLachlan and J.E. Griffin, Proc. IEEE Particle Accelerator Conference, San Francisco, 1991, p. 2826.

[15] K. Abrahamsson *et al.,* Performance of CRYRING with improved electron cooling, Proc. EPAC 94, London, 1994, p. 381.

[16] J.E. Dey and D.W. Wildman, Proc. PAC 99, New York, 1999, p. 869.

[17] P. Ausset, G. Charruau, D. DeMenezes, F.-J. Etzkorn, C. Fougeron, H. Meuth, S. Papureanu, A. Schnase, Proc. EPAC 94, London, 1994, p. 2128.

[18] M. Fujieda, Y. Mori, H. Nakayama, C. Ohmori, S. Sawada, Y. Tanabey, E. Ezura, A. Takagi, M. Toda, M. Yoshii, T. Tanabe, T. Uesugi, Proc. PAC 97, Vancouver, 1997, p. 2992.

[19] C. Ohmori, M. Fujieda, S. Machida, Y. Mori, H. Nakayama, K. Saito, S. Sawada, Y. Tanabe, M. Yamamoto, E. Ezura, A. Takagi, M. Toda, M. Yoshii, T. Tanabe, T. Uesugi, Proc. PAC 97, Vancouver, 1997, p. 2995.

[20] M. Fujieda, S. Machida, Y. Mori, H. Nakayama, C. Ohmori, Y. Sato, Y. Tanabe, T. Uesugi, M. Yamamoto, A. Takagi, M. Toda, M. Yoshii, Y. Iwashita, Proc. EPAC 98, Stockholm, 1998, p. 1796.

[21] M. Yoshii, J.M. Brennan, J. Brodowski, M. Meth, K.A. Rogers, R. Spitz, A. Zaltsman, Proc. EPAC 98, Stockholm, 1998, p. 1829.

[22] Y. Mori, M. Fujieda, K. Koba, H. Nakayama, C. Ohmori, K. Saito, Y. Satoh, Y. Tanabe, A. Takagi, Y. Toda, T. Uesugi, M. Yamamoto, T. Yan, M. Yoshii, Proc. EPAC 98, Stockholm, 1998, p. 299.

[23] C. Ohmori, E. Ezura, M. Fujieda, Y. Mori, R. Muramatsu, H. Nakayama, Y. Sato, A. Takagi, M. Toda, T. Uesugi, M. Yamamoto, M. Yoshii, M. Kanazawa, K. Noda, Proc. PAC 99, New York, 1999, p. 413.

[24] R. Muramatsu, M. Fujieda, Y. Mori, H. Nakayama, C. Ohmori, Y. Sato, A. Takagi, T. Uesugi, M. Yamamoto, M. Yoshii, M. Kanazawa, K. Noda, Proc. PAC 99, New York, 1999, p. 798.

[25] M. Fujieda, Y. Iwashita, A. Noda, Y. Mori, C. Ohmori, Y. Sato, M. Yoshii, M. Blaskiewicz, J.M. Brennan, T. Roser, K.S. Smith, R. Spitz, A. Zaltsmann, Proc. PAC 99, New York, 1999, p. 857.

[26] M. Yamamoto, M. Fujieda, Y. Hashimoto, Y. Mori, R. Muramatsu, C. Ohmori, Y. Sato, A. Takagi, T. Uesugi, M. Yoshii, Proc. PAC 99, New York, 1999, p. 860.

[27] M. Yamamoto, M. Fujieda, Y. Mori, R. Muramatsu, C. Ohmori, Y. Sato, A. Takagi, T. Uesugi, M. Yoshii, M. Kanazawa, K. Noda, Proc. PAC 99, New York, 1999, p. 863.

[28] M. Böhnke, F.-J. Etzkorn, R. Maier, U. Rindfleisch, A. Schnase, H. Stockhorst, Proc. PAC 99, New York, 1999, p. 851.

[29] Vacuumschmelze Hanau GmbH, Kerne und Bauelemente, Datenbuch 1998.

[30] A. Schnase, M. Böhnke, F.-J. Etzkorn, U. Rindfleisch, H. Stockhorst, *IKP Annual Report 1999*, Jül-3744 (Forschungszentrum Jülich GmbH, 2000), p. 150.

[31] 3M, Fluorinert Electronic Liquid FC - 77, Specialty Materials, 3M Center, Building 223-6S-04, St. Paul, MN 55144-1000, USA, see http://www.3m.com/fluids .

[32] H. Stockhorst, *IKP Annual Report 1994,* Jül-3035 (Forschungszentrum Jülich GmbH, 1995), p. 233.

APPENDIX A: THE EFFECTIVE PERMEABIILITY AND THE EFFECTIVE PERMITTIVITY OF A RESONATOR WITH MIXED AIR AND FERRITE FILL

A.1 GENERAL CONSIDERATIONS

It is important to differentiate between resonators in which the ferrite rings are separated from each other by electrically conducting material, like cooling plates, and those separated by an insulator. In the former case very little of the electric field penetrates the ferrite.

A.2 CAPACITANCE PER UNIT LENGTH

Case 1. Ferrite separation by insulator

If we consider a section of the line as in Fig. 9 with a charge (q/m) on the inner conductor then the electric field E at radius r is given by

$$E = \frac{q}{2\pi \varepsilon \varepsilon_0 r}, \tag{A.1}$$

where ε is the permittivity of the material at radius r and ε_0 the permittivity of free space.

The potential V on the inner conductor is given by

$$V = \int_{r_1}^{r_2} E\,dr + \int_{r_2}^{r_3} E\,dr \ , \tag{A.2}$$

where r_1, r_2, and r_3 are the radii of the inner conductor of the co-ax, the inner surface of the ferrite, and the outer surface of the ferrite and the outer conductor respectively.

Then

$$V = \frac{q}{2\pi\,\varepsilon_0}\left(\ln\frac{r_2}{r_1} + \frac{1}{\varepsilon}\ \ln\frac{r_3}{r_2}\right) \ . \tag{A.3}$$

The capacitance per metre is

$$C_t = \frac{2\pi\,\varepsilon_0}{\ln\frac{r_2}{r_1} + \frac{1}{\varepsilon}\ \ln\frac{r_3}{r_2}} = \frac{2\pi\,\varepsilon_e\,\varepsilon_0}{\ln\frac{r_3}{r_1}} \ , \tag{A.4}$$

where ε_e is the effective permittivity of the combined air and ferrite.

Comparing

$$\frac{\varepsilon_e}{\ln\frac{r_3}{r_1}} = \frac{1}{\ln\frac{r_2}{r_1} + \frac{1}{\varepsilon}\ \ln\frac{r_3}{r_2}} \tag{A.5}$$

gives

$$\varepsilon_e = \frac{\varepsilon}{k + \varepsilon\ (1-k)} \ , \text{ where } \ k \ = \ \frac{\ln\frac{r_3}{r_2}}{\ln\frac{r_3}{r_1}} \ . \tag{A.6}$$

Case 2. Ferrite separation by metallic cooling plates

In this case the capacitance per unit length is assumed to be just that of an air-spaced coaxial line with an outer radius r_2. This gives the capacitance C_t :

$$C_t = \frac{2\pi\varepsilon_0}{\ln\frac{r_2}{r_1}} = \frac{2\pi\,\varepsilon_e\,\varepsilon_0}{\ln\frac{r_3}{r_1}} \ . \tag{A.7}$$

From

$$\frac{\varepsilon_e}{\ln\frac{r_3}{r_1}} = \frac{1}{\ln\frac{r_2}{r_1}} \ \text{ and } \ k \ = \ \frac{\ln\frac{r_3}{r_2}}{\ln\frac{r_3}{r_1}} \tag{A.8}$$

follows

$$\varepsilon_e = \frac{1}{1-k} \ . \tag{A.9}$$

A.3 INDUCTANCE PER UNIT LENGTH

For a current i flowing in the coaxial line the inductance per unit length L_t is given by

$$L_t i = \int_{r_1}^{r_3} B\, dr = \int_{r_1}^{r_2} \mu_0 \frac{i}{2\pi r} dr + \int_{r_2}^{r_3} \mu' \mu_0 \frac{i}{2\pi r} dr \tag{A.10}$$

$$L_t = \frac{\mu_0}{2\pi} \left(\ln \frac{r_2}{r_1} + \mu' \ \ln \frac{r_3}{r_2} \right) , \tag{A.11}$$

where B is the magnetic field at radius r, μ_0 is the permeability of free space, and μ' the permeability of the ferrite.

This can be written as

$$L_t = \frac{\mu_e \mu_0}{2\pi} \ \ln \frac{r_3}{r_1} \tag{A.12}$$

where μ_e is the effective permeability of the air and ferrite.

Then

$$\mu_e \ln \frac{r_3}{r_1} = \ln \frac{r_2}{r_1} + \mu' \ \ln \frac{r_3}{r_2} \qquad \mu_e = 1 + k \ (\mu' - 1) \ . \tag{A.13}$$

The values of μ_e and ε_e will be further reduced if there are insulator-filled gaps with $\varepsilon = 1$ and $\mu = 1$ and of thickness d_2 separating the ferrite rings of thickness d_1. The effective values then become

$$\varepsilon_e = \frac{d_1}{d_1 + d_2} \ \frac{\varepsilon}{k + \varepsilon \ (1-k)} \tag{A.14}$$

$$\mu_e = \frac{d_1}{d_1 + d_2} \left(1 + k \ (\mu' - 1)\right) , \tag{A.15}$$

and μ_e remains unaltered by changing the insulator-filled caps to metal-filled gaps.

APPENDIX B: THE CURRENT AND VOLTAGE VARIATION ALONG A RESONATOR

The voltage V and the current i on the transmission line of Fig. 5 are given by the superposition of two waves in opposite directions

$$\begin{aligned} V &= A \cdot e^{j\omega(t - x/v)} + A' \cdot e^{j\omega(t + x/v)} \\ Z_0 i &= A \cdot e^{j\omega(t - x/v)} - A' \cdot e^{j\omega(t + x/v)} \ . \end{aligned} \tag{B.1}$$

The Standing Wave Ratio (SWR) of a line with a short circuit at $x = 0$ is

$$SWR = \frac{0 - Z_0}{0 + Z_0} = -1 \,, \tag{B.2}$$

so the incident and reflected waves have the same magnitude and different sign

$$A' = -A \tag{B.3}$$

$$\begin{aligned} V &= A \cdot e^{j\omega t} \; 2j \sin\frac{\omega x}{v} \\ Z_0 i &= A \cdot e^{j\omega t} \; 2\cos\frac{\omega x}{v} \end{aligned} \,. \tag{B.4}$$

For a complex definition of the voltage $V = V_g\, e^{j\omega t}$ at $x = -l$ the results are

$$A = \frac{-V_g}{2j \sin\frac{\omega l}{v}} \tag{B.5}$$

$$V = -V_g \cdot e^{j\omega t} \; \frac{\sin\frac{\omega x}{v}}{\sin\frac{\omega l}{v}} \tag{B.6}$$

$$i = jV_g \cdot e^{j\omega t} \; \frac{\cos\frac{\omega x}{v}}{Z_0 \sin\frac{\omega l}{v}} \,. \tag{B.7}$$

APPENDIX C: ESTIMATION OF REQUIRED FERRITE QUANTITY

From Section 3.2:

$$B_{rf\,\max} = \mu\mu_0 \; \frac{I}{2\pi \; r_2 \cos\frac{\omega l}{v}} \,, \tag{C.1}$$

and, as I is the current at $x = -l$, then:

$$\begin{aligned} V_g &= I\omega L \\ &= \frac{2\pi r_2 \cdot \cos\frac{\omega l}{v} \; B_{rf\,\max} \cdot Z_0 \tan\frac{\omega l}{v}}{\mu \; \mu_0} \\ &= \frac{r_2 \cdot \ln\frac{r_3}{r_1} \; B_{rf\,\max} \cdot \sin\frac{\omega l}{v} \; \sqrt{\frac{\mu_e}{\varepsilon_e} \; \frac{\mu_0}{\varepsilon_0}}}{\mu \; \mu_0} \,. \end{aligned} \tag{C.2}$$

For $\frac{\omega l}{v} < 26°$, $\sin\frac{\omega l}{v}$ approximates to $\frac{\omega l}{v}$ so that

V_g approximates to

$$V_g = \frac{r_2 \ln \frac{r_3}{r_1} \ \omega B_{rf\ \max} \cdot l \ \ \mu_e}{\mu} . \tag{C.3}$$

Vice-versa l approximates to

$$l = \frac{V_g \ \mu}{\mu_e \ r_2 \ \ln \frac{r_3}{r_1} \ \omega \ B_{rf\ \max}} . \tag{C.4}$$

By choosing B_{rfmax} (typically 0.01 T for NiZn ferrites), an estimate of the total length of ferrite required for an RF system can be made. New materials may achieve higher values of B_{rfmax}.

SERVO CONTROL OF RF CAVITIES UNDER BEAM LOADING

Alexander Gamp
DESY, Hamburg, Germany

Abstract
I begin by giving a description of the RF generator–cavity–beam coupled system in terms of basic quantities. Taking beam loading and cavity detuning into account, expressions for the cavity impedance as seen by the generator and as seen by the beam are derived. Subsequently methods of beam-loading compensation by cavity detuning, RF feedback and feedforward are described. Finally, a dedicated phase loop for damping synchrotron oscillations is discussed.

1. INTRODUCTION

In modern particle accelerators RF voltages with an extremely large amplitude and frequency range, from a few hundred volts to hundreds of megavolts and from several kilohertz to many gigahertz, are required for particle acceleration and storage.

The RF power needed to satisfy these demands can be generated, for example, by triodes, tetrodes, or klystrons. The Continuous Wave (CW) output power available from some tetrodes used at DESY is 60 kW at 208 MHz and 500–800 kW for the 500 MHz klystrons. Such RF power generators generally deliver RF voltages of only a few kilovolts because their source impedance is small compared with the cavity shunt impedance. For the TeV Energy Superconducting Linear Accelerator (TESLA) project a prototype pulsed L-band Multibeamklystron has delivered up to 10 MW peak power for 0.5 ms long pulses, and at 3 GHz 150 MW were achieved for pulses of 3 μs in length.

Typically, a tetrode has its highest efficiency for a load resistance of less than 1 kΩ ,whereas the cavity shunt impedance usually is of the order of several megaohms. This is the real impedance, which the cavity represents to a generator at the resonant frequency. It must not be confused with ohmic resistances.

Optimum fixed impedance matching between generator and cavity can be easily achieved with a coupling loop in the cavity. There is, however, the complication that the transformed cavity impedance as seen by the generator depends also on the synchronous phase angle and the beam current and is therefore not constant (as we shall show quantitatively). The beam current induces a voltage in the cavity that may become even larger than the one induced by the generator. Owing to the vector addition of these two voltages the generator now sees a cavity that appears to be detuned and unmatched except for the particular value of beam current for which the coupling has been optimized. The reflected power occurring at all other beam currents has to be handled.

In addition, the beam-induced cavity voltage may cause single- or multi-bunch instabilities, since any bunch in the machine may see an important fraction of the cavity voltage induced by itself or from previous bunches. This voltage is given by the product of beam current and cavity impedance as seen by the beam. Minimizing this latter quantity is therefore essential. It is also called beam-loading compensation, and some servo control mechanisms, which can be used to achieve this goal, will be discussed in this lecture.

2. THE COUPLING BETWEEN THE RF GENERATOR, THE CAVITY, AND THE BEAM

For frequencies in the neighbourhood of the fundamental resonance, an RF cavity can be described [1] by an equivalent circuit consisting of an inductance L_2 , a capacitor C, and a shunt impedance R_S, as shown in Fig. 1. In practice, L_2 is made up by the cavity walls, whereas the coupling loop L_1 is usually small compared with the cavity dimensions.

In this example a triode with maximum efficiency for a real load impedance R_A has been taken as an RF power generator. For simplicity we consider a short and lossless transmission line between the generator and L_1. Then there is optimum coupling between the generator and the empty (i.e. without beam) cavity for

$$N^2 = R_S / R_A = L_2 / L_1 \ , \tag{1}$$

where R_A equals the dynamic source impedance R_1. The term N is the transformation or step-up ratio.

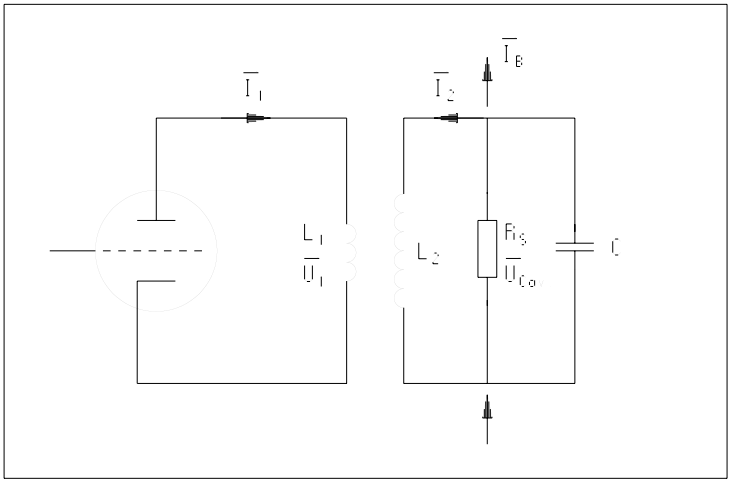


Fig. 1: Equivalent circuit of a resonant cavity near its fundamental resonance. In practice, the inductance L_2 is made up by the cavity walls, whereas L_1 is usually a small coupling loop.

Since, in general, there may be power transmitted from the generator to the cavity and also, in the case of imperfect matching, vice versa, the voltage $\vec{U}_1$ is expressed as the sum of two voltages

$$\vec{U}_1 = \vec{U}_{\text{forward}} + \vec{U}_{\text{reflected}} \ , \tag{2}$$

whereas the corresponding currents flow in the opposite directions, hence

$$\vec{I}_1 = \vec{I}_{\text{forward}} - \vec{I}_{\text{reflected}} \ . \tag{3}$$

The minus sign in Eq. (3) indicates the counterflowing currents, while voltages of forward and backward waves just add up.

So, in the simplest case, where the beam current $\vec{I}_B = 0$ and where the generator frequency $f_{GEN} = f_{CAV}$, there is no reflected power from the cavity to the generator, and $\vec{U}_1$ and $\vec{I}_1$ are identical to the generator voltage and current, respectively. One has

$$\vec{U}_{CAV} = N\vec{U}_1 \ . \tag{4}$$

Now we can derive an expression for the complex cavity voltage as a function of generator and beam current and of the cavity and generator frequency.

According to Fig. 1 the cavity voltage $\vec{U}_{\mathrm{CAV}}$ can be written as

$$\vec{U}_{\mathrm{CAV}} = L_2 \left(\frac{\dot{\vec{I}}_2 + \dot{\vec{I}}_1}{N} \right) \tag{5}$$

$$\vec{I}_2 = - \left(\frac{\vec{I}_{\mathrm{B}} + \vec{U}_{\mathrm{CAV}}}{R_{\mathrm{S}} + C\dot{\vec{U}}_{\mathrm{CAV}}} \right). \tag{6}$$

All voltages and currents have the time dependence

$$\vec{U} = \hat{\vec{U}} e^{i\omega t}. \tag{7}$$

$\vec{I}_{\mathrm{B}} = \vec{I}_{\mathrm{B}}(\omega)$ is the harmonic content at the frequency ω of the total beam current. Throughout this lecture we consider only a bunched beam with a small bunch spacing compared to the cavity filling time. In this case $\vec{I}_{\mathrm{B}}(\omega)$ is quasi sinusoidal. We also restrict the discussion to the interaction of the beam with the fundamental cavity resonance. Dedicated damping antennas built into the cavity can minimize the interaction with higher-order cavity modes.

Inserting Eq. (6) in Eq. (5) and using

$$2\pi f_{\mathrm{CAV}} = \omega_{\mathrm{CAV}} = \frac{1}{\sqrt{L_2 C}} \tag{8}$$

one finds

$$\omega_{\mathrm{CAV}}^2 \vec{U}_{\mathrm{CAV}} = \frac{1}{C} \left[\frac{1}{N} \dot{\vec{I}}_1 - \dot{\vec{I}}_{\mathrm{B}} - \frac{1}{R_{\mathrm{S}}} \dot{\vec{U}}_{\mathrm{CAV}} \right] - \ddot{\vec{U}}_{\mathrm{CAV}}. \tag{9}$$

We define

$$\Gamma = \frac{1}{2CR_{\mathrm{S}}} = \frac{\omega_{\mathrm{CAV}}}{2Q} \tag{10}$$

where the quality factor of the cavity can be expressed as 2π times the ratio of total electromagnetic energy stored in the cavity to the energy loss per cycle.

Here we would like to mention that the ratio

$$\frac{R_{\mathrm{S}}}{Q} = \sqrt{\frac{L_2}{C}} \tag{11}$$

is a characteristic quantity of a cavity depending only on its geometry.

We can rewrite Eq. (9) as

$$\ddot{\vec{U}}_{\mathrm{CAV}} + 2\Gamma \dot{\vec{U}}_{\mathrm{CAV}} + \omega_{\mathrm{CAV}}^2 \vec{U}_{\mathrm{CAV}} = 2\Gamma R_{\mathrm{S}} \left[\frac{1}{N} \dot{\vec{I}}_1 - \dot{\vec{I}}_{\mathrm{B}} \right]. \tag{12}$$

This equation describes a resonant circuit excited by the current $\vec{I} = \left(\vec{I}_1 \middle/ \left(N - \vec{I}_{\mathrm{B}} \right) \right)$. The minus sign occurs because the generator-induced cavity voltage has opposite sign to the beam-induced voltage, which would decelerate the beam. It can be shown that the beam actually sees only 50% of its own induced voltage. This is called the fundamental theorem of beam loading [2, 3].

In order to find the cavity impedance as seen by the beam we make use of Eqs. (2), (3), and (4) to express the generator current term of Eq. (12) in the form

$$\frac{1}{N}\dot{\vec{I}}_1 = \frac{1}{NR_A}\left[2\dot{\vec{U}}_{\text{forward}} - \dot{\vec{U}}_1\right] = \frac{1}{N}\left[2\dot{\vec{I}}_{\text{forward}} - \frac{\dot{\vec{U}}_{\text{CAV}}}{NR_A}\right] \quad . \tag{13}$$

This leads to a modification of the damping term in Eq. (12)

$$\ddot{\vec{U}}_{\text{CAV}} + 2\Gamma(1+\beta)\dot{\vec{U}}_{\text{CAV}} + \omega_{\text{CAV}}^2\vec{U}_{\text{CAV}} = 2\Gamma_L R_{SL}\left[\frac{2}{N}\dot{\vec{I}}_f - \dot{\vec{I}}_B\right] \quad . \tag{14}$$

With the coupling ratio

$$\beta = R_S/\left(N^2 R_I\right) \tag{15}$$

we can introduce the 'loaded' damping term

$$\Gamma_L = \Gamma(1+\beta) \tag{16}$$

and consequently, in accordance with Eq. (10), the loaded cavity Q and loaded shunt impedance are

$$Q_L = Q/(1+\beta) \quad \text{and} \quad R_{SL} = R_S/(1+\beta) \; . \tag{17}$$

In the case of perfect matching in the absence of beam, i.e. $\beta = 1$, the damping term simply doubles and Q and R_S take half their original values. This is because the beam would see the cavity shunt impedance R_S in parallel or loaded with the transformed generator impedance $N^2R_I = R_S$. Therefore we find in Eq. (14) that the transformed generator current

$$\vec{I}_G = 2\vec{I}_f / N \tag{18}$$

gives rise to twice as much cavity voltage as a similar beam current would do. Here and in Eq. (15) we assume that the transformed dynamic source impedance N^2R_I is identical to the generator impedance seen by the cavity. This is strictly true only if a circulator is placed between the RF power generator and the cavity. Without a circulator it may be approximately true if the power source is a triode. Owing to its almost constant anode-voltage-to-current characteristic the impedance of a tetrode as seen from the cavity is, however, much bigger than the corresponding R_I and therefore $R_{SL} \approx R_S$ here, where a short transmission line (or of length $n\lambda/2$, n integer) is considered.

Following Ref. [4] we write the solution of Eq. (14) in the Fourier–Laplace representation

$$\hat{\dot{\vec{U}}}_{\text{CAV}} = \frac{i\omega}{\omega_{\text{CAV}}^2 - \omega^2 + 2i\omega\Gamma_L} 2\Gamma_L R_{SL}\left[\hat{\dot{\vec{I}}}_G - \hat{\dot{\vec{I}}}_B\right] \; . \tag{19}$$

For $\Delta\omega \cdot \omega_{\text{CAV}}$ this can be approximated by

$$\hat{\dot{\vec{U}}}_{\text{CAV}} \approx \frac{R_{SL}\left[\hat{\dot{\vec{I}}}_G - \hat{\dot{\vec{I}}}_B\right]}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{\text{CAV}}}} \quad , \tag{20}$$

where $\omega = \omega_{\text{CAV}} + \Delta\omega$.

For a resonant cavity the beam-induced voltage $\vec{U}_B$, or the beam loading, is thus given by the product of loaded shunt impedance and beam current:

$$\vec{U}_{\mathrm{B}} = -R_{\mathrm{SL}}\vec{I}_{\mathrm{B}} \tag{21}$$

The ideal beam loading compensation would, therefore, minimize R_{SL} without increasing the generator power necessary to maintain the cavity voltage.

Having just discussed the impedance that the combined system generator and cavity represents to the beam, we would like to discuss the impedance *Z*, or rather admittance *Y = 1/Z*, which the combined cavity and beam system represents to the generator.

From Eqs. (1), (5), and (6) one sees [5] that

$$Y = \frac{\vec{I}_1}{\vec{U}_1} = \frac{N^2}{R_{\mathrm{S}}} + \frac{\vec{I}_{\mathrm{B}}N^2}{\vec{U}_{\mathrm{CAV}}} + \frac{N^2}{i\omega L_2}\left(1 - \frac{\omega^2}{\omega_{\mathrm{CAV}}^2}\right) . \tag{22}$$

which reduces to $Y = 1/R_{\mathrm{A}}$ for a tuned cavity without beam current in the case of $\beta = 1$.

As we are now going to show, a non-vanishing real part of the quotient $\vec{I}_{\mathrm{B}}/\vec{U}_{\mathrm{CAV}}$ will necessitate a change in β to maintain optimum matching, whereas the imaginary part can be compensated by detuning the cavity. In order to work out Re and $\mathrm{Im}(\vec{I}_{\mathrm{B}}/\vec{U}_{\mathrm{CAV}})$, we define the angle ϕ_{s} as the phase angle between the synchronous particle and the zero crossing of the RF cavity voltage. The accelerating voltage is therefore given by

$$\vec{U}_{\mathrm{ACC}} = \vec{U}_{\mathrm{CAV}} \sin\phi_{\mathrm{s}} \tag{23}$$

and the normalized cavity voltage and beam current are related by

$$\frac{\vec{I}_{\mathrm{B}}}{|\vec{I}_{\mathrm{B}}|} = \frac{\vec{U}_{\mathrm{CAV}}}{|\vec{U}_{\mathrm{CAV}}|} e^{\mathrm{i}\left(\frac{\pi}{2} - \phi_{\mathrm{s}}\right)} . \tag{24}$$

Consequently

$$\mathrm{Re}\left(\frac{\vec{I}_{\mathrm{B}}}{\vec{U}_{\mathrm{CAV}}}\right) = \left|\frac{\vec{I}_{\mathrm{B}}}{\vec{U}_{\mathrm{CAV}}}\right| \sin\phi_{\mathrm{s}} \tag{25}$$

and

$$\mathrm{Im}\left(\frac{\vec{\mathrm{I}}_{\mathrm{B}}}{\vec{\mathrm{U}}_{\mathrm{CAV}}}\right) = \left|\frac{\vec{\mathrm{I}}_{\mathrm{B}}}{\vec{\mathrm{U}}_{\mathrm{CAV}}}\right| \cos\phi_{\mathrm{s}} . \tag{26}$$

The real part of the admittance seen by the generator then becomes

$$\mathrm{Re(Y)} = \frac{\mathrm{N}^2}{\mathrm{R_S}}\left(1 + \frac{\mathrm{R_S}|\vec{\mathrm{I}}_{\mathrm{B}}|}{|\vec{\mathrm{U}}_{\mathrm{CAV}}|} \sin\phi_{\mathrm{s}}\right) . \tag{27}$$

We see that the term in the bracket describes the change of admittance caused by the beam. In order to maintain optimum coupling the coupling ratio β must now take the value

$$\beta = \left(1 + \frac{R_{\mathrm{S}}|\vec{I}_{\mathrm{B}}|}{|\vec{U}_{\mathrm{CAV}}|} \sin\phi_s\right) . \tag{28}$$

This result tells us that the change in the real part of the admittance is proportional to the ratio of RF power delivered to the beam to RF power dissipated in the cavity walls. For circular electron

machines, where the considerable amount of energy lost by synchrotron radiation has to be compensated continuously by RF power, values of $\phi_s \geq 30°$ and $\beta \geq 1.2$ are typical for high beam current and normally conducting cavities. A typical set of parameters for this case would be $R_S = 6\,\mathrm{M}\Omega$, $\vec{U}_{CAV} = 1$ MV and $\vec{I}_B = 30$ mA. This implies, of course, that for a β that has been optimized for the maximum beam current, there will be reflected generator power for lower beam intensities. If the power source is a klystron, this can be handled by inserting a circulator in the path between generator and cavity or, in the case of a tube, by a sufficiently high plate dissipation power capability.

From the imaginary part of Eq. (22) and from Eq. (26) we find that the apparent cavity detuning caused by the beam current can be compensated by a real cavity detuning (for example, by means of a mechanical plunger cavity tuner) of the amount

$$\frac{\omega}{\omega_{CAV}} = \sqrt{1 - \frac{R_S |\vec{I}_B|}{Q |\vec{U}_{CAV}|} \cos\phi_s} \,. \tag{29}$$

Expanding the square root to first order we find a cavity detuning angle Ψ

$$\tan\Psi \approx \frac{R_{SL} |\vec{I}_B|}{|\vec{U}_{CAV}|} \cos\phi_s \approx -2Q_L \frac{\Delta\omega}{\omega_{CAV}} \,. \tag{30}$$

This is essentially the ratio between beam-induced and total cavity voltage.

In order to calculate the maximum amount of reflected power seen by the generator as a consequence of beam loading we consider, for $\beta = 1$, a tuned cavity, i.e. $\omega = \omega_{CAV}$. Then, with Eqs. (25) and (26), Eq. (22) reads

$$Y = \frac{1}{R_A} \left[1 + \frac{R_S |\vec{I}_B|}{|\vec{U}_{CAV}|} \sin\phi_s + i \frac{R_S |\vec{I}_B|}{|\vec{U}_{CAV}|} \cos\phi_s \right] \,. \tag{31}$$

Solving for $\vec{U}_{refl.}$ by means of Eqs. (2) and (3) the reflected power $P_{refl.} = |\hat{\vec{U}}_{refl.}|^2 / 2R_I$ becomes

$$P_{refl.} = R_S \hat{\vec{I}}_B^2 / 8 \,. \tag{32}$$

This corresponds to half of the power given by the beam to the coupled system cavity and generator. The second half of this power is dissipated in the cavity walls. All we found is that two equal resistors in parallel dissipate equal amounts of power. As pointed out above, this is strictly true only if a circulator is placed between the RF power source and the cavity. Nevertheless, the amount of reflected power can be quite impressive. For an average DC beam current of, say 0.1 A, the harmonic current $\vec{I}_B(\omega)$ may become up to twice as large. Then, taking $R_S = 8$ MΩ, for example, we find 40 kW of reflected power have to be dissipated.

For a cavity where only the reactive part of the beam loading has been compensated by detuning according to Eq. (30), but $\beta = 1$, the reflected power is given by

$$P_{refl.} = R_S \hat{\vec{I}}_B^2 \sin^2\phi_s / 8 \,. \tag{33}$$

Summarizing the results of this section we state that the beam sees the cavity shunt impedance in parallel with the transformed generator impedance. The resulting loaded impedance is reduced by the factor $1/(1 + \beta)$. The optimum coupling ratio between generator and cavity depends on the amount of energy taken by the beam out of the RF field. The coupling is usually fixed and optimized for maximum beam current. The amount of cavity detuning necessary for optimum matching, on the other hand, depends on the ratio of beam-induced to total cavity voltage.

3. BEAM-LOADING COMPENSATION BY DETUNING

As we have shown in the previous Section, stationary beam loading can be entirely compensated by detuning the cavity, provided that the synchronous phase angle is small or zero. This is usually the case in proton synchrotrons during storage, where the energy loss due to the emission of synchrotron radiation is negligible. Here, the RF voltage is needed only to keep the bunch length short. Energy ramping also takes place at very small ϕ_s.

In the following we restrict ourselves, for simplicity, to hadron machines. Consequently $\beta = 1$, $\phi_s \approx 0$, and the generator- and beam-induced voltages are in quadrature.

There are, however, also limitations to detuning as the only means of beam-loading compensation. One is known as Robinson's stability criterion [6]. If the amount of detuning calculated by Eq. (30) becomes comparable to the revolution frequency of the particles in a synchrotron, the beam will become unstable. Another one is the finite time of, say, a second, which is needed for the tuner to react. Actually, the time scale of the cavity voltage transients, which may cause beam instabilities, is much shorter. The cavity filling time τ_{CAV} is given by

$$\tau_{CAV} = 2Q_L / \omega_{CAV} \ . \tag{34}$$

The cavity voltage rise after injection of a bunched beam with a current $\vec{I}_B(\omega_{CAV})$ can be approximated by

$$\vec{U}_B \approx R_{SL}\vec{I}_B\left(1 - e^{-t/\tau}\right) \ . \tag{35}$$

This voltage will add to the cavity voltage produced by the generator, and after a time $t \approx 3\tau$ the total cavity voltage becomes

$$\left|\vec{U}_{CAV}\right| \approx R_{SL}\sqrt{\left|\vec{I}_g\right|^2 + \left|\vec{I}_B\right|^2} \tag{36}$$

with a phase shift given by Eq. (30).

Since typical values of τ are below 100 μs and therefore much smaller than the proton synchrotron frequency in a storage ring (T_S is usually ≥ several ms), these transients will, in general, excite synchrotron oscillations of the beam with the consequence of emittance blow-up and particle loss. Additional compensation of transient beam loading is therefore necessary. This will be discussed in the following paragraphs.

In Fig. 2 a diagram of a tuner regulation circuit is shown. The phase detector measures the relative phase between generator current and cavity voltage which depends, according to Eq. (20), on the frequency $\Delta\omega$, by which the cavity is detuned. The phase detector output signal acts on a motor that drives a plunger tuner into the cavity volume until there is resonance. An alternative tuner could be a resonant circuit loaded with ferrites. The magnetic permeability μ of the ferrites and hence the resonance frequency of the circuit can be controlled by a magnetic field. This latter method is especially useful when a large tuning range in combination with a low cavity Q is required.

If proper tuner action is necessary in a large dynamic range of cavity voltages, limiters with a minimum phase shift per dB compression have to be installed at the phase detector input. Since this phase shift is decreasing with frequency, all signals should be mixed down to a sufficiently low intermediate frequency.

The signal proportional to the generator current $\vec{I}_{forw.}$ can be obtained from a directional coupler. In case the RF amplifier is so closely coupled to the cavity that no directional coupler can be installed, the relative phase between RF amplifier input and output signal can also be used to derive a tuner signal [7].

4. REDUCTION OF TRANSIENT BEAM LOADING BY FAST FEEDBACK

The principle of a fast feedback circuit is illustrated in Fig. 2. A small fraction α of the cavity RF signal is fed back to the RF preamplifier input and combined with the generator signal. The total delay δ in the feedback path is such that both signals have opposite phase at the cavity resonance frequency ω_{CAV}. For other frequencies there is a phase shift

$$\Delta\phi = \Delta\omega\delta\,. \tag{37}$$

Therefore the voltage at the amplifier input is now given by

$$\vec{U}'_{in} = \vec{U}_{in} - e^{-i\Delta\omega\delta}\alpha\vec{U}_{CAV}\,. \tag{38}$$

With the voltage gain K of the amplifier we can rewrite Eq. (20) and obtain for the cavity voltage with feedback

$$\vec{U}_{CAV} \approx \frac{K\left[\vec{U}_{in} - e^{-i\Delta\omega\delta}\,\alpha\vec{U}_{CAV}\right] - \vec{U}_B}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}}} \tag{39}$$

or

$$\vec{U}_{CAV} \approx \frac{K\vec{U}_{in} - \vec{U}_B}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}} + e^{-i\Delta\omega\delta}\,\alpha K}\,. \tag{40}$$

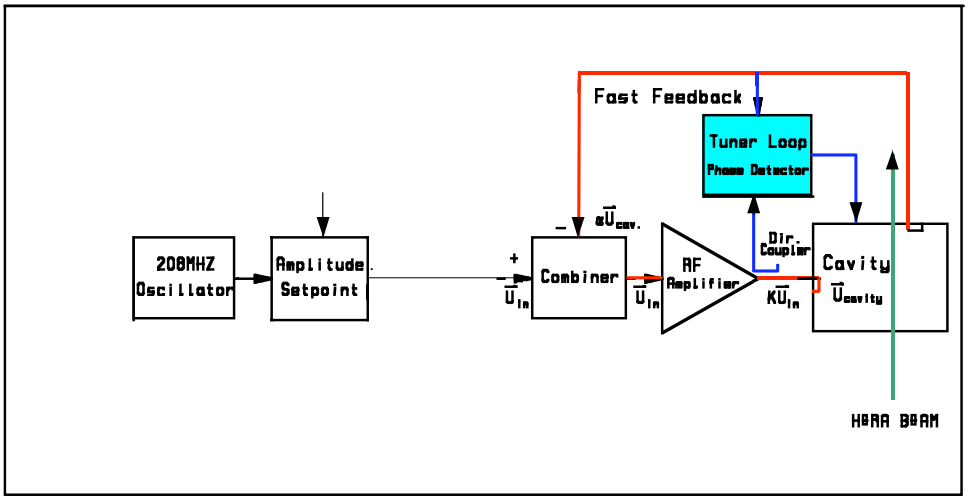


Fig. 2: Schematic of servo loops for phase and amplitude control of the HERA 208 MHz proton RF system

For $\Delta\omega = 0$ and $A_F \gg 1$ this reduces to

$$\vec{U}_{CAV} \approx \frac{\vec{U}_{in}}{\alpha} - \frac{\vec{U}_B}{\alpha K}\,. \tag{41}$$

The open loop feedback gain A_F is defined as

$$A_F = \alpha K \,. \tag{42}$$

One sees that there is a reduction of the beam-induced cavity voltage by the factor $1/A_F$ due to the feedback. This is equivalent to a similar reduction of the cavity shunt impedance as seen by the beam.

$$Z_L \approx \frac{R_{SL}}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}}} \rightarrow \frac{R_{SL}}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}} + A_F e^{-i\Delta\omega\delta}} \,. \tag{43}$$

The price for this fast reduction of beam loading is the additional amount of generator current $\vec{I}_B N$ that is needed almost to compensate the beam current in the cavity. In terms of additional transmitter power P' this reads

$$P' = R_S \hat{\vec{I}}_B^{\,2} / 8 \,. \tag{44}$$

It is the power already calculated by Eq. (32). As there is no change in cavity voltage due to P' this power will be reflected back to the generator, which has to have a sufficiently large plate dissipation power capability. Otherwise a circulator is needed. This critical situation of additional RF power consumption and reflection lasts, however, only until the tuner has reacted, and it may be minimized by pre-detuning. The generator-induced voltage is, of course, also reduced by the amount $1/A_F$, but this can be easily compensated on the low power level by increasing $\vec{U}_{in}$ by the factor 1/α as Eq. (41) suggests. The practical implications of this will be illustrated by the following example.

Let the power gain of the amplifier be 80 dB. For a cavity power of 50 kW an input power P_{IN} of 0.5 mW is thus required. This corresponds to a voltage gain of 10^4 so, for a design value of A_F = 100, α becomes 10^{-2}. Hence the power that is fed back to the amplifier input, is 5 W. In order to maintain the same cavity voltage as without feedback, P_{IN} has to be increased from 0.5 mW to 5.0005 W. This value can, of course, be reduced by decreasing α. But then the amplifier gain has to be increased to keep A_F constant. This leads to power levels in the 100 µW range at the amplifier input. All this is still practical, but some precautions, such as extremely good shielding and suppression of generator and cavity harmonics, have to be taken.

The maximum feedback gain that can be obtained is limited by the aforementioned delay time δ of a signal propagating around the loop. According to Nyquist's criterion, the system will start to oscillate if the phase shift between $\vec{U}_{in}$ and $\alpha\vec{U}_{CAV}$ exceeds ≈ 135°. A cavity with high Q can produce a ±90° phase shift even for very small $\Delta\omega$. Therefore, once the additional phase shift given by Eq. (37) has reached ± π/4, the loop gain must have become ≤ 1, i.e.

$$|A_F(\Delta\omega_{max})| \approx \frac{\alpha K}{1 + iQ_L 2\dfrac{\Delta\omega_{max}}{\omega_{CAV}}} \le 1 \tag{45}$$

where

$$\delta\Delta\omega_{max} = \pm\frac{\pi}{4} \,. \tag{46}$$

Here we assume that all other frequency-dependent phase shifts, like the ones produced by the amplifiers, can be neglected. Taking $\vec{U}_B = 0$ and inserting Eq. (45) we can solve Eq. (39) for A_F:

$$A_F = \frac{Q_L}{4 f_{CAV}\delta} \,. \tag{47}$$

This is the maximum possible feedback gain for a given δ.

A fast feedback loop of gain 100 has been realized at the HERA 208 MHz proton RF system. With a loaded cavity Q_L of ≈ 27000 the maximum tolerable delay, including all amplifier stages and cables, is $\delta = 330$ ns. Therefore all RF amplifiers have been installed very close to the cavities in the HERA tunnel.

In addition, there are independent slow phase and amplitude regulation units for each cavity with still higher gain in the region of the synchrotron frequencies, i.e. below 300 Hz. Without fast feedback these units might become unstable at heavy beam loading [8, 9] since changes in cavity voltage and phase are then correlated, as shown by Eqs. (30) and (36).

The effect of a fast feedback loop is revealed in Fig. 3, where the transient behaviour of the imaginary (upper curve) and real (middle curve) part of a HERA 208 MHz cavity voltage vector is displayed. The lower curve is the signal of a beam current monitor, which shows nicely the bunch structure of the beam and a 2 μs gap between batches of 6 × 10 bunches each. A detailed description of this measurement and of the IQ detector used is given in Ref. [10]. In this particular case the upper curve is essentially equivalent to the phase change of the cavity voltage due to transient beam loading, and the middle curve corresponds to the change in amplitude.

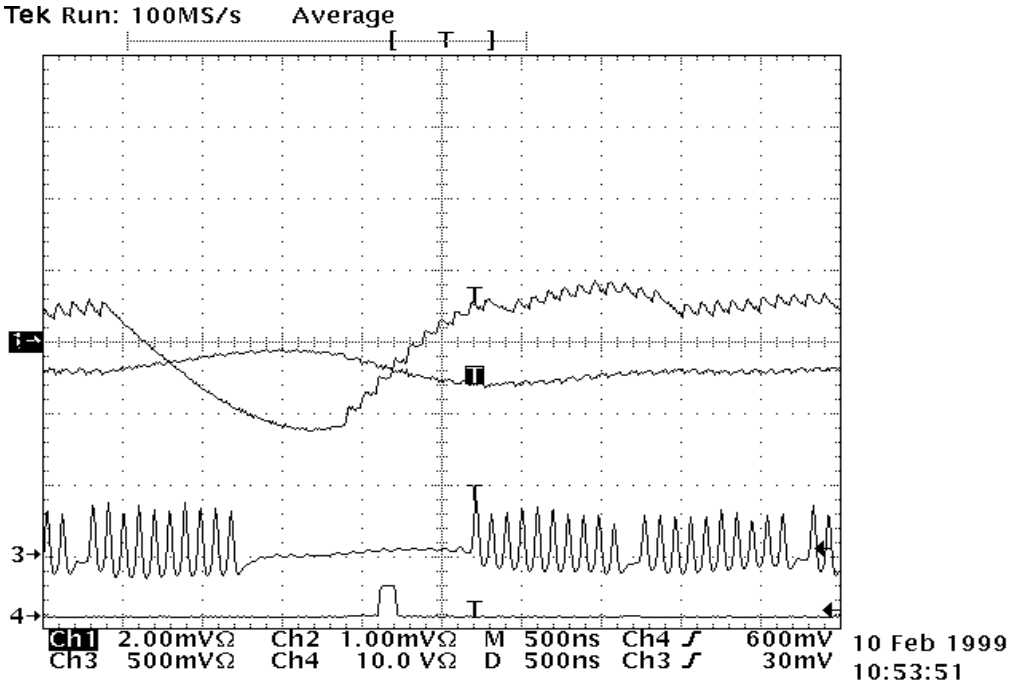


Fig. 3: Transient behaviour of the cavity voltage under the influence of fast feedback.
This figure is taken from Ref. [10].

The apparent time shift between the bunch signals and the cavity signals is due to the time of flight of the protons between the location of the cavity and the beam monitor in HERA. The transients resulting from the first two or three bunches after the gap cause step-like transients, which accumulate without significant correction. Later the fast feedback delivers a correction signal, which causes the subsequent transients to look more and more saw-tooth-like. From this one can estimate the time delay in the feedback loop to be of the order of 250 ns. After about 1 μs equilibrium with the beam is reached. Similarly, one observes in the left part of the picture that the feedback correction is still present for 250 ns after the last bunch before the gap has left the cavity. The equilibrium without beam is also reached after about 1 μs. Without fast feedback the time taken to reach equilibrium is about 100 times longer, as one would expect for a feedback gain of 100.

To summarize this Section we state that fast feedback reduces the resonant cavity impedance as seen by an external observer (usually the beam) by the factor $1/A_F$. It is important to realize that any

noise originating from sources other than the generator, especially amplitude and phase noise from the amplifiers, will be reduced by the factor $1/A_F$ because the cavity signal is directly compared to the generator signal at the amplifier input stage. Care has to be taken that no noise be created, by diode limiters or other non-linear elements, in the path where the cavity signal is fed back to the amplifier input. This noise would be added to the cavity signal by the feedback circuit.

5. FEEDBACK AND FEEDFORWARD APPLIED TO SUPERCONDUCTING CAVITIES

So far, we have only considered normally conducting cavities in a proton storage ring, where the protons arrive in the cavities at the zero crossing of the RF signal, i.e. at $\phi_s = 0^\circ$ or a few degrees.

In the following I would like to present an example of the other extreme: superconducting cavities in a linear electron accelerator where the electrons cross the cavities near the moment of maximum RF voltage, i.e. at $\phi_s \approx 90^\circ$. (Note that for linear colliders a different definition of ϕ_s is usually used, namely $\phi_s = 0^\circ$ when the particle is on crest. In this article we do not adopt this definition.)

A test facility for TESLA is currently being built at DESY. We refer to the special example of the TESLA Test Facility cavities, which are 9-cell cavities made of pure niobium. The operating frequency is 1.3 GHz.

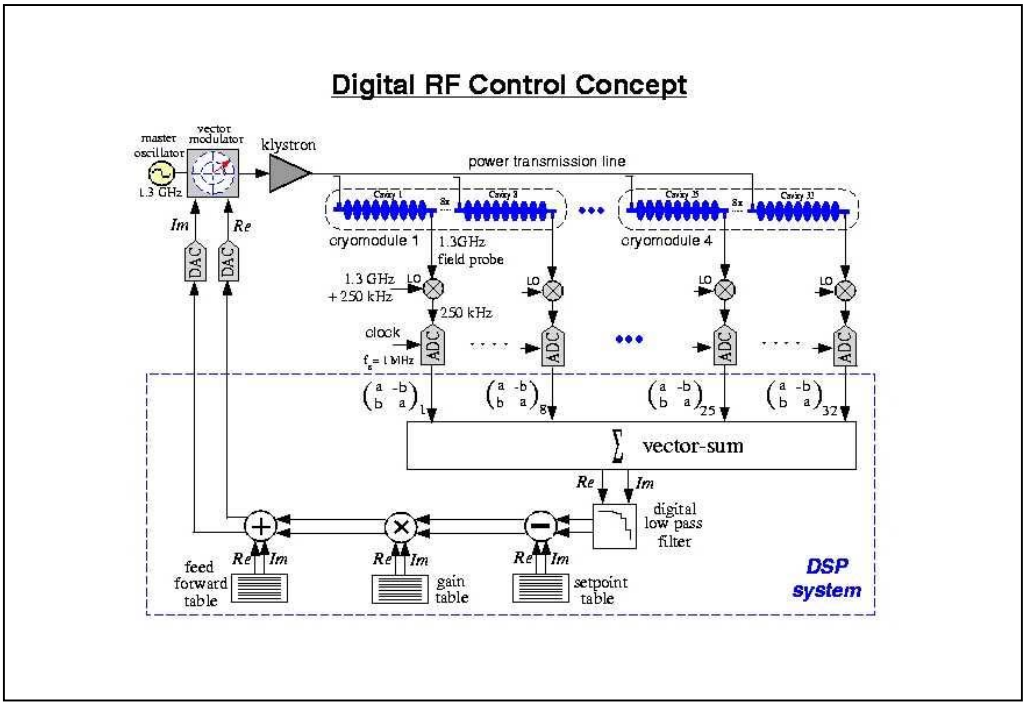


Fig. 4: Schematic of the low-level RF system for control of the TESLA Test Facility 1.3 GHz cavities. This figure is taken from Ref. [11].

The unloaded Q-value of these cavities is in the range 10^9–10^{10}, or even higher. Hence the bandwidth is only of the order of 1 Hz, and also the shunt impedance of these cavities exceeds that of normally conducting cavities by many orders of magnitude. Therefore, we have a coupling factor $\beta \approx 1000$ in this case, which also reflects the fact that the ratio of the power taken away by the beam to the power dissipated in the cavity walls is much larger for superconducting cavities than for normally conducting ones. Owing to the coupling, the nominal loaded Q-value is only 3×10^6, and the loaded cavity bandwidth seen by the beam is 433 Hz. Since in this case there is a circulator with a load to protect the klystron from reflected power, this loaded bandwidth is also seen by the RF generator, which is a klystron. Since the particles are (almost) on crest, only the real part of the admittance [Eq. (27)] seen by the generator is changed due to beam loading. This means that in this example

beam loading causes a change only in the cavity impedance seen by the generator and detuning plays no role as a means of beam-loading compensation. Therefore there is only perfect matching for the nominal beam current to which the cavity power input coupler has been adjusted.

From the circuit diagram in Fig. 4 we see that one RF generator is planned to supply up to 32 cavities with RF power. The RF power that is needed per cavity to accelerate an electron beam of 8.5 mA to 25 MeV, is close to 215 kW, hence a klystron power close to 7 MW is needed. This power is entirely carried away by the beam. In contrast to the previous example, where all the RF power was essentially dissipated in the normally conducting cavity walls, the power needed to build up the RF cavity voltage in the superconducting cavities is only a few hundred watts. A high-efficiency, 10 MW, multibeam klystron has been developed for this project. For completeness we mention that this is pulsed power, with a pulse length of 1.5 ms and a maximum repetition rate of 10 Hz. So the maximum average klystron power is 150 kW.

The RF seen by the beam corresponds to the vector sum of all cavity signals. Therefore for the RF control system such a vector sum needs to be generated. This is done by down conversion of the cavity signals to 250 kHz intermediate frequency signals that are sampled in time steps of 1 μs. Each set of two subsequent samples corresponds to the real and imaginary part of the cavity voltage vector. The vector sum is generated in a computer and is compared to a table of set point values. The difference signal, which corresponds to the cavity voltage error, acts on a vector modulator at the low-level klystron input signal. In addition to this feedback a feedforward correction can be added. The advantage of feedforward is that, in principle, there is no gain limitation as in the case of feedback. If the error is known in advance, one can program a counteraction in the feedforward table. Examples of such errors could be a systematic decrease in beam current during the pulse due to some property of the electron source, or a systematic change in the cavity resonance frequency during the pulse. This effect does indeed exist. The mechanical forces resulting from the strong pulsed RF field in the superconducting cavities cause a detuning of the order of a few hundred hertz at 25 MV/m. This effect is called Lorentz force detuning.

From Eq. (47) one might infer that, because of the large value of Q_L, the maximum possible feedback gain in this case could become significantly larger than for normally conducting cavities. However, one has to check whether there are poles in the system at other frequencies, and, at least in this case, there is a fairly large loop delay of about 4 μs caused by the 12 m length of the cryogenic modules in which the cavities are placed and by the time delay in the computer. This results in a realistic maximum loop gain of 140.

So far, we have demonstrated for up to 16 cavities that all this really does work in practice. The impressive phase and amplitude stability of 0.1 degree and 0.5% that has been reached with this feedback system is shown in Fig. 5. There one also sees that the addition of feedforward improves the amplitude and phase stability to 0.05% and 0.03 degrees respectively.

6. DAMPING OF SYNCHROTRON OSCILLATIONS OF PROTONS IN THE PETRA II MACHINE

In the preceding sections phase and amplitude control of the cavity voltage was discussed. In this last Section we would like to give an example of beam control by looking at the dedicated RF system for the damping of synchrotron oscillations of protons in the PETRA II synchrotron at DESY.

Prior to injection into HERA, protons are pre-accelerated to 7.5 and 40 GeV/c in the DESY III and PETRA II synchrotrons, respectively [13]. Timing imperfections during the transfer of protons from one machine to the next and RF noise during ramping were observed to cause synchrotron oscillations that, if not damped properly, may lead to an increase in beam emittance and to significant beam losses. Therefore a phase loop acting on the RF phase to damp these oscillations of the proton bunches is a necessary component of the low-level RF system. The PETRA II proton RF system, which consists of two 52 MHz cavities, each with a closely coupled RF amplifier chain and a fast

feedback loop of gain 50, is similar to the one shown in Fig. 2. The block diagram of the PETRA II phase loop, on which I shall now concentrate, is shown in Fig. 6.

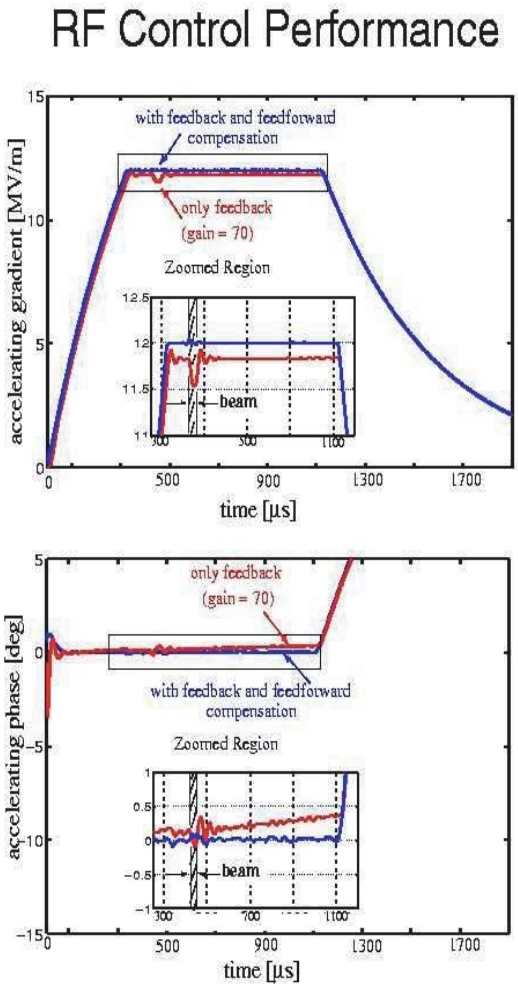


Fig. 5: Phase and amplitude stability achieved by digital feedback and feedforward in a 1.3 GHz cavity of the TESLA Test Facility. The obtained phase and amplitude stability with feedback alone is 0.5% and 0.1 degrees. With feedforward an improvement to 0.05% and 0.03 degrees was reached. This figure is taken from Ref. [12].

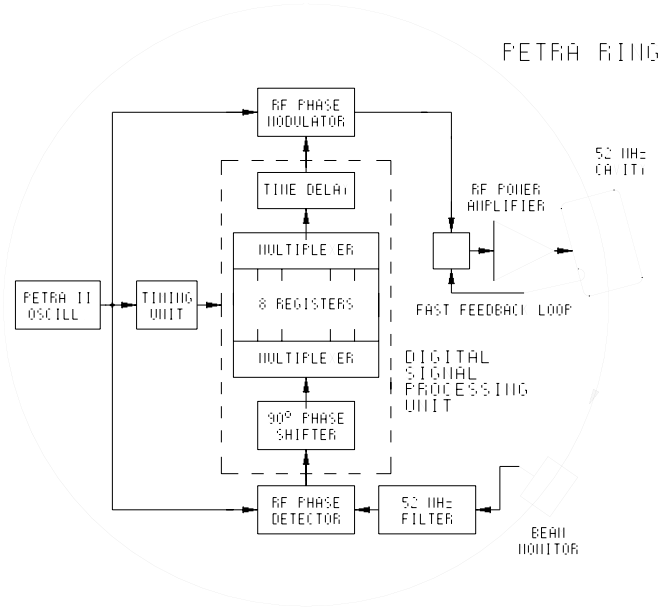


Fig. 6: Block diagram of the PETRA II phase loop. In the phase detector, synchrotron oscillations of the bunches are detected by comparing the filtered 52 MHz component of the beam with the 52 MHz RF reference source. An average phase signal for each of the eight batches of ten bunches is phase shifted by 90° with respect to the synchrotron frequency, stored in its register, and properly multiplexed to the phase modulator acting on the RF drive signal.

6.1 Loop bandwidth

The maximum number of bunches is 11 in DESY III and 80 in PETRA II, so eight DESY III cycles are needed to fill PETRA II. If synchrotron oscillations arise due to injection timing errors, all bunches of the corresponding batch are expected to oscillate coherently. Therefore one single correction signal can damp the bunch oscillations in that batch and in total up to eight such signals are needed, one for each batch. This phase loop is a batch-to-batch rather than a bunch-to-bunch feedback. Ideally, the correction of expected errors of about two degrees in the injection phase has to be switched within the 96 ns separating the last bunch of batch n from the first bunch of batch $n + 1$. Owing to the fast feedback of gain 50, the RF system has an effective bandwidth of about 1 MHz. However, it is capable of performing small phase changes of the order of 1° per 100 ns, which should be sufficient for damping synchrotron oscillations also in the multibatch mode of operation.

6.2 The phase detector

Each bunch passage generates a signal in the inductive beam monitor, also shown in Fig. 6. A passive LC filter of 8 MHz bandwidth filters out the 52 MHz component. The ringing time is comparable to the bunch spacing time as shown in Fig. 7. Amplitude fluctuations of this signal are reduced to ±0.5 dB in a limiter of 40 dB dynamic range. So the amplitude dependence of the synchrotron phase measurement between the bunch signal and the 52 MHz RF source signal is minimized. The phase detector has a sensitivity of 10 mV per degree. By inserting a low pass filter one can directly observe the synchrotron motion of the bunches at the phase detector output. This is shown in Fig. 8(a) for one batch of nine proton bunches circulating in PETRA II with a momentum of 7.5 GeV/c a few milliseconds after injection. The observed synchrotron period $T_S = 5$ ms agrees with the expected value for the actual RF voltage of 50 kV.

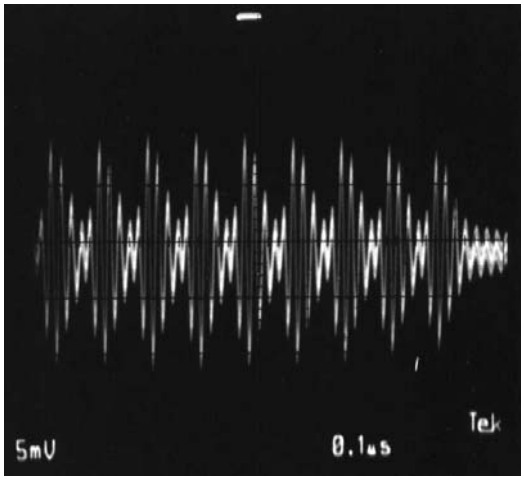


Fig. 7: Filtered signal of a batch of nine proton bunches circulating in PETRA. The bunch spacing time is 96 ns.

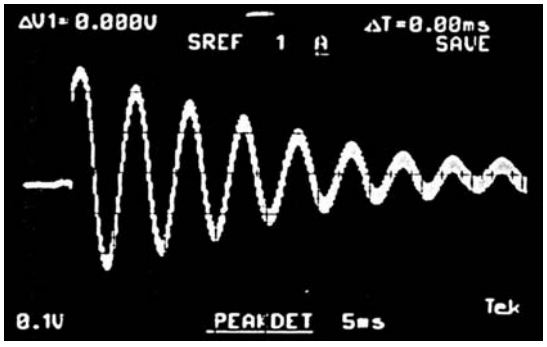


Fig. 8(a): The synchrotron oscillation measured at the phase detector output a few milliseconds after injection of a batch of nine proton bunches into PETRA II. It is smeared out by Landau damping after several periods. The damping loop is not active.

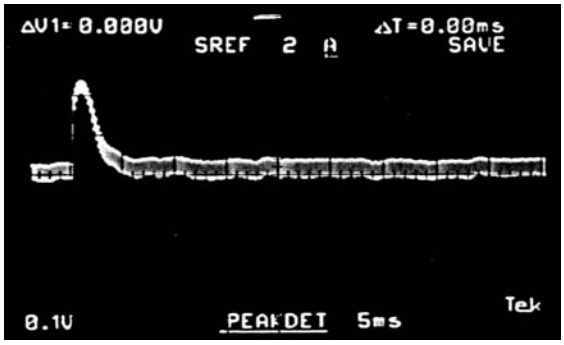


Fig. 8(b): Same as Fig. 8(a) but with the phase loop active. The synchrotron oscillation is completely damped within half a synchrotron period of 5 ms.

6.3 The FIR filter as a digital phase shifter

A feedback loop can damp the synchrotron motion if, as is indicated in Fig. 6, the synchrotron phase signal is shifted by –90° relative to the synchrotron frequency f_S, delayed properly, and fed into a phase modulator acting on the 52 MHz drive signal. The necessity of the –90° phase shift relative to f_S can be seen from the equation of damped harmonic motion $\ddot{x} + a\dot{x} + bx = 0$ with the solution $x = A\sin(\omega_S t - \phi)e^{-at}$. The damping term $a\dot{x}$ is proportional to the time derivative of the solution x, i.e. a phase shift of –90°. The correction signal will coincide with the corresponding batch in the cavity if the total delay $\delta = t_f + nT_{rev}$, where t_f is the transit time from the beam monitor to the cavity, n an integer, and $T_{rev} = 7.7$ µs is the particle revolution time in PETRA. Since $T_S \gg T_{rev}$, a delay even of more than one turn ($n > 1$) would not be critical.

Rather than using a simple RC integrator of differentiator network as a 90° phase shifter, which is not without problems [14], a more complex digital solution with a software controlled phase shift has been adopted. This is very attractive since during injection, acceleration, and compression of the bunches the synchrotron frequency varies in the range from 200 Hz to 350 Hz. In addition, storing and multiplexing the eight correction signals for each of the eight possible batches in PETRA II can also be realized most comfortably on the digital side. The phase shifter has been built up as a three-coefficient digital Finite-length Impulse Response (FIR) filter according to

$$g_\mu = \sum_{k=0}^{2} h_k f_{\mu-k} \tag{52}$$

with an amplitude response

$$H(\omega) = \sum_{k=0}^{2} h_k e^{-ik\omega T_S} \quad , \tag{53}$$

where f and g are input and output data respectively. Using the coefficients $h_0 = \frac{2}{\pi}\sin\phi$, $h_1 = \cos\phi$, $h_2 = -\frac{2}{\pi}\sin\phi$ one obtains a phase shift that, in the frequency range of interest 200 Hz $\le f_S \le$ 359 Hz, deviates by less than ±0.4 from the nominal value $\phi = -\pi/2$ in accordance with Eqs. (52) and (53). The frequency dependence of the phase shift is mainly due to the delay in the filter, which is of the order of 1 ms, i.e. two sampling periods. It can always be corrected by software, if necessary. The amplitude response is constant within a few per cent for all frequencies.

A block diagram of the filter is shown in Fig. 9. The synchrotron phase information of the eight batches is sampled at intervals T_S = 0.5 ms and passed through eight times three shift registers. The three coefficients are stored in ROMs and are appropriately combined with the phase information. So the first filter output is available after three sampling periods and is then renewed every 0.5 ms.

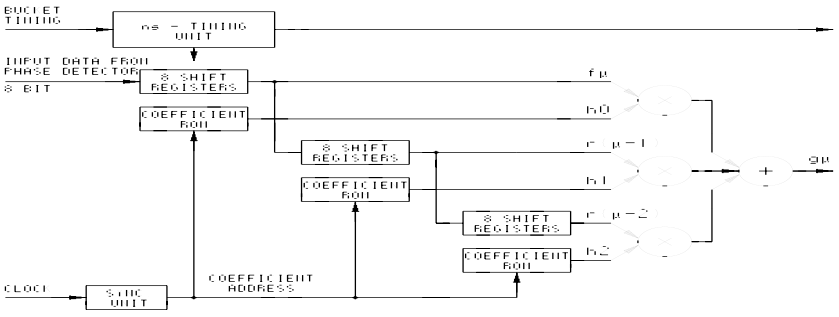


Fig. 9: Block diagram of the FIR filter. From three successive sampling periods the averaged phase signals for the eight proton batches in PETRA II are stored in shift registers and combined with the three coefficients, which are stored in ROMs. The first phase-shifted output is available after three sampling periods of 0.5 ms and is renewed every sampling period.

6.4 Performance of the phase loop

The performance of the loop is demonstrated in Fig. 8, where the phase detector output recorded by a storage scope is displayed. Complete damping of the synchrotron oscillation is achieved within less than one period. This corresponds to a damping time of less than 4 ms. If the loop is operated in the antidamping mode, the beam is lost within some milliseconds. With the loop, losses of the proton beam in PETRA II during energy ramping could be significantly reduced.

REFERENCES

[1] R.E. Collin, *Foundations for Microwave Engineering* (McGraw-Hill, New York, 1966).

[2] P.B. Wilson, CERN ISR–TH/7823 (1978).

[3] D. Boussard, CERN, SPS/86–10 (ARF) (1986).

[4] R.D. Kohaupt, 'Dynamik intensiver Teilchenstrahlen in Speicherringen', Lecture Notes, DESY (1987).

[5] A. Piwinski, DESY H 70/21 (1970).

[6] K.W. Robinson, CEA Report CEAL-1010 (1964).

[7] F. Pedersen, *IEEE Trans. Nucl. Sci.* **NS 32** (1985) 2138.

[8] D. Boussard, CERN SPS/85–31 (ARF) (1985).

[9] F. Pedersen, *IEEE Trans. Nucl. Sci.* **NS 22** (1975) 1906.

[10] E. Vogel, 'Ingredients for an RF Feedforward at HERA', DESY HERA 99–04 (1999) p. 398.

[11] S. Simrock, private communication, DESY (1999).

[12] M. Liepe, Diploma Thesis, DESY (2000).

[13] A. Gamp, W. Ebeling, W. Funk, J.R. Maidment, G.H. Rees, C.W. Planner, *in* Proc. 1st European Particle Accelerator Conference (EPAC 1), Rome, 1988, Ed. S. Tazzari (World Scientific, Singapore, 1989).

[14] A. Gamp, *in* Proc. 2nd European Particle Accelerator Conference (EPAC2), Nice, 1990, Eds. P. Marin and P. Mandrillong (Ed. Frontières, Gif-sur-Yvette, 1990).

RF GYMNASTICS IN SYNCHROTRONS

R. Garoby
CERN, Geneva, Switzerland

Abstract
The RF systems installed in synchrotrons can be used to change longitudinal beam characteristics. 'RF gymnastics' designates manipulations of the RF parameters aimed at providing such non-trivial changes. Some keep the number of bunches constant while changing bunch length, energy spread, emittance or distance between bunches. Others are used to change the number of bunches. After recalling the basics of longitudinal beam dynamics in a hadron synchrotron, this paper deals with the most commonly used gymnastics. Their principles are described as well as performance and limitations.

1. INTRODUCTION

RF systems in synchrotrons are primarily specified for beam acceleration in variable energy machines or for bunching in accumulators. At a later stage of design, and quite often after the machine is built, it frequently becomes necessary to tailor further the longitudinal beam characteristics like bunch length, energy spread, distance between bunches, number of bunches etc. 'RF gymnastics', involving the modulation of the RF parameters, are then considered to help obtain the required performance [1].

As the high-energy frontier gets higher and higher, the cost of an accelerator complex increases, as does the interest in gymnastics, which allows the adaptation of such a facility for purposes not originally foreseen.

2. LONGITUDINAL BEAM DYNAMICS

2.1 Conventions

Synchrotron radiation will not be considered, so the following analysis is only relevant for hadrons.

The longitudinal phase plane has time (or phase) as the *x*-axis and energy (or momentum) as the *y*-axis. The following variables characterize a particle:

- Charge: q
- Rest energy, energy: E_0 , E
- Speed, momentum: v, p
- Relativistic parameters: γ (γ=E/E_0), β (β=v/c)
- Revolution period in the synchrotron: T

The synchrotron parameters are the following:

- Momentum compaction factor, transition gamma: α_P, γ_T
- Parameters of the synchronous particle: E_S , v_S , p_S , γ_S , β_S , T_S . The synchronous particle is defined as the particle whose energy E_S and phase ϕ_S (measured with respect to the zero crossing with positive slope of the sinusoidal RF waveform at the lowest harmonic, h_1) are such that the particle sees the same accelerating voltage over successive turns in the accelerator.

The total voltage $V(t)$ results from the contributions of RF systems with voltages $V_1(t)$, $V_2(t)$, etc:

$$V(t) = \sum_{i=1}^{n} V_i(t) \ . \tag{1}$$

If resonant structures are used, the voltage functions are sine waves with h_i periods per revolution and a relative phase θ_i.

$$V_i(t) = \hat{V}_i \sin(h_i \omega_R t + \theta_i) \quad \text{, with} \quad \omega_R = \frac{2\pi}{T_S} \ . \tag{2}$$

2.2 Motion in the longitudinal phase plane

2.2.1 Equations of motion

The motion of particles is analysed in the frame of the synchronous particle. The *x*-coordinate is the phase difference $\Delta\phi = \phi - \phi_S$ measured at the lowest harmonic (h_1), and the energy coordinate is $\Delta E = E - E_S$ (or $\Delta p = p - p_S$). The tracked and the synchronous particles having different revolution periods, the phase difference $\Delta\phi$ changes at every revolution according to Eq. (3):

$$d\Delta\phi = 2\pi h_1 \frac{(T - T_S)}{T_S} = 2\pi h_1 \frac{\Delta T}{T_S} \ . \tag{3}$$

The rate of change of the phase is then:

$$\frac{d\Delta\phi}{dt} = \frac{2\pi h_1}{T_S} \frac{\Delta T}{T_S} \ . \tag{4}$$

In a synchrotron the relative difference in revolution period is proportional to the relative difference in momentum or energy:

$$\frac{\Delta T}{T_S} = \eta \frac{\Delta p}{p_S} = \frac{\eta}{\beta^2} \frac{\Delta E}{E_S} \ , \tag{5}$$

where $$\eta = \frac{1}{\gamma_T^2} - \frac{1}{\gamma^2} .$$

From Eqs. (4) and (5), the x component of the particle speed is given by

$$\frac{d\Delta\phi}{dt} = 2\pi h_1 \eta \frac{1}{\beta^2 T_S} \frac{\Delta E}{E_S} \ . \tag{6}$$

The y component of the particle speed is the rate of change of its energy with respect to the synchronous particle and is given by

$$\frac{d\Delta E}{dt} = \frac{q}{T_S} \left[V(\Delta\phi + \phi_S) - V(\phi_S) \right] . \tag{7}$$

2.2.2 Case of a single RF harmonic

When a single RF system is used, the voltage can be expressed as

$$V(\phi) = \hat{V} \sin\phi \tag{8}$$

and Eq. (7) simplifies into

$$\frac{d\Delta E}{dt} = \frac{q}{T_S}\hat{V}\left[\sin(\Delta\phi + \phi_S) - \sin\phi_S\right]. \tag{9}$$

The motion described by Eqs. (6) and (9) has the following first integral characterizing closed trajectories of particles oscillating around the synchronous one:

$$\frac{1}{2}\left(\frac{d\Delta\phi}{dt}\right)^2 + \frac{2\pi h_1 \eta q \hat{V}}{\beta^2 T_S^2 E_S}\left[\cos(\Delta\phi + \phi_S) + \Delta\phi \sin\phi_S\right] = \text{constant}\ . \tag{10}$$

There is a limit to the amplitude of these oscillations. The corresponding trajectory is called the separatrix, and the enclosed region is the bucket, whose area is the acceptance. The separatrix crosses the phase axis at the extreme phase elongation

$$\Delta\phi_{EXT1} = \pi - 2\phi_S\ . \tag{11}$$

The other extreme phase elongation is the solution of:

$$\cos(\Delta\phi_{EXT2} + \phi_S) + \Delta\phi_{EXT2}\sin\phi_S = -\cos\phi_S + (\pi - 2\phi_S)\sin\phi_S\ . \tag{12}$$

The extreme excursion in energy is obtained when $\Delta\phi = 0$ rad

$$\Delta E_{MAX} = \sqrt{\frac{E_S\beta^2}{\pi h_1 \eta} q\hat{V}\left[(\pi - 2\phi_S)\sin\phi_S - 2\cos\phi_S\right]}\ . \tag{13}$$

Figure 1 illustrates the case of a stationary bucket (constant B field in the main dipoles and no acceleration of the synchronous particle) below transition energy (ϕ_S = 0 rad). The separatrix extends from –π to +π radians. The speed of a moving particle inside the bucket is shown. If it is an extreme particle of a stable population, its trajectory is the contour enclosing all others. This set of particles is called a bunch and the area inside the contour is its emittance.

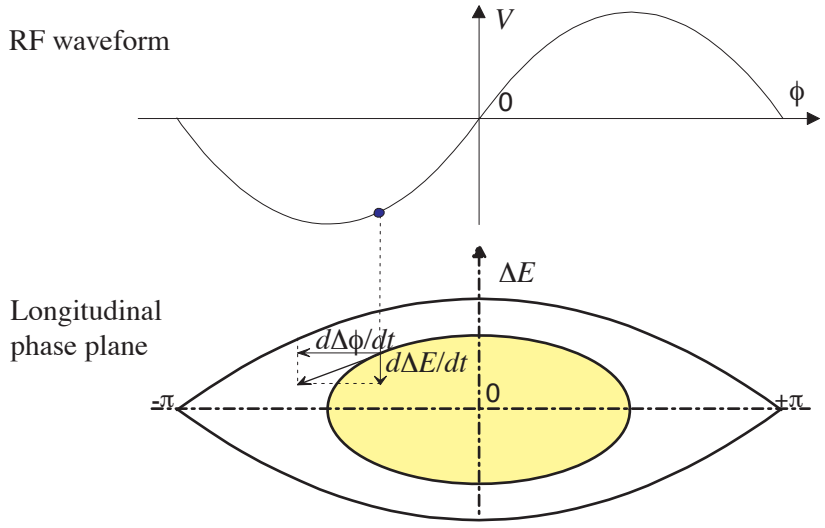


Fig. 1: Trajectories in a stationary bucket

For small amplitude of oscillation, Eqs. (6) and (9) represent a simple harmonic oscillator at the synchrotron frequency ω_S

$$\omega_S = \sqrt{\frac{2\pi h|\eta| q\hat{V}\cos\phi_S}{\beta^2 T_S^2 E_S}} \tag{14}$$

At constant emittance, the peak excursions in phase and energy scale like

$$\Delta\hat{\phi} \propto k \quad \text{and} \quad \Delta\hat{E} \propto \frac{1}{k} \quad \text{with} \quad k = \left[\frac{|\eta|}{E_S q \hat{V} \cos\phi_S} \right]^{\frac{1}{4}} \tag{15}$$

2.3 Effect of changing RF parameters

2.3.1 Adiabaticity

If the RF parameters are changed at a slow rate with respect to the smallest frequency of oscillation of the particles in the bunch, the distribution of particles is continuously at equilibrium and only depends upon the instantaneous value of these parameters. Such an evolution is called 'adiabatic'. The degree of adiabaticity is assessed with the adiabaticity parameter [2] defined as

$$\varepsilon = \frac{1}{\omega_S^2} \left| \frac{d\omega_S}{dt} \right| \tag{16}$$

A process is typically considered adiabatic when $\varepsilon < 0.1$.

2.3.2 Liouville's Theorem

The longitudinal motion that we consider is conservative (i.e. there is no energy dissipation effect like synchrotron radiation). Liouville's Theorem is therefore applicable. This states that the local density of particles in the longitudinal phase plane is always constant [3]. An implicit consequence is that any RF gymnastics is in principle reversible.

When an adiabatic process is used, this helps determine the particle distribution (or bunch shape) in the final state without having to take into account intermediate states (Fig. 2). The area occupied by particles ('emittance') is constant and always limited by a stable trajectory.

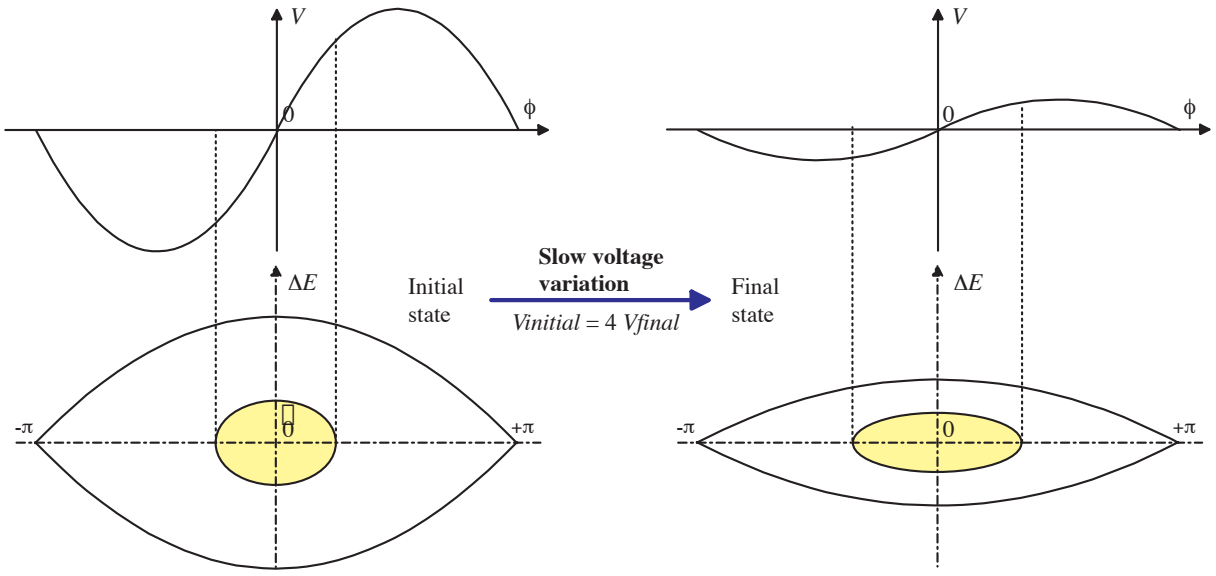


Fig. 2: Adiabatic RF voltage reduction

When a non-adiabatic gymnastic is applied, the consequences are less obvious and detailed tracking is required to evaluate the final particle distribution (Fig. 3). Although the area occupied by particles is also constant, its contour is usually not a stable trajectory in the final state. The final emittance generally has to be considered as increased to the value of the smallest area, limited by a stable trajectory that contains all particles ('macroscopic' emittance).

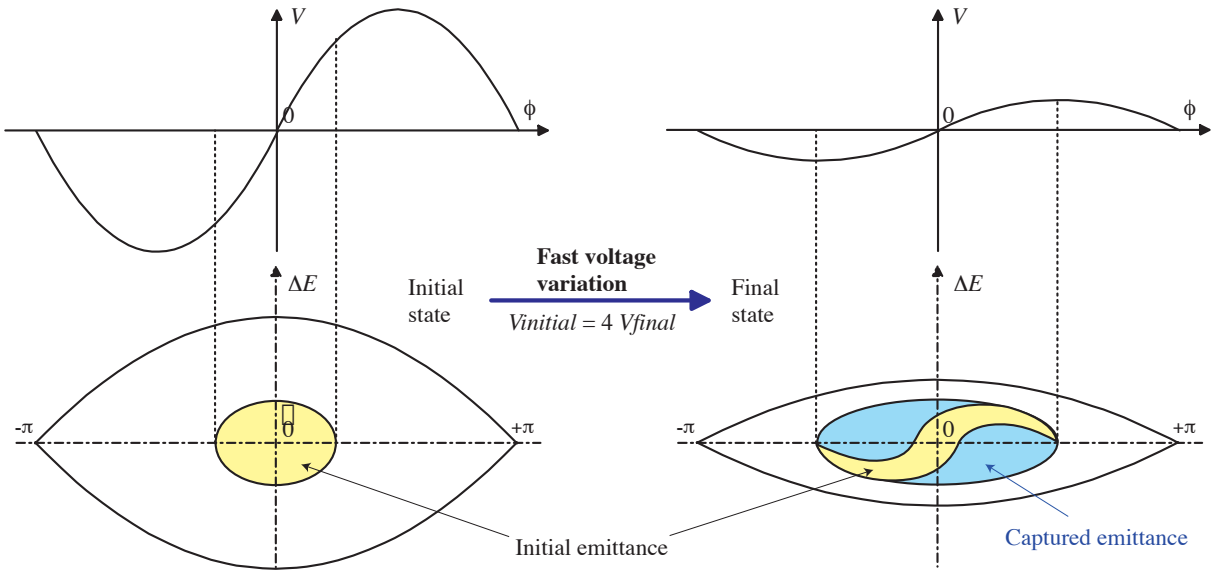


Fig. 3: Non-adiabatic RF voltage reduction

3. SINGLE BUNCH GYMNASTICS

3.1 Bunch compression

To preserve the longitudinal emittance and guarantee reproducible beam performance, the contour of the bunches entering a synchrotron must correspond to stable trajectories in the longitudinal phase plane of that machine. Such a condition is called 'longitudinal matching'. This often requires changing the ratio bunch length/energy spread of the bunches in the previous machine, and generally bunches must be made shorter. When adiabatic variation of the RF voltage cannot be used to provide the proper beam characteristics, non-adiabatic processes are applied. The corresponding gymnastics are called 'bunch compression', 'bunch rotation' or even 'phase rotation' [4, 5].

The principle (Fig. 4) is to let a bunch, initially elongated in phase, rotate in a maximum height bucket, and to eject it when it is at its shortest. Even with a single RF system, various techniques can be used for stretching the bunch:

- Reducing adiabatically the RF voltage V, the bunch length increases in proportion to $V^{-1/4}$ (Eq. (15)). This technique has the drawback of requiring a very large dynamic range in V and of becoming very slow in order to remain adiabatic at low voltages (see Eqs. (14) and (16)).
- Reducing abruptly the RF voltage, a bunch rotation is triggered, which provides, after a quarter of a turn in the phase plane, a bunch length proportional to $V^{-1/2}$. This process is more rapid and efficient than the previous one, but it is more demanding for the transient response of the cavity and the beam servo-loops.
- Switching by π radians the phase of the RF, the bunch becomes centred on the unstable phase and stretches quickly along the separatrix. This technique is also rapid and does not in principle require any voltage change, but it needs a rapid response from the RF system. The fact that the resulting bunch is tilted with respect to the phase axis implies that it will suffer more from non-linearities when rotating in the phase plane for compression.

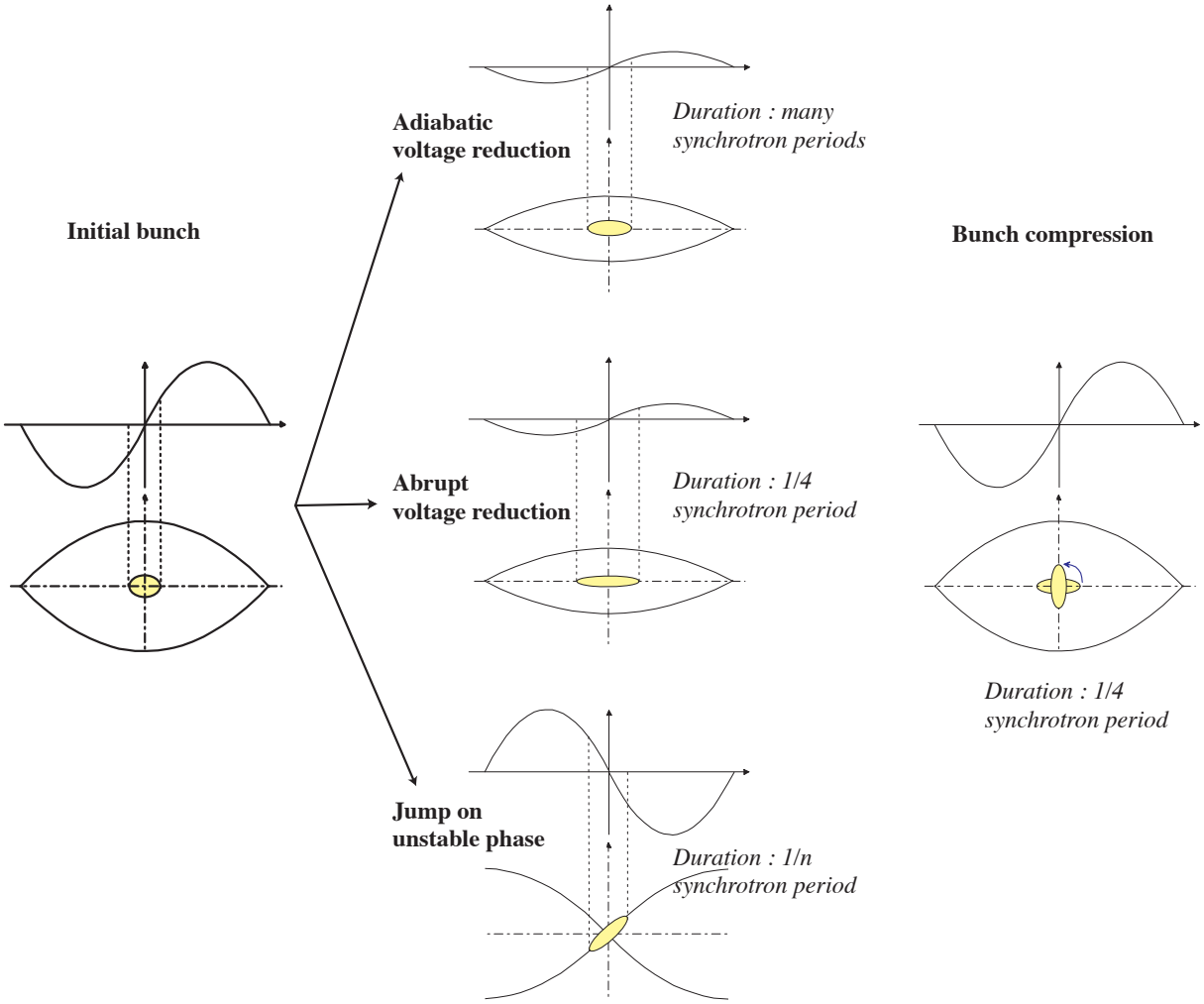


Fig. 4: Bunch compression

The performance of the compression process depends upon the bunch length and the normalized bunch emittance (ratio between emittance and acceptance) during the rotation of the elongated bunch. This is illustrated in Fig. 5, which shows the bunch at the beginning and at the end of rotation. An initially extreme particle along the energy axis (B0) becomes extreme in phase after rotation (B1) under the effect of a quasi-linear focusing voltage, approximated by the tangent at zero phase of the RF sine wave. In contrast, an initially extreme particle along the phase axis (A0) experiences a non-linear focusing voltage during rotation that is always smaller than the previous tangent, and results in a slower motion. In the time it takes for B0 to move to B1, A0 only moves to A1.

For a given normalized emittance, the minimum bunch length is obtained approximately when A1 and B1 are at the same phase. This defines an optimum initial bunch elongation, represented in Fig. 6. This figure also shows the minimum length achieved after rotation and the equilibrium length of a bunch of the same emittance in the rotation bucket. A compression efficiency can be defined as the ratio between that equilibrium bunch length and the length after rotation in optimum conditions. This efficiency is also shown.

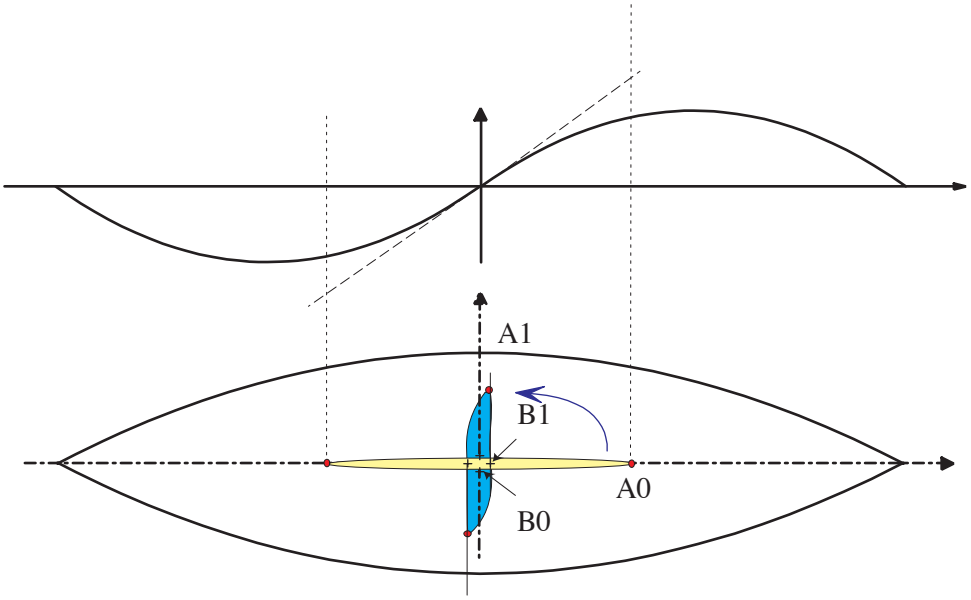


Fig. 5: Optimum bunch rotation

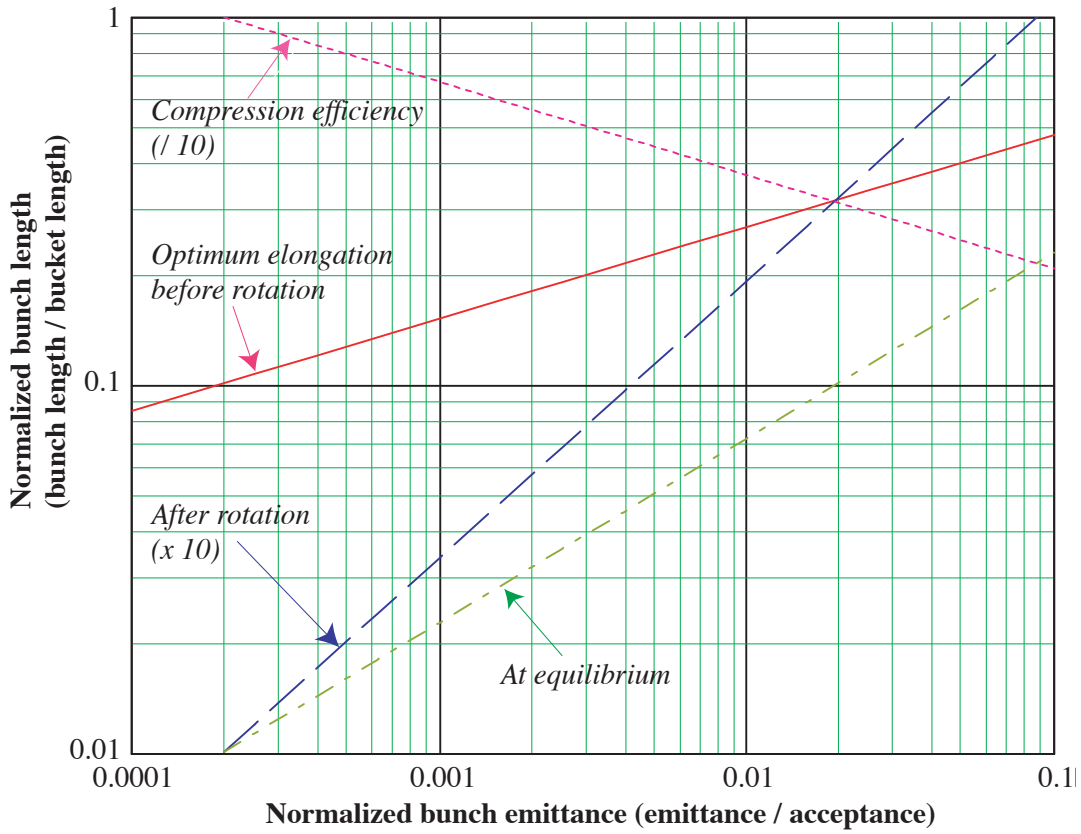


Fig. 6: Bunch rotation parameters

Numerous refinements are possible, using multiple RF harmonics or applying phase and amplitude modulations.

3.2 Longitudinal controlled blow-up

Blow-up techniques have been developed to help stabilize the beam in high-intensity accelerators. They increase the 'macroscopic' emittance in a controlled way while providing an adequate

distribution of particles with sharp edges and no tails. A typical and commonly used technique is based on the superposition of a phase-modulated high frequency (V_H, h_H) on the RF normally holding the beam (V_1, $h_1 << h_H$) [6, 7].

The high frequency phase-modulated voltage can be expressed as

$$V_H = \hat{V}_H \sin(h_H \omega_R t + \alpha \sin \omega_M t + \vartheta_H), \tag{17}$$

α being the peak phase modulation, ω_R the modulation frequency, and ϑ_H a phase constant.

This acts as a perturbation to the motion of particles in the bucket of the main RF system. Resonances can be induced that create a redistribution of density in the bunch. Large non-linearities in the motion accelerate filamentation and contribute to the fast disappearance of the density modulations induced by the high-frequency carrier. Of the different distributions that can be obtained, parabolic ones are generally preferred.

The blow-up parameters are in practice optimized either on the real accelerator or using computer simulations. The typical range of values applied in such cases is shown in Table 1.

Smaller harmonic ratio are sometimes used because of hardware constraints, and good quality blow-up can still be obtained after a longer duration.

Table 1: Typical blow-up parameters

	$\frac{\hat{V}_H}{\hat{V}_1}$	$\frac{h_H}{h_1}$	α (rad)	$\frac{\omega_M}{\omega_S}$	Duration
Typical range	0.05 to 0.2	> 10 for fast blow-up	0.8π to 1.2π	2 to 7	$\geq 10\frac{2\pi}{\omega_S}$

4. MULTI-BUNCH GYMNASTICS

4.1 Debunching–rebunching

Debunching–rebunching is the most conventional way to change the number of bunches [5, 8]. It has to take place at constant energy and hence at constant field in the main bending dipoles because of the absence of RF for a significant period of time. At the end of debunching the beam is continuous and ideally has not undergone any azimuthal modulation of the linear density of particles. Rebunching is the reverse process, during which a different RF harmonic number is used, and the beam progressively gets an azimuthal modulation of density and is finally fully bunched on the new harmonic.

Iso-adiabatic debunching is generally used to minimize longitudinal emittance blow-up. The reduction of the RF voltage from V_{I_deb} to V_{F_deb} occurs at constant adiabaticity (see Eq. (16)):

$$V(t) = \frac{V_{I\_deb}}{\left[1 - \left(1 - \sqrt{\frac{V_{I\_deb}}{V_{F\_deb}}}\right)\frac{t}{t_R}\right]^2}, \tag{18}$$

where t_R is the moment of suppression of the RF voltage after reaching the minimum controllable level V_{F_deb}. This is illustrated in Fig. 7. The process takes more time when V_{F_deb} is made smaller:

$$t_R \approx \frac{1}{\omega_S(V_{F\_deb})} \propto \frac{1}{\sqrt{V_{F\_deb}}}. \tag{19}$$

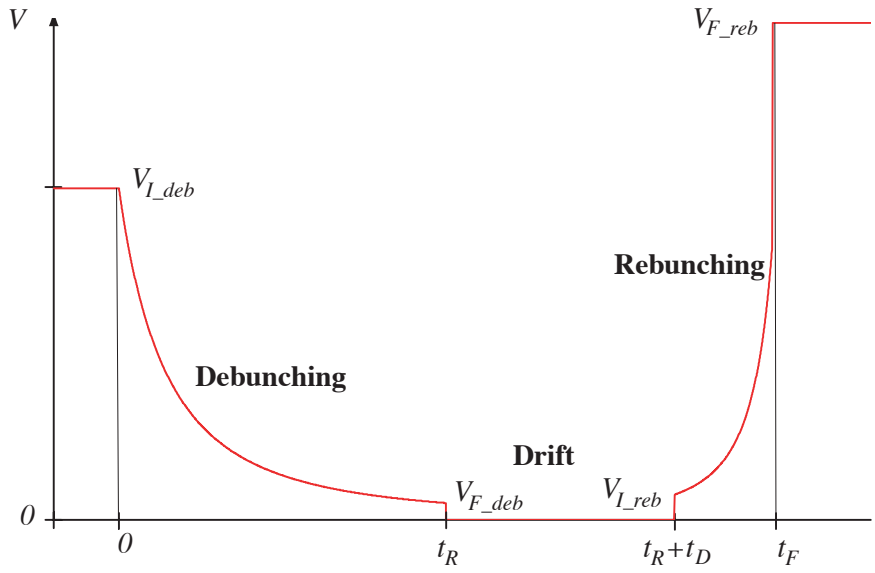


Fig. 7: Voltages for iso-adiabatic debunching–rebunching

During this voltage reduction, the bunch progressively lengthens (proportionally to $V^{-1/4}$ at the beginning according to Eq. (15)). Under the voltage V_{F_deb} the beam is generally still bunched and some time, t_D, is required without voltage for the particles to drift in azimuth and for debunching to be obtained. This results in a blow-up of the macroscopic emittance, which depends upon the normalized bunch emittance in the final bucket (as shown in Fig. 8). In the typical case, where the bunch finally completely fills the bucket, the emittance is multiplied by π/2.

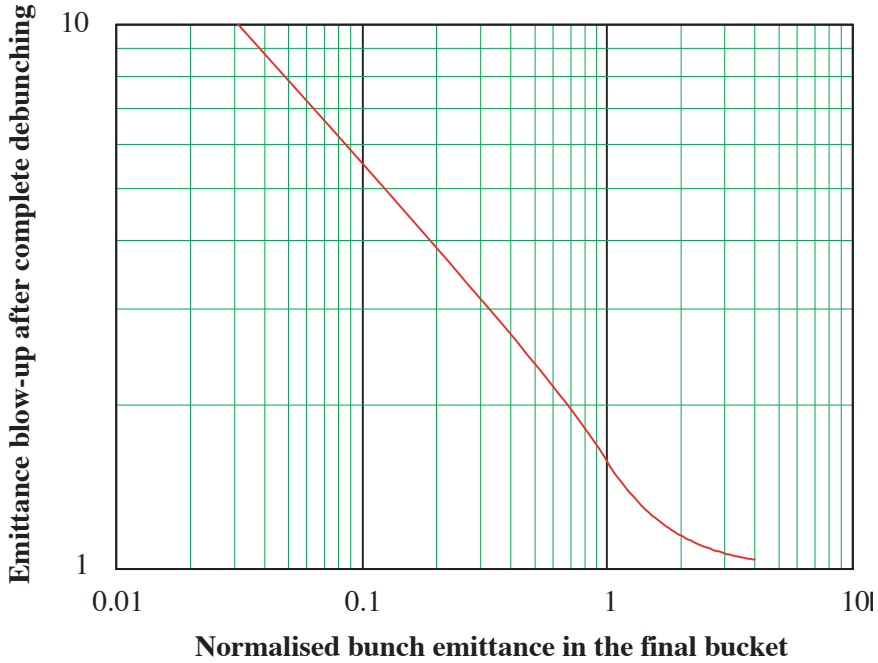


Fig. 8: Emittance blow-up after iso-adiabatic debunching

A reference debunching time can be defined as the time taken for the particles of successive bunches to begin to overlap in azimuth:

$$t_{D\_classic} = \frac{\pi - \Delta\phi}{h\omega_R |\eta| \frac{\Delta p}{p}} \tag{20}$$

where $\Delta\phi$ and Δp are the full spreads in phase and momentum of the bunch under V_{F_deb}.

A proper debunching requires $t_D >> t_{D\_classic}$.

Iso-adiabatic rebunching is generally used after debunching is completed. It is a time-reversed version of iso-adiabatic debunching, starting abruptly at V_{1_reb} and rising progressively to V_{F_reb}. Similar formulae apply.

4.2 Splitting (Merging)

Splitting is used to multiply the number of bunches by two or three and merging is the reverse process [9, 10]. Although limited in use to circumstances where such ratios are possible, it is an interesting technique compared to iso-adiabatic debunching–rebunching because it can really be made quasi-adiabatic and preserve the emittance.

Splitting bunches in two is obtained using two RF systems with an harmonic ratio of two simultaneously. The bunch is initially held by the first system (V_1, h_1) while the second (V_2, h_2=2 h_1) is stopped. The unstable phase on the second harmonic is centred on the bunch. As the voltage V_2 is slowly increased and V_1 is decreased the bunch lengthens and progressively splits in two as illustrated in Fig. 9.

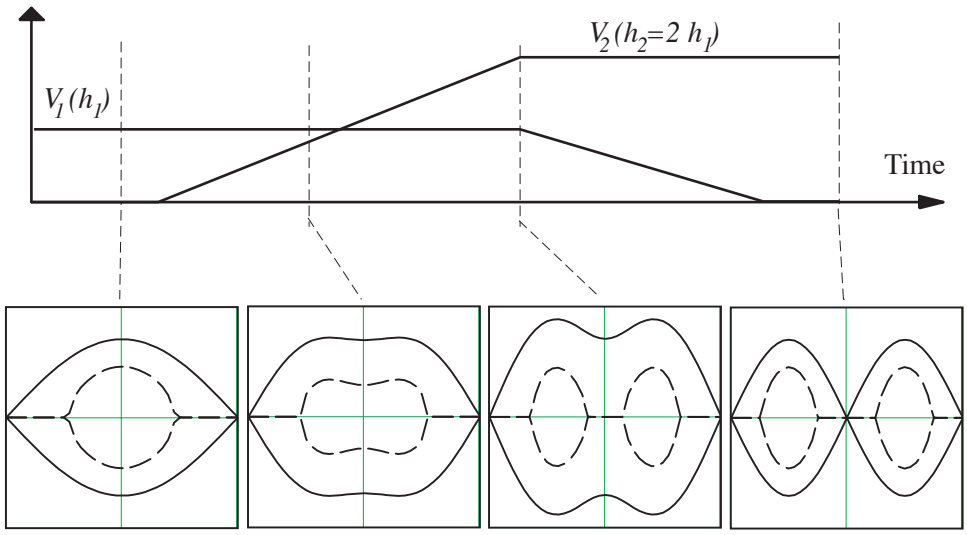


Fig. 9: Bunch splitting in two

When two separate bunches have just begun to form, the voltage on the first harmonic is V_{1_sep}. The normalized emittance at that time is defined as

$$\varepsilon_{sep} = [\text{Total emittance/Acceptance}(V_{1\_sep}, h_1)] \,. \tag{21}$$

The ratio $V_2/\ V_{1\_sep}$ at this moment is given in Fig. 10. Good results are consistently obtained for ε_{sep} ~1/3 and using voltage variations thar are linear functions of time with a total duration larger than five synchrotron periods in the bucket (V_{1_sep}, h_1). Each final bunch has half the emittance of the initial bunch, and almost no blow-up is observed.

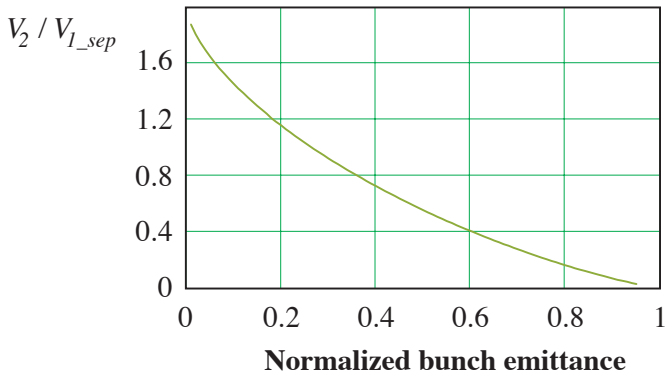


Fig. 10: V_2/V_{1_sep} at the time of bunch separation

An operational implementation of double splitting in the CERN PS is illustrated in Fig. 11. A bunch on h = 8 is split in two on h = 16 within 25 ms and no blow-up can be observed. On the left side of the same figure, the evolution of particle density in the longitudinal phase plane during the process is reconstructed using longitudinal tomography [11].

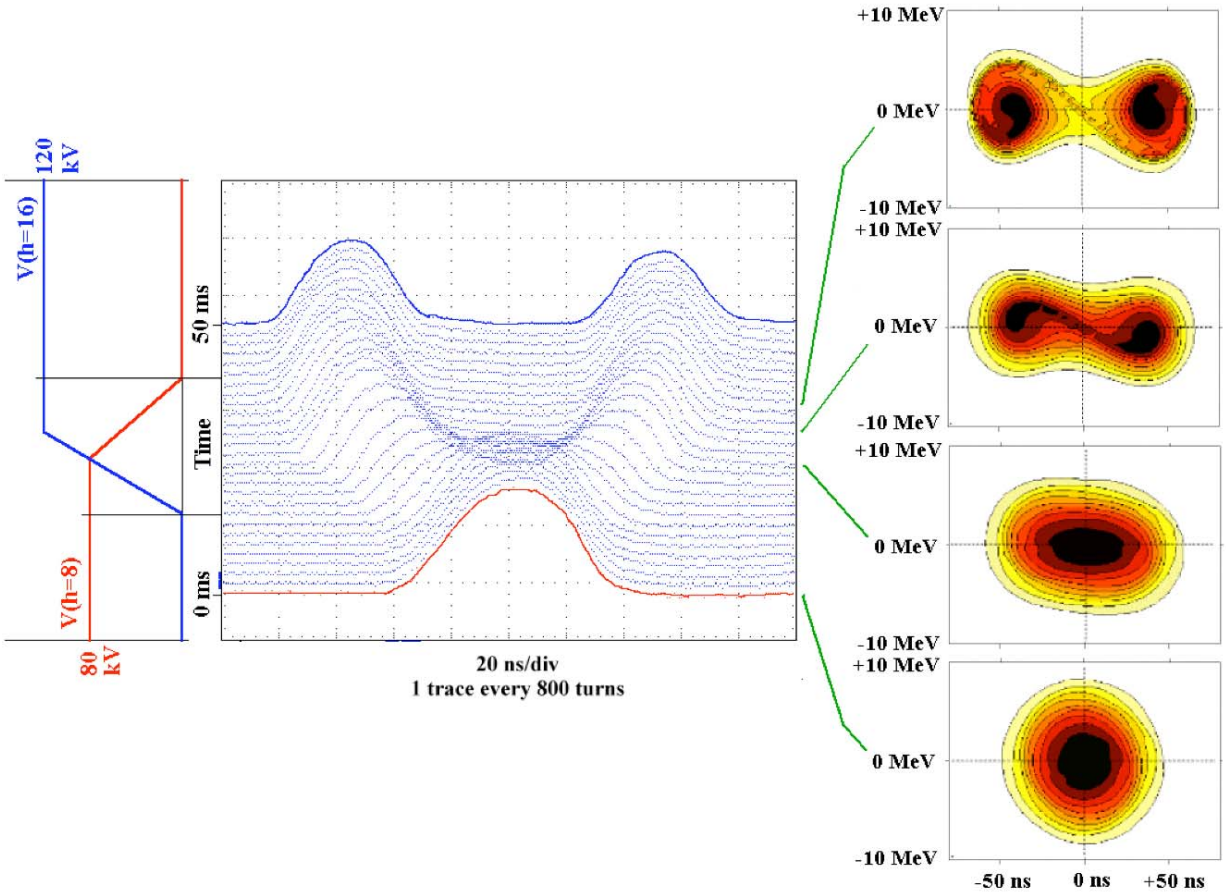


Fig. 11: Example of bunch double splitting from h = 8 to h = 16 in the CERN PS at 3.57 GeV/*c*

Splitting bunches in three has also been demonstrated using three simultaneous RF systems. The relative phases between harmonics as well as the voltage ratios must be precisely controlled for the particles to split evenly into the new bunches and longitudinal emittance to be preserved. As good results as for bunch double splitting have been achieved, and final bunches are one-third the emittance of the original one.

4.3 Batch compression

Batch compression is a process that keeps the number of bunches constant while concentrating them in a reduced fraction of the accelerator circumference [12]. When exercised at a slow enough rate it can be adiabatic and consequently preserve longitudinal emittance.

The principle is slowly to increase the harmonic number of the RF controlling the beam, as shown in Fig. 12. Starting from harmonic h_0, voltage is progressively increased on harmonic $h_1 > h_0$ and decreased to 0 V on h_0, so that harmonic h_1 finally holds the bunches. The phase on h_1 with respect to h_0 must be such that the bunches converge symmetrically towards the centre of the batch.

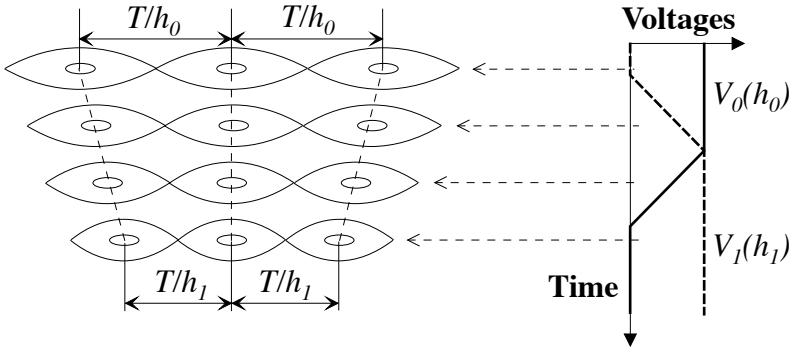


Fig. 12: Batch compression

The amount of compression achievable in a single step is limited by the need to maintain a large enough acceptance for the buckets holding the edge bunches. A consequence is that large compression factors are only obtained after multiple batch compression steps, and complicated manipulations of RF parameters are involved. A typical application is given in Fig. 13, where four bunches on h = 8 are finally brought into four adjacent buckets on h = 20: three groups of RF cavities are used, which help sweep progressively the harmonic seen by the beam from 8 to 20 in steps of two units.

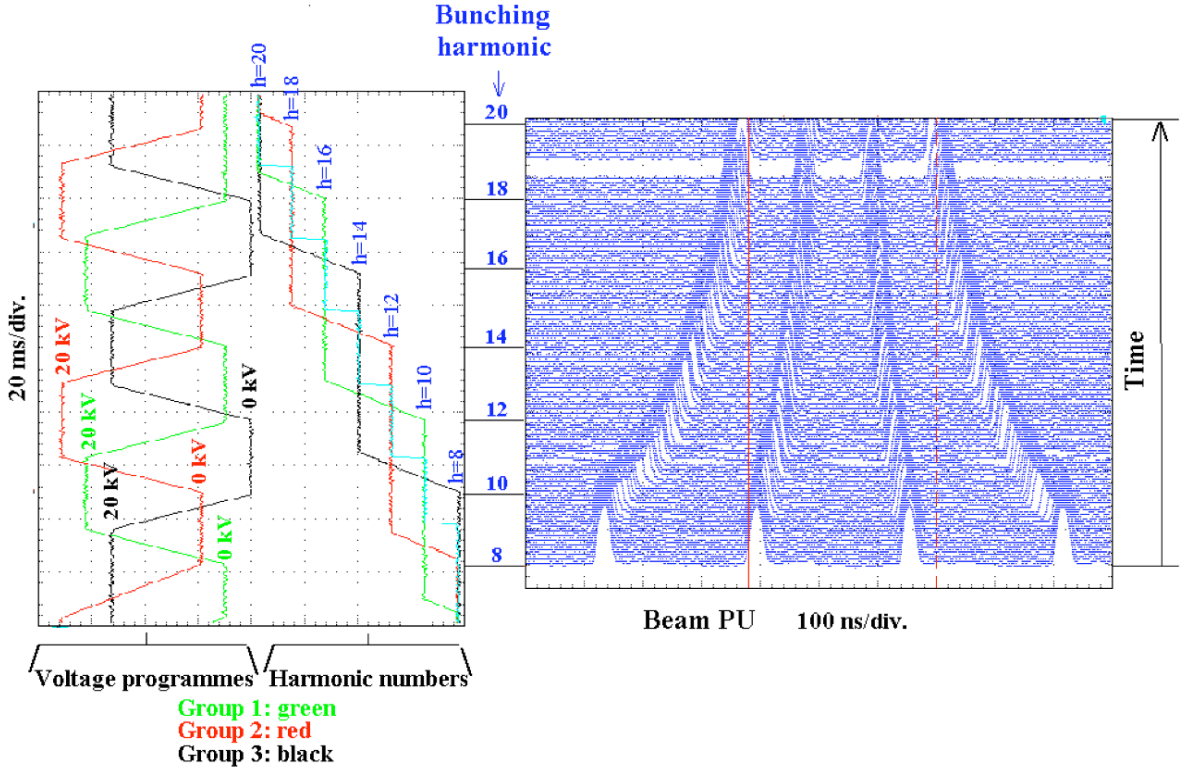


Fig. 13: Example of batch compression from h = 8 to h = 20 in the CERN PS at 26 GeV

4.4 Slip stacking

Slip stacking is used to superimpose two sets of bunches and double the bunch population [13, 14]. It is non-adiabatic and leads to large emittance blow-ups.

The principle is sketched in Fig. 14. Two different RF frequencies are simultaneously applied. If their difference is large enough ($\Delta f > 2f_s$, where f_s is the synchrotron frequency in the centre of an unperturbed bucket of one family), two families of buckets coexist, which drift towards each other because of their frequency difference. Consequently, and provided the acceptance of these buckets ($f = h_0 f_{REV} \pm \Delta f$; V_{drift}) is large enough (acceptance > 2 × emittance), the bunches drift with them and tend to slip past each other. When they are superimposed in azimuth, pairs of bunches can be captured in large buckets centred at the middle frequency ($f = h_0 f_{REV}$; $V_{capture}$).

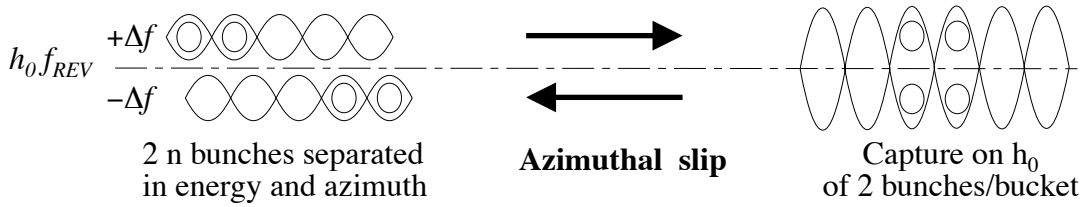


Fig. 14: Slip stacking

Although improvements can be made, like reducing the frequency difference towards the end of the process, the longitudinal contour enclosing a pair of bunches in the final bucket also contains a large area without particles. After filamentation, the macroscopic emittance is much more than doubled and longitudinal density is accordingly reduced.

5. DEBUNCHED BEAM GYMNASTICS

5.1 Barrier / isolated bucket

A single sine wave pulsing at the revolution frequency of the beam generates either an isolated or a barrier bucket, depending upon its polarity and the sign of η (Fig. 15).

In the case of the isolated bucket there is a stable ('synchronous') particle at the central zero crossing of the sine wave. Particles inside the sine wave period can be captured and execute closed trajectories around the particle. Particles outside this bucket move along the full circumference.

In the case of the barrier bucket, the central zero crossing of the sine wave is an unstable position. The stable region is limited by the other zero crossings and extends over all the circumference except the sine wave.

Such a voltage can be obtained from a wide band resonator driven by a high power amplifier or from a limited bandwidth resonator driven by a large current generator.

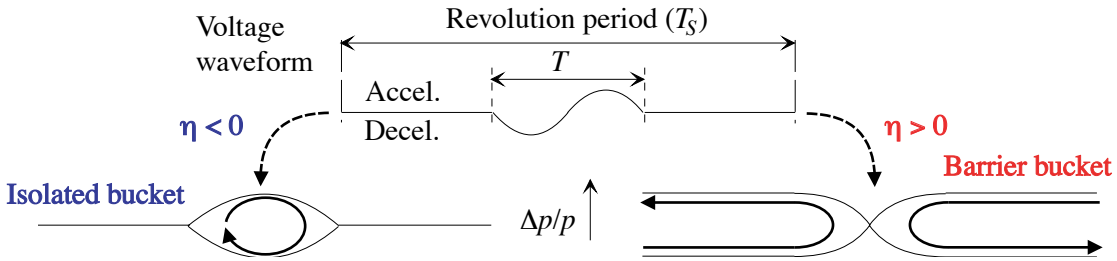


Fig. 15: Isolated / suppressed bucket

Beam dynamics is governed by the equations derived in Section 2.2. An isolated bucket is useful to capture a single bunch of small emittance in the debunched beam stack of an accumulator [8]. Barrier buckets are also typically used for high intensity accumulation, to preserve gaps without beam, and to permit lossless beam transfers [15].

5.2 Phase displacement acceleration

A debunched beam is accelerated (or decelerated) when being traversed by empty RF buckets [16]. This is due to emittance preservation for the empty volume captured by the RF buckets. The resulting change of the stack mean energy is given by

$$\Delta E_{Stack} = A_{bucket} \frac{h}{T} = A_{bucket} f_{RF} \tag{22}$$

A small voltage and a limited frequency range (of a few per cent) are sufficient, while a large beam current and emittance can be handled. However, the acceleration rate is small and the stack tends to degrade progressively as the number of traverses increases.

6. PRACTICAL IMPLEMENTATION

The possible implementation and the effective performance of RF gymnastics in synchrotrons are constrained by a number of practical limitations. In addition to those resulting from the basic hardware capabilities (number of simultaneous frequencies, minimum controllable voltages, etc.) a number of others must also be mentioned:

- Maximum duration at constant field in the dipoles. This may force the use of fast and non-adiabatic techniques or a degraded adiabaticity.
- Beam stability. The quality and reproducibility of performance of the final beam depends on the reproducibility of the initial conditions and the absence of collective beam instabilities during the process.
- Control of the RF parameters. The proper operation of the servo-loops (beam phase, radial, or synchronization loop) all along the gymnastics is critical for performance, and unavoidable transients must be minimized, with their delayed effect quickly damped. Moreover, for good performance at high beam intensity, beam loading in the RF cavities must be minimized, so local RF feedback and 'one-turn delay feedback' are often necessary.
- Variation of the dipole field during the gymnastics. This can be due to drift or ripple of the field in the main dipole, but also to changes of the orbit length induced by orbit bumps.

Solutions exist, but much time can be gained during setting up by a preliminary analysis of the likely disturbances and the direct implementation of adequate corrective measures.

REFERENCES

[1] A.W. Chao, M. Tigner (eds.), *Handbook of Accelerator Physics and Engineering* (World Scientific, 1999), p. 283.

[2] B.W. Montague, CERN 77–13, p. 63.

[3] M. Weiss, CERN 87–10, p.162.

[4] J. Griffin *et al.,* PAC 83, p. 2630.

[5] R. Garoby, CERN PS/RF/Note 93–17.

[6] V.V. Balandin *et al., Part. Accel*. 35 (1991) 1.

[7] R. Cappi, R. Garoby, E. Chapochnikova, CERN/PS 92-40 (RF).

[8] J. Griffin *et al.*, PAC 83, p. 2627.

[9] R. Garoby, S. Hancock, EPAC 94, p. 282.

[10] R. Garoby, CERN/PS 98–048 (RF).

[11] S. Hancock, P. Knaus, M. Lindroos, EPAC 98, p. 1520.

[12] R. Garoby, PAC 85, p. 2332.

[13] F.E. Mills, BNL Report AADD 176 (1971).
[14] D. Boussard, Y. Mizumachi, PAC 79, p. 3623.
[15] M. Biaskiewicz, J.M. Brennan, EPAC 96, p. 2373.
[16] E.W. Messerschmid, CERN/ISR-TH/73–31.

RADIO-FREQUENCY QUADRUPOLE LINACS

A. Schempp
Johann Wolfgang Goethe University, Frankfurt am Main, Germany

Abstract
Radio-Frequency Quadrupole (RFQ) linacs are efficient, compact, low-energy ion structures, which have found numerous applications. They use electrical RF focusing and can capture, bunch, and transmit high-current ion beams. Some recent developments and new projects such as heavy-ion injectors replacing Tandems as injectors for cyclotrons and linacs, RFQ post-accelerators, and the status of the work on high-energy implanters and small neutron sources are discussed.

1. INTRODUCTION

Injectors are combinations of an ion source, a Low-Energy Beam Transport (LEBT) system, an electrostatic pre-accelerator or an RFQ, and an intermediate section that matches the beam to a following structure, for example an IH- or an Alvarez accelerator.

The pre-accelerator defines the phase space for the following stages, in which the effective emittance will possibly increase. In the design of a high-current accelerator, the emittance and the current have to be optimized. The injector is the bottleneck because focusing forces are weak and the defocusing effects and non-linearities caused by space charge are strongest at low energies.

The development of the RFQ structure with its ability to bunch and accelerate low-energy, high-current ion beams opens new parameter possibilities for accelerator designs.

The variety of RFQ accelerators covers the full ion mass range from hydrogen to uranium, in the 5–500 MHz frequency range and duty factors up to 100%. The physics of the transport and acceleration of high-current ion beams in RFQs has been solved to such an extent that high-brilliance and high-current beams produced by ion sources and tranported in an LEBT can be captured, bunched, and transmitted with very small emittance growth by RFQs, as shown schematically in Fig. 1

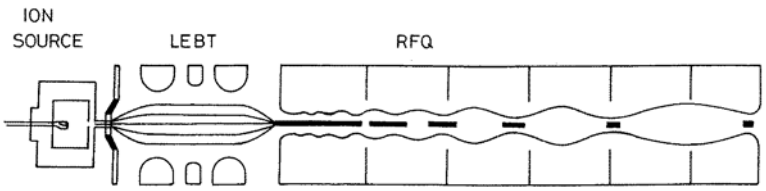


Fig. 1: Layout of an RFQ injector

Basically the RFQ is a homogeneous transport channel with additional acceleration. The mechanical modulation of the electrodes as indicated in Fig. 2 adds an accelerating axial field component, resulting in a linac structure that accelerates and focuses with the same RF fields.

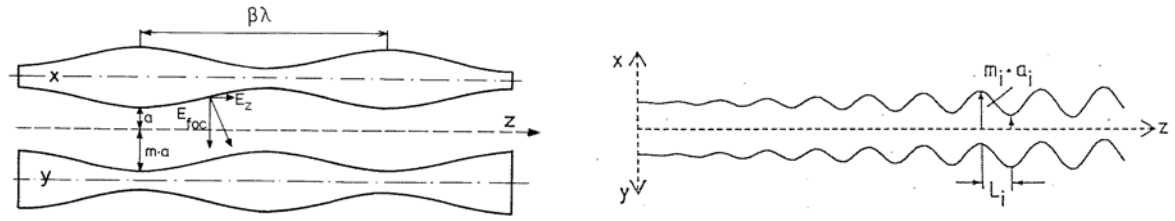


Fig. 2: RFQ electrodes

For a given injection energy and frequency the focusing gradients $G = X^* U_0/a^2$, ($X < 1$ for modulated electrodes) determine the acceptance in a low-current application. A maximum voltage U_0 has to be applied at a minimum beam aperture a if the radial focusing strength is the limiting factor. The highest possible operation frequency should be chosen to keep the structure short and compact. After the choice of U_0 and operating frequency, the 'RFQ design', the values of aperture a, modulation m, and the lengths L_C along the RFQ determine the electrode shape (pole tips), the ratio between accelerating and focusing fields, as indicated in Figs. 2 and 3, and hence the beam properties.

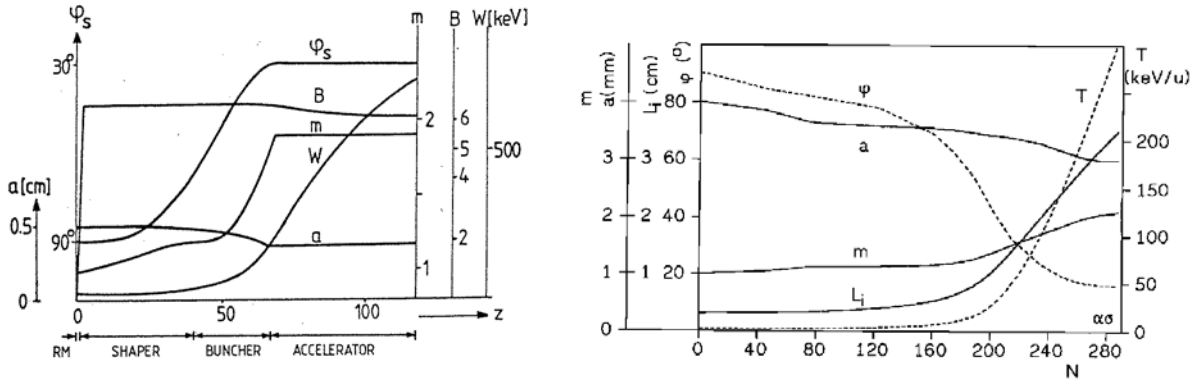


Fig. 3: Electrode designs of the first DESY RFQ and the GSI–HLI RFQ

The optimum frequency depends on many factors. In smaller projects it is the availability of the transmitters or a matching post-accelerator. Lower frequencies provide stronger focusing, less difficulties with power density and mechanical tolerances, and generally a higher current limit. Higher frequencies are favourable for compact designs with highest brilliance because the charge per bunch and the frequency jump to a final linac stage are smaller, but the currents are limited by the maximum focusing fields and sparking.

The RF structure has to generate the quadrupole fields with high efficiency and stability. The four-vane structure, which is mostly used in proton and H^- acceleration, is basically a TE_{211}-mode structure, in which the resonator has been loaded with electrodes to increase the quadrupole field, as shown in Fig. 4. The end region has to be modified to allow the magnetic fields to turn around and to shift the mode into a TE_{210} with a constant quadrupole field along the structure. Radio-frequency stability and thermal symmetry of the cavity, which can also be treated as four weakly-coupled resonators, are the reason for very tight mechanical tolerances that are a particularly limiting factor in high duty factor operation.

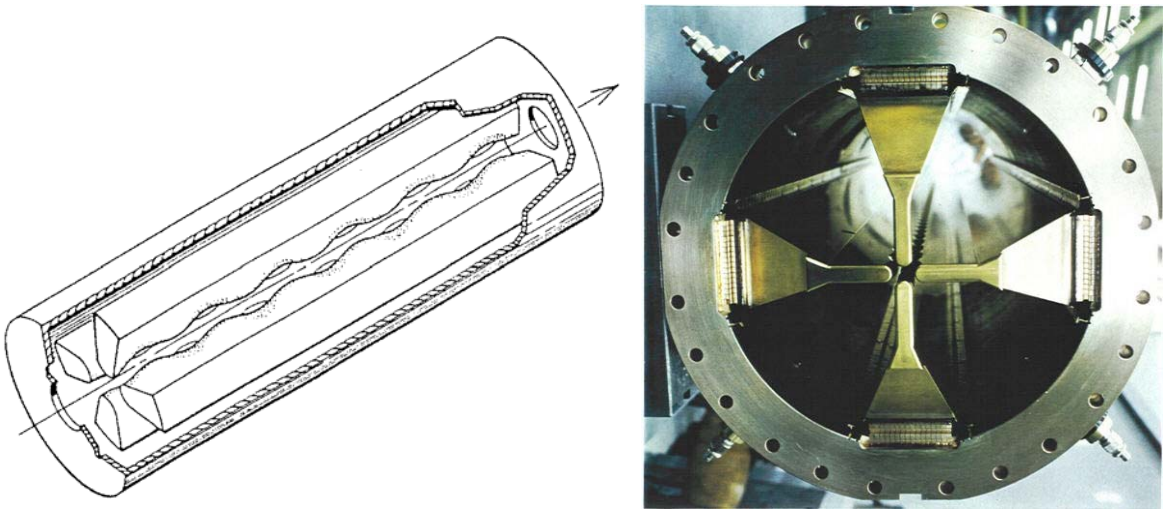


Fig. 4: The four-vane RFQ and a view of the CERN RFQ2 structure

The four-rod structure, shown in Fig. 5, consists of a linear chain of stems that form a chain of strongly coupled λ/2 resonators. By the direct connection of the electrodes with the same polarity, dipoles cannot be excited and tolerances are less stringent. The length of the stems can be changed to give compact resonators with rather low frequency that can also be used for low frequencies for low charge-to-mass-ratio heavy ions.

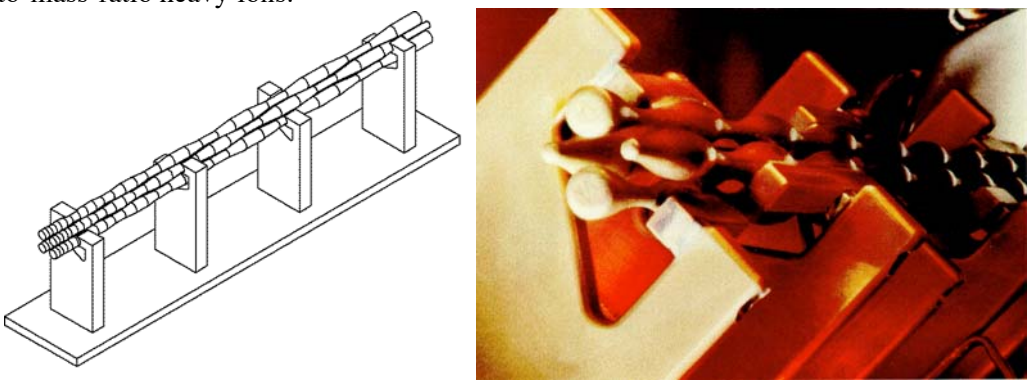


Fig. 5: The four-rod RFQ and a view of the DESY RFQ2 structure

The RFQ rods are driven by periodically arranged support stems that exite the electrodes in a TE_{210}-like mode. Inherently there are some longitudinal currents on these rods and a few per cent voltage modulation between the support stems. Like the azimuthal symmetry, this cannot be changed (improved or made worse) by tuning or heating during operation.

The rod electrodes can have a circular cross section (as used in the first low average power structures), parallel cooling channels, or a vane-tip shape to give a good approximation to the ideal two-term potential. In any case, as long as the field coefficients are calculated correctly, the electrode shape has no influence on the beam properties.

Generally, the required RF power N is independent of the frequency, whilst the acceptance and the maximum ion current are proportional to the electrode voltage and N^2, which is not a big issue in pulsed injectors with low average power. High duty factor operation is the present area of development. A first class of structures, which will be used as spallation source-linac injectors with duty factors of up to 10%, is presently being built, but more difficulties still have to be resolved for Continuous Wave (CW) RFQs.

2. RFQ APPLICATIONS

The standard application is operation as a pre-injector for an Alvarez linac feeding a synchrotron. These systems are easily matched to ion sources and RFQ designs because they have a low duty factor, which allows pulsed, high power-density operation. Examples are the injectors at BNL, DESY, CERN, IHEP and KEK. Higher duty factors have been favoured by the development of high-brilliance beams for ATS, GTA and CWDD, based on the LANL developments in structures and beam dynamics. The final version of this injector is an RFQ for 8 MeV protons, which is made of eight resonantly-coupled pieces, as shown in Fig. 6, each individually furnace-brazed, aligned, and tuned.

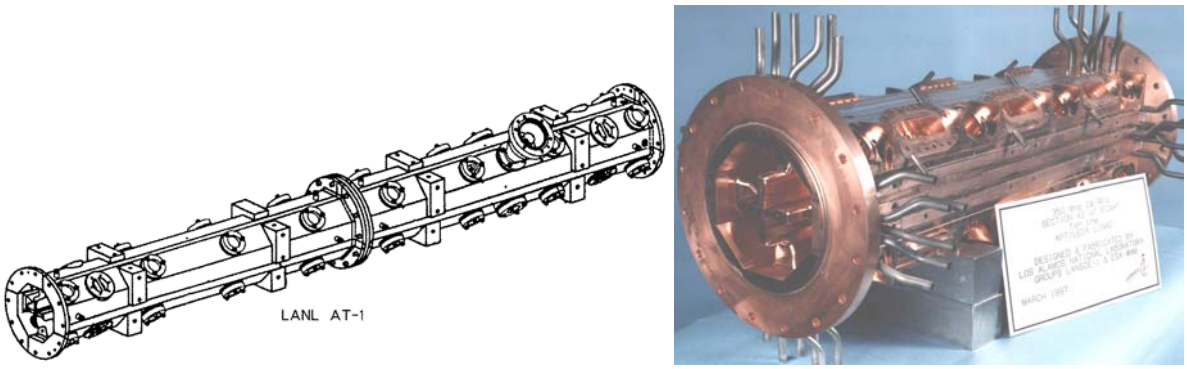


Fig. 6: The segmented CW RFQ and a view of a brazed LEDA RFQ prototype module

The LEDA RFQ is a part of the front-end demonstrator for a future accelerator-driven system, for example to transmute nuclear waste. It has successfully accelerated output beams of 100 mA in CW operation.

Somewhat smaller, but not unimportant, are projects which can be grouped as spallation neutron sources: designs for 5 MW (5–10% duty factor) have been made for H^- beams of 30–100 mA for new machines (ESS, NSNS, JHP) and the upgrade of existing high-power linacs like LAMPF and ISIS. The beam dynamics design has matured to rather small emittance growths (10–20%) which stretch the limits of simulation results. The advances in structure development are slow and reflect the limits of the technology. Present examples are the new ISIS injector and the RFQ for SNS, which uses non-resonant magnetic stabilizers as shown in Fig. 7.

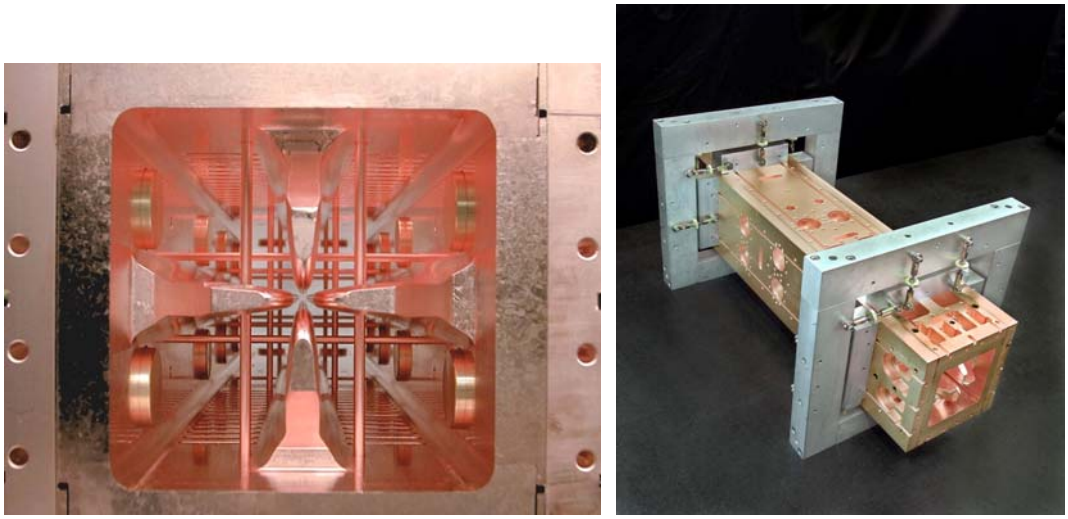


Fig. 7: The LBNL SNS RFQ

Short prototype RFQs for even higher duty-factors and CW operation have been set up in CRNL, LANL, KEK, JAERI and also in IAPF. A parallel development aims to build a 120 mA, 35 MeV, CW deuterium accelerator (IFMIF) for material testing of fusion devices.

3. HEAVY-ION ACCELERATORS

RFQs are also very attractive for low-energy, heavy-ion accelerators. They cannot replace static injectors and Van de Graaff generators in terms of energy resolution and beam quality, but are favoured for applications using high-current beams or combined with sources like an ECR, because the source can be close to ground potential and is easy to operate and service. The RFQ concept of spatially-homogeneous strong focusing proposed by Kapchinskiy and Teplyakov employs strong focusing with RF focusing that is independant of velocity, so the acceleration can start at low energy with rather short cells. This allows adiabatic capture of the DC beam from the ion source.

Heavy-ion RFQs have been built at LBL, INS, ITEP, GSI, Saclay and IAPF, for example, for atomic and nuclear physics research. They can be distiguished by the lowest specific charge they can accelerate and by the operational duty factor. Storage-ring and synchrotron injectors have a favourably low duty factor. New machines are the new HIMAC injector at NIRS, which is very similar to the TALL RFQ at INS, and the new injectors at CERN, MPI Heidelberg, and LMU Munich.

The RFQ of the CERN Pb injector is fed by a pulsed ECR source and operates at 101 MHz. It is designed to accelerate Pb^{25+} ions from 3 to 250 keV. The RFQ was built at INFN Legnaro, and is a symmetric four-rod RFQ with small vanes (Fig. 8), similar to those investigated at CRNL and IAPF.

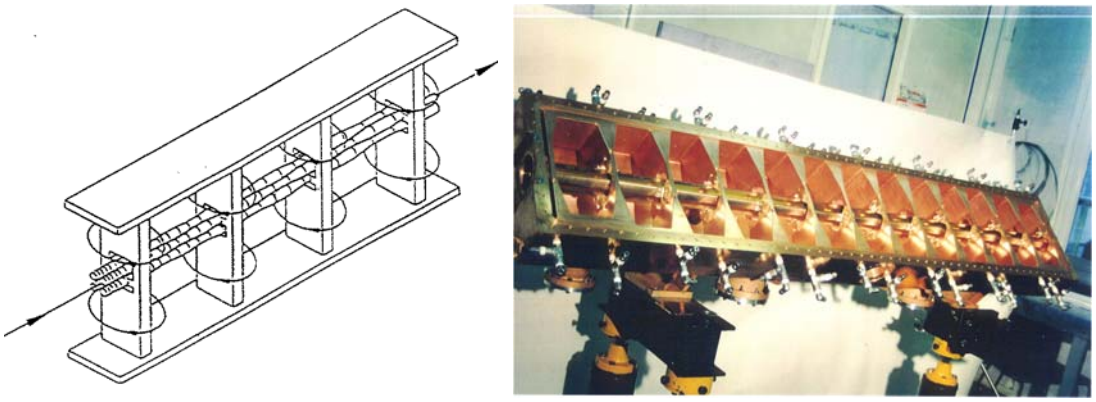


Fig. 8: The CERN heavy-ion RFQ

Less complex is the high duty factor HLI RFQ at GSI, which routinely operates at 25% duty factor. The HLI RFQ is 3 m long (designed for 108.5 MHz and U^{28+} ions $q/A = 0.117$) and accelerates from 2.5 to 300 keV/u. The high RF efficiency (figure of merit: shunt impedance) of the HLI RFQ is important for high duty factor structures in which technological problems like cooling and thermal stress control dominate, whereas this is not a major concern for synchrotron injectors.

The high-current injector at GSI has been designed for U^{4+} ions. The layout consists of an RFQ (2.2–120 keV/u, length 9.2m) for 15 mA beam current and an IH structure (0.12–1.4 MeV/u).It operates at 36 MHz, without an intermediate stripper as in the previous designs for U^{2+}, which had a spiral RFQ (27 MHz) injecting into the existing Wideröe part of the UNILAC. The IH RFQ with a special matching section for a bunched output beam (Fig. 9) consists of 10 modules (Fig. 10) with a length of 0.92 m each. It has been commissioned successfully.

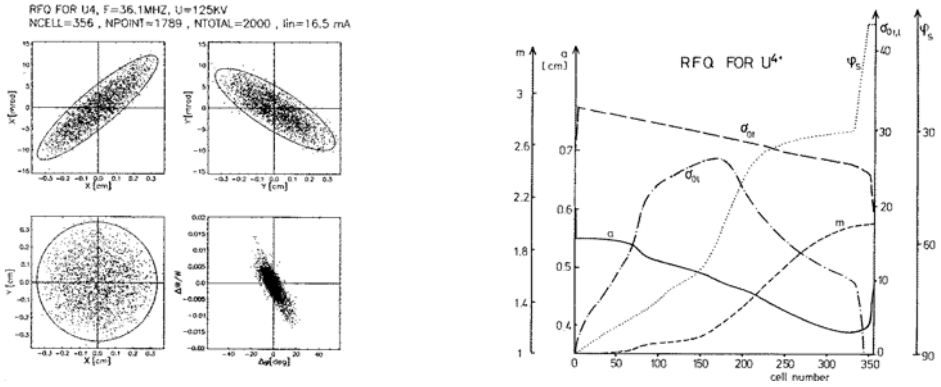


Fig. 9: The design for the GSI–HSI RFQ

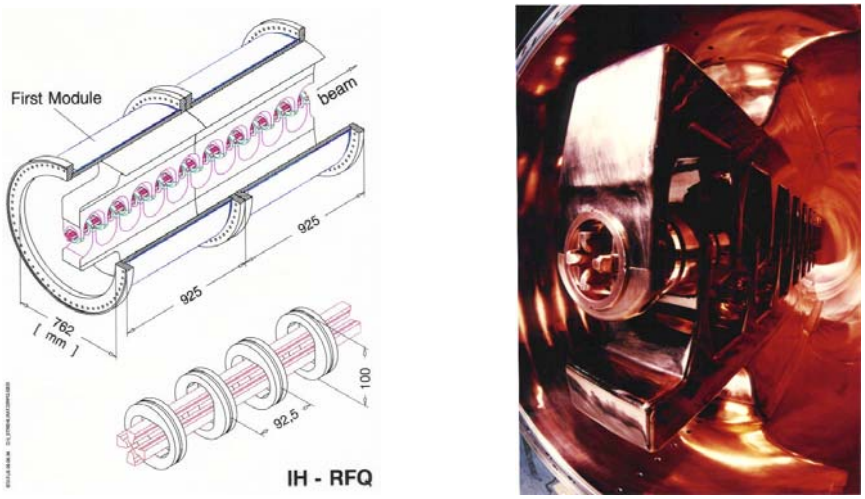


Fig. 10: The structure of the GSI–HSI RFQ

Van de Graaff Tandem machines have been the workhorse for nuclear physics research. Their limitations are low currents and low energy per nucleon for the heaviest ions, which led to the various heavy-ion RF accelerators and Tandem post-accelerators. An RFQ, although a low-velocity structure, can be applied as a post-accelerator for a Tandem installation if very heavy particles and a fixed output energy/u are not a restriction for the experiments.

For some time RFQs have been regarded as possible replacements for Tandem injectors for post-accelerators, but the first to be actually built was a superconducting heavy-ion linac bypassing the EN Tandem for uranium beams at ANL.

At MPI, Heidelberg, many experiments at the TSR storage ring were limited by the low currents delivered by the MP Tandem. A high-current injector for singly-charged light ions, consisting of a CHORDIS source, two RFQs, which are based on the GSI HLI design, and seven-gap resonators, as shown in Fig. 11, provides an increase in intensity higher by up to three orders of magnitude, and this is especially important for laser cooling experiments. A much more difficult task is to design an RFQ as the injector for a cyclotron. A first project was the conversion of the ISL at the HMI, Berlin (the former VICKSI machine), which is an isochronous cyclotron with four separated sectors. It had external beam injection with variable energy from either a CN Van de Graaff or an 8UD Tandem.

To inject into a separated-sector cyclotron, the RFQ has to provide a bunched beam at a well-defined injection energy given by the inner radius of the SSC. The operating frequency of the RFQ must be synchronized with the cyclotron frequency, which for RFQs normally means a fixed output energy per nucleon, which could be a possible solution only for fixed energy cyclotrons.

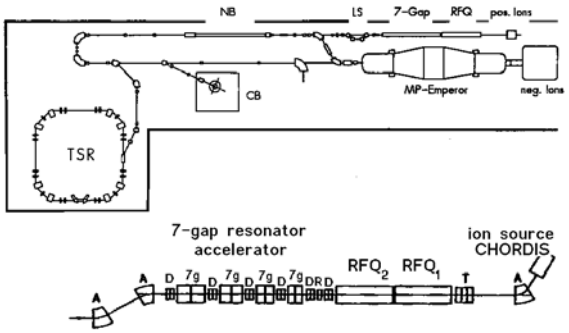


Fig. 11: The TSR injector at MPI

A fixed-velocity profile is typical for RFQs. It can only be changed by varying the cell length L or the frequency f. The possibility of changing the Wideröe resonance condition, $L = \beta_p \lambda_0/2 = v_p/2f$, has been used for RFQs with variable energy (VE RFQ). For this reason it is possible to change the output energy using the same electrode system by varying the resonance frequency of the cavity: $v_p \sim f$, $T \sim v_p^2$.

A type of RFQ resonator that can be capacitively or inductively tuned has been developed to change the frequency of the four-rod RFQ in Frankfurt. Figure 12 shows the method of tuning by means of a movable plate, which varies the effective length of the stems.

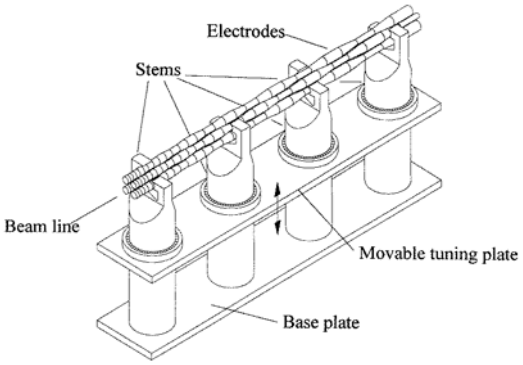


Fig. 12: The VE RFQ resonator insert

In Frankfurt the VE RFQ was first developed for application as a cluster post-accelerator at the 0.5 MV Cockroft–Walton facility at the IPNL, Lyon, France. It was designed for an input energy E_{in} = 10 keV/u and an output energy between E_{out} = 50–100 keV/u for m = 50 u. Heavy and low-velocity particles are accelerated in the second VE RFQ, with an energy range of 2–10 keV/u for singly-charged metallic clusters up to mass m = 1000 u (frequency range 5–7 MHz).

Another VE RFQ combined with an ECR ion source was built for the IKF at Frankfurt. The design values are a minimum specific charge of 0.15, an output ion energy of E_{out} = 100–200 keV/u, a maximum electrode voltage of 70 kV, and a structure length of 1.5 m. VE RFQs have a fixed ratio of output to input energy given by the length of the first and last modulation cell. This is similar to the energy gain factor of an SS-Cyclotron, and makes them well suited as injectors. To cover the energy range of 1.5–6 MeV/u, the injection energy of the ISL must be between E_{in} = 90–360 keV/u (maximum accelerating voltage U_m = 2.9 MV), at cyclotron frequencies of 10–20 MHz.

The ISL tandem injector at HMI in Berlin, as shown in Fig. 13, has been replaced by a combination of an ECR ion source on a 200 kV platform, which will produce highly charged ions with charge-to-mass ratios between 1/8 and 1/4, and a VE RFQ. To stretch the energy range the RFQ is

split into two RFQ stages. Each stage has a length of 1.5 m and consists of a ten-stem, four-rod RFQ structure. With an RF power of 20 kW per stage an electrode voltage of 50 kV is possible. In the first mode of operation both RFQs accelerate and the output energy of the cyclotron is between 3–6 MeV/u with a harmonic number of 5 for the cyclotron. For the low-energy beam only RFQ1 accelerates, while RFQ2 is detuned to transport the beam. In this mode the energy range of the cyclotron is between E_{out} = 1.5 to 3 MeV/u. The cyclotron works on the harmonic number 7. In both modes the RFQs are tuned to the eighth harmonic of the cyclotron.

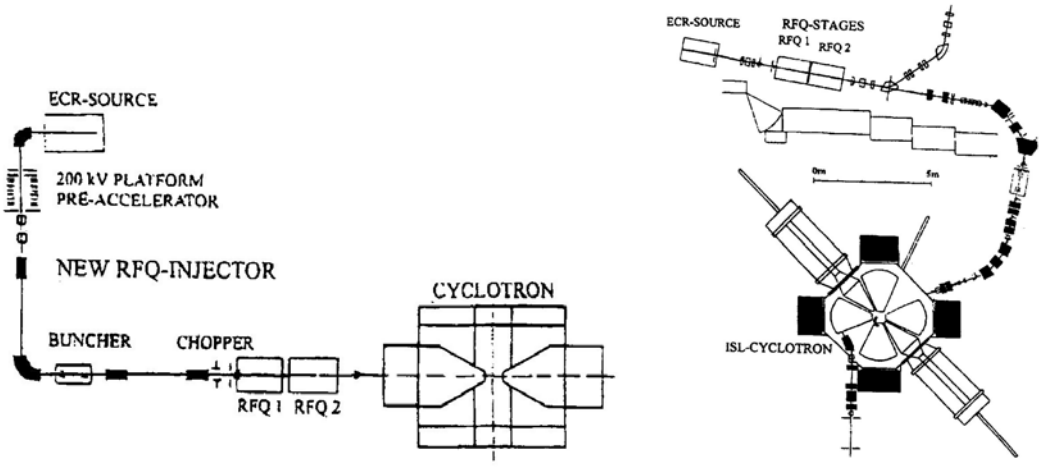


Fig. 13: The VE RFQ injector at ISL

The ECR source is mounted on the 200 kV platform formerly used for the Tandem. The vertical beam is bent through 90°, passes through the buncher–chopper system, and will be injected into the RFQs. The final matching into the RFQ will be by a triplet lens approximately 1 m in front of the RFQ to allow for diagnostics and a Faraday cup. The beam from the RFQ is transported into the injection beamline of the cyclotron, to which a rebuncher has been added to make a proper time focus for the cyclotron. The RFQ output emittance depends largely on the input conditions. For matched input beams with $\Delta E/E \leq 1$, normalized emittance $\varepsilon_n < 0.5$ mm mrad, and a bunch length $\sigma_t < 1$ ns a transmission of 100% is expected. To reach this beam quality it is necessary to have a buncher–chopper system between the ECR and the RFQ. As with the cluster linac at IPNL, there is no bunching in the RFQ. To save length and power the RFQ starts with stable phase $\varphi_S = 50°$ and at ISL the bunching is done by an existing bunching stage left over from the Tandem injector.

At RIKEN a similar problem had to be solved. The injector was formerly a variable frequency Wideröe-type linac. The 450 keV CW injector has now been replaced by an RFQ that is tunable from 8–28 MHz to match the VE linac and RRC cyclotron. An asymmetric four-rod-type RFQ, excited by one λ/4 resonator is used for this large frequency range. The electrodes must be supported by ceramic stems and, especially for the higher frequencies when the electrodes themselves are a major part of the λ/4 line, the tuning is difficult and the field distribution becomes unbalanced. This system matches the RIKEN linac very well and increases the experimental possibilities.

4. NEW DEVELOPMENTS

There are a number of studies making use of RFQ ion injectors and the proposals to build accelerators for radioactive beams seem to be of the highest interest in nuclear physics research. The typical RB facility starts from an ISOLDE type of ion source with singly-charged ions. To obtain a reasonable amount of ions, CW operation is planned. This favours superconducting structures. However, the low-energy part must use much lower frequencies for RF acceleration than is suitable for SC cavities, so room is left for NC RFQ structures to accelerate and form the beam, as shown in Fig. 14 for the ISAC-

RB accelerator at TRIUMF. The various systems can be distinguished by the heaviest mass number planned to be used and the accelerator system to be employed. For masses between 30 and 60 u, INS, TRIUMF, and ANL all have normal conducting RFQ linac-based systems.

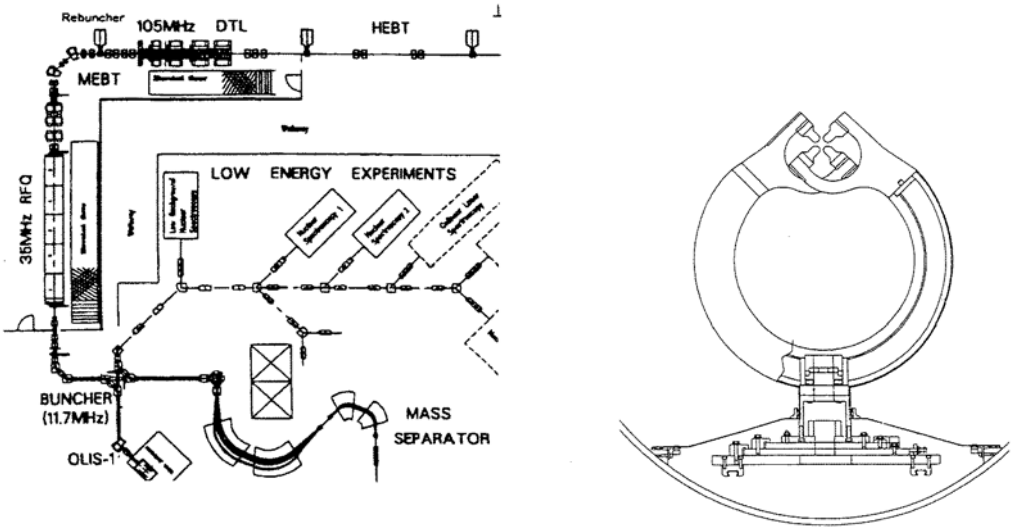


Fig. 14: TRIUMF RB accelerator

A very interesting concept is being pursued by groups from LMU Munich and MPI. It consists of a Penning ion trap followed by an EBIS ion collector–charge breeder with pulsed extraction, and a small compact pulsed accelerator of the MPI high-current injector or post-accelerator type to be installed at the CERN ISOLDE facility, as shown in Fig. 15 (see also Fig. 16).

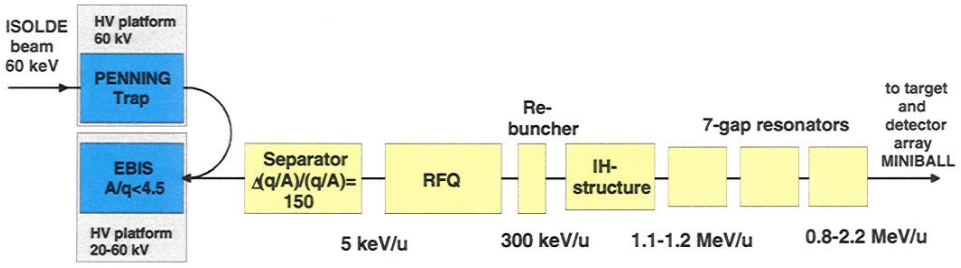


Fig. 15: The REX-ISOLDE principle

There are many applications of RFQs in industry. The first group is used for material improvement such as ion implantation in silicon, together with machines for 1 MeV boron and 6 MeV phosphorus. A second group are involved in the medical field, for example parts of PET isotope production units and medical synchrotrons such as Loma Linda and NIRS. New projects for using these therapy machines are presently under discussion at GSI, CERN, and Prague. Another application of high-current RFQs is as a compact radiation source for radiography with neutrons or resonant X-rays similar to those proposed for material detection.

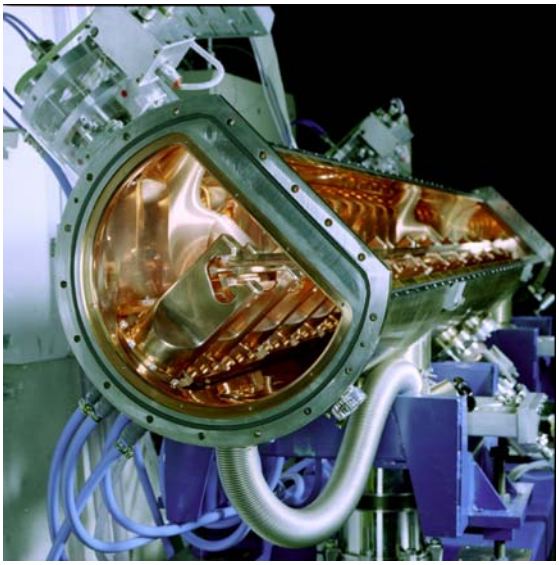


Fig. 16: The REX-ISOLDE RFQ

New developments in particle dynamics designs aim to reduce emittances while at the same time reducing the voltage and RF power requirements, but with only a minor reduction in beam quality. This would be especially important for high duty factors and industrial applications.

The matching between RFQ stages and the following accelerator stage has also been improved. The first step is optimization of the end cell to shape the transition fields. Recent designs also shape the bunch to make it appear that it has drifted longitudinally (e.g. the HSI project at GSI) and focus it longitudinally and radially (e.g. the funnel experiment at IAPF).

5. CONCLUSIONS

There are many new developments in the field of RF ion accelerators that will promote new experimental parameters for atomic and nuclear physics. Injectors into cyclotrons, Tandem replacements, and new compact ion accelerators are attractive applications of RFQs.

BIBLIOGRAPHY

I will restrict myself to the two papers from which all RFQ work started, and Accelerator School reports.

[1] I.M. Kapchinskiy and V. Teplyakov, *Prib. Tekh. Eksp.* **119** (1970) 17.

[2] K.R. Crandall, R.H. Stokes, T.P. Wangler, Linac 79, **BNL 51134** (1979) 205.

[3] M. Weiss, Proc. CERN Accelerator School: General accelerator physics, Aarhus, 1986, CERN 87–10, CERN, Geneva, 1987.

[4] A. Schempp, Proc. CERN Accelerator School: Course on RF engineering for particle accelerators, Oxford, 1991, CERN 92–03, CERN, Geneva, 1992.

[5] A. Schempp, Proc. US Particle Accelerator School, Fermilab, 1984 [AIP **C 184** (1989)].

SUPERCONDUCTING CAVITIES

G. Bisoffi
INFN, National Laboratories, Legnaro, Italy

Abstract
The main features of superconducting cavities are discussed. Their advantages with respect to the normal-conducting cavities as well as the basics of RF surface resistance are treated first. Then, design, construction and surface-treatment issues are addressed, together with the particular methods used to assess the quality factor and accelerating field of superconducting cavities. Anomalous losses, related to the surface electric and magnetic fields, are discussed in detail. A description of the present state of the art concludes the lecture notes.

1 INTRODUCTION

The interest in superconducting (SC) resonators shown by the community of elementary particle physicists is steadily increasing. Colliding beam storage rings, recirculating linacs and free electron lasers gain significantly from the high energy, high intensity and small emittance that can be achieved by using SC cavities. On the low-energy side in continuous wave (CW) mode, the SC boosters of heavy-ion Van de Graaf accelerators can maintain the excellent beam properties of their injectors, allowing them to probe nuclear structure around the Coulomb barrier between projectile and targets of all elemental species. On the application side (e.g. material studies, transmutation of nuclear wastes and subcritical reactors), SC-resonator-based accelerators of high-intensity protons (at least for the medium and high energies) were quite recently considered.

The basic aspects of SC resonators are described in this lecture. Their main advantages with respect to the normal-conducting (NC) option will be presented (Section 2) and some fundamental aspects of RF superconductivity will be recalled (Section 3). The design and construction of SC resonators (Section 4) as well as measurement techniques used to assess their performance (Section 5) will be reviewed. To conclude the lecture, the most important limitations to accelerating fields currently attainable will be summarized (Section 6) and some of the most recent advances in the field will be highlighted (Section 7). Some important ancillary topics, such as high power and high order mode couplers, cavity frequency stabilization techniques, especially with respect to microphonics and Lorenz-force detuning, can not be treated in a one-hour lecture. The interested reader should consult the more specialized literature [1].

2 ADVANTAGES OF SUPERCONDUCTING CAVITIES

The most striking advantage in using superconducting cavities is the possibility to work in a CW mode or at least at a significantly higher duty cycle than with NC cavities. Power dissipation on the wall of SC cavities is typically 10^5 times smaller than in the NC case. This is expressed in terms of the quality factor Q_0, the universal figure of merit of resonators, which is the ratio between the stored energy U and the power dissipated by the cavity P_c in one RF cycle of the resonant angular frequency ω_0:

$$Q_0 = \frac{\omega_0 U}{P_c} \, .$$

If we take the linac definition of shunt impedance per unit length, $r_a = V_c^2/P_c'$, we can easily see that the dissipated power per unit length in a linac P_c' can be expressed as:

$$P_c' = \frac{E_{acc}^2}{\left(\frac{r_a}{Q_0}\right) Q_0} \tag{1}$$

where r_a/Q_0 is the geometric shunt impedance in Ω/m, which depends basically on the structure geometry. The main attraction of superconductivity for RF cavities can be expressed in terms of P_c'. For example, an accelerating field of 5 MV/m typically corresponds to a power dissipation of $\sim$ 40 W/m for SC structures ($Q_0 = 2 \times 10^9$) versus $\sim$ 4000 kW/m for NC structures ($Q_0 = 2 \times 10^4$). The latter is not a practical value since, beyond $\approx$ 100 kW dissipation in a Cu cell, heat removal becomes a concern and problems related to the temperature (T) increase (e.g. mechanical stress, thermal expansion, outgassing) become prohibitive.

SC structures are not intrinsically limited by the cooling capability when pushing to higher E_{acc} values. They are limited rather by anomalous losses (e.g. field emission or quenches), which nowadays imposes an ultimate limit at $E_{acc} \sim 40$ MV/m (see Sections 6 and 7). Fields as high as these can also be reached with NC cavities, but only for microseconds because of the prohibitive power requirements.

Clearly, as far as the overall mains power is concerned, the situation is less advantageous than previously depicted, because of the refrigeration cost. Carnot efficiency $\eta_c = T_2/(T_1 - T_2), T_1 = 300$ K and $T_2 = 4.2$ K, as well as thermodynamic efficiency $\eta_{th} \approx 0.3$ (ratio of the compressor power needed for the ideal reversible case to the real power) must be considered, giving

$$\eta = \eta_c \eta_{th} = 4.5 \times 10^{-3} \, . \tag{2}$$

On the NC side, the efficiency of the klystron source (typically 0.5) must be considered. Eventually, an AC power saving factor of $\sim$ 200 is obtained for SC cavities, just taking cavity operation into account. This is true, for example, for heavy-ion structures, where the beam current is negligible. However, including the power necessary to accelerate, say, a 5 mA beam (and benefiting from a factor 3 higher field with SC cavities) the cavity-plus-beam power load is significantly higher and the power saving factor reduces to about 5, which is still a significant value.

The possibility of increasing inter-cell iris openings and making them rounded is another important advantage of SC cavities in many applications. The associated decrease in the cavity shunt impedance is not a concern, since it is dominated by the much larger quality factor. One advantage is that the cavity impedance Z can be typically 10 times smaller, which pushes the threshold of beam instabilities to higher currents I (the growth time of instabilities being $\tau_{inst} \propto 1/ZI$). SC cavities are also interesting for intense proton beam linacs, since beam halos are less prone to scrape on accelerator components and activate them.

3 RF SURFACE RESISTANCE AND ULTIMATE FIELD LIMITS

3.1 The RF resistance of a superconducting surface

The microwave surface resistance of normal metals is expressed by $R_s = \sqrt{\pi f \mu_0/\sigma} = 1\sigma\delta$, σ being the electrical conductivity, f the RF and δ the penetration of the EM fields (skin depth), $\delta = 1/\sqrt{\pi f \mu_0 \sigma}$. At low T and high frequency (σ increases by lowering the temperature) δ may become shorter than the mean free path of an electron. In this case the electrons spend only part of the time between collisions in the field penetrated region. Consequently the electrons become less effective in shielding the field and, somehow counter-intuitively, longer mean free paths lead to higher surface resistance than expressed above ('anomalous skin effect').

In computing the anomalous-skin-effect limit for copper at microwave frequencies and cryogenic temperatures, one can see that, although the DC conductivity increases by a factor 100, the anomalous skin effect allows only a decrease of a factor 6 in the surface resistance. This shows that it is definitely not convenient to cool an NC metal to cryogenic temperatures.

The SC regime is completely different. Following the London 'two-fluid' model [2], we can view the situation for superconductors at $T < T_c$ as the coexistence of two fluids, a superfluid of paired electrons (Cooper pairs) and the normal fluid of free electrons: a 'supercurrent' and a normal current flow in

parallel. In this situation all DC current is carried by the supercurrent and the DC resistance of a superconductor below T_c is exactly zero. In AC regime, however, the surface resistance of a superconductor is always larger than zero, albeit very small compared to NC metals. In fact, although Cooper pairs move without resistance even when exposed to RF fields (no power dissipation) they have inertia. This means that the time-varying surface magnetic field $H_s \cos \omega t$ (H_s being the magnetic field at the metal surface and ω the angular frequency) must extend beyond the surface of the material to create the forces which can provide back-and-forth acceleration of the pairs and oppose the RF surface currents. This is expressed in terms of an induced electric field

$$E \propto \frac{\mathrm{d}H}{\mathrm{d}t} \propto \omega H_s \tag{3}$$

inside the material. Once time-varying electric fields are present in the skin layer, they will act on the unpaired electrons as well and the latter can interact with the lattice and produce losses, following the anomalous skin effect. The equivalent 'superconducting skin depth' is about as large as the London penetration depth λ_L (depth of penetration of the magnetic field in a superconductor, ranging from 15 to 110 nm, depending on the particular metal), and is about a factor 100 smaller than δ.

For frequencies below 10 GHz and $T < T_c/2$, experimental data are well described by the empirical relation

$$R_s = A \, \frac{\omega^2}{T} \, e^{-\alpha \frac{T_c}{T}} + R_{res} \, , \tag{4}$$

where A and α depend on material parameters [3].

The most noticeable features of this expression are that R_s increases with the square of the RF and decreases exponentially with temperature. This suggests that high-frequency cavities (e.g. 1.5 GHz) need to be cooled below T = 2 K to give acceptably low surface resistance. The first term of this expression, coming from the two-fluid model (also called Bardeen–Cooper–Schrieffer, or BCS resistance), is in excellent agreement with experimental results for frequencies below $\sim$ 10 GHz, which is also the range of practical interest for nearly all SC cavities.

Figures showing the temperature dependence of surface resistance of niobium (Nb) at, for example, 3 GHz nicely demonstrate the existence of the residual resistance which appears in the formula above (see, for example, Fig. 1). R_{res} is temperature-independent and accounts for residual loss limitations of actual SC cavities, which will be discussed in Section 6.1. Typical values of R_{res} are 10 to 20 nΩ. R_{res} is usually the dominating resistance limitation in low-frequency low-β resonators (f = 50 to 150 MHz), for which simpler cryogenic operation at 4.2 K is sufficient.

3.2 Intrinsic field limitations in SC cavities

When a magnetic field is present at the SC to NC phase transition (as is the case when raising the EM fields in a cavity), a latent heat exists which is related to a discontinuity in entropy. As a result the transition requires some finite amount of heat input to compensate for the change in entropy. NC nucleating centres are created, beyond T_c, in an overall SC regime.

For typical SC cavity frequencies, the wave period ($\sim 10^{-9}$ s) is much shorter than the time it takes to creates the nucleation centres ($\sim 10^{-6}$ s). Therefore, it seems possible for the SC state to survive beyond H_C, up to a 'superheating' critical field H_{sh}, exceeding H_C for type I superconductors and H_{C1} for type II superconductors [2]. Nowadays H_{sh} is rated as the maximum theoretical limitation for SC resonators.

A dependence of H_{sh} on the ratio $\kappa = \lambda_L/\xi$ (where ξ is called the 'coherence length' and represents a sort of spatial extension of the Cooper pair which carries the supercurrent) was calculated [5] with a phenomenological theory of superconductivity, showing that

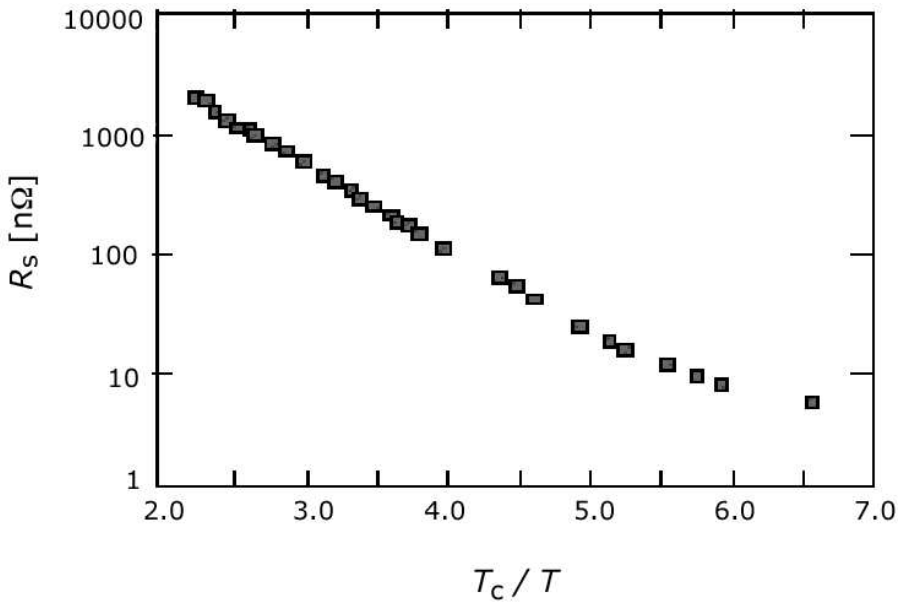


Fig. 1: Surface resistance of a 3 GHz niobium sheet cavity (RRR = 40) plotted against T_c/T. The residual resistance is 4 nΩ [4].

$$\begin{aligned} H_{sh} &\approx \frac{0.89}{\sqrt{k}} H_C \qquad \text{for } \kappa \ll 1 \ , \\ H_{sh} &\approx 1.2\, H_C \qquad \text{for } \kappa \sim 1 \ , \\ H_{sh} &\approx 0.75\, H_C \qquad \text{for } \kappa \gg 1 \ . \end{aligned} \tag{5}$$

H_C has been exceeded in some cases, while H_{sh} has never been exceeded experimentally, which seems to be an experimental indication that H_{sh} is the fundamental limit to the performance of an SC resonator. For Nb at T = 0, for example, H_C = 200 mT and H_{sh} = 240 mT.

There exists no known theoretical limitation to the peak surface field, E_{pk}, whereas the well known, often encountered, practical limitations connected with electron field emission will be discussed in Section 6.

4 DESIGN, CONSTRUCTION, AND TREATMENT ISSUES

4.1 Design

When addressing the basic design of a new SC resonator, a small flowchart of the most significant issues could be followed.

4.1.1 β and frequency

The bunch length determines an upper limit to the frequency of a resonant structure, which should be much smaller than the wavelength. To gain as much as possible from the accelerating voltage present on the various gaps of a resonant structure, the bunch should cross them while the electric field is closest to its maximum value. Consequently the resonant frequency f_0 has to be related to the inter-gap distance d and the bunch velocity βc through

$$f_0 d \approx \frac{\beta c}{2} \ . \tag{6}$$

At low values of β, the need of useful accelerating lengths ($\beta\lambda \sim 5$ to 10 cm) implies that the wavelength λ has to be large and therefore that the resonant frequency has to be small. Low-β resonators hence tend to be large in size (see two examples in Fig. 2). The condition $\beta \sim 1$ (typical of electron cavities) calls instead for frequencies in the range of 300 to 3000 MHz for accelerating lengths of 5 to 40 cm.

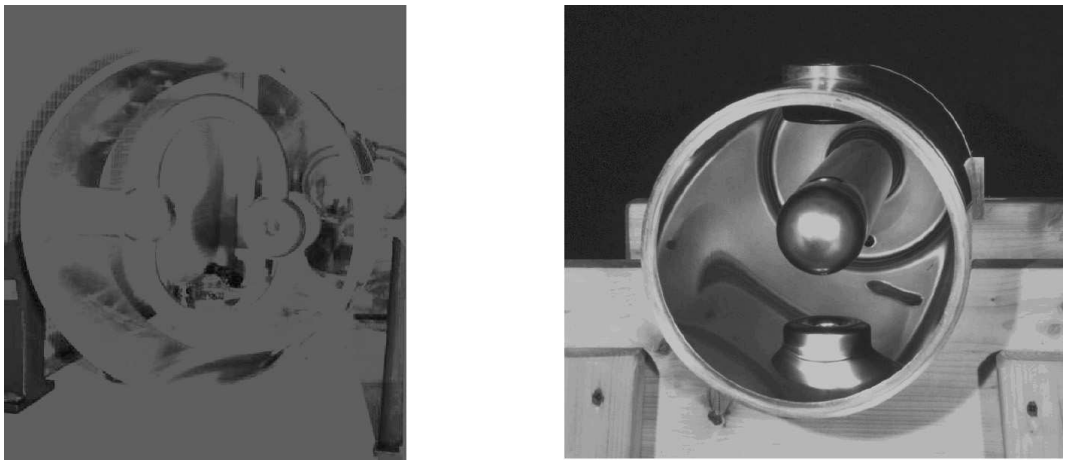


Fig. 2: Two examples of low-β SC resonators for heavy ions: a split-loop resonator (left) and a quarter wave resonator (right)

Within the somewhat broad frame mentioned above, a lower resonant frequency would imply smaller BCS resistance, calling however for a larger cryostat and higher associated fabrication costs. Larger surface cavities, moreover, increase the probability of introducing surface defects, which are a source of lower performance.

4.2 Number of gaps

As far as heavy-ion SC structures are concerned, it is convenient to build them with a small number of gaps, so as to accommodate the largest possible acceptance in the natural spread of β-values coming from the typical injectors of these machines.

This inconvenience does not arise in the design and production of SC cavities for $\beta \sim 1$, in which multigap resonators are the natural choice. They are built as a chain of N quasi-pillbox cavities, coupled through the iris openings. While the angular frequency of the lowest mode of a single simple pillbox without beam ports is $\omega_0 = 2.405\ c/a$ (a being the pillbox radius), which approximately sizes the cavity as a function of the chosen frequency, with N coupled cells the angular frequency is

$$\omega_q = \omega_0(1 + K\ \cos\ \alpha_q)\ , \tag{7}$$

where $\alpha_q = \frac{q\pi}{N}(q = 1, ...,\ N)$ and K is the coupling factor between adjacent cells, dependent on the iris diameter and geometry. Among the N fundamental resonant frequencies, it is useful to operate in the 'π' mode $(q = N)$, which has the highest shunt impedance. The field pattern of the π-mode is shown in Fig. 3.

Lower manufacturing costs, more efficient use of the accelerating voltage, smaller fringing field effects and fewer inter-cavity drift spaces are points in favour of having a *large number of coupled cells*.

A flatter electric field from cell to cell and lower overall power from the input coupler are points in favour of a *small number of coupled cells*.

The choice depends on each application.

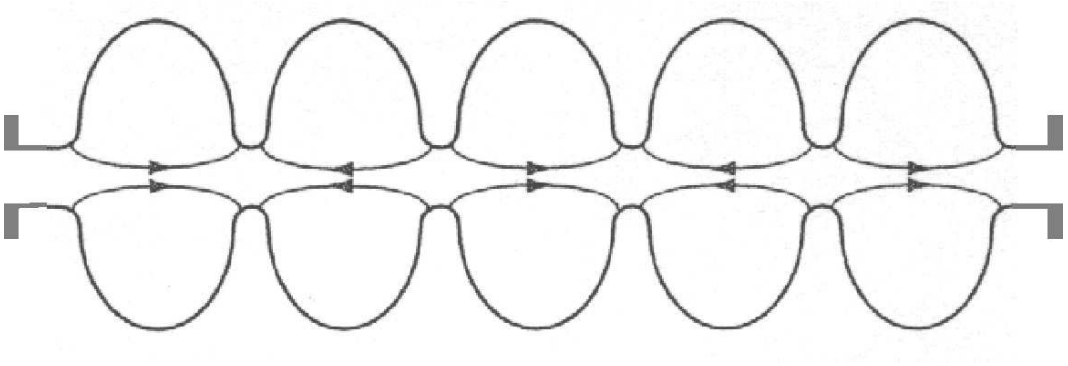


Fig. 3: π-mode of a β = 1 multicell cavity

4.3 Iris aperture

A large and rounded iris opening will increase uniformity in the field distribution from cell to cell, and will also offer a smaller impedance to the beam (implying better beam quality and larger currents).

On the other hand the shunt impedance will be lower (more refrigeration power needed) and the accelerating field will be smaller. Again the particular application will indicate the ideal choice.

4.4 Construction technologies

4.4.1 Full niobium

Once the geometric design of an SC cavity has been basically fixed, the material must be chosen, which is associated with a particular construction expertise.

Pure Nb is the most widely used material, supplied in sheets then shaped, cut and joined by electron-beam welding (EBW). This will be considered as the 'default' choice in the following discussion on construction and treatment. Niobium is a good choice since it has a quite high critical temperature (T_c = 9.2 K) and lower RF resistance, see Eq. (4). Expertise in the production of very pure (high residual resistance ratio or RRR) sheets and bars, the shaping of cavity parts and EBW has reached a high level and can be applied to almost any complex geometry with little additional R&D.

High-RRR Nb is obtained by boiling out impurities in an electron-beam furnace, in several stages, at a base pressure of 10^{-4} to 10^{-5} Torr. C, N, H and O are quite efficiently removed in this way. These are by far the worst impurities from the point of view of electron mobility. High-RRR Nb ingots are laminated in subsequent stages, annealed, and delivered by the supplier.

Deep drawing or drop-hammer-forming, using dies made out of aluminium can accomplish forming. These dies are easy to make and any aluminium sticking on the Nb surface during machining can be removed by subsequent chemical etching. Trimming Nb parts to their final pre-EBW shape should ideally be done by a computer-controlled milling machine, operated at very low speed to avoid heating up the part beyond 100°C (otherwise they would again be contaminated by gaseous impurities in the bulk). Prior to EBW, components must be degreased and lightly etched on the edges to be welded by a buffered chemical polishing (BCP) solution, made out of one part hydrofluoric acid, one part nitric acid and two parts phosphoric acid (the so-called 1:1:2 solution). In this process 5 to 20 mm material removal seems sufficient.

Typical EBW parameters are a pressure below 2 to 4 $\times$ 10^{-5} Torr in the welding chamber, and a 50 kV/ 40 mA electron-beam current at 1 cm/min velocity on the joint (usually the electron beam stays fixed and the parts are moved in the EBW vacuum chamber). Experience with EBW has shown that a full-penetration joint with a smooth 'underbead' (the welded part opposite the beam side) can be obtained by either defocusing the electron beam or by making it oscillate with an elliptical or rhombic pattern along

the joint. If EBW is performed from the 'inside' of the resonator being built, full penetration is not an issue and the EBW parameters are somewhat more relaxed.

Rough tuning to the final frequency (after carefully predicting subsequent changes due to chemical etching, resonator evacuation and temperature change) is usually performed at the end of the construction procedure. Each cell of a multicell cavity, for example, can be squeezed or stretched by gripping the irises, which also improves the field flatness as desired.

After the final geometry has been obtained, final purification of the Nb by means of solid-state gettering might be applied [6]. This can increase the RRR up to values of the order of 1000.

4.4.2 Alternatives to full electron-beam-welded niobium

Construction technologies alternative to solid Nb are discussed in the following.

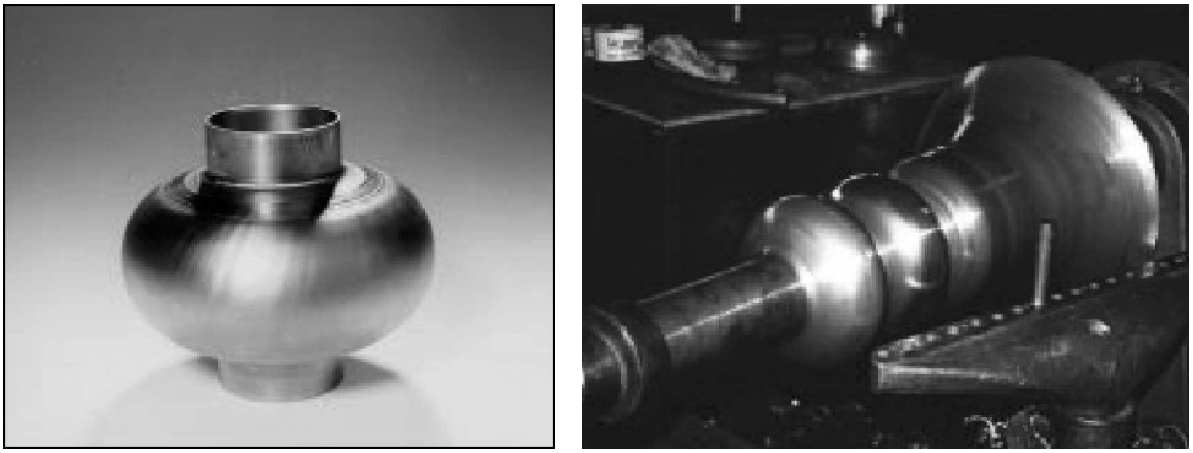
For low-β applications, for example, the residual resistance almost always dominates the BCS resistance: lower T_c and easier cryogenic operation at 4.2 K can be allowed. *Electroplating lead onto a copper substrate* is in this case a simple, less expensive choice, which can also be adapted to rather complicated geometry [7].

- Sputtering of Nb onto a Cu substrate is becoming a more and more interesting alternative to solid Nb, since performances have been demonstrated to be almost comparable. The technique has been developed with success for both $\beta > 0.8$ [8], [9] and $\beta = 0.13$ cavities [10]. One can save on the significant cost of the high-RRR Nb and profit from the excellent thermal stabilization offered by the Cu substrate, thanks to which quenches are virtually never the field limit for Nb-sputtered resonators. A value of $RRR \sim 30$ can be reached, which minimizes the BCS contribution to surface resistance. Nb-sputtered cavities still generally suffer from a non-negligible drop of the quality factor versus increasing accelerating field. It is being discussed at present whether this is related to the weak links between SC grains on the cavity surface or to problems of the Nb-to-Cu interface.
- Nb_3Sn coatings (obtained by evaporated and heated up Sn on a pure Nb surface) are extremely promising because of their higher critical temperature (T_c = 18 K). $E_{acc} \sim 15$ MV/m is the present practical limitation.
- High-T_c superconductors, such as YBCO, might be interesting if epitaxial films are made to grow on a monocrystalline substrate [11], in which case the residual resistance is not expected to increase with the RF field. However, the development of this technology is still limited to samples of very small area.
- Within the full-Nb choice, promising techniques are being developed in building seamless resonators by either hydroforming [12] or spinning [13] (Fig. 4). Reducing EBW costs seems mandatory in future high-energy accelerators, like TESLA, where a significant reduction in the cost of the whole machine could be thus obtained. Seamless resonators were also developed in Cu, then sputtered with an SC Nb layer.

4.5 Surface treatments

Coming back to full Nb, the interior of the surface needs to be chemically etched (e.g. by the same 1:1:2 solution mentioned above) so as to remove the layer damaged during the mechanical construction. It has been shown [14] that removal of 100 μm is, in most cases, appropriate. An increase of the bath temperature beyond 18 to 20°C might trap H in the material bulk, affecting the final quality factor (see Section 6.1.2): hence temperature control appears to be a very important precaution.

Complementary to BCP, electropolishing (EP) has recently gained great attention since it seems necessary to adopt it in order to push the maximum accelerating fields of multicell cavities far beyond 25 MV/m [15]. In EP, Nb is the anode, the cathode is made of Al and the electrolyte consists of a



(a) (b)

Fig. 4: Hydroforming [12] and spinning [13] represent the most advanced technologies to produce niobium resonators without electron-beam welding.

solution of H_2SO_4 and HF. The cavity is gently rotated during the treatment. The smooth mirror-like surface finishing seems responsible for the lower resistance, at grain boundaries, than that achievable with the rather rough BCP.

After polishing, rinsing with pure water (ρ = 18 MΩ-cm) for several hours is necessary. High-pressure water rinsing (HPWR) with ultra-pure water is a recent and well-established technological development. 60 to 100 bar water jets, sprayed through 0.2 mm nozzles, seem to remove the most adherent contaminants and might be beneficial to the removal of defects in regions of both high magnetic field (quenches) and high electric field (field emission).

After HPWR, the SC cavity must be dried and assembled with the utmost care in a clean environment (class 10 or 100), so that no dust is allowed to settle on the SC surface and become the cause of field emission during tests.

5 MEASUREMENT OF THE PERFORMANCE OF A SUPERCONDUCTING CAVITY

The quality factor Q gives an overall assessment of the cavity performance, since all power losses are taken into account. On some occasions, if the results need to be improved, investigations of sources of local power dissipations are possible by other means (e.g. temperature mapping [16], [17]).

For Q measurements, a resonator is equipped with an input coupler port, driving the power coming from the RF source, and a pickup port, which typically samples a small fraction of the resonator stored energy. The TEM line field is coupled to the cavity field. The further in the position of the coupler, the stronger the coupling between the field coming from the RF drive and the superimposed cavity field components in the coupler port. Since the cavity field decays exponentially in the port it is possible to adjust the position of the coupler mechanically for a critical matching of the two fields (the line sees the cavity as a perfectly matched load).

It is worth underlining that the Q_0 value of an SC cavity cannot be measured through the formula $Q_0 = f_0/\Delta f$ (f_0 being the cavity resonant frequency and Δf the resonance bandwidth) since the latter—at a Q_0 value of 10^9 or higher—could be as small as a few hertz or less. Hence the evaluation of the decay time of the cavity stored energy is the only accurate measurement technique.

5.1 Definitions

Let us recall the main parameters involved in the assessment of the quality factor, obtained from the decay time of the cavity stored energy, once the cavity is fed by rectangular forward power pulses. At RF drive switched-off, the total power lost by the cavity (P_t) is that dissipated by the cavity itself (P_c) plus that emitted from the coupler (P_e) and the pickup (P_{pu}) ports: $P_t = P_c + P_e + P_{pu}$.

While the intrinsic Q_0 of a resonator is defined as $Q_0 = \omega_0 U/P_c$ (ω being the angular frequency and U the stored energy), we can define a loaded quality factor as $Q_L = \omega_0 U/P_t$. A measurement of Q_L can be deduced by a measurement of the decay time of the cavity stored energy U, through

$$\frac{\mathrm{d}U}{\mathrm{d}t} = -P_t = -\frac{\omega_0 U}{P_t}\ , \ \text{giving}\ U(t) = U_0 e^{-\frac{\omega_0 t}{Q_L}}\ \text{and}\ \tau_L = \frac{Q_L}{\omega_0}.$$

However Q_0 (and not Q_L) is the interesting parameter in order to assess the resonator performance; we can extract it through

$$\frac{1}{Q_L} = \frac{P_c}{\omega_0 U} + \frac{P_e}{\omega_0 U} + \frac{P_{pu}}{\omega_0 U} = \frac{1}{Q_0} + \frac{1}{Q_e} + \frac{1}{Q_{pu}} = \frac{1}{Q_0}(1 + \beta_e + \beta_{pu})\ , \tag{8}$$

where the coupler and pickup quality factors Q_e and Q_{pu} are contextually defined, together with the coupling factors β_e and β_{pu} of the two antennas,

$$\beta_e = \frac{Q_0}{Q_e} = \frac{P_e}{P_c}\ \text{and}\ \beta_{pu} = \frac{Q_0}{Q_{pu}} = \frac{P_{pu}}{P_c}\ .$$

The only role of the pickup coupler is that of sampling a portion of the stored energy, which can then be calibrated to give indications on the entity of the accelerating field: it is hence reasonable to design it such that $\beta_{pu} \ll 1$. Hence in the following we shall take $\beta = \beta_e$ and consider the coupler-plus-cavity ensemble as an overall driven harmonic oscillator (at the resonant frequency of the cavity), the steady state and the transient behaviour of which can be investigated.

5.2 Response of a cavity to rectangular RF pulses

A complete treatment of the equivalent circuit and of the response of a resonator to rectangular forward pulses is given in Ref. [18], in which the switching-on and switching-off conditions are treated as particular examples.

A very useful expression of the reflected power is derived from the switching-on treatment:

$$P_r = \left[1 - \frac{2\beta}{1+\beta}\left(1 - e^{-\frac{t}{2\tau_L}}\right)\right]^2\ . \tag{9}$$

The behaviour of the reflected power during switch-on regime can be viewed on an oscilloscope, together with the pickup signal and the signal from the forward power. This gives an immediate visual assessment of the coupling strength and provides a direct measurement of its value with sufficient accuracy.

At β = 1 and at resonance (at $t = \times$) $P_r = P_e$ = 0, meaning that all forward power goes into cavity power (*critical coupling*) and the cavity is a perfectly matched load to the transmission line. This condition is normally the preferred one for Q_0 measurements. The loaded decay time τ_L can be measured by either the reflected (emitted) or the pickup signal after having switched off the forward-power rectangular pulse:

$$Q_0 = \omega_0\ \tau_L(1+\beta) = 2\omega\tau_L\ (\text{in critical coupling})\ . \tag{10}$$

5.3 Q_0 measurements require frequency locking

Given the extremely narrow bandwidth, it is necessary to keep the circuit frequency locked to the resonant frequency f_0 in the course of the measurement, since f_0 will be made to vary by the smallest mechanical vibrations.

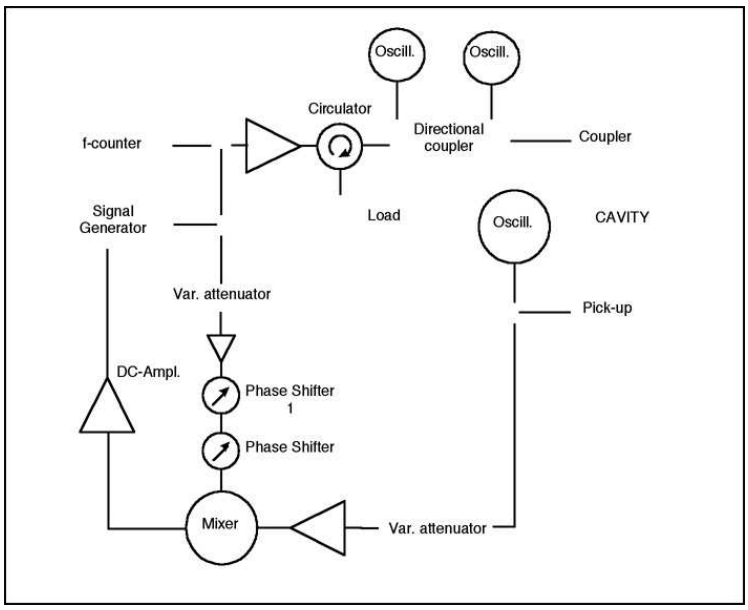


Fig. 5: Schematics of the phase-feedback control circuit for the Q_0 measurements of an SC cavity

A circuit similar to the one shown in Fig. 5 is used. A mixer compares the signal feeding the cavity (f_g) with that transmitted by the cavity. The two signals will have a phase difference θ bounded to the difference of frequency between the feeding signal and the actual resonant frequency (f_r) of the cavity:

$$tg\ \theta = \ Q(f_r/f_g)^2 \ (Q \text{ being the resonator quality factor}) \ .$$

If, at resonance, the phase difference between the two signals is set at $\pi/2$, small changes in the resonant frequency produce an output signal of the mixer which is proportional to the phase changes and, consequently, to the difference between the driving frequency and the resonant frequency. Such a signal, suitably amplified, can be used to modulate continuously the frequency output of the generator, keeping the $\omega = \omega_0$ condition locked.

5.4 Plotting quality factor versus accelerating field

As long as losses are ohmic at constant surface resistance, Q_0 is constant with increasing U (or E_{acc}), since $U/P_c = const.$

At high fields, however, losses tend to be non-ohmic. This is why it is necessary to measure Q_0 by the decay-time method at very low fields, i.e. to be sure to be in the region of purely ohmic losses.

Figure 6 shows a typical SC cavity performance curve, in which Q_0 is plotted versus the peak surface field. The significant Q_0 drop at high fields (accompanied by higher power dissipated by the cavity) is explained by field emission electron loading (see Section 6). Peak surface field (E_{pk}) and accelerating field (E_{acc}) are proportional to $\sqrt{U}$, through constants which are normally derived by appropriate simulation codes or bead pulling measurements. This defines the abscissa in the Q_0 vs. E_{acc} plot. Ordinate values (Q_0 at increasing fields) can be found by scaling the U/Pc ratio with the low-field Q_0 measured via stored-energy decay time.

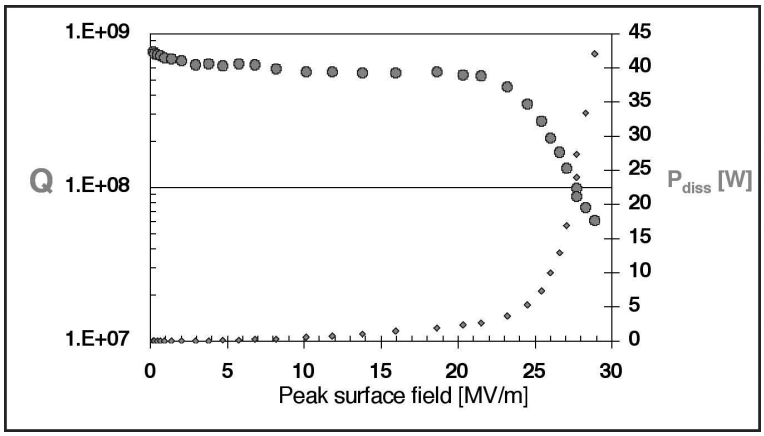


Fig. 6: Q_0 plotted against peak surface field: a typical example [19]

6 PRACTICAL LIMITATIONS TO ACHIEVABLE FIELDS

The theoretical maximum field limits, mainly correlated to the critical field of the superconductor, are not normally achievable in practical SC resonators. The various causes can be approximately grouped into (a) those which are correlated with the surface magnetic field (Section 6.1) and (b) those linked to the local electric field (Section 6.2).

Representatives of the first group are distributed anomalous losses connected to trapped magnetic flux or precipitated Nb hydrides, and local anomalous losses connected to the presence of defects on the cavity surface, driving to thermal breakdown.

Representatives of the second group are resonant field emission (also called 'multipacting') at relatively low EM fields, and non-resonant field emission at the highest fields.

6.1 Anomalous losses related to the surface magnetic field

6.1.1 Distributed losses related to a DC magnetic field

Distributed losses related to a DC magnetic field (like the Earth's magnetic field) give rise to an overall contribution to the surface resistance, which can be expressed by

$$R_{mag} = \frac{H_{ext}}{2H_{C2}} R_n \ , \tag{11}$$

R_n being the normal state resistance and H_{ext} the external magnetic field [20].

External magnetic field is expelled from the superconductor, except from the contribution given by distributed lattice defects or surface irregularities, which can act as distributed trapping sites.

If one takes the typical values for RRR = 300 Nb (H_{C2} = 2400 Oe and R_n = 1.5 mΩ at 1 GHz) the result is

$$R_{mag} = 0.3(\mathrm{n\Omega}) H_{ext}(\mathrm{mOe}) \sqrt{f(\mathrm{GHz})} \ .$$

These residual losses can be practically eliminated by providing proper shielding around the cavity and are not a fundamental problem.

6.1.2 Q-disease

Hydrogen is usually dissolved in the bulk of the newly bought material in negligible concentrations ($<$ 1 ppm wt). However, its concentration may be increased if, for example, the temperature rise during chemical etching is not kept below 20°C: in this case some ppm of H (by weight) can be present in the bulk.

Whereas at room temperature three orders of magnitude higher concentrations are necessary to form hydrides, a concentration of H of 2 ppm wt becomes critical below 150 K. Moreover, between 150 and 60 K the mobility of H is still quite significant, allowing the formation of sites of high hydrogen concentration. The part of the Nb hydride which forms on the surface exposed to RF fields can cause significant depression of the resonator Q even at quite low fields (Q-disease).

Two methods are known to prevent or avoid the problem. The first involves carefully heating the cavity up to 700–900°C to facilitate hydrogen outgassing as much as possible. The second requires cooling down the cavity more rapidly than the time necessary for H precipitation in the range 150 K to 60 K (1–2 hours are usually sufficient). Below 60 K the mobility of H becomes negligible and the new formation of hydride is no longer a concern.

6.1.3 Local defects

Local thermal breakdown starts at micrometric sites on the SC surface where the RF losses are much higher than those due to the surface resistance of the SC material.

While DC currents flow around defects, the reactive part of the impedance causes AC currents to also flow through a defect. Joule heating can locally increase the temperature. When the temperature at the defect border is higher than the critical field of the superconductor, the NC region grows in size, and for further increases of the applied EM fields a thermal instability may develop, eventually driving the whole resonator into the NC state (quench).

A quench is normally observed as a saw-tooth pattern in both the signal from the pick-up probe and the reflected signal during a Q-curve measurement carried out in critical coupling conditions. This shows that all the power is reflected once the thermal instability shows up, but goes back to its original value when the SC state is restored after some cooling of the resonator.

Typical local defects can originate from imperfections caused by EBW. Examples are microscopic holes on the welding path, and Nb material which evaporates from the weld and then condenses somewhere else on the cavity surface in the form of small spheres. Other causes are chemical residues, foreign-material inclusions (e.g. embedded abrasives, see Fig. 7) or simply micrometric particles which fall in a high-H field region during cleaning or assembling of the resonator (unless particular care is taken in these delicate phases, the number of these dust components increases).

The previously mentioned temperature-mapping technique can be applied to multicell cavities, or other cavities of simple geometry, to detect the position and the relevance of the effect of a particular defect. In fact this method is rarely adopted and the identification of the size and material of the defect can only be made by scanning-electron-microscope investigation of the suspected zone. This requires that the cavity be cut at the equator: another practice which one also tends to avoid on a production scale!

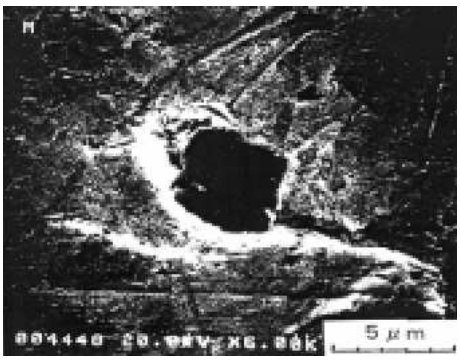


Fig. 7: Scanning-electron-microscope image of an embedded abrasive in a Nb surface [21]

6.1.4 Increasing the cavity RRR

A method often applied to increase H_{max} in the presence of small defects is to raise the thermal conductivity of Nb. As long as heat is transferred away from the location of the defect with more efficiency, it will be at higher local H-field (and corresponding E_{acc} field), such that the neighbouring SC material will exceed T_C.

The higher the RRR of the material, the higher its thermal conductivity κ. The formula

$$\kappa = \frac{RRR}{4}\left(\frac{W}{\mathrm{mK}}\right) \text{ [22]}$$

holds approximately; as a matter of fact, the correlation between electrical and thermal conductivity, which is well known for metals at room temperature, still holds true down to about 4 K, as long as electrons are still the dominant heat carriers. Below that temperature, at around 2 K, the density of phonons has a peak and lattice vibrations become the dominating phenomenon (at even lower temperatures the density of phonons also drops).

The RRR of Nb can be increased in several ways.

- Heating in very good vacuum up to $\sim 2000^\circ$C for some hours can outgas even O from the film and RRR = 1200 can be reached; however, the loss of mechanical properties and the risk to deform the resonator often make this a not very recommendable practice.
- Solid-state gettering is now widely applied: It consists in coating the exterior of a cavity with either yttrium or titanium, materials with a higher affinity to O than Nb, while warming up the resonator to 1200–1400°C to increase the mobility of O itself. Then the getter is removed by external chemical etching. RRR higher than 1000 can be obtained in this way.

With increasing RRR a linear increase was observed, over the last two decades, in the reachable accelerating gradient of SC structures, making it possible, nowadays, to exceed $E_{acc} = 30$ MV/m with $RRR \sim 300$. The main drawbacks of the RRR increase after cavity fabrication are an increase in the BCS surface resistance, due to increased electron mean free path (low-field Q-values by about a factor of two smaller should be expected), and a drop in the yield strength of the material following any of the heat treatments needed for this purpose.

6.1.5 Global heating

SC cavities based on films of either Pb or Nb deposited onto Cu substrates do not generally show any thermal breakdown limitations, due to the excellent thermal properties of the substrate.

Whether a resonator is made of an SC layer coated onto Cu or it is defect free, the ultimate observable behaviour is a global heating, supposedly driven by a large number of extremely small defects in high magnetic field regions. This can be distinguished from a local effect either via thermometry methods or via the observation of a Q-decrease with temperature increase (before any local quenches) and without photoemission (no FE).

6.2 Anomalous losses related to the surface electric field

The electric component of applied EM fields can also extract a large number of electrons from an SC surface, and they may absorb a large part of the power driven to the resonator.

Electron field emission can be either a resonant or a non-resonant phenomenon, the former being usually relevant for rather low EM fields, the latter being the practical cause of performance limitation for the largest number of SC resonators.

6.2.1 Resonant field emission (multipacting)

The resonant field emission (RFE) is a dissipative process preventing the resonator from increasing its energy in spite of the increase of the feeding power.

The phenomenon is triggered by a low-energy electron (few eV) which is released from the SC surface for some reason, such as photoemission, cosmic rays or electrons from non-resonant FE. The electron gets accelerated by the local electric field and can be bent back to the surface by the local magnetic field. On impact with the surface, this may cause the emission of other electrons, whose number depends basically on surface features, such as the emitting geometry and its coefficient of secondary-electron emission, and on the impact energy. As long as the coefficient is larger than one and the electrons hit the surface at the same EM field phase, with the same energy and approximately in the same region, the process leads to resonant multiplication of the number of electrons involved. An avalanche develops, which absorbs all the power additionally fed into the cavity, preventing the resonator stored energy from increasing and hence freezing the EM fields at the values that determine the resonant process (see Fig. 8).

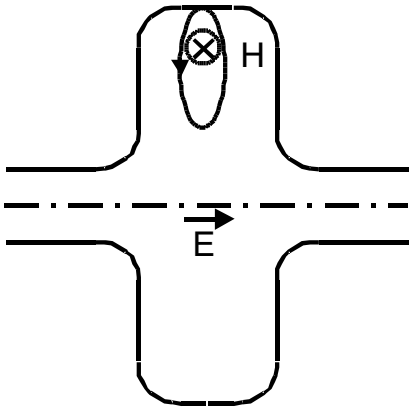


Fig. 8: Schematic of a first-order multipacting orbit. In one RF cycle the electron is bent back by the local electromagnetic fields onto the same emitting point, where it can extract secondary electrons and initiate an avalanche process.

Assuming that the electrons follow simple cyclotron orbits, the fundamental orbit frequency is $\omega_c \propto He/m$, H being the local magnetic field, e and m the electron charge and mass. To allow the resonant condition, H has to be such that emitted electrons are always in phase with the extracting EM field. The resonant condition can be satisfied clearly also for higher resonance orders: $\omega_g = n\omega_c$ (RFE of nth order).

These considerations explain why a number of RFE levels have usually to be overcome in a resonator. Of course the geometry of a cavity crucially affects the number of levels, their position versus input power and the ease with which they can be conditioned.

The coefficient of secondary-electron emission is peaked around 300–800 eV for most materials, while beyond 2000 eV it is lower than 1 for the large majority of them. That is the reason why RFE is a relatively low surface electric field phenomenon.

Ideal sites for RFE development are those regions in which the magnetic field does not vary significantly along the surface. Rounding off the equatorial region of multicell cavities to spherical (or elliptical) shape resulted in the complete elimination of multipacting, since the avalanche phenomenon is forced towards the equator line. This is where the magnetic field would allow RFE to develop, but the extracting electric field tends to zero, thus suppressing secondary electron multiplication.

Cleaning, baking, and applying gas discharges are known methods to reduce the secondary-electron coefficient. Furthermore, an RFE barrier is quite often overcome just by staying at an RFE level in high vacuum for minutes or hours (without any special discharge gas inlet).

RFE levels are easily recognisable in both steady-state and pulsed operation of the resonator, since one observes that the pickup signal on the oscilloscope remains locked at a certain level, while more power is driven from the input coupling line. One observes that the signal suddenly jumps to a higher value when the level is overcome (in some cases directly to the next level!). It is good practice to decrease the field to the same RFE level again and again until the level is actually cleaned. Computer-controlled procedures can replace the operator very efficiently and spare him this tedious work [23]. It is also advisable, whenever possible, to perform multipacting processing with the cavity in the NC state because of the larger resonator bandwidth and possible high power dissipation in the NC regime.

It should be finally mentioned that two-point multipacting has been observed when the resonance develops between two different points on the SC surfaces, e.g. mirror points around the equator in multicell cavities.

At present multipacting is not generally linked to the actual performance limitation of the cavity. It can be eliminated in multicell resonators by applying dedicated codes to find particular cavity shapes for which the phenomenon does not appear at all. However, it can still be an issue for low-β resonators, which can be described only by 3-D codes. In such a case the coarse definition of the EM fields makes it more difficult to determine the electron trajectories with sufficient accuracy. Hence there is no elegant means of trying to deal with RFE, other than to go through experimental tests. However, in most cases, RFE presents soft resonant levels, which can be overcome with properly devoted conditioning (taking maybe hours or days).

Nowadays RFE seems to play a more fundamental role in couplers, waveguide windows and coaxial lines (determining a maximum travelling-power limit) than in cavities themselves.

6.2.2 Non-resonant field emission: the phenomenon

The limit imposed by the Nb superheating field on the maximum achievable peak surface fields in β = 1 resonators is about 100 MV/m, which is still somewhat larger than the best updated performances ($\sim$ 80 MV/m). By far the most typical limiting mechanism to high fields in SC cavities is electron field emission (FE). In high electric field regions, electrons are emitted via quantum-mechanical tunnelling, which can dissipate a large part of the energy driven to the cavity. In some cases the limitation cannot be overcome, i.e. when the power absorbed by field-emitted electrons induces a quench.

The mechanism is usually explained by the modified Fowler–Nordheim (FN) equation, in which the emitted current density is given, in the RF regime, by [24]

$$J_{FN} = \frac{C}{\phi} S_{RF} (\beta_{RF} E_{RF})^{2.5} \exp\left(-\frac{B\phi^{3/2}}{\beta_{RF} E_{RF}}\right):$$

E_{RF} is the macroscopic electric field in V/m (i.e. assuming a smooth surface); ϕ is the material work function in eV; β_{RF} is the local field enhancement factor, which might be bound to the pin-like geometry of the emitting site (enhanced field emission – EFE); C is a constant which depends both on ϕ and β_{RF}; S_{RF} is the emitting area in m^2; and B is a numerical constant.

Some results show that at least half of FE sites can be well fitted with the modified FN law, whereas micron-sized particles are generally accepted to be the other most widely diffused cause of FE.

Field emitters vary substantially both in morphology and behaviour; hence, it is difficult to give a comprehensive description of the variety of experiments carried out so far on very different resonators. It is only possible to report here what are nowadays believed to be the most convincing explanations of the phenomenon.

First of all, the RF FE process has now largely been demonstrated to be similar to DC FE, except that the latter has an E^2 and not an $E^{2.5}$ dependence [25]: this resemblance induced many researchers to conduct experiments on the fundamental phenomena underlying FE in a DC regime. The onset time of

many discharges investigated in the DC regime was of the order of nanoseconds or less, with the initiation time decreasing with increasing field. These results are compatible with what would be allowed in the case of a typical SC RF cavity.

It was shown [26] that emission is certainly pushed to higher fields in a properly chemically etched surface, and that a high-RRR material and long (more than 3 h) heating at 1400°C can strongly reduce FE.

In some cases, the computed enhancement factors from FE current turned out to be much larger that the measured geometric irregularities of the surface in these regions. A micron-sized dust particle sitting on a geometric asperity was in similar cases proposed as an explanation (a multiplication of two enhancement factors $\beta_1\beta_2$ would be present). Where pure geometrical explanations fail, locally reduced work function (e.g. by contaminants) and the influence of the particle-substrate interface are proposed as a more complex explanation [26].

When the cause for FE appears clearly to be a foreign particle with poor thermal contact to the surface, thermionic emission (TE) is proposed as the explaining phenomenon [27], where the thermionic current density can be written as

$$J_{TE} = aE^2 \exp\left(-\frac{b}{E^2}\right).$$

Very revealing microscopic investigations of field emitters located by thermometry in 1.5 GHz cavities were done at Cornell [28], [29]. Emitters could be examined, at various processing stages, on the whole cavity surface (see also Fig. 9). It is deduced that the evolution of FE, increasing the applied field, consists of three phases:

- a pre-melting phase, in which FE is active, capable of degrading the Q, but where the current is still insufficient to melt the emitter;
- a melting phase, where the local E field is high enough that the current density exceeds about 10^{11} A/m^2 and the emission tip begins to melt;
- a third phase, in which the Joule losses of the emission current become so high that the emitter explodes and extinguishes.

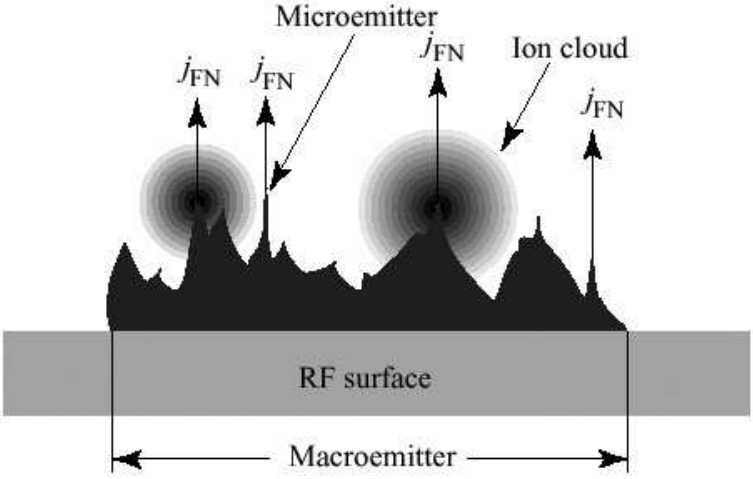


Fig. 9: Pictorial view of the macroemitter field-emission conditioning process, as proposed by J. Knobloch & H. Padamsee [29]

It is believed that initial electron emission causes, locally, sufficient heat to desorb gas, which is then ionized by the emission current. The ions are then accelerated towards the RF surface, thus reinforcing the process. The development of this large plasma would explain why melting phenomena are so much larger, in size, than the obviously small initial emitters.

6.2.3 Non-resonant field emission: the cures

Prevention is clearly the best cure for FE. This implies avoiding scratches or reducing their effect by means of a proper chemical treatment of the surface; on the other hand much effort is put into cleaning SC cavities prior to their installation, to minimize the number of particulate contaminants. The last step in cavity manipulation is an ultra-pure high-pressure rinsing with demineralized water (HPWR): this proved to be very effective in pushing the process to higher surface fields.

Before operating the cavity, RF processing can also be applied, with, for example, 60 kW/ 120 μs long pulses, allowing the necessary local current density for processing field emitters effectively to be provided. Gaseous He at a pressure around 5×10^{-5} Torr is often added to the FE conditioning recipe (He-conditioning). These processes lead to the destruction of FE sites and push the accelerating field further.

Perhaps the latest and most remarkable achievement, as far as overcoming FE is concerned, is the demonstration that 'nowadays' careful electropolishing (EP) allows the highest accelerating gradients by far to be reached and that it is superior to chemical polishing in the last treatment step. The formula for success, originally applied at KEK [15] and then adopted by other laboratories, is approximately the following: half-cell annealing (at 1440°C with Ti cage), followed by BCP and 100 μm EP. Annealing at 750°C and high-pressure rinsing complete the treatment. Acceptable drawbacks are that EP is somewhat slower than BCP (0.4 versus 1–2 μm/min) and that subsequent annealing is needed to outgas hydrogen, which is otherwise trapped in the bulk material during EP and can give rise to Q-disease.

7 STATE OF THE ART

Superconducting resonators have now been extensively adopted for the most various applications, from low-β heavy-ion linacs, to high-energy and high-current storage rings, recirculating linacs and free-electron lasers. The resonator developments are still very pronounced and this is the major reason why new projects have been funded or are proposed for funding.

I would like to recall here only those work-in-progress issues which look most striking to me.

7.1 Highest achievable gradients

The steady increase in the accelerating field has been impressive. Ten years ago $E_{acc} \sim 7$ MV/m was the limit. But then 10 MV/m was reached at TJNAF in Virginia, USA in 1992, 16.7 MV/m in 1996 and 23.9 MV/m in 1997 at DESY, Germany. While the present performances of multicell cavities seem to have settled around 25 to 30 MV/m worldwide, a recent jump in single-cell cavity field beyond 40 MV/m was obtained by means of electropolishing (pioneered at KEK, Japan and soon followed by many others). It remains to be seen whether (or when) this technology will also prompt a jump in the performance of multicell cavities.

For a very long time, the community of SC heavy-ion resonators has demonstrated its ability to reach, in a reproducible way, peak surface fields in excess of 50 MV/m (corresponding to $E_a \sim 25$ MV/m for $\beta = 1$ cavities). It should be emphasised that extreme performances have never been its main goal.

7.2 Reproducibility of excellent results

Beyond world records, it is crucial to demonstrate that the best results can be obtained with a large number of resonators, meaning that the various technological steps are under firm control.

This was demonstrated for example at TJNAF: The machine was designed for average accelerating fields of about 7 MV/m, but it has now been demonstrated that all their numerous cavities can reach fields between 25 and 30 MV/m very reliably. The same can be said with the experience gained by the TESLA test facility developments, with the additional merit that the goal of average fields between 20

and 30 MV/m has been reached through the technological transfer of the resonator production to several industries in different European countries.

7.3 New applications and new resonator shapes

The interest in intense proton accelerators for the treatment of nuclear waste, materials research and energy production has triggered the development of multicell resonators. These include modified shapes (matched for β-values from 0.5 to 0.8) [30–32] or rather old cavity shapes whose development had been left in an initial phase (spoke resonators [33] or re-entrant cavities [34], [35]). These resonators are susceptible to appreciable gradient increase in the next few years.

In the low-β cavity community, the development, construction and tests of SC RF Quadrupoles (RFQs) represent an extreme of complexity in the successful development of the technology of full Nb [19]. They feature a very large number of EBW joints and extremely high accuracy is required in the positioning of the RFQ electrodes (see Fig. 10).

7.4 Cost-reducing technologies

Last but not least, one should mention those technologies, the development of which would imply a significant cost reduction in the production of a large number of cavities. The proposed SC linear collider TESLA aims at producing 20 000 nine-cell resonators (obviously on an industrial scale): all R&D steps in the cost-cutting direction increase the probability of approval of that project. The technology of Nb sputtered onto Cu resonators is extremely interesting in order to minimize the amount of Nb used. The problems of the rather steep Q_0 drop at high fields must be solved. The development of hydroformed or spun resonators would be extremely valuable to eliminate the expensive EBW of multicell-cavity parts, once high E_{acc} values have been reliably reached. It should be emphasized that some multicell-spun resonators already exceed 30 MV/m, after significant chemical polishing (see Fig. 11). EP might also prove to be beneficial in this case.

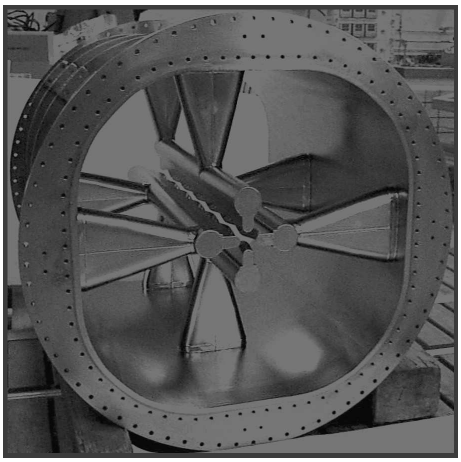


Fig. 10: The recently built and tested SC RF Quadrupole, part of the SC heavy-ion injector at INFN-LNL, exceeded a peak surface field of 28 MV/m CW and showed excellent alignment of the electrodes [19]

Hydroformed single-cell resonators very recently exceeded $E_{acc} \sim 40$ MV/m (once EP had been done at the end of the cavity treatment) [37]. Successful hydroforming of multicell cavities might follow quite soon.

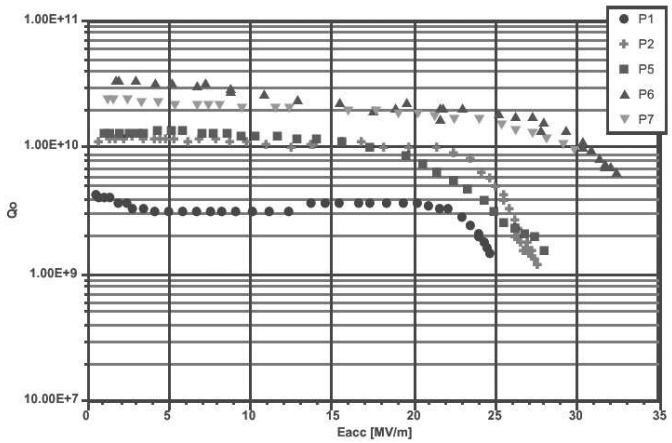


Fig. 11: Summary of the best performances achieved with five seamless monocell cavities made from high-purity niobium of $RRR \geq 250$. All Q_0 vs. E_{acc} dependencies are taken at a temperature of 2 K [36].

Acknowledgement

I wish to thank Anna Maria Porcellato for her critical revision of this manuscript.

REFERENCES

[1] H. Lengeler, CERN Accelerator School, Superconductivity in Particle Accelerators, CERN 89–04, pp. 197–229.

[2] V. Palmieri, Superconductivity, these proceedings.

[3] H. Padamsee, J. Knobloch and T. Hays, *RF Superconductivity for Accelerators* (John Wiley and Sons, New York, 1998).

[4] H. Lengeler *et al., IEEE Trans. Magn.* **MAG-21** (1985) 1014.

[5] M. Tigner and H. Padamsee, Cornell CNLS 82/553.

[6] P. Kneisel, Ti-gettering.

[7] K.W. Shepard, *Nucl. Instrum. Methods* **A382** (1996) 125.

[8] C. Benvenuti *et al.*, Proc. 8th Workshop on RF Superconductivity, Abano Terme (Padova, Italy), 6–10 October 1997, p. 1038.

[9] C. Benvenuti, N. Circelli, M. Hauer, *Appl. Phys. Lett.* **45** (1984) 583.

[10] S. Stark *et al.*, Proc. 8th Workshop on RF Superconductivity, Abano Terme (Padova, Italy), 6–10 October 1997, p. 1156.

[11] B. Bonin, CERN Accelerator School, Superconductivity in Particle Accelerators, CERN 96–03, p. 191.

[12] C.Z. Antoine *et al.*, Proc. 8th Workshop on RF Superconductivity, Abano Terme (Padova, Italy), 6–10 October 1997, p. 598.

[13] V. Palmieri, Proc. 8th Workshop on RF Superconductivity, Abano Terme (Padova, Italy), 6–10 October 1997, p. 553.

[14] P. Kneisel and B. Lewis, Proc. 7th Workshop on RF Superconductivity, Gif-sur-Yvette (France), October 1995, p. 311.

[15] K. Saito *et al.*, Proc. 8th Workshop on RF Superconductivity, Abano Terme (Padova, Italy), 6–10 October 1997, p. 553.

[16] C. Lyneis *et al.*, Proc. 1972 Proton Linear Accelerator Conference, LANL, Los Alamos (1972), 98.

[17] J. Knobloch *et al., Rev. Sci. Instrum.* **65** (1995) 3251.

[18] H. Padamsee, J. Knobloch and T. Hays, *RF Superconductivity for Accelerators* (John Wiley and Sons, New York, 1998), p. 145.

[19] G. Bisoffi *et al.*, Proc. 7th European Particle Accelerator Conference, in press.

[20] H. Padamsee, J. Knobloch and T. Hays, *RF Superconductivity for Accelerators* (John Wiley and Sons, New York, 1998), p. 174.

[21] H. Kitamura *et al.*, Proc. 8th Workshop on RF Superconductivity, Abano Terme (Padova, Italy), 6–10 October 1997, p. 667.

[22] H. Padamsee *et al.*, Proc. '1991 Electron Beam Melting and Refining Conference', November 1991.

[23] S. Canella *et al.*, Proc. 5th European Particle Accelerator Conference, Barcelona (Spain), 10–14 June 1996, p. 2109.

[24] J. Tan *et al., J. Phys. D: Appl. Phys.* **27** (1994) 2644.

[25] J. Tan, Proc. 7th Workshop on RF Superconductivity, Gif-sur-Yvette (France), 17–20 October 1995, p. 105.

[26] N. Pupeter *et al.*, Proc. 7th Workshop on RF Superconductivity, Gif-sur-Yvette (France), 17–20 October 1995, p. 67.

[27] M. Pekeler *et al.*, Proc. of the 7th Workshop on RF Superconductivity, Gif-sur-Yvette (France), 17–20 October 1995, p. 79.

[28] J. Knobloch, Proc. 7th Workshop on RF Superconductivity, Gif-sur-Yvette (France), 17–20 October 1995, p. 95.

[29] J. Knobloch and H. Padamsee, Proc. 8th Workshop on RF Superconductivity, Abano Terme (Padova, Italy), 6–10 October 1997, p. 994.

[30] C. Benvenuti *et al.*, Proc. 8th Workshop on RF Superconductivity, Abano Terme (Padova, Italy), 6–10 October 1997, p. 1038.

[31] H. Safa, Proc. 9th Workshop on RF Superconductivity, Santa Fe, NM (USA), 1–5 November 1999, in press.

[32] C. Pagani, Proc. 9th Workshop on RF Superconductivity, Santa Fe, NM (USA), 1–5 November 1999, in press.

[33] J.R. Delayen *et al.*, Proc. 1992 Linear Accelerator Conference, 24–28 August 1992, Ottawa (Canada), p. 695.

[34] P.H. Ceperly *et al., Nucl. Instrum. Methods* **136** (1976) 421.

[35] A. Pisent *et al.*, Proc. 7th European Particle Accelerator Conference, in press.

[36] P. Kneisel, V. Palmieri and K. Saito, Proc. 9th Workshop on RF Superconductivity, Santa Fe, NM (USA), 1–5 November 1999, in press.

[37] W. Singer *et al.*, Proc. 7th European Particle Accelerator Conference, in press.

CHOICE OF RF FREQUENCY

W. Pirkl *
CERN, Geneva, Switzerland

Abstract
A broad overview of effects and considerations that have a bearing for the accelerator environment is presented, and 'frequency' as a common input parameter. Aspects of beam dynamics, RF, and mechanical technology are briefly discussed to illustrate some contradictory requirements of accelerator design. While RF considerations are rarely a determining factor in that context, some trends can nevertheless be identified.

1 INTRODUCTION

In most situations the choice of RF frequency in an accelerator is determired by constraints that dictate a certain frequency or frequency band. This is particularly true for improvement programmes in existing machines. Rather than choices there are mostly contradicting trends in the choice of frequency, which will be discussed in what follows.

Most of the material presented is well known and has been published by other authors in internal reports or conference papers. In particular the scaling of cavity parameters vs frequency, cavity/beam interactions, properties of components, etc. have been the subject of extensive studies at CERN and elsewhere in the context of proposals for new machines.

This article does not pretend to present any new material, but intends to give a broad overview of a range of disjoint topics, with frequency as a common denominator.

2 REMINDER OF LONGITUDINAL PHASE SPACE PROPERTIES — RF VOLTAGE REQUIREMENTS

From a beam dynamics point of view the tendency is towards lower frequencies, as can be illustrated for the case of a synchrotron. It is assumed that the basic properties such as kinetic energy, transition energy, revolution frequency (f_{rev}) are frozen. The harmonic number h remains a free parameter and determines the operating frequency $h\cdot f_{\mathrm{rev}}$.

For a *single* stationary full bucket, bucket half-height BHH and bucket area A scale as a function of RF voltage V and harmonic number h:

$$\begin{array}{rcl} BHH \propto V^{1/2}\cdot h^{-1/2} & \text{resp.} & V \propto BHH^2 h \\ A \propto V^{1/2}\cdot h^{-3/2} & & V \propto A^2\cdot h^3 \ . \end{array}$$

Therefore, the RF voltage requirement increases linearly with the harmonic number for a given bucket *height*. Similary, the required voltage increases with the cube of the harmonic number, for a given bucket *area*. Figure 1 shows the bucket shape for different combinations of h and V.

If, however, the total longitudinal phase space area of the beam is distributed over h bunches, the required individual bucket area is reduced by a factor of h. In this case the required voltage for a given bunch area increases only linearly with the harmonic number.

The conditions are different if short bunch lengths are required. Suppose that a bunch of length σ_1, held by a voltage V_1, is adiabatically compressed to a bunch length σ_2. The required voltage V_2 in the central bucket region where the synchrotron frequency is constant is given by

* Formerly member of the CERN/PS Division. Now retired, member of the physics collaboration ASACUSA of the University of Tokyo at CERN.

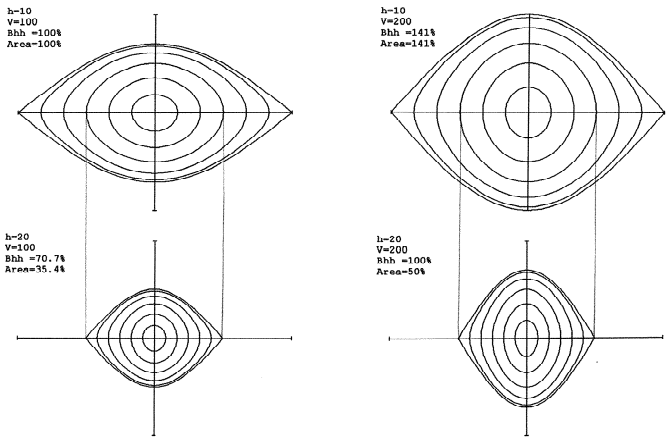


Fig. 1: Stationary bucket parameters for two different RF voltages (horizontal) and two different harmonic numbers (vertical)

$$V_2/V_1 \propto (\sigma_1/\sigma_2)^4/h \ .$$

The voltage requirement is reduced if high harmonic numbers are used, i.e. if the bucket is not unnecessarily wide. This consideration is, however, of a more theoretical nature since adiabatic bunch compression generally leads to prohibitively high RF voltages. Non-adiabatic methods like bunch rotation are then preferred.

2.1 Space charge

The usual figure of merit for evaluating space charge effects is the transverse tune shift ΔQ, which scales as $\Delta Q \propto 1/(\beta\gamma^2)$. The beam energy at the critical point, which is generally at the machine input, should therefore be as high as possible. Since the frequency in a synchrotron is proportional to β, there is a trend to extend the frequency range of the injector synchrotron to higher frequencies. While this will rarely lead to a different frequency band, it may be an important consideration for the improvement programmes of existing machines.

3 INFLUENCE OF THE TRANSVERSE PHASE SPACE — APERTURE

Frequency f and free-space wavelength λ are related by $\lambda = c_0/f$, therefore the two quantities may be used interchangeably for the argument. The electrical properties of of an RF structure with a given geometry are always determined in relation to the operating wavelength, hence dimensional constraints have a direct impact on the applicable wavelength or frequency. The cavity aperture plays a key role in this context.

Transverse phase space dynamics determine the beam size and the required aperture of the accelerating structure is determined by the radius of the good field region, i.e. the region around the beam axis where the RF accelerating field is sufficiently uniform for the application. The accelerating field tends to increase from the centre to the aperture radius and too large differences may lead to synchrotron–betatron coupling and other unwanted phenomena. Large beam tube radii and/or large good field regions consequently lead to large wavelengths, i.e. low frequencies, by essentially geometric considerations.

The trend wards to lower frequencies is enhanced by the transit time effect (see below). The gap width g should be as small as possible compared to the beam wavelength $\beta\lambda$. Therefore λ should preferably be large. In addition, large apertures increase the effective gap width, since the RF field extends deeper into the vacuum chamber than the geometric gap width. This is particularly true for

buncher cavities at the low-energy end of an accelerator, where the relative velocity β is small but the transverse dimensions of the beam are large.

Transverse focusing elements are often an integral part of the RF structure that necessitate a compromise between longitudinal and transverse performance. Examples include the drift tubes holding the focusing quadrupoles in an Alvarez structure and the electrode design in an RFQ whose modulation depth determines the ratio between longitudinal and transverse electric field strength. In general these compromises set a lower limit to the operating wavelength in some proportion to the transverse beam emittance.

Conversely, if the frequency is given by other considerations, the influence of the aperture on the remaining machine parameters is of paramount interest (see below).

4 REMINDER OF THE BASIC RF CAVITY SCALING LAW

The properties of the intrinsic cavity can be fully described by three independent parameters obtained from the electromagnetic field pattern, namely resonant frequency, stored energy W, and power losses P. The resonant frequency f_{res}, is defined by the condition that the electric and magnetic stored energy are the same, each occurring only twice during a cycle, their sum being constant at any moment. W is obtained by integrating the electric or magnetic field energy over the volume. Similarly, P is obtained by integrating of the losses over the volume and the conducting surface. W and P are normalized for a gap voltage V (more precisely the integral of the electric field strength along a defined path, generally the beam axis). These three parameters may be converted to other parameters such as the Q-factor and are in a one-to-one relation with the well-known equivalent circuit of the three lumped elements r, L, and C connected in parallel with

$$\omega_{\mathrm{res}} = 2\pi\, c_0/\lambda_{\mathrm{res}} = 2\pi f_{\mathrm{res}} = 1/\sqrt{LC}, \quad W = V^2 C/2,$$
$$\text{shunt impedance, CIRCUIT definition } r = V^2/(2P)$$
$$\text{characteristic impedance } X = \omega_{\mathrm{res}} L = 1/(\omega_{\mathrm{res}} C) = \sqrt{L/C} = r/Q$$

Please note that for historical reasons the LINAC definition of the shunt impedance R (R upper case!) is still widely used. There the cavity voltage is considered as an RMS quantity, hence the required power P for a given intrinsic cavity voltage V is $P = V^2/R$. In the preferred CIRCUIT definition of r (r lower case!) the cavity voltage is realistically taken as the peak value of a sinusoidal waveform, which then leads to $P = V^2/(2r)$. It follows that $R = 2 \times r$.

Many definitions of the shunt impedance are based on the effective cavity voltage V_{eff} seen by the beam. It is different from the intrinsic cavity voltage V because the RF field changes during the transit by time of a particle through the gap region. This is conveniently expressed by the transit time factor $TTF = V_{\mathrm{eff}}/V$, $T \leq 1$. It depends on the ratio effective gap length g to beam wavelength $\beta\lambda$ and is approximatively given by

$$TTF \approx \frac{\sin\left(\pi \frac{g}{\beta\lambda}\right)}{\left(\pi \frac{g}{\beta\lambda}\right)} \ .$$

There are three scaling laws that permit complete determination of the electric parameters if all cavity dimensions are changed (blown up) by some factor in all dimensions:

- cavity dimensions/λ_{res} = constant,
- characteristic impedance r/Q = constant,
- Q-factor $\times$ skin depth/λ_{res} = constant.

For a given cavity wall material, the skin depth is proportional to $\sqrt{\lambda}$, therefore the Q-factor increases according to the third scaling law by the same proportion. If a given geometry such as a pill-box cavity is scaled for higher frequencies or lower wavelengths, the shunt impedance $R = Q \times r/Q$

decreases in proportion to $1/\sqrt{f}$. This basic tendency is enhanced by additional loss mechanisms at high frequency, in particular surface roughness, which increases the geometrical path length of the RF currents on the surface. It appears therefore that higher frequencies would be counterproductive as far as single-cavity performance is concerned.

5 SCALING OF CAVITY PARAMETERS PER UNIT LENGTH

The picture changes if assemblies of cavities are considered. Suppose that a certain RF voltage requirement V is distributed over n cavities of the same type. Then the individual cavity voltage becomes V/n, the individual power dissipation becomes $(V/n)^2/2r$, and the total power becomes $n(V/n)^2/(2r) = V^2/(2nr)$, i.e the power needed is reduced by a factor of n with respect to a single-cavity configuration.

It therefore makes sense to see how many cavities can be stacked per unit length l, in order to evaluate the performance of a cavity assembly rather than a single cavity, in other words to normalize the cavity parameters by axial length.

Commonly used parameters used in this context are

- R': effective shunt impedance (LINAC definition) per unit length
  $R' = R \times TTF^2/l = G^2/P'$,
- TTF: transit time factor, $G = V \times TTF/l$: gradient (effective voltage per unit length)
  P': power per unit length
  r': characteristic impedance per unit length, $r' = R'/Q$.

Figure 2 gives the characteristics of a travelling-wave structure at 29 GHz.

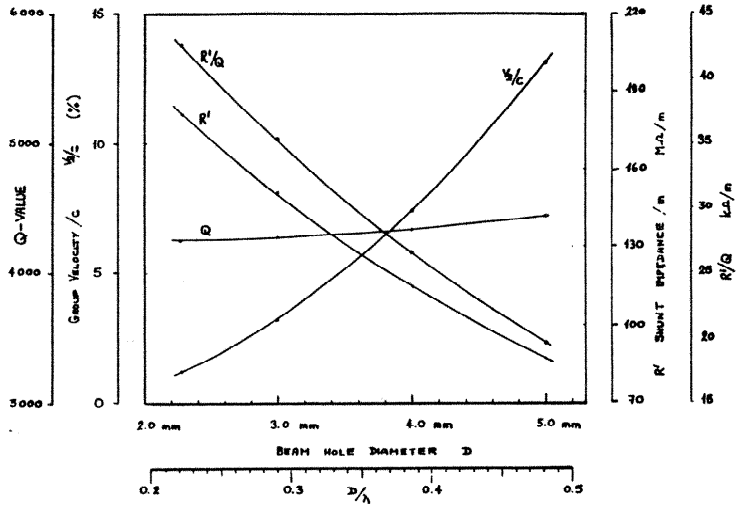


Fig. 2: R', R'/Q, Q, group velocity v_g/c as function of aperture. Reproduced from Ref. [1].

The curve for R' is approximately inversely proportional to the beam hole or aperture D normalized to the wavelength λ. This behaviour is confirmed by similar calculations at different wavelengths. It is therefore common practice to consider R' proportional to $1/(a\lambda)$ in general, or alternatively to take the quantity $R'a/\lambda$ as a figure of merit for comparisons.

A comparison of accelerating structures in different machines is given in Fig. 3.

It shows that $R'a/\lambda$ scales approximately as $\omega^{1/2}$. For cavity assemblies, the use of higher frequencies is therefore advantageous from the point of view of power consumption.

$$R' = r'Q \propto \omega^{1/2}(a/\lambda)^{-1}$$

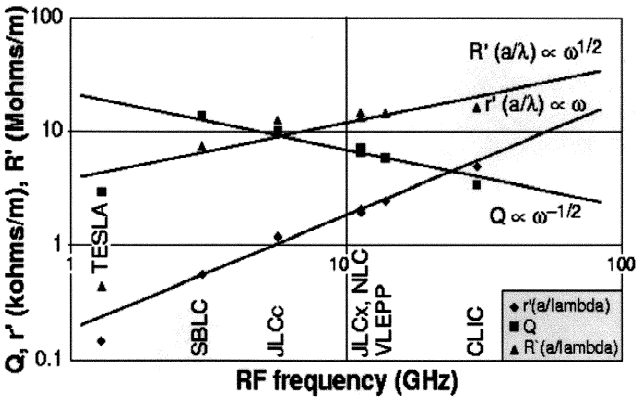


Fig. 3: Structure parameters of different machines vs frequency. Reproduced from J.-P. Delahaye et al., Ref. [2].

6 VOLTAGE HANDLING CAPABILITY

One of the most quoted figures of merit for the voltage handling capability of vacuum is the 'Kilpatrick criterion', published in *The Review of Scientific Instruments*, Vol 28, No. 10:

$$W \cdot E^2 \cdot e^{-\frac{1.75 \cdot 10^5}{E}} = 1.8 \cdot 10^{14}$$

where W is the impact energy of an electron in [eV] and E is the local field strength in [V/m].

The physical mechanism behind this empirical formula consists of two parts: first the impact of an electron of kinetic energy W on the cavity surface, second the re-acceleration of the liberated secondary electrons by the local field strength E. For DC fields without local field enhancements, $W = E \cdot d$, where d is the distance between the electrodes. As the frequency increases, a transit time factor starts to act, resulting in $W < E \cdot d$; the same is true for local field enhancements.

To solve this implicit formula, the use of a numerical root finder is necessary. Under much simplified assumptions an often used approximation is

$$E \sim 25 \cdot f^{1/2} \ \ (E \text{ in MV/m}, \ f \text{ in GHz})$$

The important fact is that *the voltage handling capabilty in vacuum increases at high frequencies*. This is the key argument for going to very high frequencies in proposals for linear accelerators/colliders at very high energies, where the necessary overall length of the accelerating structure for a given final energy becomes a determining cost factor.

The voltage handling capability is usually expressed as the net overall gradient in units of MV/m of a given configuration. It should be kept in mind that the internal cavity field strengths are higher by a factor of about two, partly due to a geometric factor (cavity length)/(gap length), partly due to local field enhancement on the edges of the gap.

The practical voltage limit is influenced by many factors such as vacuum condition, surface smoothness, surface deposits, and temperature. The Kilpatrick criterion is therefore considered more as a generally accepted figure of merit than a hard limit. It can be exceeded in practical designs by 'bravery factors' of about 1.5 (current in RFQ designs) or up to 2.5 (short pulses). It should also be mentioned that the Kilpatrick criterion is pessimistic at lower frequencies and corrections to the formula have been proposed.

One important factor is the RF pulse length. Experience shows that single-pulse voltage limits decrease with longer pulse duration. One possible mechanism for this effect is the heating of microscopic

whiskers on the cavity surface that then act as thermal emitters, with time constants in the range of microseconds to milliseconds. On the other hand, the electric breakdown is statistical by nature and the ratio (pulse length)/(average time between sparks) determines the likelihood of a breakdown.

Increased voltage breakdown limits can only be exploited up to a certain point, then thermal effects become the limiting factor (see Fig. 4).

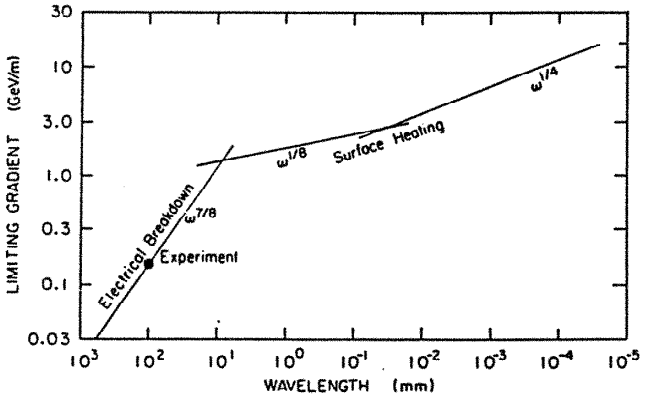


Fig. 4: Limitations on gradient as a function of wavelength due to electric field breakdown and surface heating in a SLAC-type disk-loaded structure. Reproduced from Ref. [3].

The average power density at the cavity surface increases with frequency due to the combined effects of reduced skin depth, reduced cavity dimensions and reduced shunt impedance. In addition, it is not sufficient to consider only the average power dissipation integrated over the pulse repetition time, but the thermal transient must also be taken into account. As the heat wave cannot be dissipated immediately from the outer layers to the inside of the walls, the surface attains dangerously high temperatures. Permanent damage such as roughening or cracking of the copper surface takes place above certain gradients.

7 WAKEFIELDS: ANOTHER LIMITING FACTOR TOWARDS HIGHER FREQUENCIES

Intense particle bunches create longitudinal and transverse wakefields, which dilute the time structure and/or may lead to transverse beam loss. For a given cavity geometry, the wakefield effects scale as ω^2, ω^3, or ω^4, depending on different simultaneous scaling of the beam parameters. Increasing the aperture reduces the transverse wakefield effect, as shown in Fig. 5.

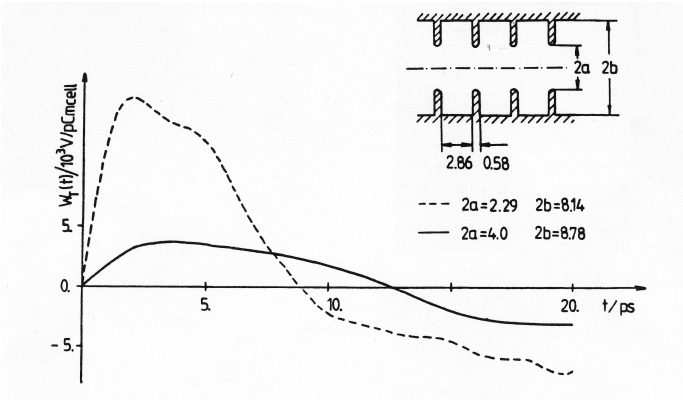


Fig. 5: Transverse wakefield for the structure given above. From Ref. [4]

Opening the aperture leads in turn to decreased shunt impedance and potentially other undesirable effects (increased group velocity, which in turn may lead to increased spark damage). The best overall compromise must take into account many contradictory requirements. This is the subject of very detailed ongoing studies in many laboratories, the details of which are beyond the scope of this article.

8 HARDWARE CONSIDERATIONS WITH LIMITED INFLUENCE ON THE CHOICE OF FREQUENCY

Under this heading there are two hardware-oriented issues, which demand different technical approaches, but which rarely play a determining role in the choice of frequency:

Superconductivity: Proof-of-principle SC cavities have been built to operate between about 80 MHz and 30 GHz, with plans to go to up to 80 GHz. Outside these rough limits, considerable R&D efforts may be necessary for new projects.

The superconducting material is mostly pure niobium, but niobium alloys and lead have also been used. The superconducting material may be either in bulk or applied in a thin sheet by sputtering or electrolysis. Simply stated, the dominant factor at high frequency is the achievement of low BCS resistance, which requires very low cryogenic temperatures. The dominant factor at low frequencies is the residual resistance of the SC conductor, which in turn requires special material properties (pure niobium is preferred). Another concern at low frequency is the large surface of the cavities, with associated difficulties with the rinsing, etc., which may lead to unusual cavity shapes.

Gradients of 44 MV/m have been attained at 1.3 GHz and the limit continues upward as more sophisticated methods of surface cleaning are developed and progress is made in materials science.

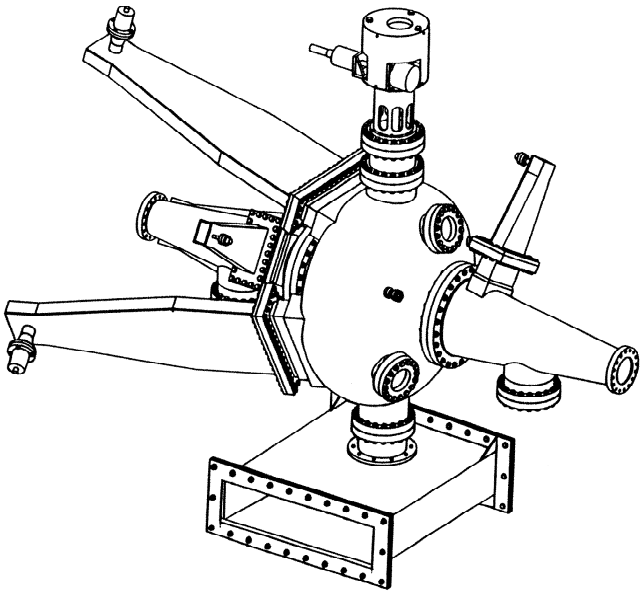


Fig. 6: Sketch of the DAΦNE RF cavity. From Ref. [5].

Higher-order-mode (HOM) damping: the operating frequency of a cavity is usually the fundamental mode at the lowest frequency, amongst a host of other unwanted resonances of higher order. To come close to the cavity designer's dream of a 'monochromatic' cavity, a either high-pass coupler, acting only on the HOMs, or a bandstop filter damping all frequencies except the fundamental, is necessary.

The technology to achieve this depends strongly on frequency, but solutions are known for all usual frequency ranges.

An example that fulfils an extreme requirement is shown in Fig. 6. The spectrum of ultrashort bunches in the DAΦNE and PEP machines extends to very high frequencies. Three waveguide couplers stretching out to the left in parallel to the beam line provide the necessary HOM damping and characterize the overall shape of the assembly.

9 RAPIDLY TUNABLE AND μ-LOADED CAVITIES

Certain applications require a rapid change of frequency. Examples include beam acceleration in rapidly cycling synchrotrons or longitudinal beam gymnastics with a change of orbit. Only a very limited range of a fixed-tuned cavity, the instantaneous (power) bandwidth, is directly usable for this purpose. Larger deviations require a dedicated fast tuning mechanism of some kind.

Slater's theorem relates the change in frequency ∂f to the corresponding change in cavity-stored energy ∂W for infinitesimally small changes:

$$\frac{\partial f}{f} = \frac{1}{2} \cdot \frac{\partial W}{W} \quad \text{or} \quad \Delta W \geq 2 \cdot W \cdot \frac{\Delta f}{f} \ .$$

This relation provides a first estimate of the total amount of stored cavity energy to be changed. For a frequency-agile cavity the overall stored energy W should therefore be kept as small as possible. The total necessary variation ΔW has a tendency to be much larger than the estimate from the above relation, since optimal power transfer from the generator to the cavity is difficult to achieve over the additional bandwidth.

The stored energy can be changed either capacitively or inductively. In practice, capacitive tuning can only be achieved by mechanical means since no static components are known to vary capacitance at high power. For inductive tuning, ferrites allow variation of their incremental permeability μ_{r} as a function a biasing magnetic field. However, their energy storage capability is rather limited, as shown in the following table

Material	Operating conditions	Stored energy/Volume [J/cm^3]
Vacuum	$E = 100$ kV/cm	440×10^{-6}
Ferrite	$B_{\mathrm{RF}} = 0.01$T, $\mu = 10$	4×10^{-6}

The active volume of an inductive tuner is therefore about two orders of magnitude larger than that of the capacitive counterpart.

9.1 Mechanically tuned cavities

Mechanically tuned cavities have the virtue of small size and low losses, but compromises between size and speed have to be made due to mechanical constraints. Therefore they are best suited to high frequency–slow speed applications. The concept of mechanically tuned resonators has gained a rather bad reputation, following negative experience with large rotatable capacitors. However, recent motion-control technology could bring a change.

9.2 Ferrite-tuned cavities

Ferrite-tuned cavities dominate the field of fast cycling machines. Ferrite cores, mostly of the nickel–zinc variety, can be arranged as integrating part of a cavity or as separate tuners. In either case provisions to apply the bias magnetic field have to be made using bias windings or external electromagnets with an

iron core. The biasing magnetic field is oriented in parallel to the RF field. Tuning speed limits occur because of intrinsic ferrite effects flipping the microscopic magnetic domains, but also because of power supply limitations and eddy current effects in the tuner walls.

The product $\mu Q f$ is a figure of merit for ferrites. It attains 200×10^9 for low induction, but generally drops for practical RF amplitudes because of eddy currents. Therefore the core losses increase much faster than U_{RF}^2. The frequency swing obtainable in practice is determined by the obtainable μ-range, in other words the highest (unbiased) μ, since the lowest obtainable relative permeability for infinitely strong bias is 1.

The highest practical RF induction in a ferrite core is around 0.01 T or 100 G. Excessive losses, non-linear effects, instabilities, and thermal runaway may start above this approximative limit. This limit is responsible for the modest capability to store/shift energy. It also sets a practical limit for the lowest operating frequency at reasonable dimensions, since for a given geometry and induction limit the applicable RF voltage is directly proportional to the frequency. Since ferrite material is brittle, care must be taken in the design of the cooling system to avoid thermal gradients that typically lead to radial cracks in ferrite rings.

Within the constraints sketched above, ferrites are available for the use in frequency range of a few hundred kilohertz to a few hundred megahertz. The demand for ferrites for switched-mode power supplies at ever increasing frequencies has led to improved ferrites up to the MHz range with slightly higher saturation limits.

9.3 Orthogonally (transversely) biased ferrite tuners

Special ferrite compositions, in particular garnets originally intended for microwave applications, rely on a different mechanism to change permeability, namely on precession effects. These ferrites are characterized by very low losses at low permeability, but require very strong biasing fields oriented orthogonally to the RF field. These ferrites are an attractive alternative for applications above about 50 MHz.

9.4 Glaseous metal cores (FINEMET$^{\mathrm{TM}}$, VITROVAC$^{\mathrm{TM}}$)

In recent years a new type of material, consisting of amorphous and galseous metals, has found increased applications. It is obtained by special thermal treatment of the basic high-μ metal to suppress the crystalline structure. Thin metal bands are then wound together with insulating layers to form rings of appreciable diameters (50 cm and more). The result is a core with very high permeability that can be used up to 5–10 MHz, with saturation limit in the kilogauss range, however with much higher losses than ferrite cores.

Two big advantages can be gained in accelerating cavities with this material.

- The obtainable cavity gradient is more than one order of magnitude higher than in ferrite cavities, i.e. several hundred kV/m compared to about 10–20 kV/m, depending on frequency. This is a decisive feature for applications where the space in the accelerator is limited. Also fixed-tuned cavities for constant low frequency can be built to deliver high RF voltage with reasonably small dimensions.
- The cavity bandwidth is inherently very large. In addition the required tuning bias field is much smaller than with ferrite cavities. The tuning system for frequency-agile cavities is therefore drastically simplified. Due to the very large instantaneous bandwidth, the implementation of multi-harmonic accelerating waveforms instead of pure sinusoids is possible with a single cavity. This is an important factor for more sophisticated applications such as beam gymnastics using barrier buckets. Last, but not least, the low cavity shunt impedance reduces the risk of instabilities due to beam–cavity interaction.

The price to pay is the much higher RF power requirement. Compared to ferrites, the rings have better thermal conductivity combined with higher possible operating temperatures and reduced susceptability to thermal gradients. The increased injected power is therefore more easily dissipated by a cooling system.

10 MISCELLANEOUS TECHNOLOGICAL CONSIDERATIONS

10.1 Availability of RF amplifier equipment

There are mainly three overlapping frequency ranges that determine the choice of RF power equipment for the main line of accelerator applications.

Figure 7 resumes the situation. The figure dates from the 1970s, but it still reflects the basic physical limitations.

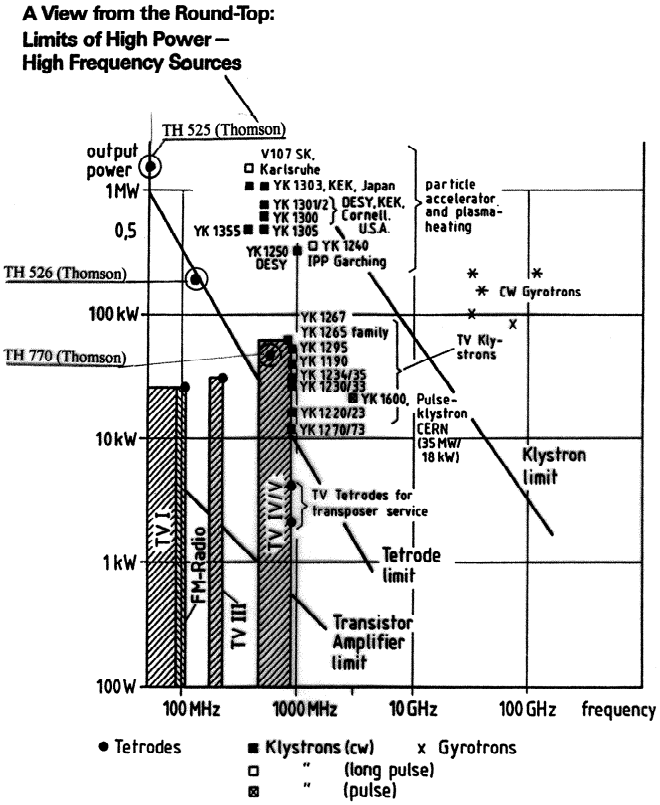


Fig. 7: Approximate limits of different RF amplifiers. From Ref.[6].

10.1.1 Up to ∼300 MHz: Electron tube technology

This is the oldest technology, used almost exclusively for communications, TV etc. until the 1950s. The amplification mechanism is based on the ideally powerless *density* modulation of an electron beam through a control grid. The RF structure thus obtained converts DC power to useful RF power. Continuous power outputs up to about 1 MW, or peak powers of about 5 MW can be obtained from single tubes.

At high frequencies, transit time effects appear that necessitate a reduction in the dimensions, which in turn leads to a reduction in power handling capabilities. In addition, the control of the electron

beam density can no longer be achieved without power losses. Parasitic capacitive coupling reduces stability margins, forcing the use of such tubes in grounded grid configurations which additionally reduce the obtainable gain.

Practical amplifications are 10–20 dB per stage, with DC /AC efficiencies at 40–80%.

A large fraction of tube manufacturers has gone out of business in the last ten years, including the firm that originally created Fig. 7. This is a sign of the outmoded technology and a general trend in telecommunications towards higher frequencies.

10.1.2 ∼300 MHz to ∼10 GHz: Klystrons

Klystrons are the preferred power sources above about 300 MHz due to their large amplification and high power capabilities. Contrary to 'gridded' electron tubes, the electron beam is initially modulated in *velocity* (rather than density) by an RF cavity. The density structure ultimately required builds up after a well-defined drift length, where the faster electrons meet and overtake the slower ones. Passive RF cavities are situated at these spots to reinforce the velocity/density modulation. Each intermediate passive cavity adds about 10 dB of gain. An output cavity finally collects the resulting RF component. The required drift distances depend on the beam wavelength $\beta\lambda$, where β represents the relative velocity of the electron beam and λ the free-space RF wavelength. β is limited by the applicable beam accelerating voltage (300–400 kV), hence the required klystron length is almost directly proportional to the RF wavelength and the required gain. This sets a low-frequency limit for these devices.

The gain is usually 40–60 dB, the efficiency 40–65 %. The output power attains several megawatt CW and tens of megawatts for short pulses. The output cavity and output ceramic window are the main power limiting factors.

10.1.3 Above ∼10 GHz : Gyrotrons, two-beam accelerators

The Gyrotron principle is based on a tubular and gyrating electron beam. This allows for the use of higher-order azimuthal modes in the output cavity to circumvent the limitations of the fundamental-mode cavities used in klystrons. Gyrotrons are mainly used for plasma heating applications, e.g ECR particle sources.

In the context of TeV linear accelerators the principle of a two-beam accelerator was proposed for the Compact LInear Collider (CLIC) working at 30 GHz that is presently under study at CERN. The low-intensity, high-energy electron beam used for physics is driven by a high-intensity, low-energy proton beam running along the main structure in a separate beam pipe. Dedicated transfer structures extract the 30 GHz component from the tightly bunched drive beam and feed it to the accelerating structures. This concept minimizes the number of active elements in in the tunnel. Ingenious schemes are used for the compression and the time-wise distribution of the drive beam (see Ref. [7]).

10.1.4 Semiconductors for all frequencies

The rapid progress in semiconductor technology that started with the transistor has led to a large family of devices that cover practically all frequencies from DC to the optical range. RF semiconductors, which are inherently low-power devices due to their small dimensions, have nevertheless made large inroads in the medium-and high-power range due to power-combining techniques. A single amplifier or generator may be composed of dozens or hundreds of individual modules that may in turn consist of many basic cells.

In view of the continuing progress in this field, it is impossible to state precise limits. The present state of the art is represented by a multi-kilowatt pulsed amplifier at X-band (10 GHz).

10.1.5 Two exotic active elements for the microwave range

There are active elements in widespread use in general microwave applications, but they are seldom found in the accelerator world. Examples are the magnetron and the TWT. The former is mass produced for microwave ovens and similar appliances, extremely cheap and powerful, but lacking the necessary linearity and phase stability. The latter is a wideband device but this main advantage is of little importance for high-energy accelerators where high Q structures of moderate passband are normally used.

10.1.6 An emerging technology: Laser acceleration

Beam acceleration by femtosecond laser beams of extremely high instantaneous power has been achieved in recent years. Ultrahigh gradients over fractions of a millimetre have been demonstrated and "acceleration gradients a million times larger" compared to classical means seem possible [8].

10.2 Radio-frequency interference (RFI), electromagnetic compatibility (EMC)

The reduction of spurious emission from RF equipment was of minor concern in the past, since most RF amplifiers on the market were intended to operate close to the transmitting antennas designed to radiate the available power with maximum efficiency into the environment. Transmitter halls in accelerators were known as electrically noisy places, where low-power equipment had to be adequately protected. In other words, emphasis was put on reduction of equipment susceptability rather than reduction of emitted fields. This situation has gradually evolved, and the meeting of applicable international regulations limiting the maximally admissible field strength has become a serious issue.

Accelerators fall into the IMS (Industrial Medical Scientific) category. Inside closed areas the interference level may be very high, up to the (medical hazard) limit of 1V/m or 120 dbμV, provided that mutual agreement between transmitter and receiver can be achieved and the limiting values outside the closed area are met.

Emission to the outside world is regulated according to frequency bands and geographic location. There are industrial frequency bands where higher than normal emissions are tolerated. Inversely there are protected bands where special lower limits apply, e.g the different air traffic NAV bands (in particular 140–160 MHz), the hydrogen window at 1542 MHz or GPS in the same L-band. However, the ever changing technology, in particular the communications revolution, leads to varying regulations. As examples the 2.4 GHz band, formerly reserved for microwave ovens, is now widely used for Bluetooth and WiFi. The 27 MHz band, formerly reserved for wireless home telephones, is now reserved for industrial remote control.

As a general recommendation, the applicable RF interference regulations and possible local restrictions (e.g. in sensitive experimental areas) should be carefully identified and followed to avoid retrofitting of already installed equipment.

10.3 Dimensional constraints

While RF structures can be scaled over a wide frequency range, there are nevertheless soft limits on either side of the spectrum that determine the practical feasibility of a design.

10.3.1 Towards low frequencies

Towards low frequencies, the sheer size of the usual resonating structures becomes a limiting factor. Beam lines in accelerators usually run at heights of 1–1.5 m, and this is at the same time the maximum practical radius for a cavity. A pure pill-box type cavity of this radius resonates at 120.80 MHz. For any lower frequency some capacitive or inductive loading becomes necessary. *Inductive* loading can be achieved by the introduction of high-μ material (ferrites of amorphous metal, see above), whereas dimensional loading (coil) is excluded due to the necessity of transmitting the resonator voltage to the

beam. *Capacitive* loading can be implemented by inclusion of materials of high dielectric constants such as ceramics, but these materials risk being destroyed by a single voltage breakdown. The usual way is to provide dimensional loading in the form of integrated capacitors, ranging from nose cones to large plates (see Fig. 8 for an extreme example).

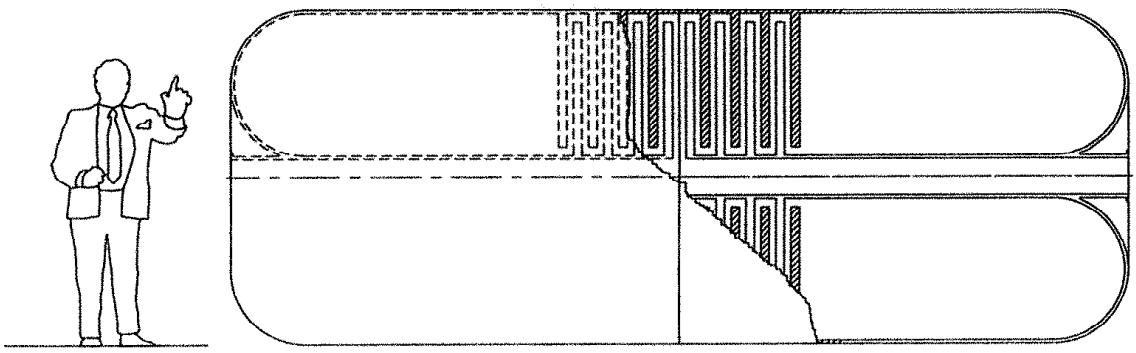


Fig. 8: Capacitively loaded high-power cavity for 500 kV gap voltage. From Ref. [9]

10.3.2 Towards high frequencies

Towards high frequencies, several phenomena come into play:

- Mechanical tolerance: structural tolerances become generally more critical for smaller dimensions of RF resonators at very high frequencies. In particular, careful alignment of the whole structure with reference to a nominal beam axis is mandatory to avoid the excitation of transverse deflecting modes. The "kick factor" $k_{1,}$ which characterizes this potentially lethal effect, scales as f^3 for a given ratio of aperture a to wavelength λ. Even if its impact is counteracted by larger aperture and other means, the required prealignment tolerances decrease with frequency. Figure 9 gives an idea of the tolerances required by different machines.
- Required surface finish: the well-known skin effect limits the penetration depth of RF fields into matter. At 10 GHz the skin depth in copper is ~ 0.65 μm, i.e. of the order of optical wavelengths, and this scales as $f^{-1/2}$. If the surface roughness is of the same order as the skin depth, the effective path length for RF currents along the surface is increased and so are the losses. Mirror-like surface finish is therefore required for cavities working above the cm-wavelength range.
- Pumping performance: the molecular conductance C_{mp} per unit length of vacuum ducts of diameter d scales as

  $$C_{\mathrm{mp}} \propto d^3 \ .$$

  Providing adequate vacuum quality becomes increasingly difficult with the smaller beam pipe diameters at very high frequencies.

11 MACHINE ENVIRONMENT: IMPACT OF PHYSICS REQUIREMENTS, NUMEROLOGY AND BUNCH TRANSFER SCHEMES

The determining factors for the choice of an RF frequency stem from the requirements of the physics experiments, together with the characteristics of existing machines in an accelerator chain that must be preserved.

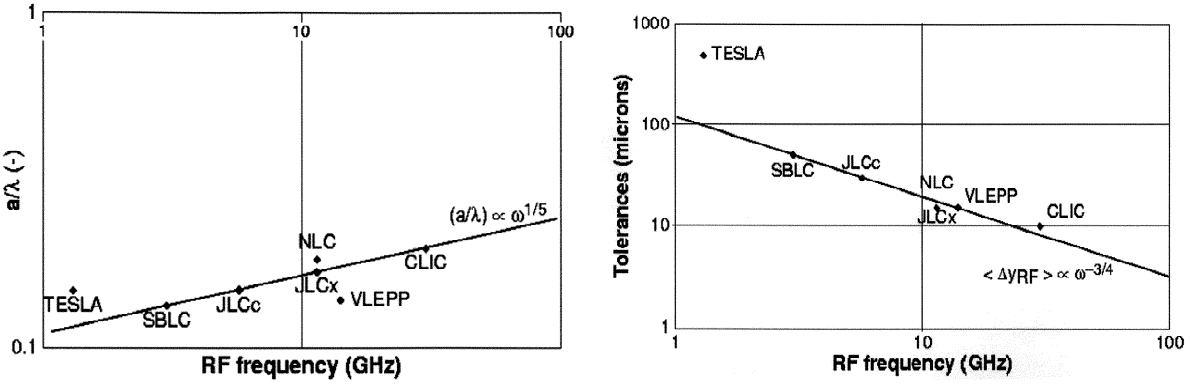


Fig. 9: Aperture and alignment tolerances in different machines

The influence of these parameters is best illustrated by the improvement programme of the PS and SPS synchrotrons at CERN for the LHC project [10]. The range of possible bunch spacings is primarily determined by luminosity requirements. This favours the concentration of the total number of available particles in a small number of bunches, leading to large bunch spacing. The characteristics of the particle detectors have also to be considered; their maximum data rate and dead time lead to another set of constraints. Within the basic acceptable range, a multiple of 5 ns must be taken to preserve the 200 MHz system of the SPS (5 ns bunch spacing). The value of $5 \times 5 = 25$ ns has finally been chosen.

Where to create the 40-MHz bunches that correspond to the 25 ns bunch spacing? In this case, it was decided to include 40 MHz and 80 MHz systems in the PS. The numerology, i.e. the ratio of the respective diameters of the three machines involved, allowed the approach of batch transfer of a number of bunches of the right spacing from in a single shot. The injection/ejection kicker rise times in the different machines determine the kicker gaps, i.e. the number of lost or missing bunches. In addition, the position of these missing bunches in the resulting bunch train has to obey certain criteria to avoid instabilities due to transient beam loading (Pacman effect). Under less favorable conditions, single bunch transfer may become necessary.

The luminosity requirements of the LHC demand a higher intensity than the four booster rings can provide in a single cycle. The PS can accept a total of eight high-intensity booster bunches (albeit at higher injection energy) and the most attractive solution was a two-batch filling scheme, whereby each batch of four bunches occupies one-half of the PS circumference. This scheme requires operating the Booster synchrotron at harmonic number $h = 1$ rather than at $h = 5$. Consequently the basic RF system was redesigned to work at one-fifth of the previous frequency.

12 CONCLUSION

Considerations of beam dynamics and accelerator physics are the key factors for the choice of RF frequency in accelerators. Questions arising from RF technology are secondary considerations. As a soft general statement it can be said that innovations in RF technology and the quest for higher beam energy push towards higher frequencies, whereas beam dynamics considerations tend to favour lower frequencies.

ACKNOWLEDGEMENTS

My thanks go to the colleagues who gave valuable hints and put their work at my disposal for this review. This includes the authors in the public domain whose results are mentioned in the general context. The original figures have been used in tribute to the contributors

REFERENCES

[1] I. Wilson, Electrical characteristics of travelling-wave disc-loaded loaded waveguide structures at 29 GHz, CERN CLIC Note 46.

[2] J.-P. Delahaye et al., Scaling laws for normal-conducting e^+ e^- linear colliders, Proceedings of LINAC Conference, Chicago, US, 1998, MO 1004.

[3] P.B. Wilson, Linear accelerators for TeV-colliders, SLAC Publication 3674 (1985).

[4] H. Henke, Wake fields in a 30 GHz Radio-Frequency Linac, CERN LEP-RF/87-36, also CLIC Note 40.

[5] R. Boni and A. Gallo, Status of DAFNE RF system, Proceedings EPAC 1994, London, p. 1826.

[6] VALVO (subsidiary of Philips), 1985.

[7] The CLIC RF power source: a novel scheme of two-beam acceleration for electron–positron colliders, CERN-99-06 (1999).

[8] D. Habs, LMU München. Private communication.

[9] W. Pirkl, Elements of RF design, The HIDIF Study, GSI-98-06 Report, Darmstadt, 1998, p. 119.

[10] The PS complex as a proton pre-injector for the LHC: design and implementation report, CERN-2000-003 (2000).

H-TYPE LINAC STRUCTURES

U. Ratzinger[*]
GSI Darmstadt, Darmstadt, Germany

Abstract

H-type cavities have been successfully developed over the last 25 years to serve for a large variety of applications in the field of ion acceleration. Radio Frequency Quadrupole (RFQ) and Drift Tube Linac (DTL) versions were designed in the H_{11}- as well as in the H_{21}-mode. The Interdigital H-type (IH) drift tube structure ($H_{11(0)}$-mode) is efficient for an energy range from 100 keV/u to 30 MeV/u, especially when combined with KONUS beam dynamics. In overall power consumption the IH-DTL can compete with existing superconducting linacs up to beam energies of around 2 MeV/u. Additionally, effective voltage gains as high as 10.7 MV/m were demonstrated in pulsed operation. The GSI High Current Linac has reached new marks for drift tube structures with $(A/q)_{max} \approx 60$ and $I_{max}/emA = 0.25 \cdot A/q$. The effective voltage gain of that 36 MHz, 80 MV IH-linac is 4.2 MV/m. The IH-RFQ (H_{110}-mode) is well suited for heavy-ion beams with $A/q \geq 10$, and the four-vane RFQ is very well established for proton and light-ion acceleration. At beam energies between 5 MeV/u and 150 MeV/u the Cross Bar H-Type (CH) drift tube structure (H_{210}-mode) also shows great potential for room temperature and for superconducting designs.

1. INTRODUCTION

H-type cavities are characterized by the direction of the RF magnetic field, which is parallel and anti-parallel with respect to the beam axis. Closed field loops are provided by connecting field lines with opposite orientation at the cavity ends.

At present RFQ and DTL versions exist for the H_{110}-mode as well as for the H_{210}-mode (Fig. 1). The Interdigital H-type drift tube structure (IH, H_{110}-mode) was mentioned as an attractive solution for proton acceleration up to 30 MeV at the 1956 CERN Symposium by J.P. Blewett [1]. Several groups have studied the capabilities of this structure [2–5] and important improvements and innovations were achieved during the design phase of many H-linac projects [6–12], including the RFQ developments at the very beginning. H-type cavities are characterized by a high capacitive load contribution of the accelerating electrodes that provide the longitudinal electric field components for beam acceleration. To minimize the electric capacitance of H-DTLs, KONUS beam dynamics was developed [11–14]. These allow the design of lens-free multigap sections transversally matched by quadrupole triplet lenses.

This article will describe a simple analytical approach suited to estimate the RF field distribution and the main geometrical and RF parameters of H-cavities. Necessary input parameters result from beam simulations using codes like PARMTEQ for RFQs and LORASR for H-DTLs. These codes allow the design of the accelerating electrodes that dominate the capacitive load of the cavities. Fairly exact cavity designs for an IH-RFQ [15] as well as complex IH-DTL cavities resulted from RF

[*] New address: IAP, J.W. Goethe-Universität, Robert-Mayer-Str. 2-4, 60325 Frankfurt, Germany

field calculations with MAFIA [16, 17]. In the latter case additional investigations of RF models were very helpful. Several aspects and details of IH-RFQs and of IH- and CH-DTLs are discussed.

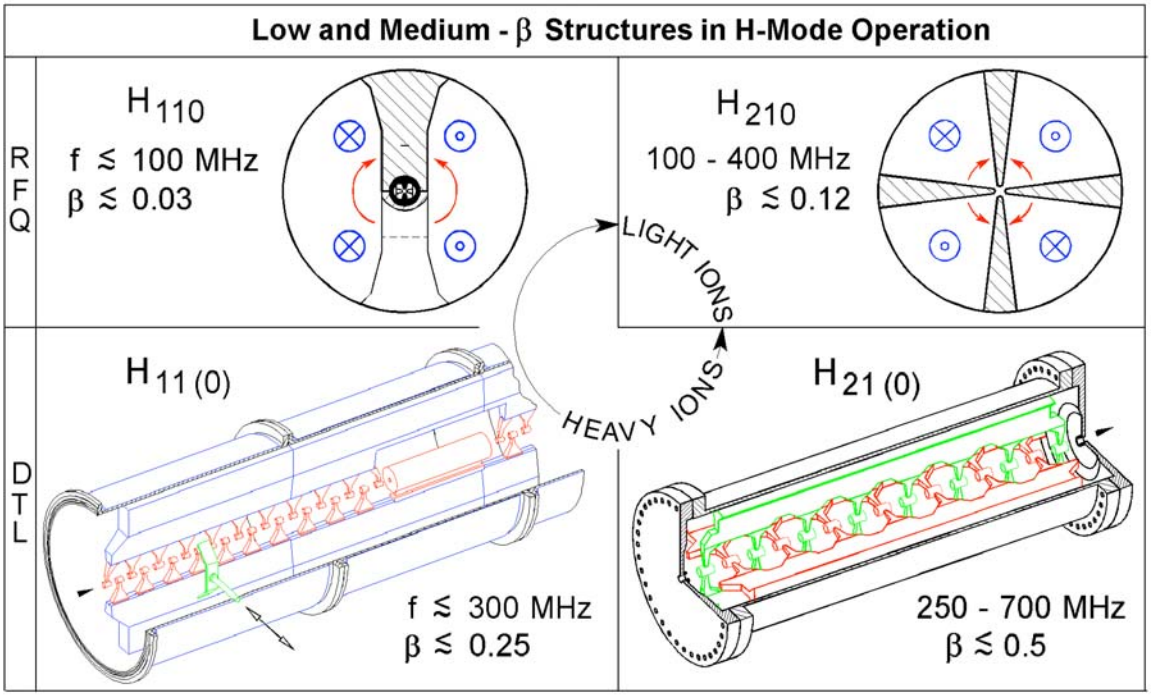


Fig. 1: H-type structure family

2. ANALYTICAL APPROACH TOWARDS H-TYPE CAVITIES

As E-modes by definition provide a dominant E-field component parallel to the beam axis, most accelerator cavities are related to that mode family. At beam velocities $\beta \lesssim 0.5$, however, H-mode cavities as shown in Fig. 1 provide excellent capabilities for ion acceleration. In RFQs the transverse electric quadrupole field components dominate, therefore the choice of H-modes is natural. For H-DTLs a number of developments were necessary to reach the present state of the art in terms of efficiency, ion current, beam quality, and reliability. Some effort is necessary to compute the characteristic parameters of H-structures. For the IH-DTL [18, 19] as well as for the four-vane RFQ [20] lumped circuit-based models were derived beforehand.

In this Section, a simplified model of the electromagnetic field distribution in H-cavities is described [14]. This allows the deduction of formulas for the estimation of all geometric and RF parameters. Optimization concepts can be deduced from these equations. Final cavity layouts, however, should be based on RF model measurements and on numerical field calculations with three-dimensional codes like MAFIA.

2.1 Field distribution

H-type accelerator structures are characterized by a very high capacitive load when compared with the corresponding mode in an empty cylindrical cavity. For example, the 202 MHz cavity shown in Fig. 2 has an averaged cavity diameter of 335 mm and length of 1.42 m. The $H_{1,1,1}$-mode of the corresponding empty cylinder has a resonance frequency of 535.3 MHz! The situation is quite different for Alvarez cavities (E_{010}-mode), where the resonance frequency of the empty cylinder is only reduced by

5–10% when inserting the drift tube structure. The frequency in H-type RFQs is degraded by factors of five (four-vane RFQ) to six (IH-RFQ) when inserting the structure.

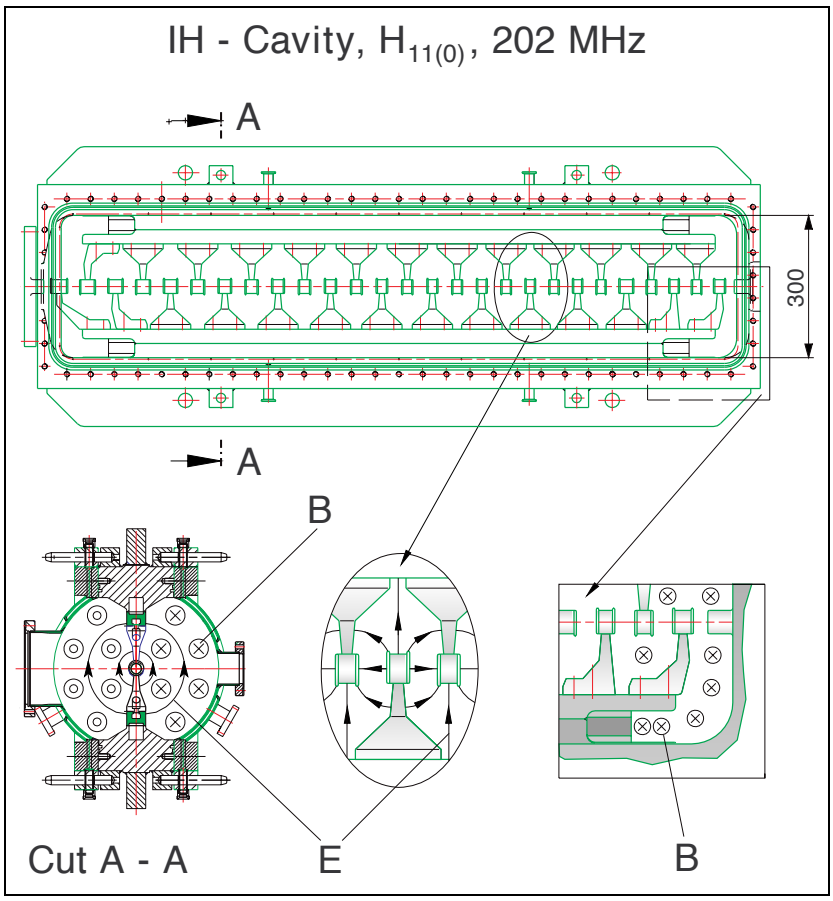


Fig. 2: Top: view of the 202 MHz drift tube structure of the CERN Pb^{25+} IH-DTL, tank 2. Bottom: a cross sectional view of this cavity as well as the electric and magnetic field distributions at characteristic regions.

This strong influence of the accelerating structure leads to a very different RF field distribution than the corresponding waveguide modes. The distribution can be characterized in good approximation in the following way:

– The cylindrical outer wall ($r = R_2$) stays on ground potential; $E\ (r = R_2) \cong 0$; the electric field is concentrated at the accelerating electrodes. Between R_1 und R_2 (Fig. 3) exists only an electric field component $E_\varphi(r)$, with the boundary condition

$$E_\varphi(R_2) = 0 \ . \tag{1}$$

– When integrating along E_φ (see Fig. 3), the induction law leads to the equation

$$\int_{\varphi_1}^{\varphi_2} E_\varphi(r) \cdot r \cdot d\varphi = \frac{d}{dt} \oint_A \vec{B}\, d\vec{A} \ ; \tag{2}$$

where A denotes the segment from r to R_2.

– The magnetic flux density B_z shows a very homogeneous distribution throughout the cavity cross sectional area between R_1 and R_2 (compare the MAFIA result shown in Fig. 4 for half the cross sectional area of a 36 MHz IH-DTL [12]):

$$B_o \cong B_{o,z} \cong const. \quad \text{for} \quad R_1 \leq r \leq R_2 \ . \tag{3}$$

While in empty cylindrical cavities the zero modes H_{mn0} are non-existent [21], they can be realized for the H-structure family discussed here by appropriate shaping of the geometry at the cavity ends (see Fig. 2 and Section 3.1.1).

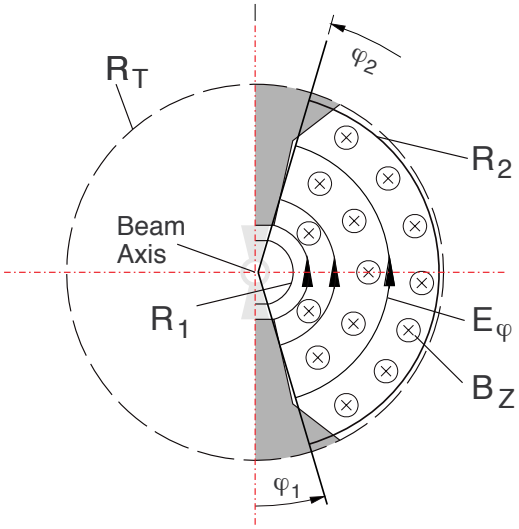


Fig. 3: Approximation of an IH-type structure one-chamber cross section by R_1, R_2, φ_1, φ_2, and field orientation

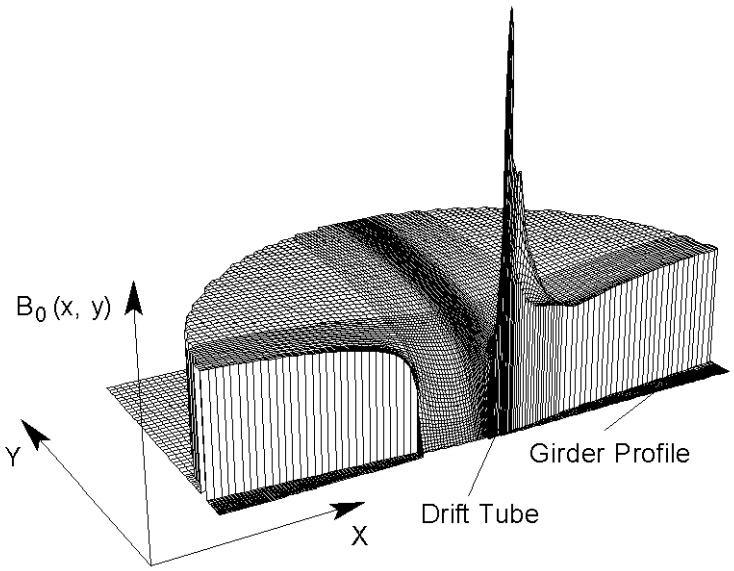


Fig. 4: Magnetic field distribution within one chamber of the 36 MHz IH-DTL, tank 2 of the GSI High Current Linac. The z-coordinate corrresponds to the stem mirror plane of a drift tube connected to the right hand girder. The peak is related to the closed field lines around the stem with conical shape.

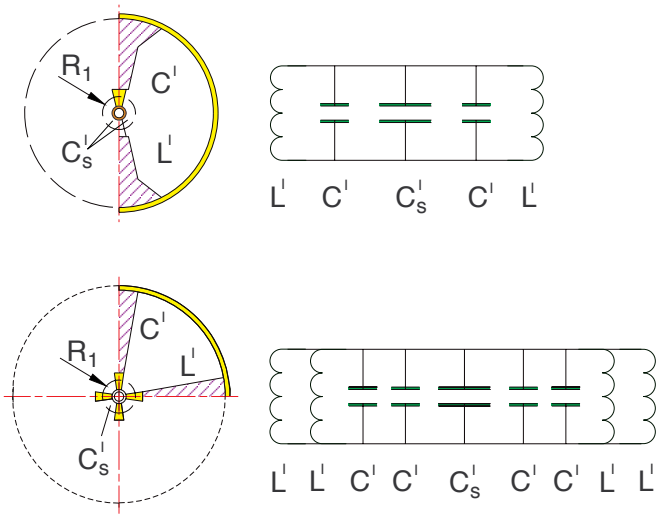


Fig. 5: Equivalent circuits connecting the capacitive and inductive load of one chamber with the corresponding parameters of the whole cavity (H_{110}- and H_{210}-modes)

The main features of H_{mn0}-type linac cavities can now be deduced by investigating the RF field distribution within a slice of the cavity that has unit length. Taking symmetry into account, only half the cross sectional area for H_{11}-modes and a quarter cross sectional area for H_{21}-modes needs to be considered. The resulting parameters can be related to the complete cavity using the simple lumped circuit schemes shown in Fig. 5.

We first divide the tank chamber into two parts. The segment with $< R_1$ is treated separately at the end of the process and refers to the accelerating electrodes and (partially) to the stems. The segment with $R_1 \le r \le R_2$ and $\varphi_1 \le \varphi \le \varphi_2$ is an approximation to the real geometry (see Fig. 4).

We next calculate $E\varphi(r)$. The RF field components are described by

$$E_{\varphi}(r,t) = E_0(r) \cdot e^{i\omega t}; \qquad B_z(t) = B_0 \cdot e^{i\left(\frac{\pi}{2}+\omega t\right)} \quad . \tag{4}$$

Eq. (2) with

$$\Phi(r,t) = B_z(t) \cdot \int_r^{R_2} \int_{\varphi_1}^{\varphi_2} r_1 dr_1 d\varphi \tag{5a}$$

results in

$$-\dot{\Phi}(r,t) = -i\omega\, B_z(t)\ \ (R_2^{\,2} - r^2)\ \ (\varphi_2 - \varphi_1)/2\,. \tag{5b}$$

The corresponding voltage is given by the integral

$$U_i(r) = -\dot{\Phi}\,(r,t) = \int_{\varphi_1}^{\varphi_2} E_0(r) \cdot r \cdot d\varphi \quad . \tag{6}$$

Equating gives the result

$$e^{i\omega t} E_0(r) \cdot r \cdot (\varphi_2 - \varphi_1) = e^{i\omega t} \cdot \omega \cdot B_0 \cdot (R_2^{\,2} - r^2)(\varphi_2 - \varphi_1)/2; \tag{7}$$

$$E_0(r) = \frac{\omega\, B_0 (R_2^{\,2} - r^2)}{2\, r} \quad . \tag{8}$$

We next calculate the stored magnetic field energy per chamber unit length,

$$W'_m = \frac{B_0^2}{2\mu_0} \int_{R_1}^{R_2} \int_{\varphi_1}^{\varphi_2} r\, dr\, d\varphi = \frac{B_0^2}{4\mu_0}\left(R_2^{\,2} - R_1^{\,2}\right)\left(\varphi_2 - \varphi_1\right) \; ; \tag{9}$$

and for the electric field

$$W'^{*}_{el} = \frac{\varepsilon_0}{2} \int_{R_1}^{R_2} \int_{\varphi_1}^{\varphi_2} E_0^2(r) \cdot r\, dr\, d\varphi \; ; \tag{10a}$$

$$W'^{*}_{el} = \frac{\varepsilon_0 \omega^2 B_0^2}{8} \left[R_2^{\,4} \left(\ln \frac{R_2}{R_1} - \frac{3}{4} \right) + R_2^{\,2} R_1^{\,2} - \frac{R_1^{\,4}}{4} \right] \left(\varphi_2 - \varphi_1\right) \; . \tag{10b}$$

While the magnetic field energy within r < R_1 can be neglected for the moment, the electric contribution is estimated by assuming a constant field strength $|E_0| = |E_0(R_1)|$ for this volume. Co m-parisons with MAFIA simulations have shown that this method produces good results as soon as the accelerating structure is included, but that it underestimates the corresponding field contribution when no electrodes are mounted.

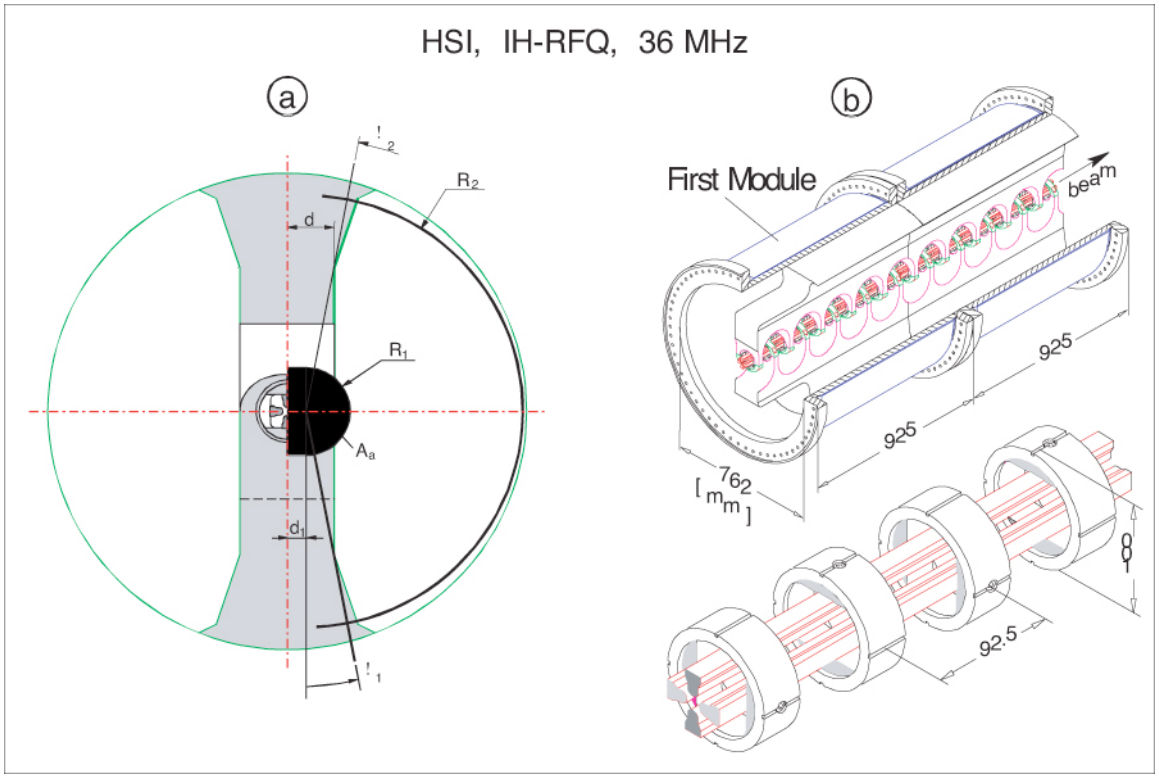


Fig. 6: IH-RFQ: (a) equivalent chamber geometry; (b) the first two of the ten modules from the 9.4 m GSI IH-RFQ, and illustration of the electrodes and carrier rings.

To gain more flexibility in approximating a given cavity geometry, the area within r < R_1 is matched by introducing a factor F (see Fig. 6).

$$A_a = F \cdot \frac{R_1^{\,2}}{2} \cdot \left(\varphi_2 - \varphi_1\right) \; . \tag{11}$$

For the electric field energy, neglecting R_1^4-terms, this finally gives

$$W'_{el} \cong \frac{\varepsilon_o \omega^2 B_o^2}{8} \left[R_2^4 \left(\ln \frac{R_2}{R_1} + \frac{2F-3}{4} \right) + R_2{}^2 R_1{}^2 (1-F) \right] (\varphi_2 - \varphi_1). \tag{12}$$

We can now estimate the resonant frequency of the cavity when the part with $r < R_1$ is omitted. The balance of field energies now leads to the resonance frequency by equating Eqs. (9) and (12):

$$\omega_T = 2c \cdot \left(\frac{R_2{}^2 - R_1{}^2}{R_2{}^4 \left(2 \ln \frac{R_2}{R_1} + \frac{2F-3}{2} \right) + 2 R_2{}^2 R_1{}^2 (1-F)} \right)^{1/2} . \tag{13}$$

It should be noted that ω_T does not depend on $(\varphi_2 - \varphi_1)$. By reducing $(\varphi_2 - \varphi_1)$ the inductivity is decreased in the same way as the capacitance is increased. The capacitive load per unit length fulfils the equation

$$W'_{el} = \frac{1}{2} C' \cdot U_0{}^2 \; ; \quad \text{with} \quad U_o = E_o(R_1) \cdot R_1 \cdot (\varphi_2 - \varphi_1) \; . \tag{14}$$

With Eq. (12) this results in

$$C' = \varepsilon_o \frac{R_2{}^4 \left(\ln \frac{R_2}{R_1} + \frac{2F-3}{4} \right) + R_2{}^2 R_1{}^2 (1-F)}{\left(R_2{}^2 - R_1{}^2 \right)^2 \cdot (\varphi_2 - \varphi_1)} . \tag{15}$$

Including the resonance condition $\omega^2 = (L' \cdot C')^{-1}$, one gets from Eqs. (13) and (15)

$$L' = \frac{\mu_o \left(R_2{}^2 - R_1^2 \right) (\varphi_2 - \varphi_1)}{2} \tag{16}$$

Now follows the calculation of the complete cavity parameters by including the real installations around the beam axis (equivalent magnitudes for the capacitive and inductive load within R_1 are determined). This step has to be done separately for every type of accelerating structure.

The capacitance of so-called mini-vane quadrupoles (Fig. 7) can be estimated by the semi-empirical formula [22]

$$C'_Q / pF/m = \frac{39.37}{\cosh^{-1}((1+R_o/\rho)/\sqrt{2})} + \frac{31.05}{R_o/\rho - 0.414} + 25.28 \ln \left(1 + \frac{h}{a+\rho} \right). \tag{17}$$

In the case of the IH-RFQ, the stem structure leads to an additional capacitive load (see Section 2.1.2).

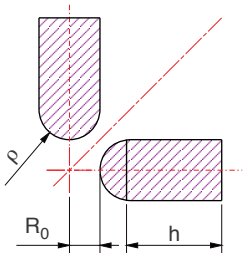


Fig. 7: Illustration of the main parameters defining a mini-vane array

For drift tube structures with diameter ratios $\varnothing_a / \varnothing_i > 1.5$ a semi-empirical formula for calculating the capacitive load is given by Ref. [23]. As H-type DTLs profit significantly from minimizing $\varnothing_a$, the recommended range is $\varnothing_a / \varnothing_i \leq 1.4$ for that geometry—including the dipole field correction in the case of IH-structures (Section 3.4)—and the following concept is useful:

The capacitance of a massive cylinder with radius r_a and length d_e, shielded by two parallel, conducting planes with distance $g_e/2$ is given in good approximation [24] by

$$C^* \approx \frac{2 \cdot \varepsilon_o \pi \, r_a^2}{g_e} \left[2 + \frac{4 \, g_e}{\pi \, r_a} \left(\ln 2 + F_d(x) \right) \right], \tag{18}$$

where F_d is defined by Eq. (19a).

When the shielding planes are positioned in the gap centres, the mirror images produce the neighbouring cell (Fig. 8). The gap voltage is twice the voltage between tube and shielding plane, therefore the capacitance of two $\beta\lambda/2$-cells with $\beta\lambda = 2\,(g + d)$ corresponds to $C^*/2$.

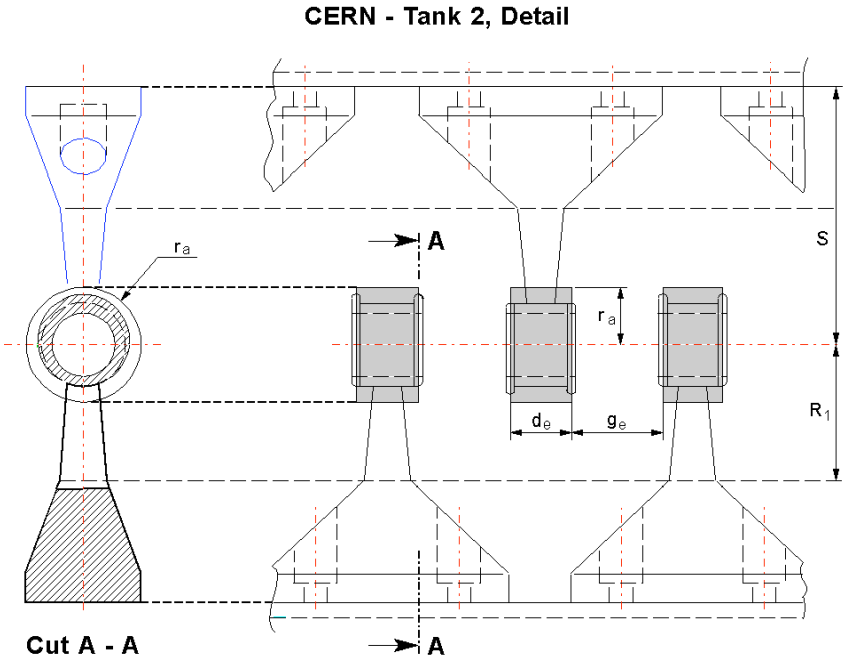


Fig. 8: IH-type drift tube structure and equivalent geometry to calculate capacitive load

The capacitive load per unit length of the drift tube structure results in

$$C'_D = \frac{\varepsilon_0 \pi r_a^2}{2g_e(g+d)} \left[2 + \frac{4g_e}{\pi r_a} \left(\ln 2 + F_d(x) \right) \right] ; \tag{19}$$

with $$F_d(x) = (1+x)\ln(1+x) - x \ln x \; ; \quad x = \frac{d_e}{2g_e} \; . \tag{19a}$$

The following examples will demonstrate the application of this formalism to realized H-cavities.

2.1.1 Four-vane RFQ (H_{210})

The modulated vane tips are treated by Eq. (17), see also Fig. 9 for an image of the RFQ. To reduce the stray field contribution h = 0 is used. As the total capacitance for r < R_1 is included in this geometry one has additionally to insert F = 0. The cavity parameters are then given by

$$C'_{4Vane} = 4C' + C'_Q \; ; \tag{20}$$

$$L'_{4Vane} = \frac{L'}{4} \cdot \left(\frac{R_2^{\,2}}{R_2^{\,2} - R_1^{\,2}} \right) ; \quad \text{and} \tag{21}$$

$$\omega_{4Vane} = \omega_T \cdot \left(\frac{4C'}{4C' + C'_Q} \right)^{1/2} \cdot \left(\frac{R_2^{\,2} - R_1^{\,2}}{R_2^{\,2}} \right)^{1/2} . \tag{22}$$

Good results were obtained when comparing to realized structures with the parameter choice

$$R_1 \cong 2 \cdot (\rho + R_0) \, . \tag{23}$$

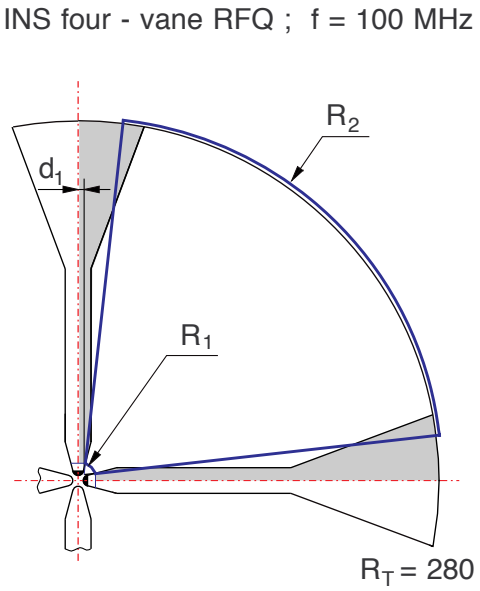


Fig. 9: Equivalent chamber geometry of a four-vane RFQ

2.1.2 IH-RFQ (H_{110})

The mini-vane capacitance is described by Eq. (17). Additionally, the stem structure has to be included. The capacitance C_D between neighboured stems can be estimated by Eq. (19). The contribution C_{QD} between the vane carrier ring and the oppositely charged pair of mini vanes can be deduced by treating each array ring-electrode as a segment of a cylindrical capacitor. Ultimately one gets

$$C'_{IHQ} = 2C' + C'_Q + C'_D + C'_{QD} \; ; \tag{24}$$

$$L'_{IHQ} = \frac{L'}{2} \cdot \left(\frac{R_2^{\,2}}{R_2^{\,2} - R_1^{\,2}} \right) \; ; \text{ and} \tag{25}$$

$$\omega_{IHQ} = \omega_T \cdot \left(\frac{2C'}{2C' + C'_Q + C'_D + C'_{QD}} \right)^{1/2} \cdot \left(\frac{R_2^{\,2} - R_1^{\,2}}{R_2^{\,2}} \right)^{1/2} \; ; \tag{26}$$

Fig. 6 shows the geometry approximation corresponding to the GSI 36 MHz IH-RFQ [25], as well as a three-dimensional view of two of the ten cavity modules. The individual contributions to the capacitive load of that tank are estimated to be

$$2C' = 11.84\,pF/m \; ; \quad C'_Q = 108.5\,pF/m \; ; \quad C'_D = 43.2\,pF/m \; ; \quad C'_{QD} = 22.9\,pF/m \, .$$

2.1.3 IH-DTL (H_{110})

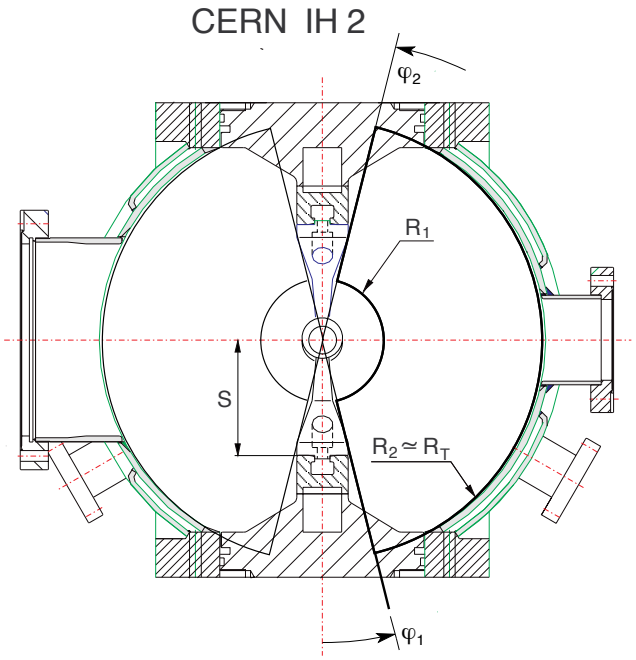


Fig. 10: Equivalent chamber geometry for an IH-DTL

The capacitive load of the drift tubes is estimated by Eq. (19) and the choice of a 'typical' cell geometry for the investigated cavity. The results for the IH-DTL are

$$C'_{IHD} = 2C' + C'_D \; . \tag{27}$$

The capacitance ratios for cavities with slim drift tubes are in the range

$$0.6 < C_D{}' / 2C' < 2.0; \tag{27a}$$

$$L'_{IHD} = \frac{L'}{2} \cdot \left(\frac{R_2{}^2}{R_2{}^2 - R_1{}^2} \right); \tag{28}$$

$$\omega_{IHD} = \omega_T \cdot \left(\frac{2C'}{2C' + C'_D} \right)^{1/2} \cdot \left(\frac{R_2{}^2 - R_1{}^2}{R_2{}^2} \right)^{1/2} . \tag{29}$$

Figure 10 shows an example of a choice of equivalent chamber segment, and Table 1 summarizes parameters for a number of realized cavities.

Table 1: Cavity parameters and measured quantities Z_0, R'_{p0}, Q, with resulting loss factors c_2 for constructed cavities

Machine **H-Type**	**GSI-HLI** **IH-DTL**	**CERN-LINAC 3** **IH-DTL, Tank 2**	**GSI-HSI** **IH, DTL, Tank 2**	**GSI-HSI** **IH-RFQ**	**INS-Tokyo** **four-vane RFQ**
f/MHz	108.4	202.06	36.1	36.1	100
R_2/mm	367	172.5	1017	345	277.4
R_1/mm	60	47.5	120	70	14.6
R_0/mm	–	–	–	5.8	4.1
ρ/mm	–	–	–	4.9	4.1
S/mm	130	90	280	140	–
r_a/mm	16.9	20.2	44	–	–
ϕ_a/mm	24	29	56	–	–
ϕ_i/mm	18	22	46	–	–
Δ/mm	1.5	1.5	5.0	–	–
$\overline{\beta}$	0.045	0.072	0.049	0.0128	0.0137
$Z_0/M\Omega/m$	650	278	390	–	–
$R'_{p0}/k\Omega m$	–	–	–	620	225
Q	20000	12550	39000	13000	11000
N_S	26	14	26.5	–	–
$\sum_j N_{L,j}$	5	0	3	–	–
F_L	0.67	1	0.8	–	–
c_1	1	1	1	1.13	–
c_2	1.47	1.66	1.47	1.31	1.51

2.1.4 CH-DTL

This structure has so far been investigated by MAFIA simulations [26]. Investigations of RF models and prototype cavities are planned in a next step. The corresponding equations for this structure are:

$$C'_{CHD} = 4C' + C'_D; \tag{30}$$

$$L'_{CHD} = \frac{L'}{4} \cdot \left(\frac{{R_2}^2}{{R_2}^2 - {R_1}^2} \right); \text{ and} \tag{31}$$

$$\omega_{CHD} = \omega_T \cdot \left(\frac{4C'}{4C' + C'_D} \right)^{1/2} \cdot \left(\frac{{R_2}^2 - {R_1}^2}{{R_2}^2} \right)^{1/2}. \tag{32}$$

2.2 Acceleration efficiency

The efficiency of drift tube structures is characterized by the resulting effective voltage gain, while the main criterion for RFQs is the achieved vane voltage. As the corresponding parameters are not interpreted in the same way everywhere the used definitions are given by Eqs. (33)–(35) for DTLs and by Eq. (36) for RFQs.

Shunt impedance:

$$Z_0 = \frac{\left(\int_0^l \left| E_{o,z}(z) \right| dz \right)^2}{P \cdot l}; \tag{33}$$

transit time factor:

$$T = \frac{\left| \int_0^l E_{o,z}(z) \cos \omega t(z) dz \right|}{\int_0^l \left| E_{o,z}(z) \right| dz}; \tag{34}$$

effective shunt impedance:

$$Z_{eff} = Z_0 \cdot T^2; \text{ and} \tag{35}$$

specific shunt resistance

$$R'_{p0} = \frac{U_0^2 \cdot l}{P}. \tag{36}$$

The term U_0 denotes the gap voltage and vane voltage amplitude, respectively, while P denotes the RF wall losses. The dimension of Z_0 and Z_{eff} is Ω/m, and R'_p has the dimension Ωm.

The numerical calculation of RF wall losses for complex geometrical structures is not trivial. In addition, contact losses at RF joints and the inclusion of locally enhanced surface current densities along corners etc. are difficult to include. In practice, good results have been obtained by applying the following concept:

– take the parameters from a known cavity of the same type as the one to be calculated;

– derive the equivalent geometrical parameters R_2, R_1, φ_2, φ_1 for both cavities;

– the ratio of shunt impedances between both cavities can then be calculated by using Eq. (50) for DTLs and Eq. (55) for RFQs.

These equations are now derived following in part the concepts of Refs. [5] and [6]. The transversally directed surface current is related to the magnetic field by

$$I_0' = H_o = \frac{1}{\mu_o} \cdot B_o \ . \tag{37}$$

The current path length in the corresponding cavity sector is given by

$$d_I = R_2 \ (\varphi_2 - \varphi_1 + 2c_1) \ . \tag{38}$$

The losses within the region $r < R_1$ are estimated by assuming a homogeneous surface current density from $r = 0$ to $r = R_1$ along virtual surfaces that close the sector under consideration. The value c_1 allows us to regard effective path extensions for 'long' drift tubes and for mini-vanes, and c_1 covers the range

$$1 \le c_1 \le 1.2 \ . \tag{39}$$

The ohmic resistance of a sector length dz is then given by

$$dR \ = \ \frac{R_2(\varphi_2 - \varphi_1 + 2\,c_1)\,c_2}{\kappa \cdot \delta \cdot dz} \ , \tag{40}$$

with skin depth

$$\delta \ = \left(\frac{1}{2} \omega \mu_o \kappa \right)^{-1/2} \ , \tag{41}$$

and a loss coefficient of

$$c_2 = c_c \cdot c_g \ ; \tag{42}$$

where c_c stands for a reduction in conductivity caused by surface quality, tank flanges and RF joints; and c_g allows us to include the additional losses caused by the cavity ends, especially by the girder and vane undercuts, with their high current densities.

The power losses within a cavity sector amount to

$$dP_s \ = \frac{1}{2} \left(I_0' \cdot dz \right)^2 \cdot dR \ ; \tag{43}$$

and with Eqs. (37) and (40)

$$dP_s \ = \frac{1}{2} \left(\varphi_2 - \varphi_1 + 2c_1 \right) c_2 R_2 H_o^2 \left(\frac{\mu_o \omega}{2\kappa} \right)^{1/2} dz \ . \tag{44}$$

The corresponding losses within the complete cavity cross section are given by

$$dP = N_m \cdot dP_s \tag{45}$$

with $N_m = 2$ for H_{11}-cavities and $N_m = 4$ for H_{21}-cavities.

The voltage amplitude is estimated using Eq. (6) with r → 0

$$U_o = \frac{1}{2}\,\mu_o \omega H_o R_2^2\ (\varphi_2 - \varphi_1)\,. \tag{46}$$

Now the shunt impedance and the R'_p value, respectively, can be deduced for all H-type structures discussed here.

2.2.1 Drift tube structures

From Eq. (33) one gets for N_G gaps with gap voltage amplitude U_0 along cavity length l

$$Z_0 = \frac{N_G{}^2 \cdot U_o^2}{P \cdot l}\,. \tag{47}$$

The distance between gap centres corresponds to βλ/2. In non-relativistic approximation one can calculate a velocity parameter β, which is valid for the cavity centre, from given values β_i and β_f at the cavity ends, by the formula

$$\overline{\beta} = \left(\frac{\beta_i^3 + \beta_f^3}{2}\right)^{1/3}\,. \tag{48}$$

It is assumed that the drift tube structure covers the whole cavity length: the number of gaps within l is estimated by

$$N_G \cong \frac{2 \cdot l}{\overline{\beta} \cdot \lambda} = \frac{\omega \cdot l}{\pi \cdot \overline{\beta} \cdot c}\,; \tag{49}$$

Finally, using Eqs. (45)–(49) the shunt impedance results in

$$Z_0 \cong \frac{\mu_o{}^{3/2}\ \kappa^{1/2}\ R_2^3\ (\varphi_2 - \varphi_1)^2\ \omega^{7/2}}{\sqrt{2}\,\pi^2\ c^2 N_m\ (\varphi_2 - \varphi_1 + 2c_1)\,c_2\ \overline{\beta}^2}\,. \tag{50}$$

At a typical operating temperature of 30°C with $\kappa_{Cu} \sim 56 \cdot 10^6\ S/m$ one gets

$$Z_{0,Cu}/\Omega/m \cong 8.4 \cdot 10^{-24}\ \ \frac{R_2{}^3\ (\varphi_2 - \varphi_1)^2 \omega^{7/2}}{N_m \cdot\ (\varphi_2 - \varphi_1 + 2c_1)\ c_2\ \ \overline{\beta}^2}\ \ ; \tag{51}$$

We can now calculate Z_0 exclusively from geometric cavity parameters by including Eqs. (13) and (29) with $N_m = 2$, and Eq. (32) with can be used to calculate ω.

The quality factor of the cavity can be calculated using the definitions

$$P = \frac{\omega\, W}{Q}\,; \qquad P = \frac{N_G^2 \cdot U_0^2}{Z_0 \cdot l}\,; \quad \textit{and} \qquad W = \frac{1}{2}\,C_t^{'} \cdot l \cdot U_0^2\ ;$$

and inserting Eq. (49). The result is

$$Q \cong \frac{\pi^2 \cdot C_t' \cdot Z_0 \cdot \overline{\beta}^{\,2} \cdot c^2}{2\omega}\,. \tag{52}$$

Table 1 (above) shows all relevant parameters for realized structures. By comparing the measured values one gets the following range for the loss coefficient c_2.

$$1.3 \le c_2 \le 1.7 \ . \tag{53}$$

Typical transit time values, Eq. (34), are within the range 0.8–0.84 for an averaged gap/period to length ratio of 0.5. This results in effective shunt impedances

$$0.64 \cdot Z_0 \le Z_{eff} \le 0.70 \cdot Z_0 \ . \tag{54}$$

Acceleration efficiencies of operated IH-structures, including the synchronous RF phase angle Φ_s from the beam dynamics design, are plotted in Fig. 11.

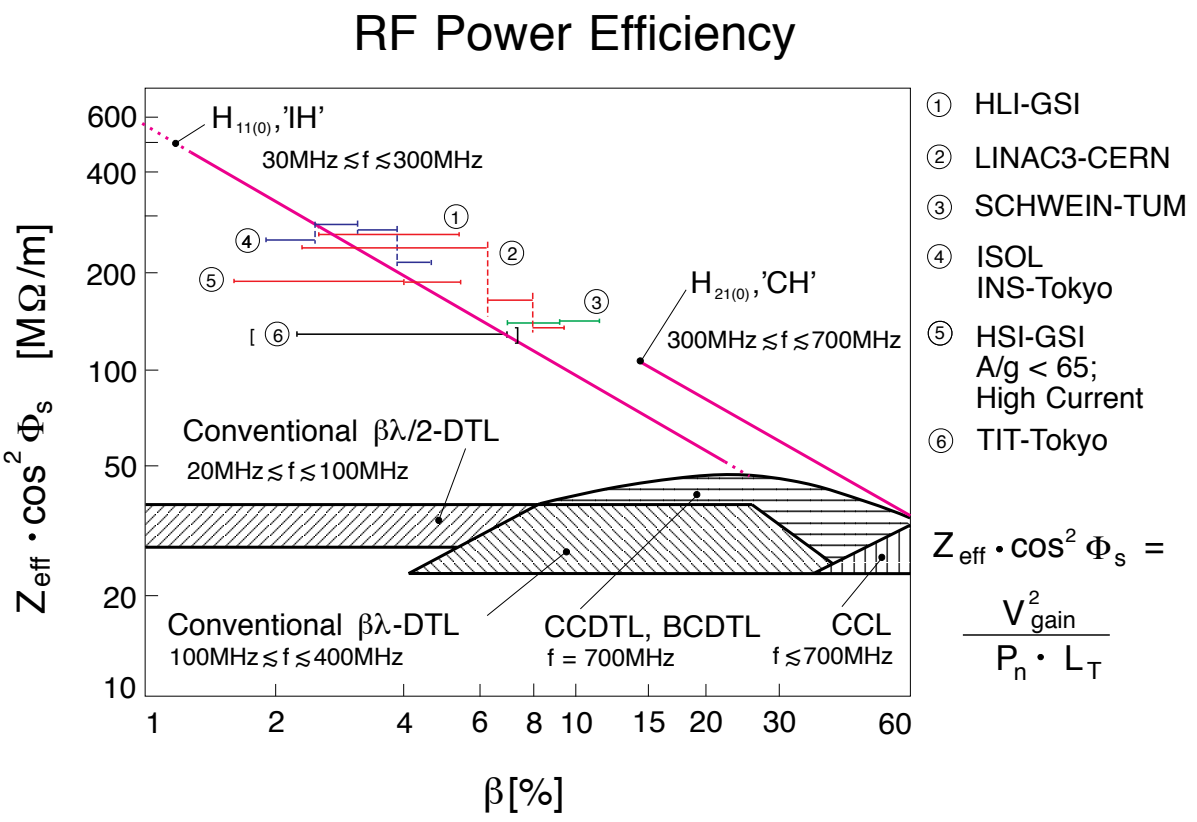


Fig. 11: RF power economy of multicell structures. The advantage of H-type cavities, especially at low beam velocities, is evident. Linacs no. (3) and (4) do not include focusing elements inside the cavities; no. (6) uses conventional beam dynamics (Φ_s = -30°) with quadrupoles in every second drift tube (Φ_a = 100 mm). The 700 MHz CCDTL and BCDTL structures were recently proposed at Los Alamos [27].

2.2.2 RFQ structures

Starting from Eq. (36), R'_{po} is given by

$$R'_{p0} = \frac{\mu_o{}^{3/2} \kappa^{1/2} R_2^3 (\varphi_2 - \varphi_1)^2 \omega^{3/2}}{\sqrt{2} N_m (\varphi_2 - \varphi_1 + 2c_1) c_2} \ , \tag{55}$$

when inserting Eqs. (44)–(46). Again, using $\kappa_{Cu} \sim 56 \cdot 10^6 \, S/m$ at a typical operating temperature of 30°C, one gets

$$R'_{p0}/\Omega m \;=\; 7.45\cdot 10^{-6}\;\frac{R_2{}^3(\varphi_2-\varphi_1)^2\omega^{3/2}}{N_m(\varphi_2-\varphi_1+c_1)\;c_2}\;. \tag{56}$$

The geometric parameters of the cavity are sufficient to calculate R'_{po} by obtaining ω from Eqs. (13) and (26) with $N_m = 2$ and Eq. (22) with $N_m = 4$, respectively.

The quality factor for RFQs results in

$$Q = \frac{\omega\;C'_t\,R'_{p0}}{2}\;. \tag{57}$$

The parameters of two RFQs that have been constructed are contained in Table 1 (above).

3. DESIGN PRINCIPLES FOR H-CAVITIES

Section 2 explained why a capacitive load reduction of the accelerating structure is especially effective in H-type linacs. This fact has provoked the development of several structure-specific solutions. The examples given here relate to the IH-DTL, but—with the exception of Section 3.4—the concepts can be applied to other H-structures as well.

One main step in the design of efficient H-type DTLs is the application of KONUS beam dynamics to create long, quadrupole-free drift tube sections with small outer drift tube diameters [11–14]. Additionally, a number of specific technical solutions have been developed and are described below.

3.1 Adjustment of the voltage distribution along the cavity

Two powerful concepts are now described that allow the tuning of the cavity with only modest technical effort.

3.1.1 Undercuts

This geometry at the cavity ends (compare Figs. 2 and 12) is most suited to creating the zero mode. The short circuit action of the end walls with respect to H-modes is suppressed by the geometry if the undercuts are large enough.

A lumped circuit equivalent is shown in Fig. 12(a). The centre part of the cavity is represented by C_M and L_M, and the corresponding quantities of one cavity end are denoted by C_E and L_E, and depend on the undercut size. The magnetic field that penetrates the undercuts was measured for a 52-cell IH-DTL in dependence on the window depth L_I (using metallic and dielectric beads [28]).

Figure12(b) shows the conditions for L_I = 60 mm (zero mode) as well as for smaller and larger window sizes. Figure 12(c) plots the magnetic field strength distribution along three paths marked in Fig.12(a) at a fixed window depth.

Figure 13 demonstrates the influence of the undercut depths L_I at the left-hand cavity end on the gap voltage distribution of a 52-cell structure. It becomes obvious that, besides the zero mode, ramped gap voltage distributions can be also be created by this tuning method.

A disadvantage, however, is that considerable RF power is lost with large undercuts because of the high local magnetic surface fields. Therefore many DTL designs prefer to keep the windows below ‘zero mode size’ and to add a second technique for tuning the voltage distribution (described below).

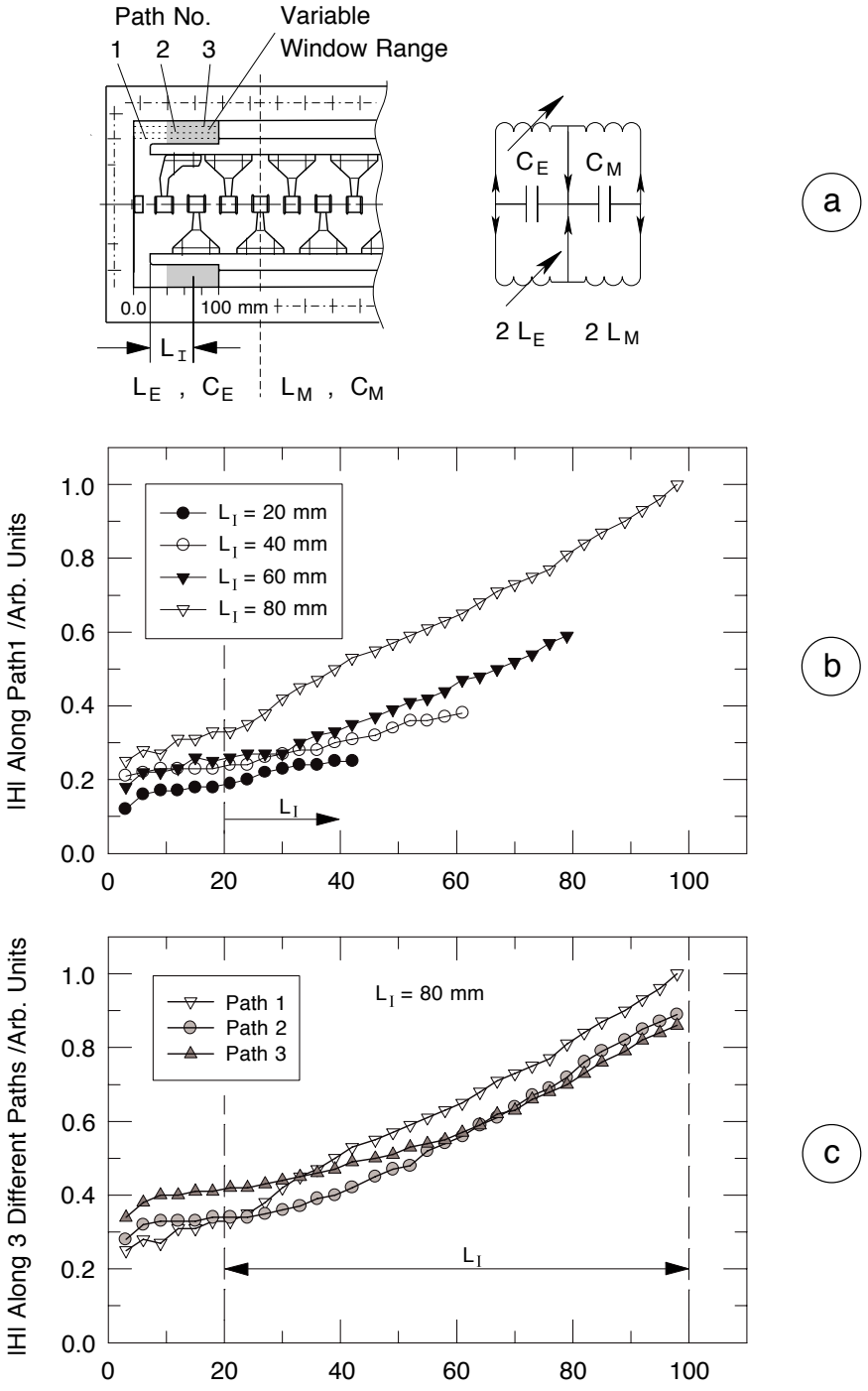


Fig. 12: (a) Functional principle of undercuts; (b) influence of the undercut size on the magnetic field distribution; and (c) field distribution within the undercut cross section for one selected geometry.

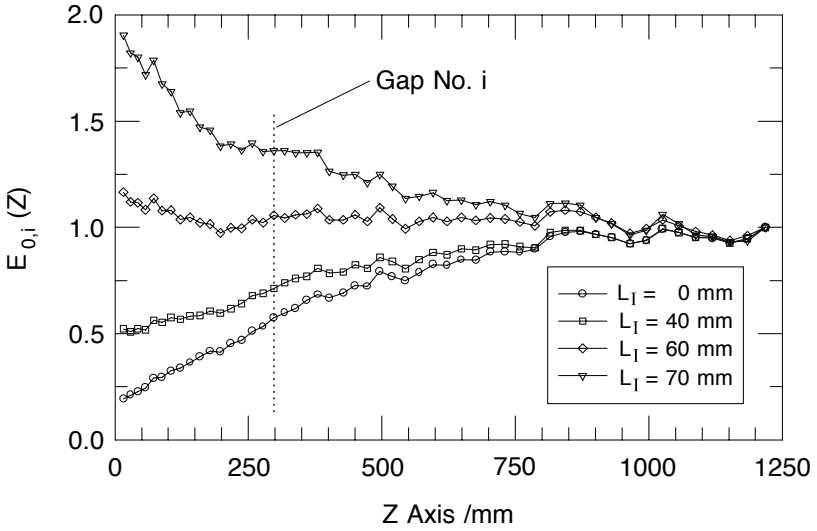


Fig. 13: Variation of the gap voltage distribution along a 52-cell IH-type cavity when changing the undercut depth L_I (see Fig. 12(a)) at the left-hand cavity end. The distributions are normalized at an identical field level in the final gap.

3.1.2 Capacitive load distribution along the drift tube structure

Local gap voltage changes in the range ±20% can be achieved by varying the drift tube geometry along the resonator. The options for this are:

– variation of the gap-to-periodic length ratio;

– variation of the drift tube diameters.

As a consequence one gets a modulation of the capacitive load. To understand the influence on the voltage distribution we start with a zero mode cavity.

Local increase of the drift tube capacitance between z_1 and z_2, with $z_2 - z_1 = l_c$ from the initial value C' to $C'(1+\delta_c)$, results in a change of the cavity frequency

$$\omega_f = \omega_i \left(1 - \frac{\delta_c \cdot l_c}{2 \cdot l_R}\right) . \tag{58}$$

The parameters: charge density, voltage amplitude and surface current, are related by

$$Q'_0(z) = C'(z) \cdot U_0(z) , \tag{59}$$

$$I'_0(z) = \omega \cdot C'(z) \cdot U_0(z) , \text{ and} \tag{60}$$

$$U_0(z) = L' \omega \, I'_o(z) . \tag{61}$$

If one still assumes a constant voltage distribution, the corresponding surface current results in

$$I'_{0,f}(z) = I'_{0,i}(z) \left(1 - \frac{\delta_c \cdot l_c}{2 \cdot l_R}\right) ; \text{ for } z \le z_1 \text{ and } z \ge z_2 ; \tag{62}$$

$$I'_{0,f}(z) \;=\; I'_{0,i}(z)\left(1-\frac{\delta_c \cdot l_c}{2 \cdot l_R}\right)(1+\delta_c);\;\; \text{for} \;\; z_1 < z < z_2\,. \tag{63}$$

This current distribution leads, with Eq. (61), to a voltage increase within l_c:

$$U_{0,f}(z) \;=\; U_{0,i} \;\; \text{for} \;\; z \le z_1 \;\text{and}\; z \ge z_2\,, \tag{64}$$

$$U_{0,f}(z) \;=\; U_{0,i}(1+\delta_c);\;\; \text{for} \;\; z_1 < z < z_2\,, \tag{65}$$

leading to a further current increase in this region, etc.

This simple model explains why the voltage can be increased locally by an increase in the corresponding capacitive load. It can not explain the mechanism quantitatively.

In reality not only transverse, but also longitudinal components of the electric surface currents are induced, as indicated by Fig. 14, which shows that distortions of the capacitive load along a cavity tuned to the H_{110}-mode induce admixtures of $H_{11\ell}$-modes. The most frequently used tuning steps are shown in Fig. 15. Besides the zero-mode needed for RFQs, there is a range of gap voltage distributions indicated in Fig. 15(a), which is used for DTLs in practice. An example of an $H_{11\ell}$ dispersion curve is plotted in Fig. 16.

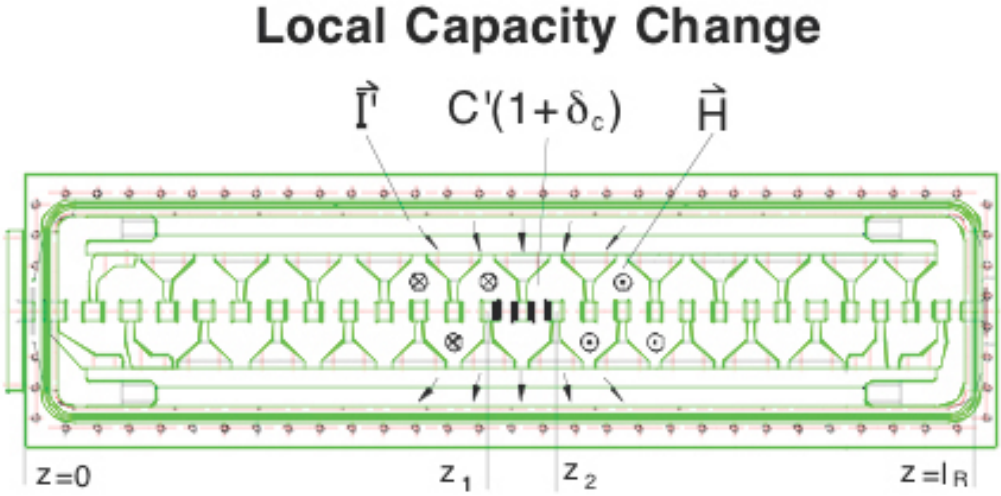


Fig. 14: Electric currents and magnetic field lines induced by increasing the drift tube capacitance locally.

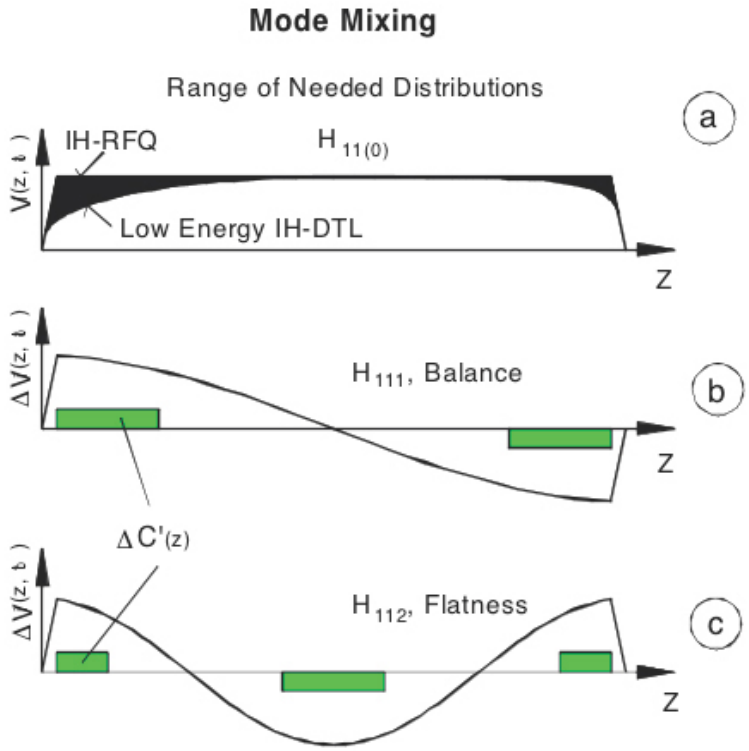


Fig. 15: Top: range of typical voltage distributions along multicell IH-type structures; bottom: superpositioning of $H_{11\ell}$-modes, induced by local variations of the capacitive load.

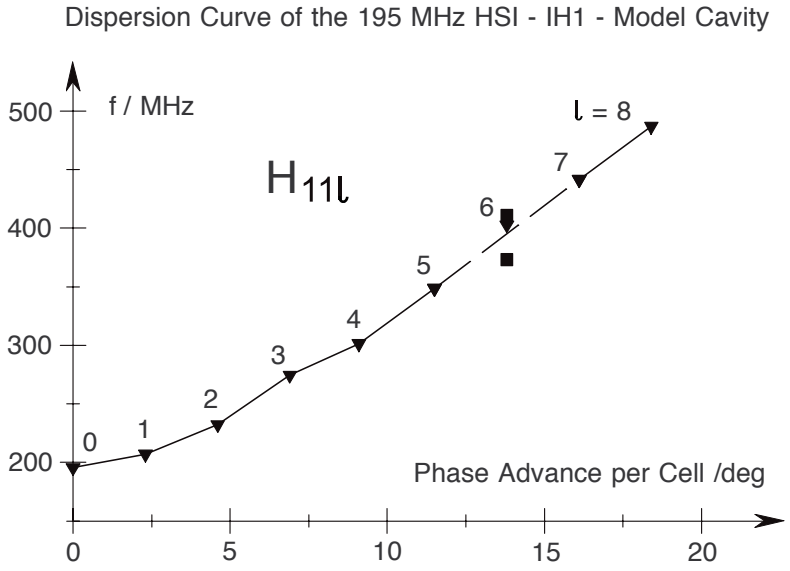


Fig. 16: Measured dispersion diagram for $H_{11\ell}$-modes of a cavity with $\varnothing/l = 0.2$, containing 53 gaps and three triplet lenses.

3.2 Cavity internal lenses

At the DTL front end the voltage amplitude and the corresponding gap number within one KONUS period is relatively small, according to beam dynamics rules. In many cases the resulting geometries correspond to cavity length-to-diameter ratios considerably below 1.0. Such short cavities can be realized. However, the housing of two or more short KONUS periods within one cavity has the following advantages:

– The number of RF amplifier chains required is reduced and the linac cavities can be designed for one RF power class by assigning the corresponding number of KONUS sections to each cavity.

– There is a reduction in low-level RF equipment.

– Simplification of the accelerator structure; short H-cavities are not trivial with respect to RF and mechanical design aspects;

  - intertank sections need some care during design and construction as well as during installation and alignment; and

  - the costs of vacuum equipment for short tanks are higher.

Compared with the advantages mentioned above, the extra efforts involved in housing quadrupole triplets within a cavity are not significant.

Remark: Only if continuous flexibility of the beam energy is needed must one use short cavities. One example of achieving continuous energy variability above 150 keV/u and applying the KONUS beam dynamics is described in Ref. [16].

We now describe two proven concepts for the integration of quadrupole triplets into IH-cavities.

3.2.1 RF-grounded lens housing

At cavity diameters below 1 m and corresponding RF frequencies above 75 MHz, it is desirable to connect the voluminous lens housing to RF ground potential. This allows a match with the capacitive load of the neighbouring drift tube structures. The concept was developed and realized for the 108 MHz High Charge State Injector at GSI [11]. Figure 17 shows the drift tube structure of that cavity as well as a cross sectional cut along the circular stem of one lens, including the adjusting device. The triplet housings are water cooled. The stems are connected to the cavity end by bellows, and the adjusting devices are screwed to the robust main frame. This concept is attractive for cavities with high duty cycle operation, while at low duty cycle the lenses can be connected directly to the cavity without any risk of mechanical displacements during operation.

As the mechanical tolerances of quadrupole lenses are typically ± 0.1 mm, the lens stems should be oriented vertically and point towards the bottom. At the same time the lens is connected to RF ground potential when the accelerating structure is mounted in the horizontal plane (Fig. 17). Matching with the drift tube structure is achieved by the following:

– the distances between girder and lens housing (the left- and right-hand gaps act as series capacitance) cause a capacitive load close to that along the adjacent drift tube structures;

– the two neighbouring drift tubes are at opposite RF potential; RF currents along the lens housing are driven by that array, while the lens stem is transporting no net current if the gap capacities on both sides are equal.

The total length of the j'th lens housing is roughly given by

$$L_{L,j} \cong N_{L,j} \cdot \beta_j \cdot \lambda - \overline{g}_j \quad . \tag{66}$$

The reduction factor of shunt impedance by installing *j* triplet lenses in a cavity, Eq. (50), is given by

$$F_L = \left(1 - \frac{\sum\limits_j N_{L,j}}{N_S}\right)^2 \left(1 - \frac{\sum\limits_j N_{L,j} \cdot \dfrac{R'_S - R'_L}{R'_S}}{N_S}\right)^{-1} . \tag{67}$$

The notation is explained in Appendix A. At moderate tank sizes ($\varnothing \leq 1$ m, $l \leq 4$ m) a well-proven technique is to divide the tank into an upper and a lower shell, plus the centre frame containing the girders as well as the drift tube structure (Fig. 17). Metal sealings are used throughout for all joints (silver, copper, aluminium).

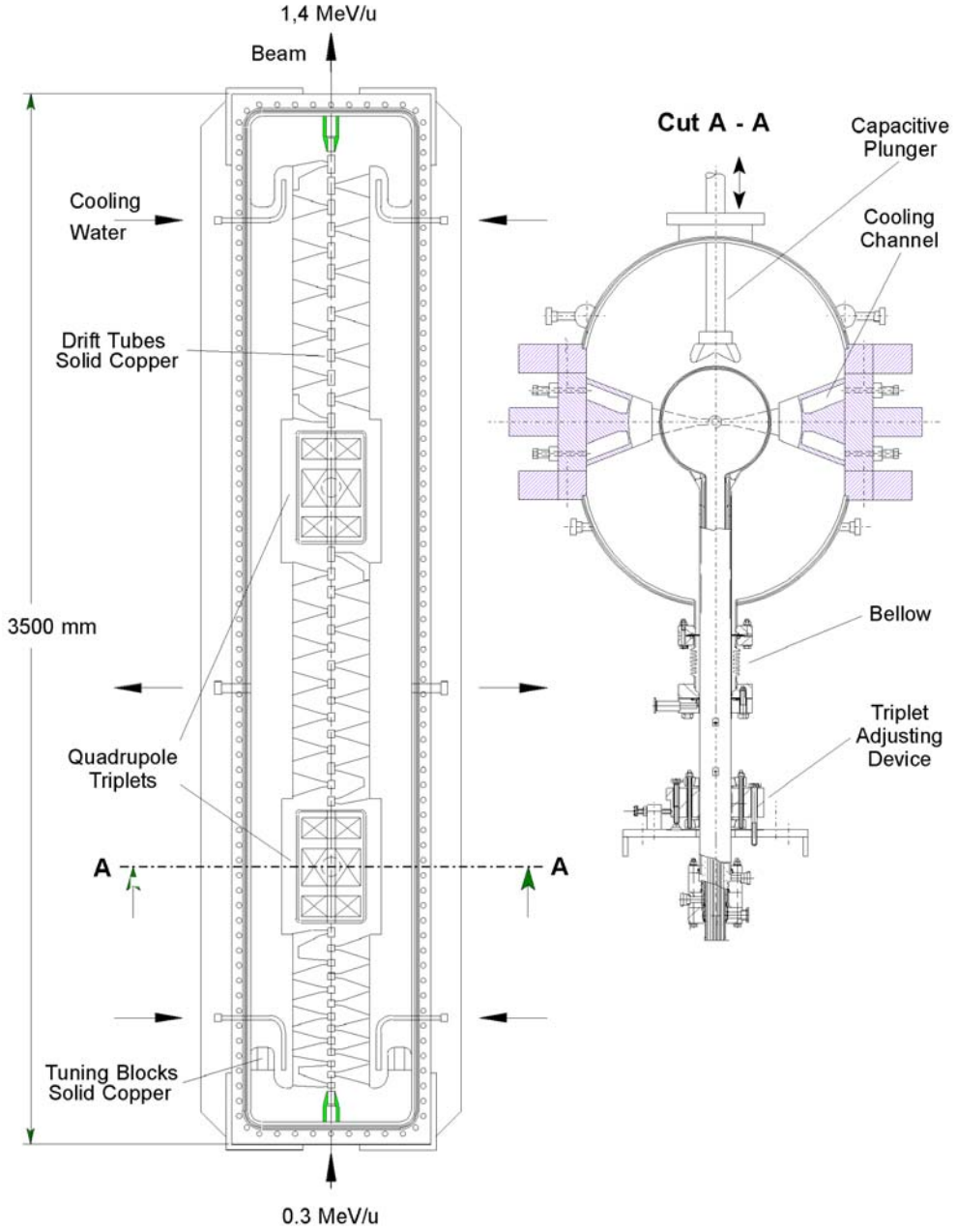


Fig. 17: View of the 108 MHz IH-cavity of the GSI High Charge State Injector containing two quadrupole triplets. Cut A-A shows that the stem causes RF grounding of the corresponding lens housing.

3.2.2 *Lens housing on RF potential*

In the case of large cavity diameters it is preferable to install all accelerating and focusing elements in the vertical plane to achieve sufficient mechanical stability. Moreover, when mounting all lenses of a cavity on top of the bottom girder this girder acts like an optical bench. A flat distribution of the capacitive and inductive load is achieved by creating deep recesses in the top girder and by elongating the lens supports close to the lens ends (Fig. 18). The length of a lens housing j is now roughly given by

$$L_{L,j} = \left(N_{L,j} + \tfrac{1}{2}\right) \beta_j \cdot \lambda - \overline{g}_j \quad , \tag{68}$$

and the shunt impedance when compared to Eq. (50) is reduced by the factor

$$F_L = \left(1 - \frac{\sum_j N_{L,j}}{N_S}\right)^2 \left(1 - \frac{\sum_j \left(N_{L,j} + \frac{1}{2}\right) \frac{R'_S - R'_L}{R'_S}}{N_S}\right)^{-1} \quad . \tag{69}$$

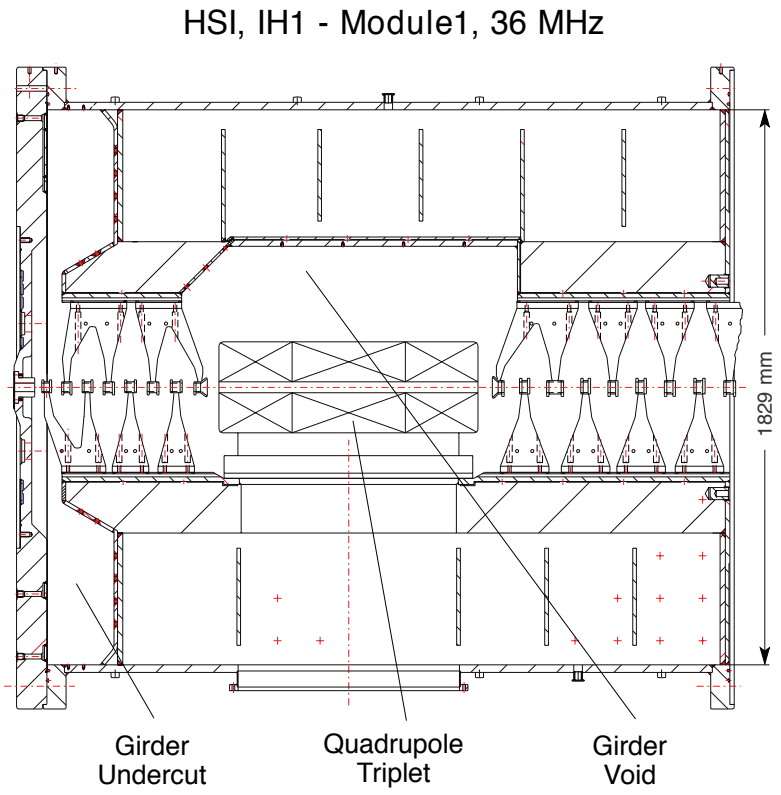


Fig. 18: First module of the 36 MHz High Current IH-DTL, containing one quadrupole triplet mounted on the bottom girder. The pronounced top girder recess is necessary to match the capacitive load.

It should be noted that the length of focusing elements is also an issue if they are mounted outside the cavity. But then the shunt impedance is not affected when the definition as given by Eq. (33) is used.

Each of the two GSI 36 MHz IH-cavities consists of four cylindrical modules, which are bolted together and sealed by aluminum wires. This design is recommended for large tanks.

Remark: The need to integrate lenses into CH-type structures is of less importance as this structure is more often used at higher beam energies. Nonetheless, it is assumed that the concepts described above are applicable to CH-structures as well.

3.3 Electric dipole compensation of the gap fields

IH-type DTLs contain a remaining contribution from the transverse electric field orientation—from stem to stem across each gap. The dipole orientation is inverted from gap to gap and causes either a parallel displacement of the beam or an oscillation around the beam axis (if the beam is injected at an angle relative to the beam axis, or if the voltage of the first and last gap equals just half the voltage of the intermediate gaps).

In non-relativistic approximation the parallel displacement of an axially injected beam after passage of two gaps and of two drift tubes is given by

$$Y = 2\,Y_g + Y_d\,;\ \text{with} \tag{70}$$

$$Y_g \cong \frac{V_{eff}\cdot\cos\Phi\cdot q\cdot tg\delta_{eff}\cdot g}{2m_0c^2\cdot\beta^2}\ ;\ 0.7\,\delta_0 \le \delta_{eff} \le 0.8\,\delta_0\ ;\ \text{and} \tag{71}$$

$$Y_d = d\cdot\alpha(1)\,;\ \alpha(1) \cong \frac{2Y_g}{g}\ ; \tag{72}$$

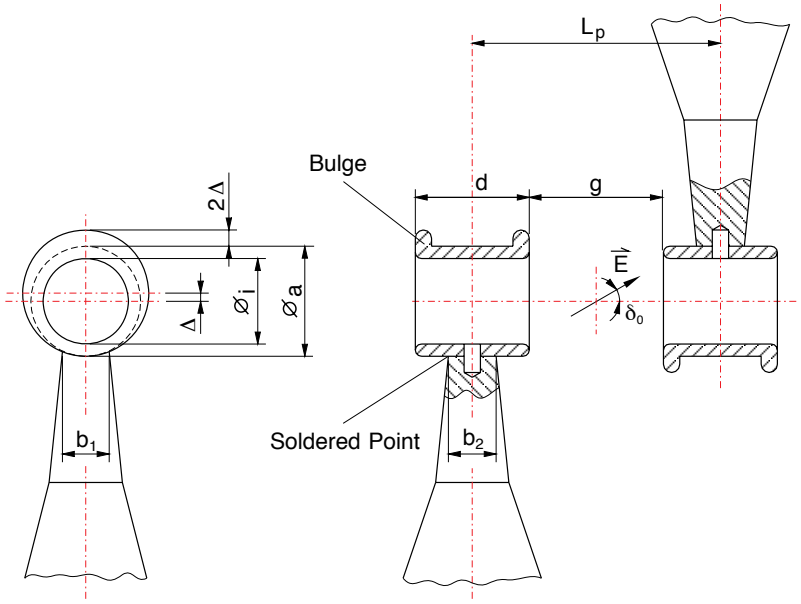


Fig. 19: Drift tube geometry for IH-type structures, reducing the electric dipole content and symmetrizing the field around the beam axis

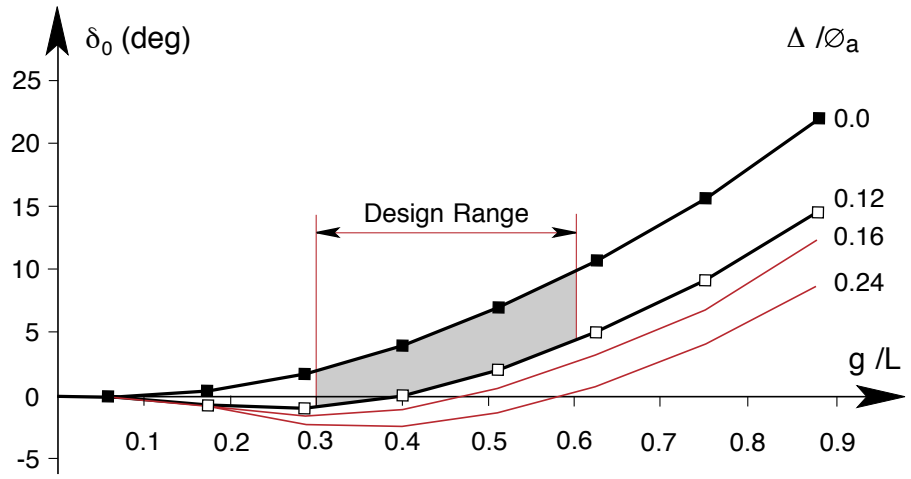


Fig. 20: Angle δ_0 between electric field and beam axis at the gap centre as a function of g/L_p and of the bulge geometry. The plot is valid for the case $\varnothing_a / L_p = 0.48$, $\varnothing_a / \varnothing_i = 1.4$ [29].

Compensation of the dipole component at the low energy end of each linac section corresponding to the same resonance frequency was achieved by developing the drift tube geometry illustrated by Fig. 19. Only at these low energy regions can the conditions

$$g/L_P \leq 0.6\,; \quad \varnothing_a/L_P \geq 0.5 \tag{73}$$

be fulfilled fairly well. Figure 20 shows the δ_0-dependance on the parameters g/L_p and Δ. For the other parts the remaining steering effects have to be investigated, and sufficient steering power within the intertank sections has to be ensured.

A second advantage of the proposed drift tube geometry (Fig. 19) is the improved axial gap field symmetry: the pronounced field maxima within the aperture are shifted radially outwards by the bulges.

3.4 Voltage gain per unit length

The effective voltage gain of an IH-DTL with slim drift tubes as described above is surprisingly high. At the 202 MHz tank 2 of CERN Linac 3 (Fig. 2), an effective voltage gain of up to 10.7 MV/m between end flanges was achieved for 500 μs pulse length and 0.83 Hz repetition rate [30]. Only a few days were available to perform these tests, with voltage levels of up to a factor of 1.82 above the regular operating value (Fig. 21). At the maximum level, about 14% of the RF forward power was absorbed by dark currents. With 1300 kW forward power, 16.5 MV in effective voltage gain were achieved within a total tank length of 1.54 m. Voltage gains of up to around 9 MV/m in the dark current contribution remain in the 1% region. Figure 22 shows MAFIA results for the gap with maximum surface fields. Reasons for the high sparking resistance of the structure when compared, for example, with an Alvarez cavity are assumed to be explained by:

– multipacting at low power levels only (typically in the P ~ 100 W region);

– no high field regions with homogeneous electric field distribution (slim tubes); and

– low stored field energy (modest damages from sparking).

High voltage gains per metre are useful for KONUS structures as they provide sufficient longitudinal focusing power.

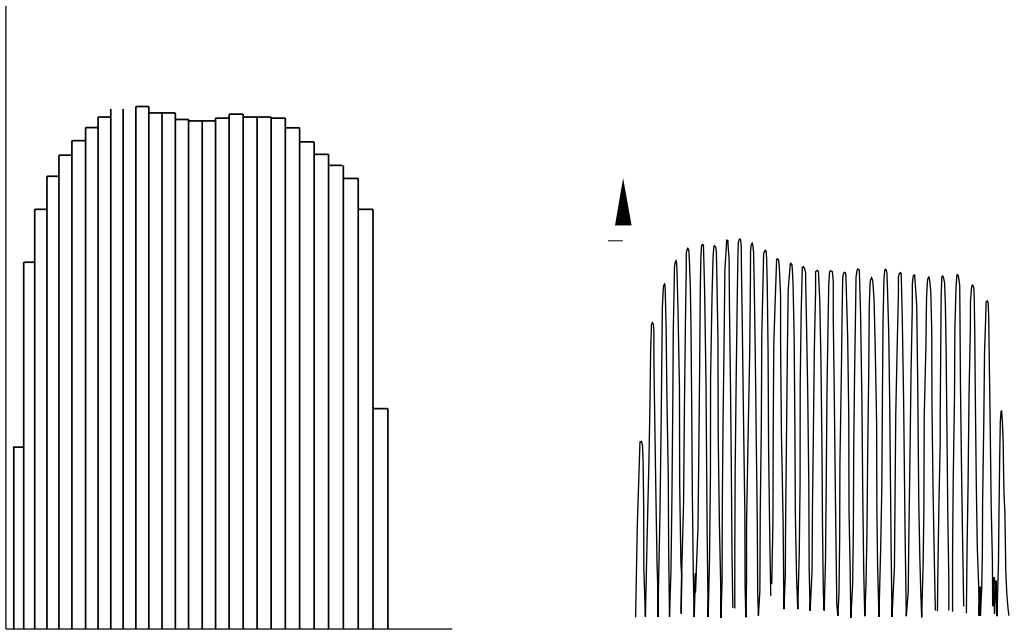


Fig. 21: Gap voltage distribution and on-axis field distribution along the 202 MHz CERN Pb-Linac, tank IH2 (see Fig. 2) at maximum amplitude.

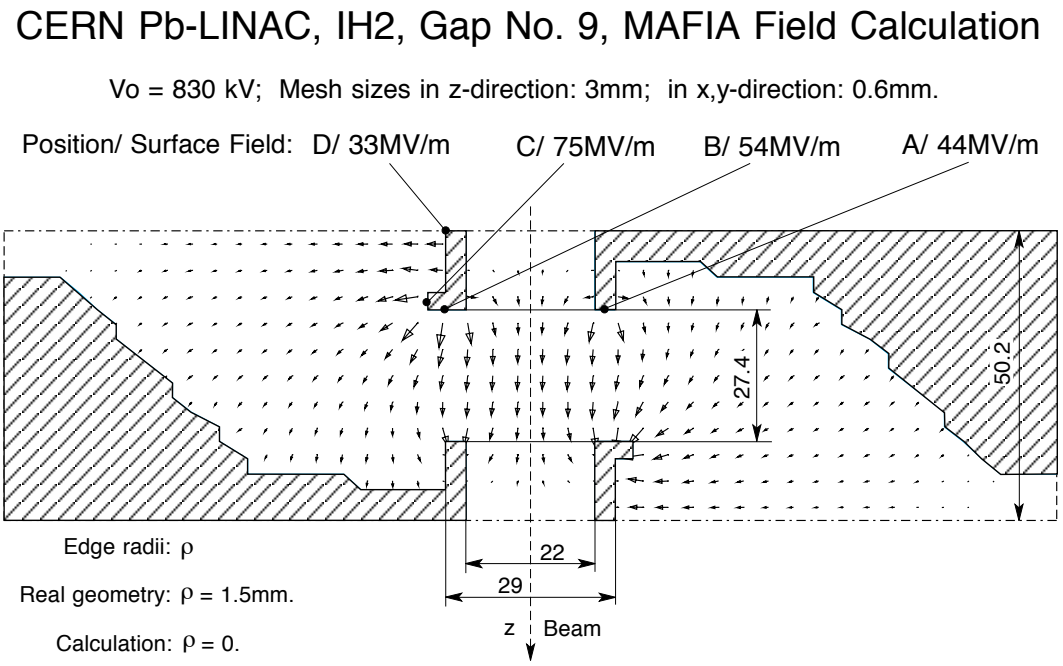


Fig. 22: Calculated electric field distribution at gap no. 9 for the case illustrated in Fig. 21.

4. POTENTIAL OF THE CH-DTL

Analytical estimations and investigations with MAFIA have shown that drift tube structures up to $\beta \leq 0.5$ become accessible with the CH-DTL mode. Operating frequencies can be as high as 700 MHz. A parallel, two-directional investigation is envisaged:

– the development of room temperature prototype cavities; and

– the design and construction of a superconducting prototype cavity [31].

Preliminary beam dynamics investigations with LORASR [32] show that this structure is especially attractive for p and light-ion acceleration including high (100 mA) beam currents. It can be used from the RFQ exit (typically 5–7 MeV/u beam energy) up to injection into a Coupled Cavity Linac (CCL). It is assumed that high voltage gains—similar to IH-DTLs—can be realized. Because of the stem array, no transverse dipole components are contained in the gap field of that structure.

5. CONCLUSIONS

H-type modes are by definition desirable in building RFQs. The IH-RFQ and four-vane RFQ together cover the whole frequency range of interest. By the use of slim drift tubes with adequate stem geometries, matched cavity ends, and application of the KONUS beam dynamics, efficient beam acceleration of up to $\beta \sim 0.25$ by the IH-DTL and $\beta \sim 0.5$ by the CH-DTL can be ahieved. Because of the high capacitive load, the magnetic field distribution becomes very homogenous and realization of the zero mode is possible. These consequences allow for an analytical approach to calculate the RF field distribution.

Three-dimensional codes, together with RF models, are very helpful in the design of H-type structures. Four-vane RFQs were very successful from the beginning. They are the only choice at frequencies above 250 MHz. A number of IH-DTLs are successfully used in routine heavy-ion acceleration operations. The CH-DTL may become the missing link between RFQ and CCL for proton linacs. This structure also provides good mechanical and RF conditions for the development of superconducting CH-cavities.

REFERENCES

[1] P. Blewett, Proc. Symposium in High-Energy Accelerators and Pion Physics, CERN, Geneva, Switzerland, 1956, p. 162.

[2] P.M. Zeidlitz, V.A. Yamnitskii, *J. Nucl. Energy,* **C4** (1962) 121.

[3] V.A. Bomko, E.I. Revutskii, *Sov. Phys. Tech. Phys.* **9** (1965) 973.

[4] M. Bres *et al., IEEE. Trans. Nucl. Sci.* **NS-16/3** (1969) 372.

[5] J. Pottier, *IEEE Trans. Nucl. Sci.* **NS-16/3** (1969) 377.

[6] E. Nolte *et al., Nucl. Instrum. Methods* **158** (1979) 311.

[7] I.M. Kapchinskiy, Proc. LINAC Conf., Seeheim, Germany, 1984, GSI-84-11, p. 43.

[8] J.M. Potter *et al.,* Proc. PAC Conf., San Francisco, USA, 1979, (*IEEE Trans. Nucl. Sci.* **NS-26,** 1979, 3745).

[9] U. Ratzinger *et al., Nucl. Instrum. Methods Phys. Res.* **A263** (1988) 261.

[10] Y. Oguri, E. Arai, T. Hattori, *Nucl. Instrum. Methods Phys. Res.* **A235** (1985) 7.

[11] U. Ratzinger, Proc. LINAC Conf., Newport News, Virginia, USA, 1988, CEBAF–Report–89–001, p. 185.

[12] U. Ratzinger, R. Tiede, *Nucl. Instrum. Methods Phys. Res.* **A415** (1998) 229.

[13] U. Ratzinger, Proc. IEEE Particle Accelerator Conf., San Francisco, USA, 1991, p. 567.

[14] U. Ratzinger, Habilitationsschrift, Goethe-Univ. Frankfurt, 1998.

[15] K. Kaspar, U. Ratzinger, Proc. EPAC Conf., Sitges, Spain, 1996, vol. 3, p. 1973.

[16] R.E. Laxdal, P. Bricault, Proc. LINAC Conf., Geneva, 1996, CERN 96-07, p. 435.

[17] B. Krietenstein *et al.,* Proc. PAC Conf., Vancouver, Canada, 1997, p. 2645.

[18] T. Weis, H. Klein, A. Schempp, Proc. LINAC Conf., Seeheim, Germany, 1984, GSI 84-11, p. 417.

[19] K. Satoh, S. Kamohara, E. Arai, *Nucl. Instrum. Methods Phys. Res.* **A287** (1990) 353.

[20] T.P. Wangler, Proc. LINAC conference, Seeheim, Germany, 1984, GSI-84-11, p. 332.

[21] O. Zinke, H. Brunswig, *Lehrbuch der HF-Technik,* 2nd ed. (Springer-Verlag, 1973).

[22] I. Ben-Zvi *et al.,* Proc. LINAC Conf., Albuquerque, USA, 1990, LA-12004-C, p. 73.

[23] E. Müller and H. Klein, *Nucl. Instrum. Methods Phys. Res.* **A224** (1984) 17.

[24] H. Höft, *Passive elektrische Bauelemente* (Dr. A. Hüthing Verlag, 1977).

[25] U. Ratzinger, K. Kaspar *et al., Nucl. Instrum. Methods Phys. Res.* **A415** (1998) 281.

[26] J.H. Billen *et al.,* Proc. LINAC Conf., Tsukuba, Japan, 1994, p. 341.

[28] R. Popescu-Tiede, diploma thesis, IAP, Goethe Universität Frankfurt, 1993.

[29] J.-C. Nanon, CERN-PS/HI/Technical Note 92-03.

[30] J. Broere *et al.,* Proc. LINAC Conf., Chicago, USA, 1998 ANL-98/28, p. 771.

[31] R. Eichhorn, U. Ratzinger, Proc. 9th Workshop on RF Superconductivity, Santa Fe, USA, 1999.

[32] U. Ratzinger, Proc. GSI Scientific Report, 1998, GSI 99-01, p. 166.

APPENDIX A: DEFINITION OF PARAMETERS

a	aperture radius in RFQs
A	atomic mass number
A_a	geometric area (see Fig. 9(a))
$B, B_0, B_{0,z}$	magnetic flux, amplitude, amplitude of z-comp.
c	velocity of light in vacuum
c_1	loss factor for extra paths along electrodes
c_2	loss factor regarding surface quality, RF joints, cavity ends
C'	capacitance per unit length in one chamber, without electrodes
C'_D	capacitance per unit length of a drift tube structure and of the stems with carrier rings for IH-RFQs, respectively
C'_Q	capacitance per unit length of a quadrupole, consisting of mini vane
C'_{QD}	capacitive contribution 'stems with carrier rings' — 'mini vanes' in IH-RFQs
C'_t	total capacitance per unit length for a cavity
d	drift tube length
d_e	d without rounded ends
d_I	path length of the electric current
$E, E_0, E_{0,z}$	electric field, amplitude, amplitude of z-comp.
f	resonance frequency
F	ratio of geometric areas
F_L	reduction factor of Z_0
g	gap length
$\overline{g}$	average value of two neighbouring gap lengths
g_e	$g + d - d_e$
h	mini vane geometric parameter (see Fig. 6)
H, H_0	magnetic field, amplitude
I	electric current
I'	electric surface current, A/m
$I'_{0,i}, I'_{0,f}$	amplitude of I' before and after a modification
l	cavity length
ℓ	longitudinal mode index
L_I	depth of undercuts (see Fig. 12(a))
L'	chamber inductance per unit of a H-cavity without electrodes
L_L	length of lens housing
L_P	cell length $\beta\cdot\lambda/2$
L_T	tank length between end flanges
m	mode index with respect to φ
n	mode index with respect to r
N_G	gap number of a cavity

$N_{L,j}$	cell number in units of βλ/2 corresponding to the length of lens no. j, including both gap half-lengths
N_S	cavity length in units of βλ/2
P, P'	RF power losses, per unit length
P_s	RF power losses within one resonator chamber
q	ion charge state
Q	quality factor
Q'_0	amplitude of electric charge per unit length
r_a	equivalent outer drift tube radius (see Fig. 8), including the contribution of stems within R_1
R	ohmic resistance
R'_L	ohmic res. per unit length of a cavity at the position of internal lenses
R'_S	ohmic res. per unit length of a cavity along the accelerating sections
R'_{p0}	specific shunt resistance of RFQs, defined by Eq. (36)
R_0	geometric vane parameter (see Fig. 7): averaged aperture radius of a modulated electrode
$R_1, R_2; R_T$	inner and outer radius of the equivalent segment to approximate a cavity chamber cross section; outer chamber radius
T	transit time factor
U, U_0	voltage amplitude
$U_{0,i}, U_{0,f}$	voltage amplitude before and after a modification
Y_g, Y_d	transverse beam excursions in IH-DTLs along gaps and drift tubes
z_0	shunt impedance of DTLs (Eq. 33)
Z_{eff}	effective shunt impedance (Eq. 35)
$\alpha(1)$	deflection angle of the beam caused by a dipole field component after passage of one gap
β	relativistic parameter
$\overline{\beta}$	β-value in the centre of a multicell structure (Eq. 48)
δ	skin depth
δ_C	variation of the capacitive layer
δ_{eff}	reduction of δ_0 when integrating along the whole gap field
δ_0	angle of the electric field direction against the beam axis at the gap centre
Δ	bulge parameter (see Fig. 19)
ε_0	absolute permittivity
κ	electric conductivity
μ_0	absolute permeability
ρ	vane parameter (see Fig. 7)
Φ	magnetic flux
ω	2 π f
ϕ_i, ϕ_a	inner and outer drift tube diameter

RECENT DEVELOPMENTS IN THE USE OF ACCELERATORS FOR RADIATION THERAPY

E. Pedroni
Paul Scherrer Institute,
Villigen, Switzerland

1 INTRODUCTION

This is a write-up of a seminar given in the framework of the CERN School for accelerator specialists in Darmstadt (April 2000). The intent of the presentation is to give a brief overview of the use of accelerators for cancer therapy. It focuses mainly on the utilization and the respective merits of different types of accelerators for radiation therapy (RT), but not on the accelerator technology itself. The first half of the work contains a superficial introduction to RT. In the second half the emphasis is put on the most recent developments in this field. For time reasons we do not discuss other medical applications of accelerators, for example isotope production or the use of synchrotron light. For a comprehensive historical review we refer to some excellent textbooks and to the specific literature [1],[2].

A major intention of the presenter's own work (the advanced use of proton therapy) is to show that research in accelerator technology continues to play an important role in the progress of medicine.

2 THE ROLE OF RADIATION THERAPY IN THE MANAGEMENT OF CANCER

Cancer is the second major cause of death (after cardiovascular diseases) in developed countries. The total cure rate is nowadays about 45% (of all cancer patients). This is the result of a slow but steady improvement recent decades.

For about two-thirds of the patients the disease is still well localized within a specific region of the body when the patient is confronted with the diagnosis of cancer. For these patients the chances of cure using a local therapy are reasonably high.

2.1 Cancer therapy modalities

The following therapeutic strategies are presently available:

- **Surgery**

Whenever it is possible a radical surgical excision of the disease is the preferred therapeutic choice. The earlier the diagnosis and the smaller the tumour, the better the chances of a good therapeutic outcome. In this context screening plays an important role in the early detection of the disease. Surgery is the most successful therapy and it accounts for 22% of all cures (this includes early cases, small tumours and skin cancer).

- **Radiation therapy (RT)**

This is the second choice. RT is used with a curative intent when the tumour is inoperable but still well localized. Radiation therapy alone accounts for 12% of all cures, and a further 6% in combination with surgery.

- **Chemotherapy**

When the disease has already spread to the whole body (with distant metastases) the chances of cure are correspondingly lower. Chemotherapy is then used with the intent of eliminating microscopic distant metastasis. This accounts for the remaining 5% of the cure rate.

- **New biological methods**

The largest effort in cancer research today is being made by the big pharmaceutical companies on developments based on new biological sciences, mainly with the aim of controlling the distant spread of the disease. At present, the highest hopes are placed in immuno and genetic therapies. Systemic therapies usually have to cope with the unwanted side effects of the drugs, which are distributed to the whole body. The difficulty of transporting the drug into the core of the tumour and the extremely large number of cancer cells involved in a solid tumour represent major problems for the utilization of systemic therapies for the elimination of advanced solid tumours. For these reasons the new biological methods still aim, in the first place, at the inactivation of isolated cancer cells (the microscopic spread of the disease).

It is likely that, in the foreseeable future, surgery and radiation therapy will continue to play a major role in the control of the primary solid tumour. Whilst waiting for big breakthrough from genetic technologies, it is reasonable to continue to improve traditional established methods, such as surgery and radiotherapy.

2.2 Radiation therapy: a local treatment

It is important to note that the different types of therapy are often complementary and not necessarily exclusive. Many recent successes in cancer management are in fact based on the combined use of different modalities. In addition to the use of radiation with a direct curative intent, radiotherapy is also used for palliation (release of pain, prolongation and maintenance of the quality of life) and for prophylactic purposes (sterilization of localized zones with a suspected microscopic invisible spread of the disease).

Nowadays about two-thirds of all cancer patients receive radiation therapy alone or in combination with other modalities.

Radiation therapy is a local treatment and as such it will continue to play a very important role in the fight against cancer in the future. About 15% of all patients die from failure to control the primary tumour. In Germany alone the cancer registry indicates about 340 000 new cases per year. A total of 210 000 patients still die from their disease [3]. Additional progress in radiation therapy, even by a few per cent, could have a positive impact on the quality of life of a large number of patients. Improvements, as discussed below, could be achieved by using more advanced treatment techniques (for example intensity modulated radiation therapy) and/or by using unusual types of radiation sources (such as external therapy with protons or light ions).

3 RADIATION THERAPY CONCEPTS: AN OVERVIEW

The action of radiation on a biological target is a chain of effects, initiated by the beam. Through ionization (physics) we have the production of radicals (chemistry) causing reparable and irreparable errors in the genetic code (DNA) at the cellular level (biology). For cancer therapy the most important endpoint is to stop the capability of the cancer cells to replicate and thus prevent the tumour from growing further. Unfortunately, not only the cancer cells are damaged by radiation. The dose burden on healthy tissue is the other important effect to consider when planning a radiation treatment. The dose in the tumour is usually chosen as a function of the tolerance of the surrounding healthy tissue. There a delicate balance between the chances of a cure and treatment morbidity due to acute and late reactions.

The following are important concepts in radiation therapy:

- **Therapeutic ratio**

A necessary condition for success of treatment is an inherent higher sensitivity to radiation of the cancer cells compared to the healthy cells. The most important mechanisms are the capability of the healthy cells to repair radiation damage spontaneously and for the healthy tissue to repopulate the irradiated organs with new non-irradiated cells.

- **Fractionation**

Fractionation is the repeated application of the treatment with a fractionated dose at time intervals long enough for repair and repopulation mechanisms to take place (the assumed minimum interval is of the order of 6 h or more). In this way the effects of a positive therapeutic ratio are repeatedly applied and the chances of the outcome are correspondingly improved.

Radiation therapy is typically applied 5 days per week over 5–6 weeks. The fraction dose is typically 2 Gy and the total dose applied is around 50–70 Gy. It is generally believed that the cure is reached by the inactivation over the course of the full treatment of each single cancer cell in the body. Eventually the immune system of the body contributes to the cure at the level of the very last few remaining cancer cells. Usually the dose is chosen as a balance between the chances of eradicating the tumour and the risk of producing complications in the healthy tissue. The desired precision for the prescription of the dose and for the homogeneity of the dose distribution within the target volume is (ideally) of the order of a few per cent (altogether the combined errors should be less than 5%).

- **Physical selectivity (dose localization)**

The success of RT is based on the ability to confine the dose delivery to a small region of the body, which contains the whole tumour. Ideally one would like to deposit only the minimum necessary dose inside the target volume and zero dose outside. In practice this is technically impossible. The most recent strategy is to confine the dose as strictly as possible to the target volume by shaping the dose in all three dimensions to conform to the shape of the target volume (3D dose conformation).

Historically, most of the improvements in RT were achieved by increasing the dose localization using different technologies. This includes computer-based treatment planning, better target delineation using modern diagnostics, advanced beam delivery treatment techniques (conformal therapy), and the use of exotic types of radiation.

- **Radiation quality: the dependence on Linear Energy Transfer (LET)**

The efficacy of only a given type of radiation in destroying the reproductive capability of a cell does not only depend on the applied dose (the energy per mass). The mechanism details of the deposition of the ionization at a microscopic level play an important role as well. The LET describes the density of ionization around the track of the primary ionizing particles measured at the typical size of the cell (of the DNA). The 'dose quality' (expressed by the LET) depends on the type and on the energy of the ionizing particles and has a strong influence on the biological effects. (This is a major subject of radiobiology.)

- **Conventional therapy: low LET radiation as the reference**

The most commonly used sources for external radiation therapy are photon and electron beams delivered by electron linacs. This technology is now substituting cobalt sources almost entirely. Radioactive sources today are often used for interstitial therapy (using afterloading techniques and radioactive implants). All these conventional sources are very similar in their radiobiological effects and represent the standard reference in radiobiology (low LET radiation).

- **High LET radiation: the exotic type of radiation**

The higher cell-killing efficiency of high LET radiation is usually characterized by a radiobiological efficiency (RBE) value higher than 1 (of the order of 2 to 5). The RBE is defined as the ratio of the dose of non-conventional radiation compared to the standard (photons) necessary to obtain the same biological effect. The highest cell-killing efficiency is found at LET values of the order of 100 keV per micron. Above this value the effect drops again (overkill).

Another important radiobiological quantity is the so-called oxygen enhancement ratio (OER). Cells with a deficit of oxygen are usually more resistant to radiation than well-oxygenated cells. The OER expresses this difference. The oxygen-deficient cells are suspected to be the cause of tumour relapse when the tumour is treated with a low LET radiotherapy (an attempt to get rid of this effect is

through the use of special drugs, the so-called radiosensitizers). The negative effect represented by a high OER value at low LET disappears gradually with increasing LET. High LET radiation is therefore potentially more efficient for treating large radio-resistant tumours, when the poor vascularization of the tumour implies the presence of anoxic tumour cells as a potential cause for a treatment failure.

A good classical textbook on radiobiology can be found in Ref. [4].

4 CLASSIFICATION OF ACCELERATOR-BASED RADIATION SOURCES

An excellent introduction to the field of radiotherapy, mainly from the point of view of charged particles therapy, can be found in Refs. [5],[6].

4.1 Low LET beams

The most commonly used low LET radiation sources are:

- **Photons and electrons**

Photons and electrons are classified as low LET radiation beams (the LET distribution is below 20 keV per micron). This radiation type represents the standard used in hospitals. Conventional therapy is discussed in more detail in Section 5.

- **Protons**

Protons are considered to be biologically similar to photons. They are potentially interesting for their superior physical selectivity. The significance of proton therapy in the world is slowly increasing. This is where accelerator technology is expected to contribute most in the future (see Section 6).

4.2 High LET beams

Classified as high LET beam therapies, with a clearly different radiobiological behaviour than conventional therapy, are the following radiation beams:

- **Neutron therapy**

Neutron therapy has been an active field of research over the last two decades. The depth dose curve behaviour for the neutron is exponential with depth (similar to photons). Its physical selectivity is therefore similar to photon therapy. A dozen facilities worldwide are applying neutron therapy (mainly using a dedicated cyclotron). The clinical results were partially encouraging but not enough to justify the installation of new dedicated facilities. The most cited success of neutron therapy has been achieved for the treatment of salivary gland tumours. Owing to its superior physical selectivity, light ion therapy is now slowly replacing neutron therapy.

- **Pion therapy**

Pion therapy is mentioned here only for historical reasons. High LET beams were explored in the 1980s at the so-called pi meson factories at Los Alamos, U.S.A., Triumf, Canada, and at PSI (formerly SIN) in Switzerland. The trials were abandoned mainly due to the difficulty of obtaining beams with a high enough intensity. Another reason was the superiority of the physical selectivity of light ion beams: pions show an excellent behaviour with respect to the dose profile in depth, but multiple Coulomb scattering spoils the quality of the penumbra in the lateral direction of these beams.

- **Light ions (mainly carbon)**

Therapy with high LET radiation is now being proposed again, particularly in the context of light ion therapy. The rationale for the use of ion beams resides in the combination of the high LET with an excellent physical selectivity. The world-wide situation of ion therapy is described in more detail in Section 7.

- **BNCT**

A special role—due to the potential selectivity at a cellular level of the drug (for example antibodies) carrying boron—is played by Boron Neutron Capture Therapy (BNCT) [7]. Accelerator technology could play a role in the production of thermal neutrons with dedicated cyclotrons, instead of nuclear reactors.

5 CONVENTIONAL THERAPY WITH LINACS

5.1 Standard therapy

Today the use of electron linacs represents the state of the art in radiation therapy (low LET radiation with electrons and photons). One should not forget interstitial radiotherapy with radioactive sources and therapy with radioactive isotopes. Linacs can be found in every major hospital and their total number is probably in the order of many thousands worldwide. The electron beam energy can be chosen in the range of 6–25 MeV. The accelerator size is typically less than 1 m long and is mounted on the head of a rotating gantry. The electron beam is bent by a magnetic channel in the direction of the patient. The photon beam is created by bremsstrahlung of the electrons impinging on a metallic target in the head of the gantry. Figure 1 shows a photograph of a modern linac, a Varian machine at the Radiotherapy Department of the Triemli Hospital in Zurich.

Fig. 1: Photograph of a modern electron linac (courtesy of U. Schneider)

The photons are emitted from the head of the gantry with a homogeneous flux in the solid angle used for the treatment (this covers target sizes of maximally 40 cm at a distance of 1–1.5 m from the source, the so-called source-to-isocent distance). The dose in the patient arises from the secondary electrons created by the photo- and Compton effect. After a build-up region in the first few centimetres below the patient's skin, the dose falls exponentially with depth (see Fig. 2). In order to deposit more dose into the target volume than the surrounding tissue, it is necessary to apply the beam from several directions. For this the gantry is rotated around the patient table. The patient remains immobilized in his treatment position on the treatment couch (preferably in the supine position). An alternative to

the delivery of single static dose fields is the use of a continuous rotation with the beam switched on (arc therapy). The individual dose fields are shaped in the direction transverse to the beam by using individually shaped collimators (designed by the treatment planning programme). So-called Multi-Leaf Collimators (MLC), with a dynamic computer-controlled movement of the leaves, are nowadays slowly replacing fixed collimators. Through the convergent superposition of the dose fields on the target volume and by distributing the entrance and exit doses over several organs, it is possible to apply the necessary (homogeneous) dose to the target volume, whilst maintaining the burden on the healthy tissue below the tolerance limit. The data used for defining the target volume in modern radiotherapy are nowadays obtained directly from computer tomography.

After removal of the bremsstrahlung target from the beam path, the electron beams can be used directly for electron therapy. The depth dose curve of an electron beam entering the patient shows a shallow bump with a limited depth of penetration. The electron range is only several centimetres depending on the initial energy of the beam. Electron therapy is used for treating superficial tumours. Another important application of electron therapy is for intra-operative applications (open body irradiation during surgery, mainly for sterilizing the region where the tumour has been removed).

The accelerators used in hospitals are based on very advanced technology (standing wave electron linacs) and are produced commercially (this has an important impact on cost). The beam is pulsed (with typically 1000 pulses of a few microseconds duration per second). This kind of technology is definitively replacing other, historical types of radiation sources used in the past (X-rays, cobalt sources, and betatrons).

5.2 Most recent developments in conventional therapy: Intensity-Modulated Radiation Therapy (IMRT)

The use of sophisticated beam delivery techniques, support by computer technology, and the information gained with modern diagnostic techniques (Computer Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emisson Tomography (PET)) have been at the origin of the progress achieved in RT in the last decade. The dynamic use of multi-leaf collimators and the control of patient positioning using portable imaging devices are the most important technical developments today [8].

The common goal of all developments is to give enough (homogeneous) dose to the target volume whilst reducing as much as possible the dose outside the target volume. Another possibility is to choose where to deposit the unwanted dose (selection of gantry angles). The most promising but also the most challenging approach is to shape the geometrical distribution of the dose such that it conforms as much as possible to the three-dimensional shape of the target volume (3D conformal therapy).

The dynamic use of multi-leaf collimators offers additional possibilities. The most interesting is to apply the dose with a non-uniform distribution of the photon fluence. The flux is varied across the cross section of each dose field. This is IMRT. The superposition of intentionally non-homogeneously shaped dose distributions can result in a homogeneous dose distribution of superior quality (with a higher degree of conformity, especially in the case of target volumes with concavities). This is made possible by the use of advanced optimization algorithms in treatment planning using modern computer technology. The optimization of the delivery of radiation using multiple beam ports is a typical ‘inverse problem’ with a strong analogy to CT. With CT one uses multiple projections (from many angles) to reconstruct complex density images. With IMRT one uses complex dose projections to produce more ideal dose distributions.

Figure 2 shows an IMRT example of the dose distribution for a tumour near the base of the skull. The calculation has been made with the treatment planning package of PSI (courtesy of T. Lomax). The dose is delivered using nine photon fields. Each field is modulated in intensity. One recognizes immediately that the optimization algorithm avoids sending photons towards the brain stem in each of the fields. The dose delivered to this sensitive organ is reduced considerably in this way. The availability of a large amount of degrees of freedom and the strength of the mathematical methods make it possible to

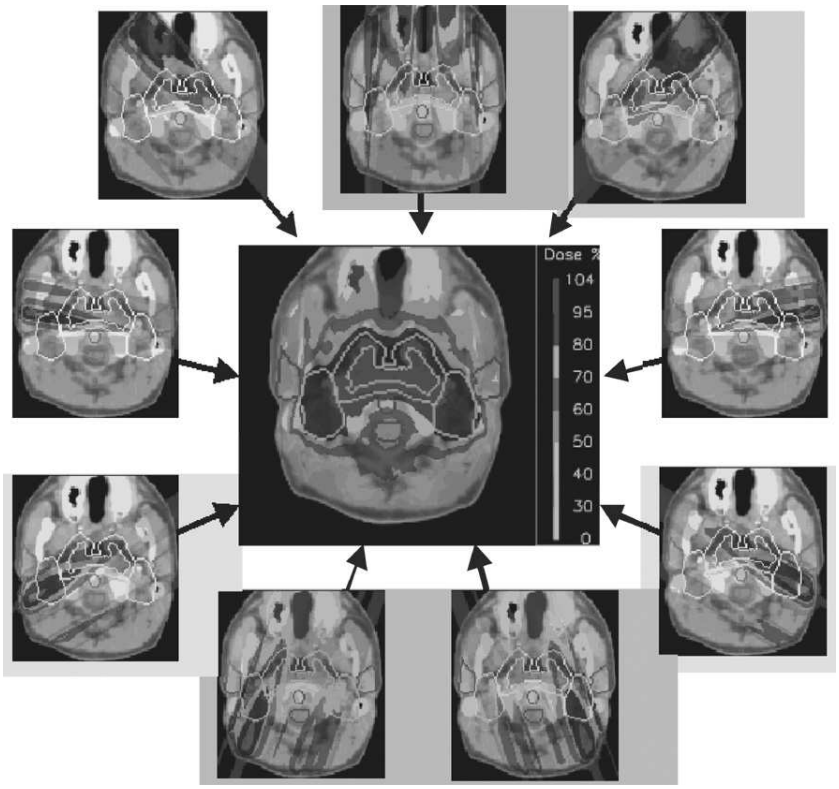


Fig. 2: An example of intensity-modulated radiation treatment planning with photons. Through the addition of nine fields it is possible to construct a highly conformal dose distribution with good dose sparing in the region of the brain stem (courtesy of T. Lomax).

produce a very good dose distribution, shaped in all three dimensions to conform precisely to the target volume with a sufficient homogeneity.

The obvious but most tedious way of applying intensity-modulated photon beam delivery is by using absorbing filters of variable thickness in the path of the beam (individually shaped filters). A more modern approach is to change the position of the leaves of an MLC under computer control as a function of time during beam delivery: this is what is usually meant by IMRT [9]. Another way is to use an array of individual beam shutters in one lateral direction and to work on the dose deposition slice by slice and angle by angle in the other directions. This is called tomotherapy [10]. Tomotherapy is where the analogy with CT is the closest, also from the point of the beam delivery (full rotation and slice by slice beam delivery).

The fashionable keywords in medical physics nowadays are ‘inverse planning’, ‘intensity modulation therapy’, ‘tomotherapy’ and ‘computer-controlled multi-leaf collimators’. These ideas are however not very new. A first example of computer-optimized dynamic beam delivery of conformal radiation goes back as early as 1981 with the pion therapy project of PSI. Pion therapy was delivered to about 500 patients using a dynamic beam delivery technique. The pions were applied simultaneously from 60 concentric pion beams. The patient was moved under computer control along a three-dimensional path inside a cylinder bag filled with water. The degrees of freedom for the optimization algorithm were chosen by changing the velocity of the movement of the hot spot of the dose (fixed at the iso-centre of the machine). From the point of view of the technology this was similar to tomotherapy, but 20 years in advance [11]. Unfortunately, the spot used for pion therapy was rather large (5 cm FWHM). This was because of the large size of the target used for the production of the pions, a problem related to the difficulty in obtaining a sufficiently high number of pions for therapy. Pion and neutron therapy trials

were necessary experiments, which would have to be repeated today if they had not already been done. It is now difficult to prove how these early developments at physics laboratories had an influence on later developments in hospitals. Applied research activities at scientific laboratories are often too far ahead in time to have a direct impact on contemporary life. But the methods are often rediscovered in a slightly modified form at a later time. Wide acceptance is achieved only at the moment of their availability on a large scale through industrial production. Many professionals working in hospitals are now convinced that thanks to this recent technological progress in beam delivery, photons will soon be competitive enough to beat proton therapy. Can IMRT make proton therapy obsolete? This is probably the new crucial question for all centres investigating the potential of proton therapy.

6 PROTON THERAPY

6.1 The Bragg peak

The advantage of proton therapy is given by its superior physical selectivity. Protons have a well-defined penetration range in biological tissue and they deposit the maximum of their energy in the region where they stop. This gives rise to the so-called Bragg peak. Figure 3 shows the dose deposition of a monoenergetic proton beam as a function of the depth. This must be compared with clinical photons, which have a characteristic exponential fall-off of the dose. Protons also offer the possibility to localize the dose as a function of the depth and not only in the lateral direction. Compared to photons and using similar techniques, one can expect to achieve with protons a general reduction of the integral dose outside the target volume by a factor of 2 or 3 (additional dose sparing).

Another potential advantage is given by the electric charge of the particle beams, which creates the possibility to scan the beam by using magnetic scanning techniques (this is probably a more practical alternative to multi-leaf collimators).

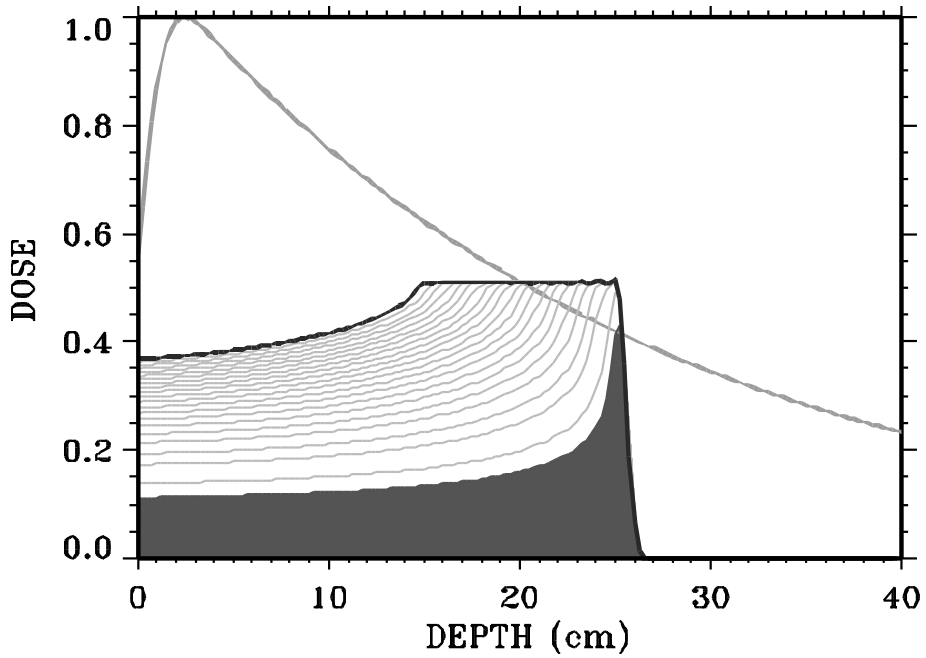


Fig. 3: Comparison of the depth dose profiles of proton and photon beams. The photon dose falls off exponentially with depth. A mono-energetic proton beam is characterized by the presence of the Bragg peak in the region where the protons stop. Through the superposition of many proton beams of different residual range it is possible to deposit a homogeneous dose Spread-Out Bragg Peak (SOBP) in the regions of the tumour (in this case from 15 to 25 cm depth). One can recognize from the picture the potential of dose sparing of the protons in the entrance and exit regions of the beam.

Protons are expected to produce superior results for the treatment of large tumours of complex shape, where a significant reduction of the dose outside of the target volume is clinically desirable. On the other hand, the disadvantage of proton therapy is given by the large size of the accelerator and of the beam lines needed for the transport of the beam. The maximum energy of the beam is chosen at around 230–250 MeV (corresponding to a penetration depth of 33–37 cm in water). Owing to the high magnetic rigidity of the beam the minimal bending radius that can be applied to the beam is around 1.3–1.5 m using conventional magnets close to the saturation limit.

6.2 Beam delivery techniques for protons

For a more detailed discussion on beam delivery we refer to Ref. [12].

The main techniques are:

- **Passive scattering**

Figure 4 shows schematically the principles at the base of the beam delivery with passive scattering.

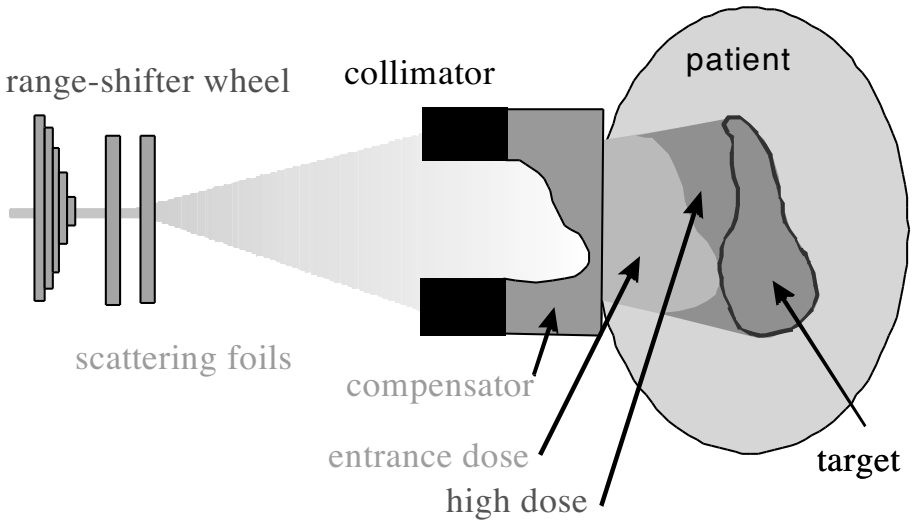


Fig. 4: Elements necessary for beam delivery using passive scattering techniques. Scatter foils are used to produce a homogeneous proton flux. A range shifter wheel is used to produce a SOBP of given dose homogeneity thickness (at a given depth). These hardware elements are selected as a function of the beam energy and target size. Individually shaped collimators are used to shape the dose laterally. A compensator shifts the edge of the dose distribution to conform to the distal side of the target volume. The fixed range of modulation implies an unnecessary deposition of dose at the 100% level outside the target volume.

This technique is the traditional established beam delivery method. The proton beam is scattered by material in the beam ahead of the patient in order to produce a homogeneous flux of protons in the solid angle used for the irradiation (in close analogy with the point source of radiation of conventional therapy). The dose is then shaped in the lateral direction using individual collimators.

A fast spinning wheel of variable thickness (range shifter wheel) introduces a variable amount of absorbing material in the beam as a function of time. The resulting modulation of the proton range is chosen so as to produce a homogeneous region of the dose in depth (SOBP). The SOBP 'thickness' must be chosen to cover the whole target volume. With protons one can deposit a homogeneous dose distribution from just a single beam direction. This is already an improvement compared to photon therapy, where the dose homogeneity can be achieved only by using several beam directions.

An individual compensator bolus can optionally be added to this set-up. The variable thickness of the bolus is machined to shift the distal edge of the dose field to conform more closely to the deepest side

of the target volume. This improves the degree of conformity of the dose distribution. All the necessary hardware must be adapted and in part created individually for each single field.

This method produces by default a homogeneous dose field of fixed SOBP thickness in depth (fixed range modulation). Without additional devices the passive scattering method is not capable of delivering IMRT methods. One could envisage implementing IMRT with protons by using an additional multi-leaf collimator. In this case one would not make use of the potential to modulate the dose in depth by changing dynamically the proton range (variable range modulation).

The most elegant, flexible, and efficient method for providing inverse planning with protons (IMPT) is by magnetic beam scanning.

- **Beam scanning**

  Figure 5 shows the basic principles of beam scanning technology.

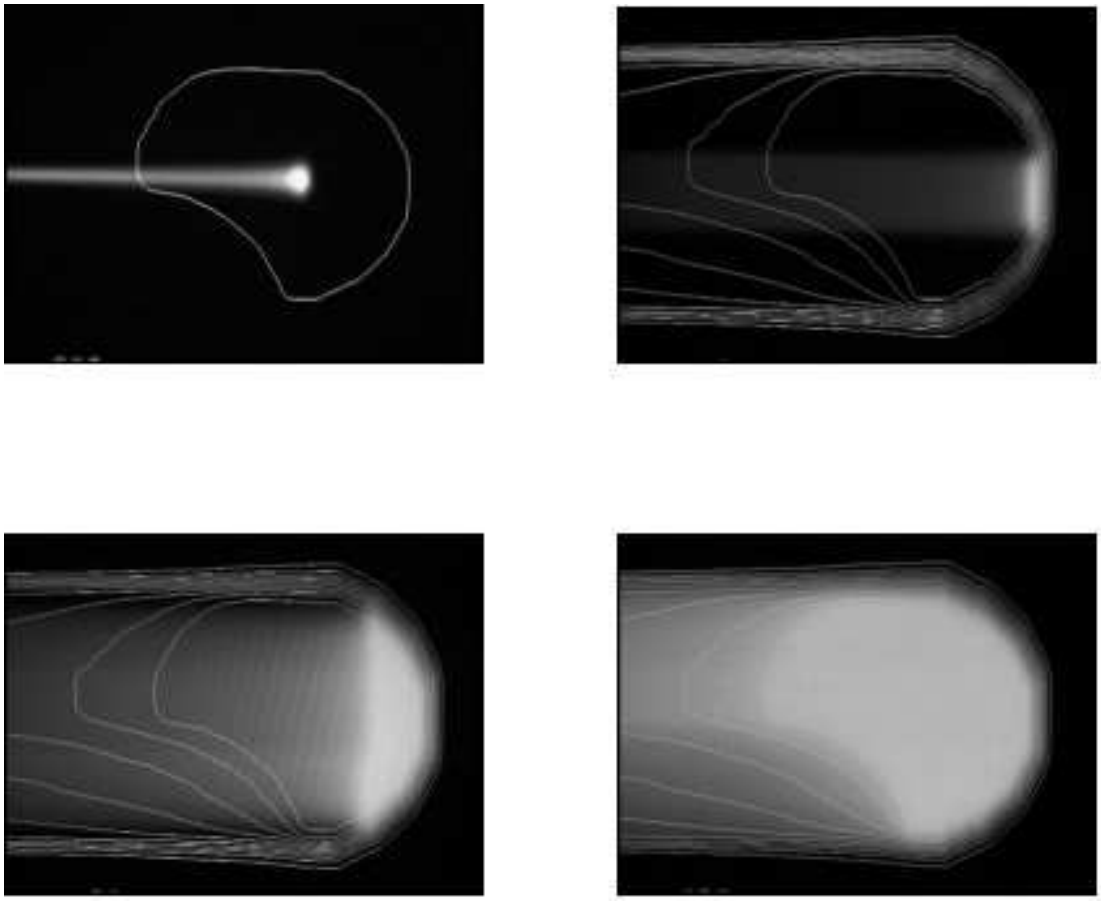


Fig. 5: Basic principles used for beam scanning with protons. Through the delivery of individual proton pencil beams one can shape the distribution of the dose in three dimensions at will, directly under computer control.

In this case the proton pencil beam coming from the accelerator is delivered directly to the patient. Individual pencil beams are added under computer control to provide an individually shaped dose distribution with maximal conformity with the target volume (using variable modulation of the range).

In the lateral direction the beam is usually scanned by magnetic deflection in the beam line ahead of the patient. The modulation in depth is achieved by changing the range of the protons dynamically. A high conformity is achieved by changing the dosage and the position of each pencil beam individually under computer control.

At present the proton facility of the PSI is the only one capable of delivering proton therapy using a dynamic beam scanning technique [13]. The design of the PSI scanning system is based on maximal simplicity without compromising for the flexibility of the dose delivery.

6.3 IMPT: an example

Figure 6 shows, as an example, the potential use of the spot scanning technique for delivering Intensity Modulated Therapy with Protons (IMPT) (courtesy T. Lomax of PSI). With only four modulated fields one can deliver a highly conformal dose to the primary target and a reduced dose to the affected lymph nodes (the secondary target) with a maximal sparing of the organs at risk (brain stem and parotid glands). All this can be designed and delivered simply under computer control, without the need for specific hardware.

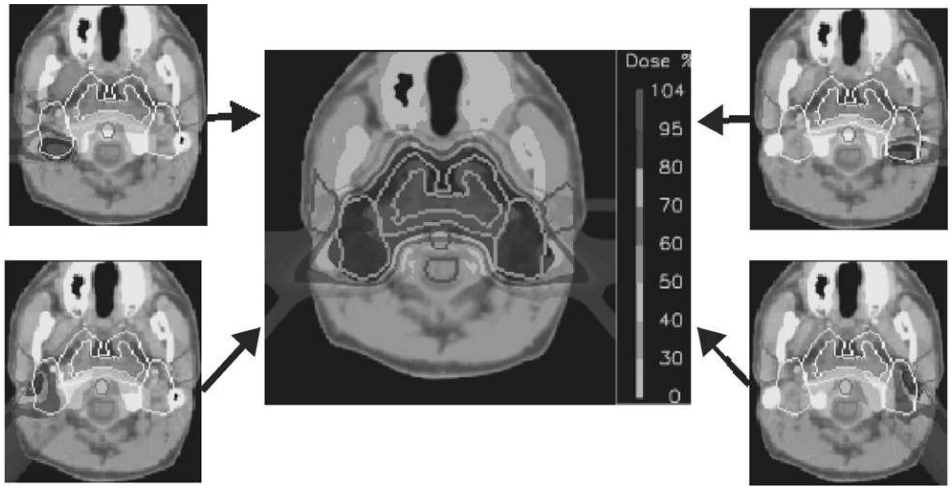


Fig. 6: Example of intensity modulated therapy with protons. A high degree of conformity is achieved using a low number of dose fields. The advantage compared with photons is the general reduction of dose burden outside of the target volume (courtesy of T. Lomax).

6.4 Proton gantries

In order to be able to apply the beam from many directions onto the supine patient, it is desirable to apply the dose using a rotating beam line. This implies the use of a proton gantry. All new dedicated facilities today are designed with gantries.

The proton gantries constructed up to now include that at Loma Linda University, the first place in the world where a proton gantry was developed. The facility started operation in 1991. The second design is the compact gantry at PSI (patient treatments started in 1996). A third type of gantry was recently built at Kashiwa (Japan), where operation started in 1999. A very similar gantry design is ready to go into operation in Boston (USA) and a fourth gantry type is being assembled in Tsukuba (Japan).

Typical gantry designs can be assigned into the following categories:

- **Cork-screw gantry (example: Loma Linda University, USA) [14]**

This gantry is dedicated to beam delivery with passive scattering. The diameter of the rotating structure is 12 m. The double bending of the beam line (first into and then inside a flat rotating structure) with a cork-screw design makes the gantry more compact longitudinally at the expense of a longer beam line.

- **Compact eccentric gantry (PSI, Switzerland) [12],[15]**

This gantry in dedicated to (Cartesian) beam scanning and IMPT. Part of the beam scanning is performed before bending the beam towards the patient. The eccentric mounting of the patient table reduces further the gantry diameter which, at only 4 m, is the most compact of all the known designs. Figure 7 shows the PSI proton gantry.

Fig. 7: Photograph of the PSI proton treatment room with the head of the compact gantry dedicated to proton beam scanning

An improved version of this design with the patient table mounted at the isocentre is anticipated for use in hospitals.

- **Barrel gantries (examples: Kashiwa, Japan, and Boston, USA)**

Beam delivery by passive scattering is necessarily performed after bending the beam towards the patient. The space needed for spreading the beam must be added to the gantry radius. The gantry diameter is therefore at least 10 m. A large gantry radius is the price of maintaining the possibility to use the passive scattering technique on the gantry. A large throw gantry can be used for beam scanning. The NPTC crew in Boston is planning to implement beam scanning on their gantry at a later stage.

6.5 Proton facility design

The investments for proton therapy equipment are high. For this reason all dedicated proton facilities are designed with an accelerator serving several treatment rooms simultaneously. The first dedicated proton therapy facility in a hospital was installed at the Loma Linda University Hospital in the USA.

Figure 8 shows, as an example, the layout of the new facility at the Massachusetts General Hospital in Boston, expected to become operational this year.

6.6 Choice of accelerator

The energy of the protons is in principle low enough to allow the use of several types of accelerator. The accelerator types presently used in dedicated facilities are the following:

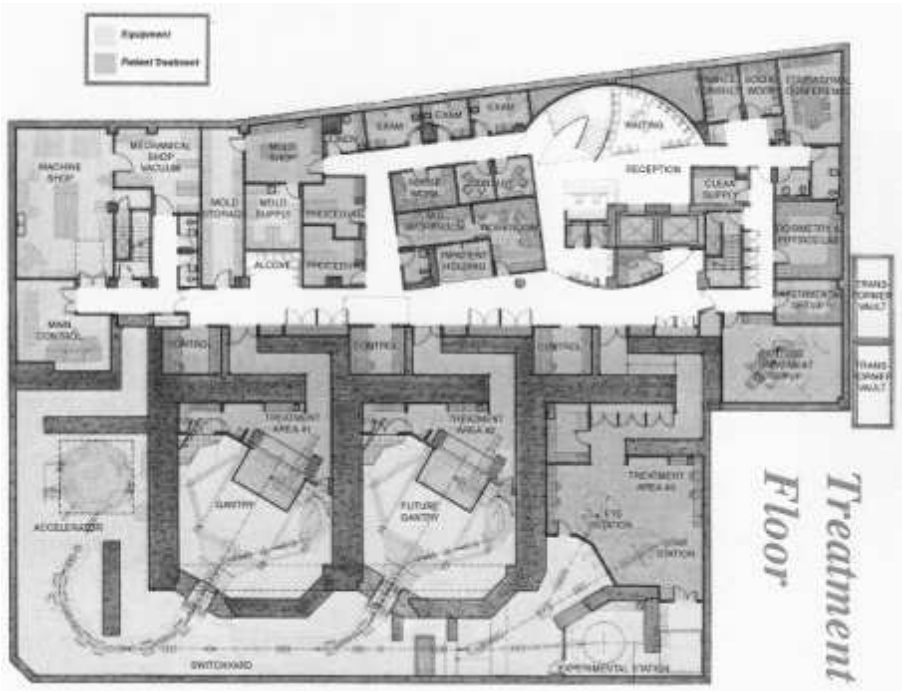


Fig. 8: Layout of the NPTC proton therapy facility at the Massachusetts General Hospital in Boston USA (courtesy of J. Flanz)

- Cyclotrons (examples: Kashiwa and Boston)
- Synchrotrons (examples: Loma Linda University and Tsukuba University)
- Linacs (a proton linac with a sufficiently high repetition rate of the pulse frequency to be used for beam scanning is under development in Rome, Italy) [16]
- Synchrocyclotrons (Orsay, Uppsala)

At present, the major competition concerning the choice of accelerator is between cyclotrons and synchrotrons. The main advantage of a proton synchrotron is the variable choice of the beam energy extracted from the machine. The main disadvantage is the pulsed nature of the beam, which is not well suited to beam scanning. One can, however, overcome parts of the problem by providing a stable slow extraction of the beam.

The advantage of the cyclotron is the high duty factor of beam (DC beam)— which is well-suited beam scanning, the high proton current, and the inherent stability of the beam (including a possible easy control of the beam intensity at the ion source). The active modulation of the intensity at the ion source could also be used as an aid for the scanning of the beam. The main disadvantage is the fixed energy, which requires the use of a degrader followed by an analysing beam line. This implies a higher activation of components in the initial region of the facility.

There have been several other propositions:

- **H^- synchrotrons [17],[18]**

The advantage of this approach is the expected easy extraction of the beam from the ring by foil stripping. The idea is to provide a separated extraction for each (short) beam line used for feeding the protons into several treatment rooms, which are ordered radially around the synchrotron ring. The radius of the ring is quite large (the magnetic bending must be maintained low to avoid magnetic stripping of the negative ions).

- **Separated sector cyclotrons [19]**

By having a large separation between the sectors using superconducting magnets, one could provide variable beam energy extracted from different orbits. This would combine the advantages of a cyclotron with the advantage of a variable energy machine.

- **Superconducting cyclotrons [20]**

Possible advantages of superconducting cyclotrons are the reduction in the size, the lower power consumption, and possibly the better efficiency of extraction. In the same context there have been propositions to rotate a very compact cyclotron on a gantry. The idea has already been successful implemented, but at a lower beam energy, in the Detroit Facility for Neutron Therapy [21]. Whether this is a good solution for proton therapy is a matter for debate.

- **Variable energy cyclotron using H_2^+ sources?**

The costs for the accelerator are similar to the costs for the beam line switchyard and to the cost of one single treatment room. The accelerator is an important component in the economics of a proton facility. The reliability of the system is a major factor in the specification list of the system.

6.7 List of proton and ion facilities in the world

A regularly updated list of links to the existing charged particle therapy centres of the world can be found at our home page [15].

Table 1 shows the latest update of the statistics of patients treated with charged particle beams (taken from *Particles*, the journal edited at the Harvard cyclotron on behalf of the charged particle therapy community) [22]. This list includes patients treated with light ions and pions. Ion therapy will be discussed in Section 7. Table 2 gives a list of the centres proposing new dedicated facilities for proton or ion therapy.

6.8 Indications for proton therapy

One of the most important indications for proton therapy is the treatment of ocular melanomas. This method was developed at Harvard in Boston. The treatment of ocular melanomas was introduced in Europe for PSI in 1985 (it was actually the great success of this programme and the previous experience with pions that made it possible for PSI to start the construction of a new gantry project for treatment of deep-seated tumours). Up to now, more than 3000 patients have been treated at PSI for ocular melanomas. In Europe the method used for ocular treatment is now also applied at other places: UK (Clatterbridge), Sweden (Uppsala), France (Nice and Orsay), and Germany (Berlin).

The best known example of superior results with protons on deep-seated tumours was demonstrated at the Harvard cyclotron with treatments of tumours close to the base of the skull (chordomas and chondrosarcomas).

Well-documented clinical results were obtained at Tsukuba on liver cancer. Loma Linda is today the only dedicated hospital-based facility that has been operational for many years. LLUMC has recently shown the capability of treating about 1000 patients per year. The majority of these treatments are prostate cancers. This is becoming a true speciality of this centre. Loma Linda has clearly shown that proton therapy is feasible on the basis of commercial and scientific criteria.

There are many other possible therapies waiting to be explored. Proton therapy should be applied in all situations where conventional therapy is encountering difficulties and where a reduction in treatment morbidity is desirable (for example pediatric tumours). PSI is trying to contribute to the field with the development of compact gantries dedicated to beam scanning techniques capable of delivering proton therapy for 'inverse proton planning' in order to compete with the most advanced photon beam delivery methods in the future.

Table 1: World-wide charged particle patient totals January 2000 (from Ref. [22])

Institution	**Country**	**Type**	**Date first Rx**	**Date last Rx**	**Recent patient total**	**Date of total**
Berkeley 184	CA, USA	p	1954	1957	30	
Berkeley	CA, USA	He	1957	1992	2054	June 1991
Uppsala	Sweden	p	1957	1976	73	
Harvard	MA, USA	p	1961		8372	Dec. 1999
Dubna	Russia	p	1967	1974	84	
Moscow	Russia	p	1969		3100	Dec. 1998
Los Alamos	NM, USA	π^-	1974	1982	230	
St. Petersburg	Russia	p	1975		1029	June 1998
Berkeley	CA, USA	Heavy ion	1975	1992	433	June 1991
Chiba	Japan	p	1979		96	Oct. 1996
TRIUMF	Canada	π^-	1979	1994	367	Dec. 1993
PSI (SIN)	Switzerland	π^-	1980	1993	503	
PMRC, Tsukuba	Japan	p	1983		629	July 1999
PSI (72 MeV)	Switzerland	p	1984		3014	Dec. 1999
Dubna	Russia	p	1987		43	Dec. 1999
Uppsala	Sweden	p	1989		215	Oct. 1999
Clatterbridge	UK	p	1989		960	Dec. 1999
Loma Linda	CA, USA	p	1990		4726	Dec. 1999
Louvain-la-Neuve	Belgium	p	1991	1993	21	
Nice	France	p	1991		1350	June 1999
Orsay	France	p	1991		1522	Sept. 1999
N.A.C.	South Africa	p	1993		341	Dec. 1999
MPRI	IN, USA	p	1993		34	Dec. 1999
UCSF CNL	CA, USA	p	1994		246	Dec. 1999
HIMAC, Chiba	Japan	Heavy ion	1994		473	Sept. 1998
TRIUMF	Canada	p	1995		55	Dec. 1999
PSI (200 MeV)	Switzerland	p	1996		41	Dec. 1999
GSI Darmstadt	Germany	Heavy ion	1997		46	Dec. 1999
Berlin	Germany	p	1998		105	Dec. 1999
NCC, Kashiwa	Japan	p	1998		18	Dec. 1999
					1100	pions
					3006	ions
					26104	protons
				TOTAL	30210	All particles

Table 2: Proposed new facilities for proton and ion beam therapy, January 2000 (from Ref. [22])

Institution	**Country**	**Type**	**1st Rx**	**Comments**
INFN-LNS, Catania	Italy	p	2000	70 MeV; 1 room, fixed horiz. beam.
NPTC (Harvard)	MA, USA	p	2000	At MGH; 230 MeV cyclotron; 2 gantries + 2 horiz.
Hyogo	Japan	p, ion	2001	2 gantries; 2 horiz; 1 vert; one 45°; under construction.
NAC, Faure	South Africa	p	2001	New treatment room with beam line 30° off vertical.
Tsukuba	Japan	p	2001	270 MeV; 2 gantries; 2 fixed (research); under construction.
CGMH, Northern Taiwan	Taiwan	p	2001?	250 MeV synchrotron/230 MeV cyclotron; 3 gantries, 1 fixed.
Wakasa Bay	Japan		2002	Multipurpose accelerator; building completed mid. 1998.
Bratislava	Slovakia	p, ion	2003	72 MeV cyclotron; p; ions; +BNCT, isotope prod.
IMP, Lanzhou	China	C-Ar ion	2003	C-ion from 100 MeV/u at HIRFL expand to 900 MeV/u at CSR; clin. treat.; biol. research; no gantry; shifted patients.
Shizuoka Cancer Center	Japan		2002?	Synchrotron 230? MeV; 2 gantries; 1 horiz; funded.
Erlangen	Germany	p	2002?	4 treatment rooms, some with gantries.
CNAO, Milan & Pavia	Italy	p, ion	2004?	Synchrotron; 2 gantries; 1 fixed beam room; 1 exp. room.
Heidelberg	Germany	p, ion	2005?	
AUSTRON	Austria	p, ion	–	2 p gantry; 1 ion gantry; 1 fixed p; 1 fixed ion; 1 exp room.
Beijing	China	p	–	250 MeV synchrotron.
Central Italy	Italy	p	–	Cyclotron; 1 gantry; 1 fixed.
Clatterbridge	UK	p	–	Upgrade using booster linear accelerator to 200 MeV?
TOP project ISS Rome	Italy	p	–	70 MeV linac; expand to 200 MeV?
3 projects in Moscow	Russia	p	–	Including 320 MeV; compact, probably no gantry.
Krakow	Poland	p	–	60 MeV proton beam.
Proton Development NA Inc.	IL, USA	p	–	300 MeV protons; therapy & lithography.
PTCA, Tenet HealthSystem	USA	p	–	Several systems throughout the USA.

Close to 10 hospital-based proton therapy facilities are approved for construction or already available in the USA and Japan. There are several propositions for dedicated facilities in Europe but none seems to be starting at this time.

7 ION THERAPY

7.1 Rationale for light ion therapy

As stated in Section 3, the main advantage of ion therapy is expected from the use of the high LET character of the beam. High LET should have a positive impact on the treatment of very radio-resistant tumours and could be used to eliminate potentially surviving anoxic cells (by virtue of the low OER).

Concerning the physical selectivity, ions have excellent properties, well suited for precision therapy. This is due to the large inertial mass of the ions: multiple Coulomb scattering and range straggling effects are very much reduced compared to protons. A factor adding confusion, however, is the problem of the fractionation of the projectiles. Together with the strong variation of RBE with depth, this makes the dosimetric characterization of ion beams a difficult task.

A major practical disadvantage of this therapy is the higher magnetic rigidity of the beam by a factor of 3 compared to protons (energies up to around 400 MeV per nucleon are necessary for the treatment of deep-seated tumours). This makes the accelerator and the beam lines more bulky and therefore ion therapy more expensive than proton therapy. Owing to the high magnetic rigidity of the beam only synchrotrons are taken into consideration for the acceleration of clinical ion beams.

High LET is, on the other hand, not expected to do better in every clinical situation. What matters in the end are the differential effects between cancer cells and healthy tissue. The cell repair capability of the healthy cells and the amplification of the therapeutic ratio with the fraction number are known to be less pronounced with high LET than with low LET (high LET is for this reason usually applied with a lower number of fractions). Eventually one expects more late reactions. The use of high LET does not automatically imply, as often stated, a generic 'radiobiological advantage'. There could be disadvantages as well. Whether the use of high LET radiation is indicated or not depends on how much the therapeutic ratio improves compared with conventional therapy for a given clinical situation (type of tumour, organs at risk, oxygen content of the tumour). Ion therapy represents new radiobiological ground and the use of high LET is still in a very experimental stage. This is the risk but also the big challenge of this new type of therapy.

The conditions under which high LET radiation is expected to bring superior results under are for the treatment of slowly growing radio-resistant tumours. This concerns a limited but significant number of patients. According to estimates of the German Cancer Research Institute in Heidelberg (DKFZ) about 8000–11000 patients could benefit from proton and ion therapy in Germany alone. According to these figures this would account for the need of typically one or two ion machines and several proton machines in each large country of the Western world.

7.2 Ion beam facilities

Most of the pioneering work on ion therapy was carried out at the Bevalac at Berkeley, USA. Unfortunately, in 1992, for financial reasons, the Department of Energy withdrew its support for the project (mainly because of the high costs of running an old physics accelerator).

In Chiba, Japan, the Institute of Radiological Sciences took over the pioneering role by creating a dedicated facility for ion therapy. This rather expensive facility has two synchrotrons on top of each other. It has been in operation since 1994. The beam is delivered with the passive scattering technique through several fixed horizontal and vertical beam lines.

In Germany ion therapy was started at GSI in 1997. The most important feature of the beam delivery system developed at GSI is the very advanced beam scanning technique [23]. The raster scanning

method of GSI is done on a dedicated fixed horizontal beam line. Transverse scanning is performed by a double magnetic deflection of the beam. The shaping of the dose is achieved by adapting the speed of scanning to the beam intensity delivered by the synchrotron during a slowly extracted beam spill. The energy of the beam is varied on a pulse by pulse basis by changing the settings of the whole synchrotron after completion of the painting of a monoenergetic dose layer. The dose is shaped in all three dimensions through the delivery of many differently shaped iso-energy layers. This is probably the most advanced example of the integration of the accelerator into the beam delivery to the patient. The GSI (with ions) and the PSI (with protons) are the only places in the world where charged particle therapy is applied dynamically using scanning beams. The use of the raster scanning technology on a dedicated ion gantry has been proposed for a dedicated hospital ion facility at the DKFZ in Heidelberg.

In Japan another ion facility (dedicated to both proton and ion therapy) is being assembled near Kobe (the Hyogo facility).

In Europe there are several propositions for ion therapy: the Tera project in Italy, the Austron project in Austria, and the ion-proton facility in Heidelberg. None of these have been approved at the time of writing.

8 CONCLUSIONS

Radiotherapy represents an important instrument in the fight against cancer and is a field of continuous evolution. In the field of conventional therapy we expect to be able to observe, in the near future, significant progress using very advanced beam delivery techniques with photons (IMRT). Similar more advanced developments are also possible with charged particle beams (IMPT). The combination of the new technologies with the physical advantages of charged particles is expected to produce superior results. In this future scenario new technical developments are necessary for both sides to remain competitive. This is why we believe that beam scanning will soon become the generic beam delivery method for proton and ion beam therapy. This strategy must be considered when designing accelerator and beam delivery systems for future planned dedicated facilities.

Radiotherapy with charged particles is the field that can probably profit most from accelerator technology. In the near future the accelerator will no longer be considered a separate entity for the delivery of beam, but will be more and more directly involved with the task of delivering the dose safely, reliably, and precisely to the patient.

Acknowledgements

The author would like to thank Tony Lomax and Uwe Schneider for providing essential material for this presentation, and T. Bhringer for his careful reading of the manuscript.

REFERENCES

[1] W.H. Scharf, *Biomedical Particle Accelerators* (American Institute of Physics Press, New York, 1994).

[2] W.H. Scharf, O.A. Chomicki, Medical Accelerators in Radiotherapy, *Phys. Med.* **18** (1995).

[3] Krebsatlas für Deutschland, DKFZ, Heidelberg, http://www.dkfz-heidelberg.de /epi/Home_d/Programm/AG/Praevent/Krebshom/main/deutsch/frame.htm

[4] E. Hall, *Radiobiology for the radiologist*, Fourth Edition (Lippincott, Philadelphia, 1993).

[5] U. Amaldi and B. Larsson (Eds.), Hadrontherapy in Oncology, *Excerpta Medica*, ICS 1077 (Elsevier, Amsterdam, 1994).

[6] U. Linz (Ed.), *Ion Beams in Tumor Therapy* (Chapman & Hall GMBH, D–69469 Weinheim, 1995).

[7] B. Larsson and J. Crawford (Eds.), Advances in neutron capture therapy, *Excerpta Medica*, ICS 1132 (Elsevier, Amsterdam, 1997).

[8] A. Wambersie and R.A. Gahbauer, Medical Applications of Electron Linacs, CERN 96–02.

[9] T. Bortfeld et al., Realization and verification of three-dimensional conformal therapy with modulated fields, *Int. J. Radiat. Oncol. Biol. Phys.* **30** (1994) 899.

[10] T.R. Mackie et al., Tomotherapy: a new concept for the delivery of dynamic conformal radiotherapy, *Med. Phys.* **20** (1993) 1709.

[11] E. Pedroni, Therapy planning system for the Sin-Pion therapy facility, *Treatment Planning for External Beam Therapy with Neutrons*, (Eds.) G. Burger, Urban & Schwarzenberg, 1981 p. 60–69.

[12] E. Pedroni, p. 434 of Ref. [5] and E. Pedroni, p. 213 of Ref. [6].

[13] E. Pedroni et al., The 200 MeV proton therapy project at the Paul Scherrer Institute: conceptual design and practical realization, *Med. Phys.* **22** (1995) 37.

[14] J.M. Slater et al., *Int. J. Radiat. Oncol. Biol. Phys.* **8** (1988) 1499.

[15] The PSI medical division home page: http://www1.psi.ch/www_asm_hn/asm_home_page. html

[16] E. Picardi et al., Progetto del TOP Linac, ENEA, Centro Ricerche Frascati, Rome. RT/INN/97/17.

[17] Private communication: R.L. Martin, ACCTEK Associates, 901 S. Kensington, LaGrange, IL 60525, USA.

[18] Private communication: V.S. Khoroshkov, Medical facility project development, ITEP, Moscow (1990).

[19] H. Jungwirth, A superconducting Ring-cyclotron for proton therapy, Proc. 14th Int. Conf. on Cyclotrons and their Applications, Cape Town (South Africa), 1995.

[20] H. Blosser et al., National Superconducting Cyclotron Laboratory, Michigan State University, Report MSUCL–760.

[21] H. Blosser, Medical Cyclotrons, *Physics Today*, p. 70 (1993).

[22] Particles, The newsletter of the Proton Therapy Co-operative Group (PTCOG) Editor: J. Sisterson, Harvard Cylotron, Cambridge Ma, USA. Online edition:
http://neurosurgery.mgh.harvard.edu/hcl/ptles.htm

[23] T. Haberer, Magnetic scanning system for heavy ion therapy, *Nucl. Instrum. Methods* **A330** (1993) 296.

STOCHASTIC COOLING AND RELATED RF COMPONENTS

F. Nolden
Gesellschaft für Schwerionenforschung, Planckstraße 1, D-64291 Darmstadt, Germany

Abstract
This paper describes some of the pick-up and kicker structures that are currently in use for stochastic cooling. The emphasis is on the general physical properties of the devices.

1 QUALITATIVE INTRODUCTION TO STOCHASTIC COOLING

Stochastic cooling is a powerful method of decreasing the phase space volume of coasting beams. Invented by Simon van der Meer in 1969, the first successful cooling experiments were performed in the 1970s at CERN. Important physics experiments such as the detection of the W and Z bosons at CERN and the top quark at Fermilab depended crucially on the stochastic cooling of antiproton beams which were needed in colliders like the SPS and the Tevatron. An extensive bibliography can be found in Dieter Möhl's lecture at the 1993 CERN accelerator school [1]. Nowadays, the scope of stochastic cooling has been extended to different particles and different laboratories, such as proton cooling at COSY [2] and heavy ions at GSI (Darmstadt) [3]. A radioactive beam facility with stochastic cooling is planned at RIKEN in Tokyo [4].

The principal idea of stochastic cooling is to make use of the information inherent in the fluctuations of a coasting beam, in order to derive a correction signal which is applied to the beam very shortly after detection. These fluctuations are unavoidable because beams consist of a finite number of particles. The signal is detected by pick-up electrodes, processed in order to extract the desired information (momentum deviation or transverse emittance), amplified and transmitted to kicker electrodes, where the properties of the beam are changed in order to decrease the phase space volume. The kick is applied in the same revolution as the signal is detected. The principal layout of a stochastic cooling system is shown in Fig. 1. For an introduction to the physics of stochastic cooling, the reader should refer again to Ref. [1].

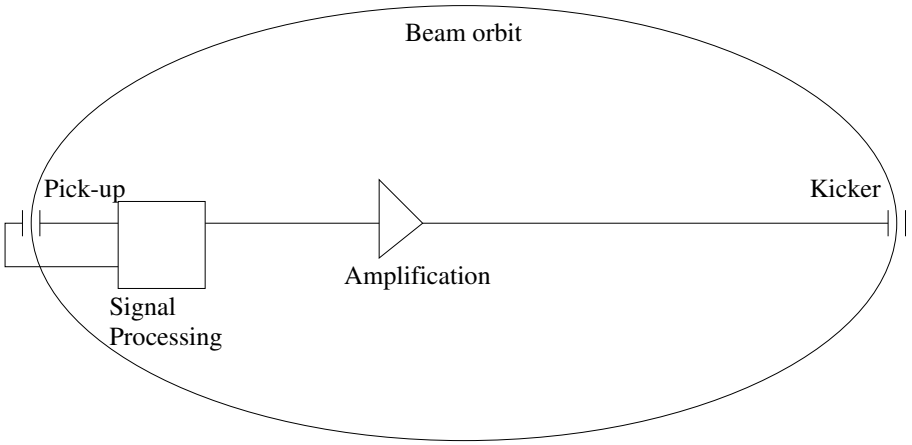


Fig. 1: Principal layout of a stochastic cooling system.

An important figure of merit for the quality of the signal as well as the efficiency of the cooling system is the temporal resolution of signal detection. The higher the resolution, the smaller are the parts of the beam to be observed simultaneously and the more informative is the fluctuating signal. Because of the uncertainty relation between the time and the frequency domains, a high temporal resolution is

equivalent to a large frequency bandwidth W. Indeed, it can be shown rather easily that the optimum cooling rate $1/\tau$ of a stochastic cooling system is $2W/N$, where N is the number of particles in the beam.

In order to get reasonable cooling times, practical stochastic cooling systems are usually run with bandwidths in the range 0.5–8 GHz. They are therefore based on microwave technology. Inside and outside the beam chamber, careful impedance matching is a necessary prerequisite for effective cooling. Fast signal transmission is performed either through air-filled coaxial lines, or even, if long transmission lengths are used, by optical systems [5]. Large signal amplifications > 100 dB are common and broad-band power amplifiers with good linearity needed. Travelling wave tubes are used at Fermilab, whereas amplifiers based on gallium arsenide power FET's developed for the first time at CERN [6] are common in the European laboratories. The very large bandwidths > 1 GHz are realized using several sub-bands.

Another important issue is the signal-to-noise ratio at the input port of the first (low noise) preamplifier as well as the noise figure of the preamp. The signal is proportional to the sensitivity of the pick-up structure and the noise is proportional to the effective temperature of the pick-up-preamplifier system. Expensive low-temperature systems have been built for antiproton and proton cooling systems. Even for heavy ions with their intrinsically large Schottky signals (see below) thermal cooling is being considered at GSI for second-generation machines, where very short cooling times are envisaged for rare secondary nuclear species with short half-lives.

The purpose of this paper is to present a few of the common pick-up and kicker devices and, with more emphasis, to explain some general physical relationships. It will be shown that reciprocity enables one to consider pick-up and kicker structures using a common formalism based on the notion of sensitivity. Sensitivity will be defined for the longitudinal response and this treatment will be generalized to betatron motion.

The simplest way to define sensitivity is to look at the fields excited by a kicker and calculate the integrated longitudinal field a particle experiences during passage. Later, we will show how the reciprocity theorem (or rather, the assumption of reciprocity) can be used to describe the signal induced by a particle inside a pick-up structure which is built identically. In order to calculate transverse kicks, we will apply the Panofsky–Wenzel theorem.

For a different view of the subject, the reader is referred to related lectures by C.S. Taylor [7], F. Caspers [8], G. Lambertson [9], or J. Bisognano and C. Leeman [10]. Many of the ideas presented here originate from these texts.

2 KICKERS

2.1 Sensitivity

It will be useful in the following chapters to perform some of the calculations with the vector potential $\vec{A}$, because it turns out to be very straightforward to formulate the Panofsky–Wenzel theorem with $\vec{A}$. The electric RF field is $\vec{E} = -\partial\vec{A}/\partial t$. We shall use the coordinates s, x, y for the longitudinal, horizontal, and vertical directions, and neglect effects of orbit curvature. Indeed, most of the pick-ups and kickers are installed in straight sections (an exception is the system at the ESR storage ring at GSI). Let us assume short structures where the variations of x and y are negligible during one passage. The signal at the entrance port of the device, located at s_0, can be described by a voltage $V(t)$. Inside the kicker structure there is an electromagnetic wave propagating with (positive) signal velocity v_{S}. For large particle velocities the wave is usually chosen to be counter-propagating. It gives rise to a longitudinal field component

$$A_s(s,x,y,t) = a(s,x,y)V(t \pm \frac{s-s_0}{v_{\mathrm{S}}}) \,, \tag{1}$$

where the plus sign is used for co-propagating and the minus sign for counter-propagating waves. If the velocity of the particle is v, then the integrated field seen by the particle becomes

$$\int ds A_s(s,x,y,t+\frac{s-s_0}{v}) = \int d\tau S(x,y,\tau)V(t-\tau) = S(x,y,t) * V(t) \ , \tag{2}$$

where we have introduced the sensitivity

$$S(x,y,\tau) = v_{\rm eff} a(s_0 - v_{\rm eff}\tau, x, y) \ . \tag{3}$$

The effective velocity is defined by

$$\frac{1}{v_{\rm eff}} = \frac{1}{v} \mp \frac{1}{v_S} \ . \tag{4}$$

The last equation of Eq. (2) shows that the vector potential can be calculated from the sensitivity via a convolution. Hence, with the help of the Fourier convolution theorem Eq. (A.1.8), it is easy to switch to the frequency domain where products of the Fourier transforms of S and V can be used. The reader should verify the fact that $S(x,y,t)$ is a dimensionless number. It is easy to calculate from Eq. (2) the integrated electric field:

$$\int ds E_s(s,x,y,t+\frac{s-s_0}{v}) = -\frac{\partial S(x,y,t)}{\partial t} * V(t) \ . \tag{5}$$

We therefore define a longitudinal electric sensitivity via

$$S_{\|}(x,y,t) = -\frac{\partial S(x,y,t)}{\partial t} \ . \tag{6}$$

In order to descibe spectra, later on we will need the Fourier transform $S_{\|}(x,y,\Omega)$ of $S_{\|}(x,y,t)$, which is again a dimensionless quantity.

In order to calculate the transverse kicks, we define the transverse sensitivities

$$S_x(x,y,t) = \frac{\partial S(x,y,t)}{\partial x} \tag{7}$$

$$S_z(x,y,t) = \frac{\partial S(x,y,t)}{\partial z} \ . \tag{8}$$

2.2 Electrostatic approximation

For highly relativistic beams, the electric near fields of a beam particle are disk-shaped and the sensitivity can be calculated using electrostatic field models. For kickers, Lambertson [9] has shown that the transverse variation of the accelerating voltage $U = \int E_s ds$ is governed by the equation

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{1}{(\beta\gamma c)^2}\frac{\partial^2 U}{\partial t^2} = 0 \ , \tag{9}$$

which looks very much like a two-dimensional Poisson equation in the relativistic limit $v \to c$. Note the appearance of the factor $1/(\beta\gamma)$ which can be attributed to the relativistic shrinking of the Coulomb field. As a typical application of Eq. (9) we calculate the field of a stripline electrode.

2.3 Stripline or quarter-wave electrode

2.3.1 Geometry of the device

A stripline electrode consists of an interruption of the upper or lower part of the vacuum chamber of width w and length L (see Fig. 2). The longitudinal electric field only penetrates into the vacuum chamber at the two ends of the device. Let us use the electrostatic approximation to derive a common model of the sensitivity. In the following, we examine this structure in some detail in order to illustrate some typical problems and applications.

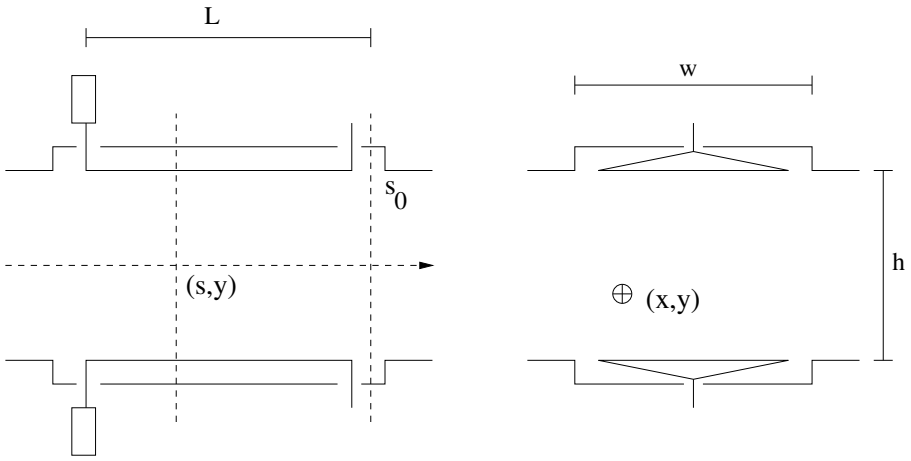


Fig. 2: Schematic sketch of a quarter-wave electrode.

2.3.2 Solution of the electrostatic problem

In order to solve the electrostatic problem, we use a simple two-dimensional model which is depicted on the left in Fig. 3. The plates of width w are symbolized by the thick lines around $x = 0$. They are at a potential V, whereas the rest of the x-axis and the parallel walls at $y = h$ are at zero potential. In order to compute the potential inside the vacuum chamber, we apply a conformal mapping. The reader who does not wish to go into the details should skip the following derivation and continue with the result in Eqs. (19) and (20). The prescription for the conformal mapping is adapted from Ref. [11]. We use complex variables $z = x + iy$ and a conformal mapping to the plane $\zeta = \xi + i\eta$ via

$$\zeta = g(z) = \exp\left(\frac{\pi z}{h}\right) . \tag{10}$$

A geometrical sketch of this mapping is shown in Fig. (3). The z-plane is shown on the left-hand side, the ζ-plane on the right-hand side. The shaded areas are potential-free. The (complex) potential in the

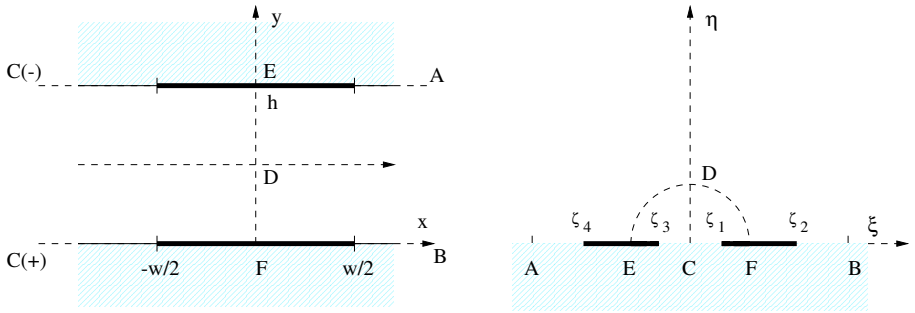


Fig. 3: Electrostatic model of quarter-wave electrode and conformal mapping.

ζ-plane is the superposition of the potentials from the upper and lower electrodes:

$$\Phi(\zeta) = \Phi_{\text{lower}}(\zeta) \pm \Phi_{\text{upper}}(\zeta) . \tag{11}$$

The positive sign is valid if the device is used as a longitudinal pick-up or kicker, i.e. if the upper and lower plates are operated in the sum mode. For operation in the difference mode (vertical pick-ups or kickers), the negative sign is valid. The constituent potentials are

$$\Phi_{\text{lower}}(\zeta) = -\frac{Vi}{\pi} \log \frac{\zeta - \zeta_1}{\zeta - \zeta_2} \tag{12}$$

$$\Phi_{\text{upper}}(\zeta) = -\frac{Vi}{\pi} \log \frac{\zeta - \zeta_4}{\zeta - \zeta_3} \tag{13}$$

with

$$\zeta_1 = g(-w/2) \tag{14}$$
$$\zeta_2 = g(w/2) \tag{15}$$
$$\zeta_3 = g(-w/2 + ih) \tag{16}$$
$$\zeta_4 = g(w/2 + ih)\ . \tag{17}$$

Because the mapping is conformal, the potentials in the z- and ζ-planes are equal, and the electric potential is

$$V(x,y) = \mathrm{Re}\left(\Phi[g(x+iy)]\right)\ . \tag{18}$$

The calculation of $V(x,y)$ from the complex functions involved is somewhat lengthy. The result is again the superposition of the two potentials

$$V(x,y) = V\sigma(x,y) = V\left(\sigma_{\mathrm{lower}}(x,y) \pm \sigma_{\mathrm{upper}}(x,y)\right)\ , \tag{19}$$

with

$$\sigma_{\mathrm{lower,upper}}(x,y) = \frac{1}{\pi}\arctan\left[\frac{\sin(\pi y/h)\sinh(\pi w/2h)}{\cosh(\pi x/h) \mp \cos(\pi y/h)\cosh(\pi w/2h)}\right]\ , \tag{20}$$

where the negative sign applies to the lower plate and the positive sign to the upper one. In the mid-plane, both potentials are equal:

$$\sigma_{\mathrm{lower}}(x,h/2) = \sigma_{\mathrm{upper}}(x,h/2) = \frac{1}{\pi}\arctan\left[\frac{\sinh(\pi w/2h)}{\cosh(\pi x/h)}\right]\ . \tag{21}$$

In the sum mode, the maximum voltage in the mid-plane is found at $x = 0$ and is $2V$ if $w \gg h$.

2.3.3 Time domain versus frequency domain picture

If an RF voltage is applied at the upstream end of the device, a wave with velocity v_{S} will propagate down the electrode and dissipate in the terminating resistor (see Fig. 2). The electric fields inside the vacuum chamber, due to a positive voltage between the electrode and the ground plane, point in opposite directions at the two ends. In order to obtain a net accelerating effect, the length of the electrode must be adapted to the beam and signal velocities in such a way that the beam particle arrives at both ends of the device at the right moment, when the wave amplitude has changed sign. We assume that the interaction time between the beam and the field at the end of the device is short compared to the characteristic time $2t_e = L/v_{\mathrm{eff}}$ (refer to Eq. (4) for the definition of v_{eff}). Then we can approximate the interaction by delta functions in the time domain and obtain for the sensitivity

$$S_{\parallel}(x,y,t) = \sigma(x,y)\left(-\delta(t) + \delta(t-2t_e)\right)\ . \tag{22}$$

The origin of time is defined such that the exciting voltage $V(t)$ is measured at the upper port of the electrode. In order to get the response in the frequency domain, one has to take the Fourier transform of Eq. (22). This is given by

$$S_{\parallel}(x,y,\Omega) = -2i\sigma(x,y)\sin(\Omega t_e)e^{-i\Omega t_e/2}\ . \tag{23}$$

Up to a phase the response is therefore essentially sinusoidal with a maximum at the frequency

$$f_{\mathrm{max}} = \frac{\Omega_{\mathrm{max}}}{2\pi} = \frac{1}{4t_e}\ . \tag{24}$$

If $v_{\mathrm{S}} = v = c$, then $f_{\mathrm{max}} = 4c/L$. The stripline electrode is therefore also called the quarter-wave electrode.

2.3.4 *Superelectrodes*

A superelectrode is generated by connecting two quarter-wave electrodes in series, with a delay $2t_e$ between the exit port of the first and the entrance port of the second one (the dashed line in Fig. 4).

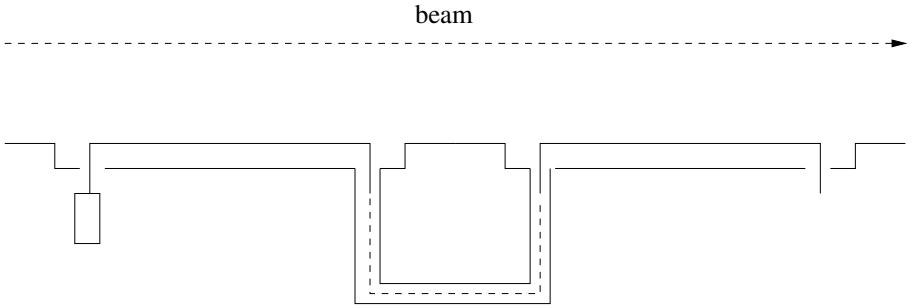


Fig. 4: Principal sketch of the lower half of a superelectrode kicker with delay line.

The sensitivity is then

$$S_{||}(x,y,t) = \sigma(x,y)\Big(-\delta(t) + \delta(t-2t_e) - \delta(t-4t_e) + \delta(t-6t_e)\Big) \tag{25}$$

with the frequency response

$$S_{||}(x,y,\Omega) = -4i\sigma(x,y)\sin(\Omega t_e)\cos(2\Omega t_e)e^{-3i\Omega t_e/2}\,. \tag{26}$$

At mid frequency, this is twice the value of a normal electrode, but the bandwidth is smaller, which can be attributed simply to the increased duration of the accelerating pulse train. Adding even more electrodes is even more obstructive to the bandwidth and has therefore never come into practice in stochastic cooling systems. Instead, power combination is used, as will be discussed later on.

2.3.5 *Finite β correction*

As can be seen from Eq. (9), the electrostatic approximation no longer holds if the beam velocity $v = \beta c$ is substantially smaller than c. As the electrostatic aproximation is intimately related to the relativistic shrinking of the Coulomb field, a generalization of the material presented above can be found in Ref. [12], where the point-like interactions in Eqs. (22) and (25) are replaced by the typical temporal variation of the Coulomb field at finite β,

$$\delta(t) \mapsto \frac{\left[1 + (t/\Delta t_{\rm pulse})^2\right]^{-3/2}}{\Delta t_{\rm pulse}}\,. \tag{27}$$

At a distance d from the particle, the finite length of the pulse is $\Delta t_{\rm pulse} = d/(\beta\gamma c)$. The Fourier transform of the right-hand side of Eq. (27) is

$$\tilde{f}(\Omega) = \Omega\Delta t_{\rm pulse}\, K_1(\Omega\Delta t_{\rm pulse})\,, \tag{28}$$

where K_1 is a modified Bessel function. In order to make the finite β correction effective, the idealized frequency response of Eqs. (23) and (26) must be multiplied by $\tilde{f}(\Omega)$. The effect of the correction is always a reduced sensitivity compared with the case $v = c$. Some estimates are given in Ref. [12]. This simple correction comes close to the results of sophisticated field theoretical models [13].

2.4 Planar loops

Planar loops [14] have been used extensively in the Fermilab stochastic cooling systems. They are fabricated completely with the microstrip line technique. The ground plane of the microstrip plate is cut in order to interrupt the image current of the particle [see Fig. 5]. On the opposite side, where the microstrips are situated, there is a tapered connection [the dashed trapezoid in Fig. 5] from the 50 Ω signal line. A hole is drilled through the dielectric in order to produce a galvanic connection between the end of the tapered microstrip and the 'electrode'. The electrode itself can be interpreted as a 100 Ω structure. On the other side of the 'electrode', a terminating resistor is soldered on the connection using standard microstrip technology. The structure is rather similar to the usual quarter-wave electrodes, but more easily constructed. However, it depends on the UHV requirements, whether a suitable dielectric can be found. For a mathematical analysis of planar loops, refer to Ref. [14].

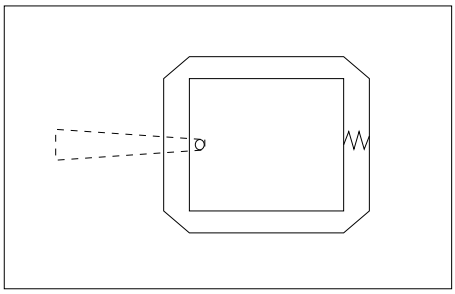


Fig. 5: Schematic view of a planar loop structure.

2.5 Slotted structures

In a slotted structure, a rectangular waveguide is coupled to the vacuum chamber by a regular array of slots (see Fig. 6). The effect of the slots is not only to couple the beam to the structure, but it also decreases the wave velocity inside the waveguide. The addition of an internal conductor transforms the waveguide into a coaxial line. Such devices were conceived by L. Faltin at CERN in the 1970's [15]. At Fermilab, slotted waveguides are used in the 4–8 GHz debuncher upgrade [16]. For a discussion of these and other structures, the reader is referred to C.S. Taylor's lecture on the same subject at the first CAS [7].

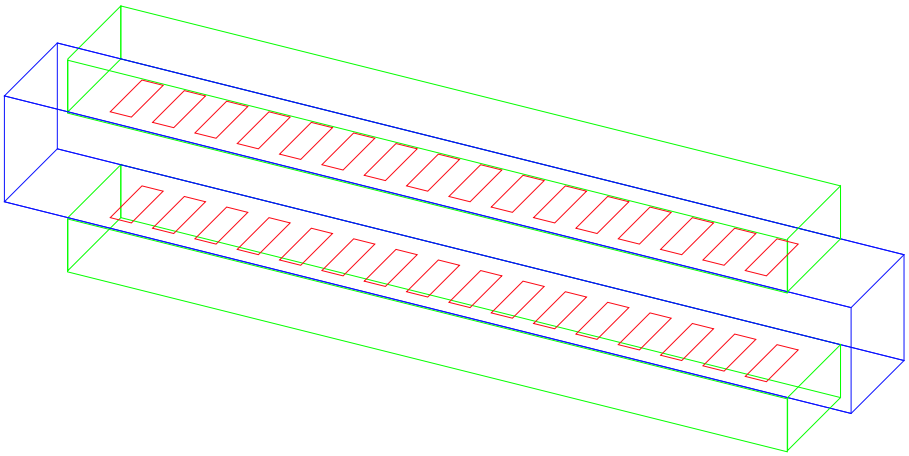


Fig. 6: Schematic view of a slotted waveguide. By inserting an inner conductor inside the waveguide, the principal layout of a Faltin type structure arises.

2.6 Signal combination

Whenever distributed pick-up or kicker structures such as striplines or planar loops are used in stochastic cooling systems, the sensitivity of one such element is seldom sufficient in practice. Therefore one uses several electrodes connected to standard microwave power combiners in order to increase the signal-to-noise ratio on the pick-up side, and the efficiency of turning power into accelerating voltage on the kicker side. A schematic view is shown in Fig. 7. The length of the connecting cables must be adapted

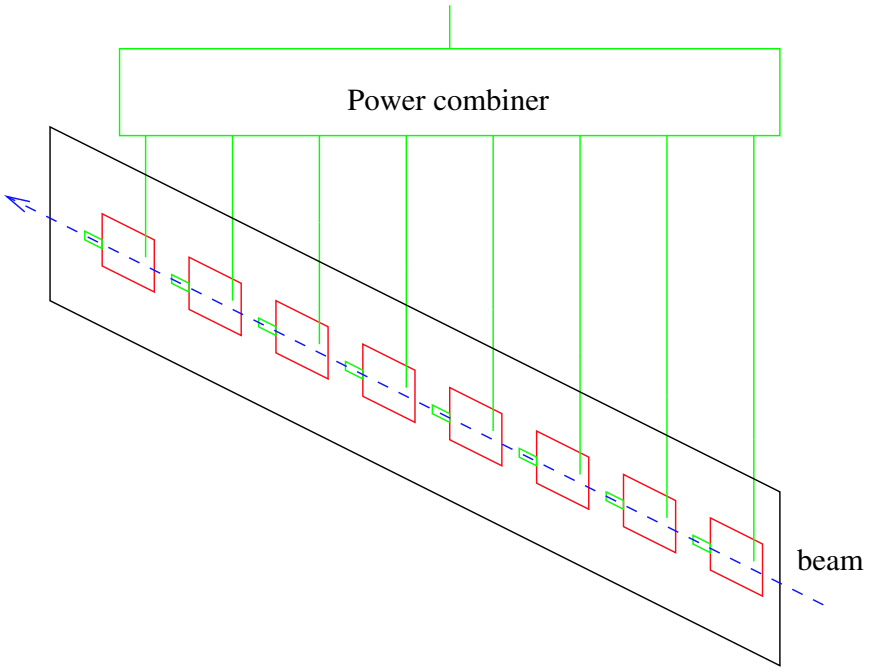


Fig. 7: Schematic view of power combination from distributed pick-up electrodes.

to the beam velocity in order to yield a synchronous addition for signals from the individual pick-ups. Neglecting losses, the output voltage of an n-fold power combiner is

$$V_{\text{out}} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} V_j \ , \tag{29}$$

where the V_j are the voltages at the input ports. The average power is

$$P_{\text{out}} = \frac{1}{nZ_{\text{line}}} \sum_{j,k=1}^{n} \langle V_j V_k \rangle \ . \tag{30}$$

If the cable lengths are adjusted correctly, the input voltages are equal and one gets $P_{\text{out}} = nP_{\text{single}}$, where P_{single} is the beam power from a single pick-up electrode. But the noise voltages from the individual terminating resistors, V_j and V_k, are uncorrelated for $j \neq k$, and there is no increase of the output noise power in comparison to the individual noise. (This would violate the second law of thermodynamics!.) Therefore one gets an optimum increase of the signal-to-noise ratio by a factor n.

3 RECIPROCITY

Pick-ups and kickers are usually constructed in the same way. This makes sense because of reciprocity. The reciprocity theorem of antenna theory can be proved directly from Maxwell's equations. For this proof, the reader is referred to Ref. [10]. Here we limit ourselves to the physical contents of the theorem. This refers to a given volume with 'metallic' boundaries on which the tangential electric field is

perpendicular to the magnetic field. The pick-up case is treated with fields $\vec{E}_p$, $\vec{H}_p$ and a beam current density $\vec{j}_p$. Fields $\vec{E}_k$ and $\vec{H}_k$ apply to the kicker, and there is no driving current density in this case. The connection between the pick-up and kicker fields is only given by the fact that they rely on equal boundaries. For illustrative purposes, imagine a geometry as in Fig. 8. This figure is adapted from Ref. [10]. The boundaries are sketched by thick lines, there are open surfaces $S_1 \ldots S_4$ for the signal and terminating resistors, and the surfaces S_{c1}, S_{c2} symbolize the boundary between the 'electrode' and the the rest of the vacuum chamber. The reciprocity theorem, stated as an integral equation, relates the pick-up and kicker fields by

$$\int_{\partial\mathcal{V}} dS \left(\vec{E}_p \times \vec{H}_k - \vec{E}_k \times \vec{H}_p \right) = \int_{\mathcal{V}} \vec{E}_k \cdot \vec{j}_p \ . \tag{31}$$

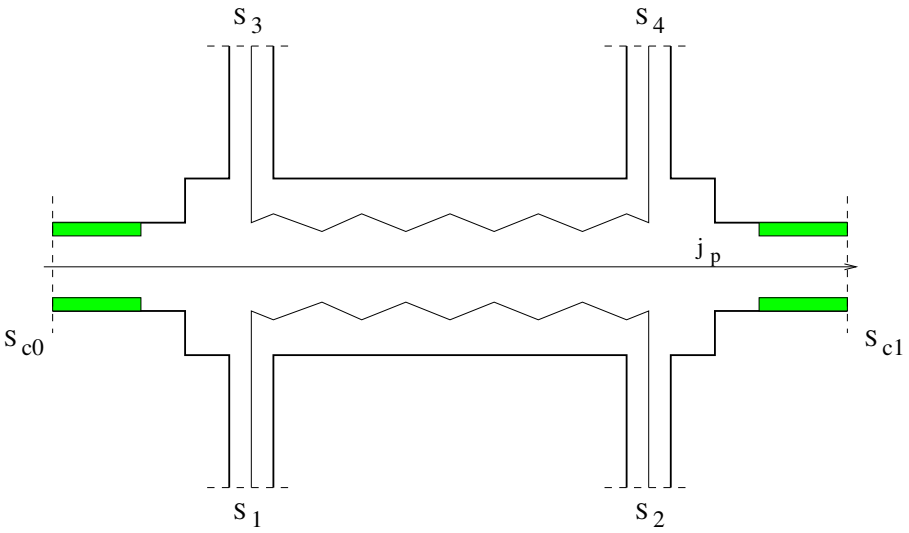


Fig. 8: Schematic view of a pick-up or kicker structure relevant to the discussion of reciprocity.

In this equation, $\mathcal{V}$ is the volume bounded by $\partial\mathcal{V}$. So far everything can be derived strictly from Maxwell's equations. The surface integrand on the left-hand side of Eq. (31) vanishes everywhere except on the open surfaces described above.

An important assumption is that the vacuum chamber boundaries S_{c1} and S_{c2} can be placed so far away from the electrode that their contribution to the surface integral can be neglected. If these propagating modes are damped by attenuating material inside $\mathcal{V}$, the contribution of the vacuum chamber modes is negligible, too. Dampers of this kind are symbolized by the filled areas at the ends of the structure in Fig. 8. The second assumption is that the fields at the ports $S_1 \ldots S_4$ can be described by TEM waves, as they exist in coaxial cables. Then the fields can be desribed using RF voltages V_k or V_p. In Ref. [10] it is shown that in certain circumstances where an outcoming wave $\vec{E}_k$ and an incoming wave $\vec{E}_p$ are considered, the reciprocity theorem can be reduced to the equation

$$V_p(\Omega) = \frac{-Z_{\text{line}}}{2V_k(\Omega)} \int_{\mathcal{V}} \vec{E}_k(x,y,s,\Omega) \cdot \vec{j}_p(x,y,s,\Omega) \ , \tag{32}$$

where Z_{line} is the line impedance of the coaxial line. We shall show that Eq. (32) can be related to the definition of sensitivity in Eqs. (5) and (6). The point-like current density of a single particle on the right-hand side of Eq. (32) can be represented by delta functions:

$$\vec{j}_p(x,y,s,t) = e\vec{v}\delta(x-x_0)\delta(y-y_0)\delta(s-s_0-\beta ct) \ . \tag{33}$$

We have assumed that $\vec{v}$ points essentially in the s direction and that the change in x and y can be neglected during the flight of the particle through the structure. Equation (33) describes just a single passage. In Eq. (32), we need the Fourier transform of the current density:

$$\vec{j}_p(x, y, s, \Omega) = e\frac{\vec{v}}{v}\delta(x - x_0)\delta(y - y_0) \exp\left(\frac{-i\Omega\,(s - s_0)}{v}\right) . \tag{34}$$

The integral on the right-hand side of Eq. (32) yields

$$\int_{\mathcal{V}} \vec{E}_k(x, y, s, \Omega) \cdot \vec{j}_p(x, y, s, \Omega) = \int ds E_{k,s}(x_0, y_0, s, \Omega) \exp\left(\frac{-i\Omega\,(s - s_0)}{v}\right) . \tag{35}$$

This is the Fourier transform of

$$\int ds E_{k,s}\left(x_0, y_0, s, t - \frac{s - s_0}{v}\right) .$$

Comparing with Eq. (5) and using the Fourier convolution theorem, we conclude that

$$V_p(\Omega) = \frac{-qeZ_{\text{line}}}{2} S_{\parallel}(x_0, y_0, \Omega) . \tag{36}$$

This equation shows that the sensitivity we defined in Eq. (3) to describe the action of a kicker is also useful for the description of pick-ups, if both are built identically. The factor 2 in the denominator can be explained qualitatively in the following way: A kicker is usually excited by a counterpropagating wave. On the other hand, a particle passing through a pick-up produces waves in both directions, of which only one is used because of the need to synchronize the wave with the particle velocity.

In the time domain, the signal due to a single passage of the particle is, according to Eq. (36):

$$V_p(t) = \frac{-qeZ_{\text{line}}}{2} S_{\parallel}(x_0, y_0, t) . \tag{37}$$

We shall use this result in the following chapter. An approach to reciprocity, where far-fields are also taken into account, was presented by F. Caspers and G. Dôme in Ref. [17].

4 SIGNAL FROM A CIRCULATING BEAM

4.1 Simplified one-particle signal

In Eq. (36) we have learned how to calculate the pick-up spectrum for a single passage through the device. In a beam coasting in a storage ring, each particle passes through the pick-up periodically. In order to get an easy-to-use mathematical description, we approximate the particle arrival at the pick-up by an infinite train of delta functions:

$$f(t) = \sum_{n=-\infty}^{\infty} \delta\left(t - t_0 - nT\right) . \tag{38}$$

Here t_0 denotes the time of arrival at revolution number $n = 0$. This time is different for each particle in the beam. T is the revolution time, which is normally distributed around a mean revolution time with a certain spread which is proportional to the momentum spread of the beam. The Fourier spectrum of $f(t)$ is also an infinite train of delta functions, but this time in the frequency domain:

$$f(\Omega) = \omega_{\text{rev}} \sum_{m=-\infty}^{\infty} \exp\left(i\Omega t_0\right) \delta\left(\Omega - m\omega_{\text{rev}}\right) , \tag{39}$$

where $\omega_{\mathrm{rev}} = 2\pi/T$ is the (angular) revolution frequency. The time t_0 is expressed by a phase in the spectrum. In order to understand the Fourier pair in Eqs. (38) and (39) properly, it helps to note that the infinite number of frequencies in Eq. (39) is due to the infinitely sharp pulses in the time domain. The reader who is interested in what happens if one or both of these idealized assumptions are dropped, may find it interesting to follow Feynman's rule 'The same equations have the same solutions', and re-examine a well-known analogy in optics: the diffraction pattern of a diffraction grating, which is the Fourier transform of the optical transmission function of the grating. In our case, the number of diffraction slits is mapped to the number of revolutions, whereas the role of the transmission through a single slit is played by the sensitivity $S(x, y, t)$. Working out the details of the analogy is left as an exercise to the reader.

The signal of a pickup with a circulating particle can be modeled in the time domain as the convolution of $f(t)$ with Eq. (37). Therefore the spectrum of a single circulating particle becomes:

$$V_p(\Omega) = \frac{-qeZ_{\mathrm{line}}\omega_{\mathrm{rev}}}{2} \sum_{m=-\infty}^{\infty} \exp\left(i\Omega t_0\right) S_{\|}(x_0, y_0, \Omega)\delta\left(\Omega - m\omega_{\mathrm{rev}}\right) \ . \tag{40}$$

The spectrum consists of sharp lines at each harmonic of the revolution frequency, with an amplitude that is given by the sensitivity, and a phase that depends on the position of the particle.

4.2 One-particle signal including betatron motion

In deriving Eq. (40) we have tacitly assumed that at each revolution, the particle passes through the pick-up at the same position in x and y. However, this assumption does not hold in general, because particles in a storage ring perform betatron oscillations. The pick-up produces a stroboscopic view of these oscillations because it does not observe the transverse motion directly, but can detect the position only revolution by revolution. The position at the nth revolution can be written

$$x_n = x_0 + A_x \sin\left(2\pi n Q_x + \mu_x\right) \tag{41}$$

$$y_n = A_y \sin\left(2\pi n Q_y + \mu_y\right) \ . \tag{42}$$

The quantities A_x and A_y denote the amplitude of the oscillations, and the quantities Q_x and Q_y are the number of oscillations per turn, a very important quantity for the beam dynamics in a storage ring. μ_x and μ_y quantify the phase of the oscillation during revolution number $n = 0$. Usually the beam is centred with respect to the electrode, but in the case of horizontal motion there can be a systematic shift x_0 which is caused by the fact that the particle momentum differs from the design momentum. This effect is called dispersion. For the vertical motion, the effect of dispersion can usually be neglected. In order to calculate the pick-up spectrum with arbitrary sensitivity $S(x, y, t)$ including betatron motion, there is an elegant mathematical method [10]. It begins with a wavelength decomposition of $S_{\|}$, i.e. we calculate the spatial Fourier transform of $S_{\|}$:

$$\tilde{S}_{\|}(k_x, k_y, t) = \int dx\, dy\, S_{\|}(x, y, t)\, e^{-i(k_x x + k_y z)} \tag{43}$$

or, alternatively

$$S_{\|}(x, y, t) = \frac{1}{(2\pi)^2} \int dk_x\, dk_y\, \tilde{S}_{\|}(k_x, k_y, t)\, e^{i(k_x x + k_y y)} \ . \tag{44}$$

The advantage of this representation is that the oscillating quantities x and y appear in an exponential, which can be developed in a series of Bessel function, using the well-known identity familiar to RF engineers from the spectrum of a frequency-modulated signal:

$$\exp(ia\sin x) = \sum_{l=-\infty}^{\infty} J_l(a)\, e^{ilx} \ .$$

Inserted in Eq. (44), this yields

$$S_{\|}(x_n, y_n, t) = \sum_{l_x=-\infty}^{\infty} \sum_{l_y=-\infty}^{\infty} S_{\|}^{(lx,ly)}\left(A_x, A_y, t\right) \exp\left[2\pi i n \left(l_x Q_x + l_y Q_y\right)\right] \tag{45}$$

with the important sensitivity coefficients

$$S_{\|}^{(lx,ly)}\left(A_x, A_y, t\right) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} dk_x\, dk_y\, J_{l_x}\left(k_x A_x\right) J_{l_y}\left(k_y A_y\right) \tilde{S}_{\|}(k_x, k_y, t) e^{ik_x x_0 + i(l_x \mu_x + l_y \mu_y)} \,. \tag{46}$$

These coefficients depend only on constants of the motion and on time.

Let us derive the signal from a single circulating particle. Using the reciprocity theorem Eq. (37) and the pulse train from Eq. (38) yields:

$$V_p(t) = \frac{-qeZ_{\text{line}}}{2} \sum_{l_x=-\infty}^{\infty} \sum_{l_y=-\infty}^{\infty} S_{\|}^{(lx,ly)}\left(A_x, A_y, t\right) \exp\left[2\pi i n \left(l_x Q_x + l_y Q_y\right)\right] * f(t) \,. \tag{47}$$

The exponential in Eq. (47) produces a modulation of the delta peaks in $f(t)$, which leads to a new sequence of sidebands. Indeed, if we insert $f(t)$ from Eq. (38) and transform the exponential in Eq. (47) according to

$$\begin{aligned} &\sum_{n=-\infty}^{\infty} S_{\|}^{(lx,ly)}\left(A_x, A_y, t\right) \exp\left[2\pi i n \left(l_x Q_x + l_y Q_y\right)\right] * \delta(t - nT - t_0) \\ = &\sum_{n=-\infty}^{\infty} S_{\|}^{(lx,ly)}\left(A_x, A_y, t - t_0\right) * \exp\left[i\omega_{\text{rev}} t \left(l_x Q_x + l_y Q_y\right)\right] \delta(t - nT) \,, \end{aligned}$$

then the Fourier transform of this expression becomes

$$\omega_{\text{rev}} \sum_{m=-\infty}^{\infty} S_{\|}^{(lx,ly)}\left(A_x, A_y, \Omega\right) e^{i\Omega t_0} \delta\Big(\Omega - \omega_{\text{rev}}\left(m + l_x Q_x + l_y Q_y\right)\Big)$$

with the spectral coefficients

$$S_{\|}^{(lx,ly)}\left(A_x, A_y, \Omega\right) = \int_{-\infty}^{\infty} dt\, S_{\|}^{(lx,ly)}\left(A_x, A_y, t\right) e^{-i\Omega t} \,. \tag{48}$$

Therefore the spectrum consists of lines at each revolution harmonic with a series of betatron sidebands at a distance $\left(l_x Q_x + l_y Q_y\right) \omega_{\text{rev}}$:

$$V_p(\Omega) = \sum_{m=-\infty}^{\infty} \sum_{l_x=-\infty}^{\infty} \sum_{l_y=-\infty}^{\infty} V_{m,l_x,l_y}(\Omega)\, \delta\left(\Omega - \omega_{\text{rev}}\left(m + l_x Q_x + l_y Q_y\right)\right) \tag{49}$$

$$V_{m,l_x,l_y}(\Omega) := \frac{-qeZ_{\text{line}}\omega_{\text{rev}}}{2} S_{\|}^{(lx,ly)}\left(A_x, A_y, \Omega\right) \exp\left(i\Omega t_0\right) \,. \tag{50}$$

This is the general expression for a pick-up with arbitrary sensitivity including betatron motion. Equations (49) and (50) resemble the simplified expression [Eq. (40)], except that the sensitivity has been replaced by the more complicated coefficients [Eq. (46)], and that all the betatron sidebands have emerged.

In order to see how the formalism presented above works, let us investigate Eq. (46) for the case where the betratron amplitudes A_x, A_y are small compared to the physical dimensions of the electrode. The Fourier transform $\tilde{S}(k_x, k_y, t)$ is negligible for large wave numbers $k = 2\pi/\lambda$ where the wavelength λ is much smaller than these physical dimensions. Therefore only small arguments $k_x A_x$ and $k_y A_y$ of

the Bessel functions contribute significantly to Eq. (46), and we can use the approximation $J_l(w) \approx w^l/(2^l l!)$. This yields factors k^l in front of the $\tilde{S}$ in Eq. (46). Using Eq. (A.1.7), we get the approximate coefficients

$$S_{\|}^{(lx,ly)}(A_x, A_y, \Omega) \approx \frac{A_x^{l_x} A_y^{l_y}}{(2i)^{l_x+l_y}\, l_x!\, l_y!} e^{i(l_x\mu_x + l_y\mu_y)} \left.\frac{\partial^{l_x}\partial^{l_y} S(x,y,\Omega)}{\partial x^{l_x}\partial z^{l_y}}\right|_{x=x_0,y=0}, \quad l_x \geq 0, l_y \geq 0\,. \tag{51}$$

For negative indices, there is a general relation that follows from the properties of the Bessel functions:

$$S_{\|}^{(-l_x,l_y)} = (-1)^{l_x} S_{\|}^{(lx,ly)}, \; S_{\|}^{(l_x,-l_y)} = (-1)^{l_y} S_{\|}^{(lx,ly)}\,. \tag{52}$$

In particular, we get the approximate amplitudes at the revolution harmonics and the first betatron sidebands

$$V_{m,0,0}(\Omega) \approx \frac{-qeZ_{\text{line}}\omega_{\text{rev}}}{2} e^{i\Omega t_0} S(x_0, 0, \Omega) \tag{53}$$

$$V_{m,\pm1,0}(\Omega) \approx \frac{\mp qeZ_{\text{line}}\omega_{\text{rev}} A_x}{4i} e^{i\Omega t_0 \pm i\mu_x} \left.\frac{\partial S(x,y,t)}{\partial x}\right|_{x=x_0,y=0} \tag{54}$$

$$V_{m,0,\pm1}(\Omega) \approx \frac{\mp qeZ_{\text{line}}\omega_{\text{rev}} A_y}{4i} e^{i\Omega t_0 \pm i\mu_y} \left.\frac{\partial S(x,y,t)}{\partial y}\right|_{x=x_0,y=0}\,. \tag{55}$$

In the case of small betatron amplitudes, we have rediscovered Eq. (40) for the signal at the revolution harmonics. The signal at the horizontal [Eq. (54)] and vertical [Eq. (55)] betatron sidebands is proportional to the betatron amplitude and the slope of the sensitivity.

After having shown that the general expression for the sensitivity coefficients, Eq. (46), yields rather simple expressions in the case of small betatron amplitudes, we recall how the coefficients $S_{\|}^{(lx,ly)}$ are calculated in general:

1. Take the spatial Fourier transform $\tilde{S}$ of the sensitivity.
2. Multiply by the weights $J_{l_x}(k_x A_x)$ and $J_{l_y}(k_y A_y)$.
3. Calculate the inverse transform at the point $x = x_0, y = 0$.
4. Take into account the initial betatron phases μ_x and μ_z.

The coefficients $S_{\|}^{(lx,ly)}$ are a generalization of the approximate relations Eqs. (53)–(55) to the case that the betatron motion becomes so large that it begins to feel the non-linearities of the sensitivity. Indeed, it can be shown that Eqs. (53)–(55) are exact if the sensitivity is linear. As the formalism relies on Fourier transforms, it will be easy to implement on a computer as soon as useful sensitivity models are available. It is useful for the geometrical design of the electrodes. During the design procedure, the aim will be to produce reasonable sensitivities even for the largest betatron amplitudes expected.

It can be seen from Eq. (51) that the signals at the higher-order sidebands $|l_{x,y} \geq 2|$ are directly proportional to higher derivatives of the sensitivity and higher powers of the betatron amplitude. For the practical application of stochastic cooling, they should never play a role. However, they are observed (for example at GSI) in spectra of beams with collective transverse instabilities.

4.3 Spectral density and total power

4.3.1 Schottky signal

The total signal from the beam is the superposition of the single particle signal, Eq. (49), of all particles circulating in the storage ring. It is very important to distinguish between bunched beams and coasting beams. Here we treat only the latter, because stochastic cooling is only switched on after the injected beam has been debunched.

In a coasting beam the particles are distributed randomly. This randomness holds for the time t_0 [compare Eq. (38)] which measures the particle position during a given revolution, and the initial betatron

phases μ_x and μ_z [see Eqs. (41) and (42)]. The distribution in $\omega_{\rm rev}$, and in the betatron amplitudes A_x and A_y at the pick-up, can be expressed by a distribution function $\Psi(\omega_{\rm rev}, A_x, A_y)$ which is normalized to the number of particles N in the ring. Because the signal consists of contributions with random phases due to t_0, μ_x and μ_z, the output voltage from a pick-up disappears on the average, but fluctuates around zero. Therefore the average power is non-zero.

The signal from a coasting beam is called a Schottky signal. In 1918 W. Schottky published a paper [18] in which he showed that the current in semiconductors is never exactly constant, but that there are statistical fluctuations which he attributed to the fact that the current consists of individual electrons with a finite elementary charge that are randomly distributed.

The signal from the beam fluctuates, as well. We pointed out in the introduction to this paper that these fluctuations are the very basis of stochastic cooling. In order to describe the signal quantitatively, we briefly recall some common definitions of signal theory [19].

4.3.2 Basic definitions from signal theory

The Schottky signal from a coasting beam is a stochastic voltage $V(t)$. It is characterized statistically by its autocorrelation function, which is defined as the expectation value of the product of V at times t and $t+\tau$. $V(t)$ is a representative of a *stationary* process because the autocorrelation function does not depend on t (This is no longer true for a bunched beam!). Therefore the autocorrelation can be written in the form

$$\langle V(t)\, V(t+\tau)\rangle = R(\tau)\ . \tag{56}$$

The spectral density C is the Fourier transform of the autocorrelation function

$$C(\Omega) = \int_{-\infty}^{\infty} d\tau\, R(\tau) e^{-i\Omega\tau}\ . \tag{57}$$

One should bear in mind that C is defined as a *voltage* density. For a stationary process the Wiener-Khintchine-Theorem holds:

$$\left\langle \tilde{V}(\Omega)\, \tilde{V}^*(\Omega')\right\rangle = 2\pi\, C(\Omega)\, \delta(\Omega-\Omega')\ . \tag{58}$$

In this notation, $\tilde{V}(\Omega)$ is the Fourier transform of $V(t)$ and * denotes the complex conjugate.

The average total electric power is

$$< P >= \frac{1}{Z_{\rm line}} \left\langle V(t)^2\right\rangle = \frac{R(0)}{Z_{\rm line}} = \frac{1}{2\pi Z_{\rm line}} \int_{-\infty}^{\infty} d\Omega\, C(\Omega)\ . \tag{59}$$

4.3.3 Schottky spectral density

The Schottky spectral density C_p at the output port of a pick-up electrode can be derived from the single particle signal Eq. (49) using the distribution Ψ and the assumption of a smooth distribution in t_0, μ_x and μ_y, as discussed above. In Eq. (49) we saw that each particle produces signals at frequencies

$$\Omega = (m + l_x Q_x + l_y Q_y)\, \omega_{\rm rev}\ . \tag{60}$$

The mathematical description of the power density can be complicated due to the possibility of Schottky band overlap. Let us call the maximum width of the distribution in revolution frequencies $\Delta\omega_{\rm rev}$. Then, if $m\Delta\omega_{\rm rev}/\omega_{\rm rev} > 1$, the extreme parts of the frequency distributions begin to overlap for neighbouring harmonics. Already, at lower harmonics, betatron sidebands overlap with revolution harmonics, and with one another.

To be complete, it should be noted that in the following we neglect chromaticity, which would introduce a correlation between ω_{rev} and Q_x or Q_y. For the Schottky power density, we get

$$C(\Omega) = \sum_{m,l_x,l_y} \int dA_x dA_y \Psi \left(\frac{\Omega}{|m + l_x Q_x + l_y Q_y|}, A_x, A_y \right) \left| V_{m,l_x,l_y}(\Omega) \right|^2 . \tag{61}$$

The sum has to be taken over all bands that contribute to Ω and one has to integrate over betatron amplitudes. Let us derive from Eq. (61) some useful approximations. If

1. the approximation Eq. (53) holds for all betatron amplitudes, and
2. if $S_{m,0,0}$ does not depend on ω_{rev}, and
3. Schottky bands do not overlap,

then the following proportionality holds for the integrated power $P_{m,0,0}$ over one band at harmonic m:

$$P_{m,0,0} \propto N q^2 . \tag{62}$$

Under these assumptions, the Schottky power in one band is proportional to the number of particles and to the square of the charge state. It can serve to measure the beam current, if it is properly calibrated. Obviously, the signal-to-noise ratio of heavy ion beams is intrinsically much higher than for singly charged particle beams, if one assumes identical electrodes and the same number of particles. The signal from a beam of fully stripped uranium ions is 39 dB higher than the one from a corresponding (anti)proton beam.

For the power in the betatron sidebands $P_{m,\pm1}$ there is a similar relation if one assumes $S_{m,\pm1,0}$ or $S_{m,0,\pm1}$ to be constant and that they can be described by Eq. (54) or Eq. (55), respectively:

$$P_{m,1,0} \propto N q^2 \left\langle A_x^2 \right\rangle \tag{63}$$
$$P_{m,0,1} \propto N q^2 \left\langle A_y^2 \right\rangle . \tag{64}$$

The integrated sideband power is therefore a useful diagnostic for the mean betatron amplitude.

4.3.4 Noise spectral density

The noise power of a source at temperature T_{eff} in a frequency interval Δf can be approximated by

$$P_{\mathrm{noise}} = k_{\mathrm{B}} T_{\mathrm{eff}} \Delta f . \tag{65}$$

The noise spectral density C_{noise} inside Δf is constant:

$$C_{\mathrm{noise}}(\Omega) = \frac{1}{2} Z_{\mathrm{line}} k_{\mathrm{B}} T_{\mathrm{eff}} . \tag{66}$$

The factor $1/2$ stems from the necessity to integrate over both positive and negative values of Ω, when one deals with Fourier transforms.

4.4 Transverse pick-up designs

For a good betatron signal the sensitivity should be zero for a centred beam and approximately linear up to the highest betatron amplitudes to be expected [compare Eqs. (54) and (55)]. A vertical pick-up, for example, could consist of two stripline electrodes operated in the difference mode. A horizontal pick-up would be obtained by turning this device around by $90°$. If large horizontal apertures are required, a more effective solution consists of two pairs of stripline electrodes. Each pair is operated in the sum mode, and then one takes the difference of the sum signals. Such a geometry is sketched in Fig. 9.

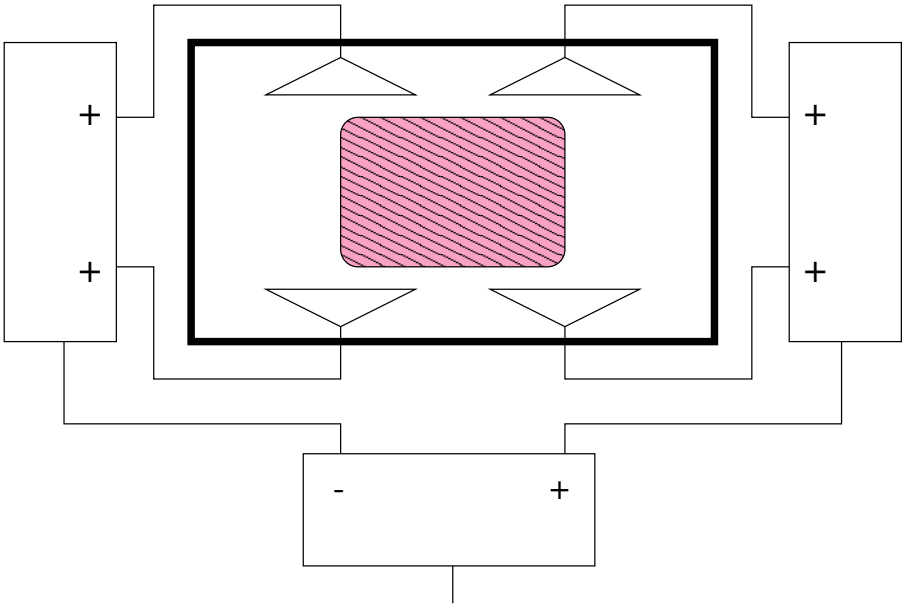


Fig. 9: Schematic sketch of horizontal pick-up. The useful beam aperture is shaded.

4.5 Momentum signal

4.5.1 General requirements

The momentum correction signal at the kicker should carry information about the deviation δp of the particle momentum from the design momentum p_0. Ideally it should be proportional to $\delta p/p_0$. Two methods have been developed to extract this information. The Palmer method relies on dispersion, whereas the notch filter method operates completely in the frequency domain.

4.5.2 Palmer pick-ups

In the Palmer method, the momentum correction signal is derived from the position of the particle as it passes through the pick-up. The dependence of position on momentum is called dispersion. It can be written in the form

$$x_0 = D\frac{\delta p}{p_0}\,, \tag{67}$$

where D is a quantity measured in metres and is called the dispersion function. It is caused by the fact that the orbit curvature in a dipole magnet is proportional to momentum. In a storage ring, D varies around the ring. Therefore, an appropriate location must be chosen for the Palmer pick-up. Palmer pick-ups are transverse pick-ups like the one shown in Fig. 9. The signal can be described approximately by Eq. (53). The proportionality to $\delta p/p_0$ is caused by the x_0 dependence of the sensitivity. In order to get a clean signal from the device, the horizontal betatron amplitude A_x at the location of the Palmer pick-up should be small compared to the dispersion amplitude x_0.

4.5.3 Thorndahl notch filters

The pick-up for the notch filter method has constant sensitivity. The signal is derived by electronic post-processing. It relies on the fact that the revolution frequency depends on momentum:

$$\frac{\delta\omega_{\rm rev}}{\omega_{\rm rev}} = \eta\frac{\delta p}{p_0}\,, \tag{68}$$

where η is a quantity called frequency dispersion. It is explained in every introductory text on longitudinal beam dynamics, because it is important for the description of what happens to a bunch in an RF

field.

In a notch filter the signal is split in a power splitter. The two components pass a short and a long delay line and are then fed into to a 180° hybrid. If the time difference of the signal transmission in the

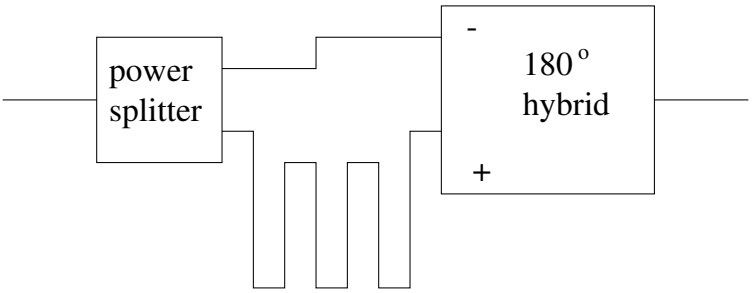


Fig. 10: Schematic sketch of a notch filter.

two cables is t_{notch}, the output voltage will take the form

$$V_0 \left(e^{i\Omega t_{\mathrm{notch}}} - 1\right) = 2iV_0 e^{i\Omega t_{\mathrm{notch}}/2} \sin \frac{\Omega t_{\mathrm{notch}}}{2} \,. \tag{69}$$

If $t_{\mathrm{notch}}/2$ is made equal to the revolution frequency, the transmission will be periodic with the revolution frequency and exhibit zeros at all harmonics of the revolution frequency. This is due to destructive interference between the waves from the short and long cables. These zeros are often called notches. The device is also known as a comb filter. The quality of the cancellations depends on the attenuation of the two waves in the two cables which must be equalized over the whole useful bandwidth. The transmission S_{21} of the notch filters fulfils the requirements of a momentum signal for stochastic cooling in a frequency interval of half a revolution frequency at most. Therefore, the frequency spread of the Schottky signal should be smaller than that at the highest frequency used in the cooling system. The decision to use notch filters in stochastic cooling systems has an important impact on the ion optical design of the cooler ring, because of the constraints on the frequency dispersion η. The appplication of notch filters for momentum cooling was invented by L. Thorndahl and G. Carron at CERN in the 1970s [20]. It played a major role in the successful stochastic cooling of antiproton beams.

An important advantage of notch filters is the suppression of thermal noise around the notches. As the length of the long cable is usually of the order of a few hundred metres, notch filters tend to be sensitive to thermal variations. A modern solution to this problem has been worked out at COSY [21]. They replace the long cable by a fibre-optic, which is connected to an optical modulator and demodulator. The optical devices are less sensitive to thermal variations, and the electric length can be tuned by movable mirrors.

5 TRANSVERSE KICKS

5.1 Panofsky–Wenzel theorem

Transverse kicks to particles moving in an RF field are described by the Panofsky–Wenzel theorem [22]. To every newcomer in the field, this theorem appears to come as a surprise, because it maintains that transverse electric fields are not needed to describe the transverse kick. A proper understanding of the Panofsky–Wenzel theorem is a necessary prerequisite for the design of transverse kicker electrodes. At first we present a straightforward derivation where the electric and magnetic fields are used directly. The equations needed are the Lorentz force equation

$$\frac{d\vec{p}}{dt} = qe\left(\vec{E} + \vec{v} \times \vec{B}\right) \tag{70}$$

and Faraday's law of induction

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \,. \tag{71}$$

It will soon become clear that the magnetic component of the Lorentz force is as important as the electric one in the derivation. Remember that the fields seen by the particle depend on the space coordinates and on time, i.e. they must be written as $\vec{E}(\vec{r}(t), t)$ and $\vec{B}(\vec{r}(t), t)$. Then the time derivative of the electric field is

$$\frac{d\vec{E}}{dt} = \frac{\partial \vec{E}}{\partial t} + \left(\vec{v} \cdot \vec{\nabla}\right) \vec{E} \,. \tag{72}$$

Because we are dealing with sinusoidally varying fields, we investigate the partial time derivative of the Lorentz force, making use of Eqs. (71) and (72):

$$\begin{aligned} \frac{\partial}{\partial t}\left(\vec{E} + \vec{v} \times \vec{B}\right) &= \frac{\partial \vec{E}}{\partial t} - \vec{v} \times \left(\vec{\nabla} \times \vec{E}\right) \\ &= \frac{\partial \vec{E}}{\partial t} + \left(\vec{v} \cdot \vec{\nabla}\right) \vec{E} - \vec{\nabla}\left(\vec{v} \cdot \vec{E}\right) \\ &= \frac{d\vec{E}}{dt} - \vec{\nabla}\left(\vec{v} \cdot \vec{E}\right) \,. \end{aligned} \tag{73}$$

We assume sinusoidally varying fields $\vec{E} = \vec{E}_0(\vec{r}) \exp i\Omega t + \text{c.c.}$ and $\vec{B} = \vec{B}_0(\vec{r}) \exp i\Omega t + \text{c.c.}$, where c.c. denotes the complex conjugate. With Eq. (73), the integrated change of momentum in a kicker over a length $(s_1 - s_0)$ can be written

$$\Delta\vec{p} = \frac{qe}{i\Omega}\left[\vec{E}(s_1) - \vec{E}(s_0) - \frac{1}{v}\int_{s_0}^{s_1} ds \vec{\nabla}\left(\vec{v} \cdot \vec{E}\right)\right] + \text{c.c.} \tag{74}$$

The velocity change in an RF kicker is usually small, and therefore $\vec{v}$ has been assumed to be approximately constant in the integrand. The entrance and exit points may be moved to a place where the electric field can be neglected. For a kick in the x-direction, Eq. (74) yields

$$\Delta p_x = \frac{iqe}{\Omega v}\int_{s_0}^{s_1} ds \left(v_x \frac{\partial E_x}{\partial x} + v_y \frac{\partial E_y}{\partial x} + v\frac{\partial E_s}{\partial x}\right) + \text{c.c.} \tag{75}$$

In most practical field configurations, the terms $v_x \partial E_x/\partial x$ and $v_y \partial E_y/\partial x$ can be neglected because the velocity points mainly in the s direction. Then

$$\Delta p_x = \frac{i}{\Omega}\frac{\partial \Delta E}{\partial x} \,, \tag{76}$$

where ΔE is the change of energy in the kicker. This relation looks very unfamiliar at first sight. The following comments should clarify the physical meaning of Eq. (76):

- The transverse kicks are proportional to the transverse gradient of the longitudinal electric field.
- No transverse fields are needed for a complete description of the transverse kick.
- Even if a particle passing through a transverse kicker on-axis does not feel an accelerating voltage, off-axis particles do.
- For a given transverse deflection, the necessary transverse field gradient of the longitudinal electric field is proportional to the RF frequency.
- There is a 90° phase difference between the maximum transverse kick and the maximum (off-axis) energy change.
- In the low-frequency limit, the longitudinal electric fields do not play a practical role.
- As Panofsky and Wenzel pointed out in their concise original paper, there can be no transverse deflection in a pure TE field.

Tracing back the derivation of Eqs. (73)–(76), it becomes obvious that the effective transverse kick is produced by a component of the magnetic force. Because of the law of induction, it can be described in terms of the electric field. There is no net effect of the actual electric force because there is a second component of the magnetic force which adds up to the electric field to produce a total time derivative of $\vec{E}$, the effect of which vanishes under suitable boundary conditions.

A more elegant, perhaps somewhat less instructive way to derive the Panofsky–Wenzel theorem makes use of the vector potential $\vec{A}$ and the relations $\vec{E} = -\partial\vec{A}/\partial t$, as well as $\vec{B} = \vec{\nabla} \times \vec{A}$ in the Lorentz force term. Doing so, there is no need to operate with the partial time derivative of the Lorentz force. Instead of terms like $1/(i\Omega)\partial E_s/\partial x$ [see Eq. (75)] the simpler, equivalent terms $\partial A_s/\partial x$ turn up. The reader should try and derive this as an exercise. This is the main reason behind defining the sensitivity via the vector potential in Eq. (2). Definitions in the time domain as Eqs. (7) and (8) would have become too cumbersome without using the vector potential.

It is rewarding to compare the order of magnitude of the transverse kick to the longitudinal kick for off-axis particles. Let the beam be centred with respect to the transverse kick electrode and let the longitudinal kick to particles in the centre vanish. Let the transverse kick be constant inside an aperture Δx. According to the Panofsky–Wenzel theorem, a transverse kick by an angle $\Delta x'$ must be accompanied by an off-axis energy change ΔE. Both are related by

$$\Delta x' = \frac{\Delta p_x}{p_0} = \frac{1}{i\Omega m\beta\gamma c}\frac{\Delta E}{\Delta x} \,. \tag{77}$$

On the other hand,

$$\frac{\Delta E}{m\gamma c^2} = \frac{\Delta E}{E_0} = \beta^2\frac{\Delta p}{p_0} \,. \tag{78}$$

Combining Eqs. (77) and (78) one gets an easy-to-use relation between the transverse and longitudinal kicks:

$$\frac{\Delta p}{p_0} = \frac{i\Omega\Delta x}{\beta c}\Delta x' \,. \tag{79}$$

For $\Omega/2\pi = 1$ GHz, Δx = 10 mm, and $\beta = 1$, we get $\Omega\Delta x/\beta c = 0.2096$, i.e. a kick by 1 mrad would be accompanied by a change in relative momentum of 2×10^{-4} at a beam radius of 10 mm. With Eq. (79) this relation can easily be scaled.

6 BEAM DYNAMICS WITH KICKS

A quantitative description of the beam response to the voltages applied at the kicker(s) is beyond the scope of this lecture because of the complicated beam dynamics involved. At this point, a few general remarks should be sufficient. The reader is advised to consult the literature. As already mentioned, a good introduction can be found in Ref. [1]. The ambitious reader interested in the physics of fluctuating signals should consult Ref. [23].

A particle circulating in the ring can get into resonance with the applied kicker signal. In the long term, it will be most sensitive to signals at the frequencies given by Eq. (60), i.e. at the same frequencies that it produces a signal at a pick-up. For the longitudinal response, the most important effect occurs at the frequencies $m\omega_{\mathrm{rev}}$. It is proportional to the kicker sensitivity $S_{\parallel}^{(0,0)}$. Neglecting the effect of betatron oscillations, this is roughly the quantity $S_{\parallel}$.

The change of the betatron amplitudes occur mainly because of the signal at the betatron sidebands $(m \pm Q_x)\,\omega_{\mathrm{rev}}$ and $(m \pm Q_y)\,\omega_{\mathrm{rev}}$. Because of the Panofsky–Wenzel theorem, it is proportional to $\partial S/\partial x$ and $\partial S/\partial y$ [see Eqs. (7) and (8)].

APPENDIX - FOURIER TRANSFORMS AND NOTATION

The Fourier transform is defined as a transformation between the time domain and the frequency domain. In this paper, we consistently use the definitions from Ref. [19]. Time is denoted by t and angular frequency by Ω. The original function and its Fourier transform are denoted by the same letter, e.g. $V(t)$ and $V(\Omega)$. This is mathematically not quite correct, but saves a lot of nasty indices. A localized exception to this rule is the spatial transform $\tilde{S}$ used in Eqs. (43)–(46). The Fourier transform is defined as

$$f(\Omega) = \int_{-\infty}^{\infty} dt\, f(t) \exp(-i\Omega t) \tag{A.1.1}$$

The Fourier inversion theorem yields

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt\, f(\Omega) \exp(+i\Omega t) \tag{A.1.2}$$

If $f(t)$ and $f(\Omega)$ are related by a Fourier transform, we write

$$f(t) \leftrightarrow f(\Omega) \tag{A.1.3}$$

The following theorems are used in this paper

1. the linearity theorem
$$af(t) + bg(t) \leftrightarrow af(\Omega) + bg(\Omega) \tag{A.1.4}$$
2. the shift theorem
$$f(t-a) \leftrightarrow f(\Omega) \exp(-i\Omega a) \tag{A.1.5}$$
3. the modulation theorem
$$f(t) \exp(i\omega_0 t) \leftrightarrow f(\Omega - \omega_0) \tag{A.1.6}$$
4. the theorem on derivatives
$$\frac{d^n f(t)}{dt^n} \leftrightarrow (i\Omega)^n f(\Omega) \tag{A.1.7}$$
5. and, perhaps, the most important of them all, the convolution theorem
$$\int_{-\infty}^{\infty} d\tau\, f(t) g(t-\tau) \leftrightarrow f(\Omega) g(\Omega) \tag{A.1.8}$$

The convolution is often written in short-hand form

$$\int_{-\infty}^{\infty} d\tau\, f(t) g(t-\tau) := f(t) * g(t) \tag{A.1.9}$$

A very useful identity concerns the convolution with a delta function

$$\delta(t-a) * g(t) = g(t-a) \tag{A.1.10}$$

Most often the quantities $f(t)$ in the time domain refer to physically measurable quantities , i.e they are represented by real numbers. The negative frequency components of the Fourier transform of a real function $f(t)$ can easily be calculated from the positive frequency components: $f(-\Omega) = f^*(\Omega)$.

REFERENCES

[1] D. Möhl, CERN Report 95–06, 587–664.

[2] R. Stassen, U. Bechstedt, J. Dietrich, K. Henn, A. Lehrach, R. Maier, S. Martin, D. Prasuhn, A. Schnase, H. Schneider, H. Stockhorst, R. Tölle, Proc. EPAC 98, 553–555.

[3] F. Nolden, B. Franzke, A. Schwinn, F. Caspers, Proc. EPAC 98, 1052–1054.

[4] T. Katayama, S. Watanabe, Y. Batygin, N. Inabe, K. Ohtomo, T. Ohkawa, M. Takanaka, T. Tanabe, M. Wakasugi, I. Watanabe, Y. Yano, K. Yoshida, Proc. EPAC 98, 529–531.

[5] R.J. Pasquinelli, Proc. PAC 99, 1094–1096.

[6] G. Carron, F. Caspers, L. Thorndahl, CERN Report 85–01.

[7] C.S. Taylor, CERN Report 92-03, 458–473.

[8] F. Caspers, Proc. Joint-US-CERN-Japan International School on Frontiers of Accelerator Technology, 3–9. Nov. 1994, World Scientific.

[9] G. Lambertson, AIP Conf. Proc. **153** (1987) 1414–1442.

[10] J. Bisognano, C. Leemann, AIP Conf. Proc. **87** (1982) 584–655.

[11] E. Durand, Electrostatique, Tome II, Masson et Cie, Paris 1966.

[12] F. Nolden, K. Beckert, F. Caspers, B. Franczak, B. Franzke, R. Menges, A. Schwinn, M. Steck, Nucl. Instrum. Methods A **441** (2000) 219–222.

[13] P. Raabe, VDI Fortschrittberichte, Reihe 21 (Elektrotechnik), Nr. 128 (1993).

[14] J. Petter, D. McGinnis, J. Marriner, Proc. PAC 89, 636–638.

[15] L. Faltin, Nucl. Instrum. Methods **148** (1978) 449–455.

[16] D. McGinnis, Proc. PAC 99, 1713–1715.

[17] F. Caspers, G. Dôme, Proc. EPAC 94, 1208–1210.

[18] W. Schottky, Ann. Phys **57** (1918) 541.

[19] A. Papoulis, Signal Analysis, McGraw-Hill 1977.

[20] G. Carron, L. Thorndahl, internal report CERN/ISR–RF/78–12.

[21] U. Bechstedt, J. Dietrich, K. Henn, A. Lehrach, R. Maier, D. Prasuhn, A. Schnase, H. Schneider, R. Stassen, H. Stockhorst, R. Tölle, Proc. PAC 99, 1701–1703.

[22] W. Panofsky, W. Wenzel, Rev. Sci. Inst. **27** (1956) 967.

[23] S. Chattopadhyay, CERN 84–11.

PRACTICAL INSTRUCTIONS FOR THE RF AND MICROWAVE MEASUREMENT TUTORIAL

F. Caspers
CERN, Geneva, Switzerland
G. Hutter
GSI, Darmstadt, Germany

Abstract
For this practical tutorial using (vector) network and spectrum analysers detailed instructions are given on how to operate these sophisticated instruments, avoiding the need for reading heavy instruction manuals. The tutorial is basically split into two parts, namely network analysis and spectrum analysis. In the network section the reflection and transmission characteristics of passive and active one- and two-ports are measured with a particular emphasis on resonators and cavities as well as calibration procedures. Spectrum analysers are used to examine amplitude- and frequency-modulated (AM, FM) signals, signal strength measurement, non-linear behaviour of amplifiers via intermodulation and the evaluation of noise and noise-figure properties.

1. INTRODUCTION

Radio-frequency (RF) spectrum analysers (SPA or SA) can be found in virtually every control room of a modern particle accelerator. They are used for many aspects of beam diagnostics, including Schottky signal acquisition and RF observation. We discuss here only the application of classical super-heterodyne SPAs and not systems based on the acquisition of time-domain traces and subsequent Fourier transform (FFT analysers). Such a super-heterodyne SPA is very similar (in principle) to any AM or FM radio-receiver. The incoming RF signal is moderately amplified (sometimes with adjustable gain) and then sent to the RF port of a mixer. A mixer is a non-linear element (containing one or more fast diodes) and acts like an analogue multiplier. The output signal of a local oscillator (LO) is connected to the LO input of the mixer and superimposed on the RF. As a consequence we obtain (non-linear) mixing products at the sum and difference frequency between RF and LO which appear at the intermediate frequency (IF) port of the mixer. This IF signal is subsequently band-filtered and sent to the vertically deflecting plates of the cathode ray tube (CRT) in the SPA (or via a demodulator to the loudspeaker in a radio-receiver). The word ‘super-heterodyne’ is a mixture of Latin and Greek meaning adding (super) a different (hetero) force (dyne), where the different force is no other than the LO signal, superimposed (added) in a non-linear fashion on the RF signal.

A network analyser can often be found in a control room but its real home is the RF laboratory. There are two kinds of instrument, namely the Scalar and the Vector Network Analysers (SNA and VNA, respectively). The SNA can only display the magnitude of some property (e.g. reflections or transmission coefficient) versus frequency. It basically consists of a tunable RF generator and a power detector. Often SPAs with a tracking generator (tracking to the actual frequency shown in the display) are used for scalar network analysis. However, with its additional phase display capability and much higher dynamic range with a simple power detector, the VNA offers a much wider range of applications than the SNA. The higher dynamic range (around 100 dB) is also linked to the fact that VNAs apply the mixing concept mentioned above. We are only using VNAs in this tutorial, but if a spectrum analyser with tracking generator is available one may also try using it as a scalar analyser. A typical RF laboratory application of a VNA is the measurement of the scattering parameters (S-parameters) of some device under test (DUT). In this tutorial we will perform most of the standard measurements on objects such as cavities, complex impedances including non-ideal resistors,

capacitors and inductors, filters and amplifiers. It will also be possible to do some 'fun' experiments (not described in this paper), e.g. using a coaxial waveguide transition as an antenna and building a simple radar system.

2. EXPERIMENTS WITH THE VECTOR NETWORK ANALYSER

2.1 General information about the VNAs used here

We use three different models of the HP8753 VNA, namely types B, C, and E. All have similar functions and frequency, power, and dynamic range, but use different software and have slightly different control menus. So, for example, the desired output power of the HP8753C is obtained by applying an external attenuator to an output power from –10 dB to +20 dB. For the HP8753D an output power range is selected (which internally switches attenuators) and then the desired output power is chosen. However, the instructions given here are also 'in principle' applicable to comparable instruments from other manufacturers.

The HP8753 has two kinds of control buttons: hard keys on the front panel, which are used to set the kind of measurement, the frequency range and other functions, and soft keys on the right-hand side of the display. In the following descriptions bold letters are frequently used to denote hard keys while normal letters denote soft keys: **MEAS,** REFL., FWD S_{11} (A/R), etc.

The HP8753 analysers have two display channels, which can be displayed in parallel in two modes: split screen and common screen. Each channel can be processed using mathematical functions, memory, data/memory display and other features.

Only one channel at a time responds to control inputs. Changes to some settings act only on the active channel, whereas frequency range and other principal parameters remain unchanged for both channels. For critical measurements the noise level can be improved either by averaging or reducing the IF bandwidth or by increasing the output power to the maximum allowed level. The VNAs have powerful marker functions to simplify search operations such as 3 dB bandwidths, maxima and minima and desired power levels. These can be combined with special mathematical functions to allow the calculation of the *Q*-factor of a resonator, for example.

All analysers used in the tutorial have a tome-domain option, which can be found under **SYSTEM**, TRANSFORM MENU. A frequency range up to 2 GHz with 801 data points is selected for these measurements and **FORMAT,** MORE, REAL is used to adjust the display format. Note that the frequency range is automatically adapted to the values required for the transformation (chirp–Z type) when using the 'low pass' time-domain mode.

The 'power-sweep' function for a fixed Continuous Wave (CW) frequency can be applied for measuring the 1 dB compression point of an amplifier.

All new measurements are started with a **PRESET** of the VNA, and the desired type of measurements (S_{11}, S_{21}, or other) are to be keyed in. Then the frequency range is set and the measurement is calibrated with the proper choice of the calibration kit, for example N 50 Ω or 3.5 mm for SMA, depending on the connector required to connect the DUT.

2.2 Becoming acquainted with the HP 8753 (first VNA lesson)

First of all, it is very important to define the kind of measurement (e.g. reflection, transmission) to be performed.

The VNA provides the following basic choices:

MEAS S_{11} (A/R) REFL.: FWD	for impedance, VSWR, etc. at port 1.
S_{21} (B/R) TRANS.: FWD	for attenuation or gain, phase, group delay etc. between ports 1 and 2 (not from 1 to 2!).

S_{12} (A/R) TRANS.: REV — for attenuation, gain etc. between ports 2 and 1.

S_{22} (B/R) REFL.: REV — for impedance, VSWR, etc. at port 2.

The notation A/R and B/R refers to the coaxial connections between the S-parameter test set and the actual VNA. These connections are externally accessible for certain instruments and permit special configurations. Here the letter A stands for the input to channel A and similarly B denotes the input of channel B. R is the reference channel. The instrument will essentially display (in complex formats such as real and imaginary or magnitude and phase) the (complex) ratios A/R or B/R. The S-parameter test basically set contains two directional couplers and several remotely controlled RF relays in order to establish the required connections between the two S-parameter ports (port 1 and port 2 at the front panel) and the connections A, B, R, and signal source. FWD stands for 'forward' measurement and REV for 'reverse'.

The desired frequency range must be set once the type of measurement has been defined. The analysers start with their full range sweep, which is likely to be too high for measuring lumped elements like resistors and capacitors.

The following possibilities are available for setting frequencies:

either

START and

STOP

or

CENTER and

SPAN

Having set the frequency range, the influence of the mismatch in the connecting cables between the ports of the VNA and the DUT should be eliminated, as well as the impact of generator mismatch and finite directivity of internal couplers of the VNA. In addition, the reference plane will be moved to the end of the cable. In other words, the typically applied application for a reflection measurement (open, short, load) eliminates errors resulting from generator and cable mismatch, electrical length and losses, as well as finite coupler directivity. Of course the imperfections mentioned above cannot be suppressed by themselves, but by inferring the results from the calibration measurements a software routine allows for numerical correction.

The calibration procedures will be described in more detail in the first measurements, but one should always remember that *the calibration can only de done AFTER having defined the frequency range* (except for the interpolation mode when using frequency ranges smaller than the calibrated range)

– As an exercise, **PRESET** the instrument (push the green button on the HP8753).

  The VNA starts with full range sweep, CH1 active in S_{11} mode with logarithmic magnitude (log mag) and reference level at 0 dB (REF). Dial a frequency span from 0.5 to 5 MHz using the Start and Stop controls.

  Set start frequency to 0.5 MHz — **START; 0.5; M/u.**

  Set stop frequency to 5 MHz — **STOP; 5; M/u.**

– Calibrate S_{11} (port 1) with the reference plane at the end of the cable. Typical RF cables are equipped with M-type (male) N-connectors on either end. Hence components in the calibration kit have F-type (female) N-connectors. During the calibration procedure a question will appear on the screen of the instrument asking which component of the calibration kit is being used. It should be

remembered that the kind of calibration is defined for the type of connector you are doing the calibration on (i.e. the end of the cable), NOT for the calibration-kit component itself!

By the way, the term N for the type of cable connector we are using here has its historical roots in the word navy, since this kind of connector was first used by the US Navy more than 50 years ago. The frequently applied competitor, the BNC connector, comes from the same shop (Bayonet Navy Connector). But the BNC has much lower performance than the N-connector.

- Now follow the menu on the calibration page and attach the open, short, and load as requested by the system.
- Don't forget to confirm that the calibration is done; you may want to store it for later use.
- For very critical and accurate measurements you may calibrate the system using the trace average function, but this is rather time consuming. However, you may reduce the IF bandwidth from 3 kHz (standard setting) to 100 Hz (in the average menu) and note the difference.
- As a next step select in the display menu the split screen display and then activate channel 2. For channel 2 we are measuring now in transmission from port 1 to port 2, i.e. S_{21}. Connect the end of the cable where you just did the S_{11} calibration to port 2 of the VNA.
- Carry out a 'response' calibration for transmission (not full two port) and store the result.

Now you are well prepared to measure simultaneously some DUT in reflection and transmission with a calibrated system. As a DUT you can use the 10 Ω resistor in a SUCO (blue) box. This DUT contains a simple 10 Ω carbon resistor between the inner conductors of the input and output connector. You can check now up to what frequency this test object looks like a pure resistor and what kind of parasitic effects occur. Despite the fact that the geometrical length of the DUT is just a few centimetres you will already see a considerable inductive component at 5 MHz. The free space wavelength of 5 MHz amounts to 60 m! For the next exercise disconnect the cable from port 2, attach the DUT and terminate the open connector of the DUT with a short (from the calibration kit).

- Using channel 1, select as display format the Smith Chart and look at the read-out of the marker on top of the screen: it gives a reading of the resistance and inductance of the lumped resistor. Which read-out changes significantly with frequency? Discuss the results.
- Exchange the blue test box for another one, look at the results in different formats, and turn the dial to measure at different frequencies.
- Now you may try a further method to measure the frequency-dependent complex impedance of the 10 Ω test box (leaving the DUT connected to the cable to port 1 with a short at its end). The instrument provides a conversion menu that allows direct display from the S_{11} measurement of the real and imaginary parts of the DUT's complex impedance as a function of frequency. This kind of conversion is possible from both the reflection and the transmission measurements. Try both techniques and discuss the reason for possible discrepancies. Play with the electrical-delay correction in the transmission-type test. How can you explain the (of course unphysical) negative real parts for certain electrical-delay settings?

2.3 Demonstration of calibration effectiveness

- Display the locus of S_{11} of a 25 Ω DUT in the Smith Chart for the frequency range 600–1200 MHz after having done a reflection calibration. This 25 Ω DUT may be made by using two 50 Ω terminations and a coaxial T-piece to connect them 'in parallel'.
- Now produce a severe generator mismatch with another coaxial T-piece inserted at port 1, between port 1 and the coaxial cable. The open port of this T-piece will be terminated with another 50 Ω load. In this configuration we have artificially modified the generator impedance from about 50 Ω to around 25 Ω.

– As a next step perform the usual open, short, load calibration at the end of your (say 1 m long) test cable. Reconnect the 25 Ω DUT at the end and display its characteristic in the Smith Chart. You should not now be able to see any significant difference from the result found in the first step.

– Then insert the triple-stub tuner between the end of the calibrated test cable and the 25 Ω load and by adjusting the length of the three stubs (cut and try method) try to achieve an impedance match ($S_{11} = 0$) around 900 MHz.

Sometimes one is not really sure whether a certain calibration procedure, in particular in reflection, has been done well. There are many possible ways of making a mistake and an independent cross-check is desirable. A rather sensitive test consists of connecting a short piece of coaxial cable with its end left open or shorted and then displaying the reflection coefficient in the Smith Chart. In case of a poor calibration you may find the locus of S_{11} exceeding the boundary of the Smith Chart, which would only be valid for an active element with a reflection coefficient larger than unity. For a passive device this kind of negative resistance response is impossible and indicates a calibration or display error. However, in the past reflection amplifiers were used for special applications (certain parametric amplifiers or negative resistance devices containing tunnel diodes).

2.4 Time-domain measurements

When using the time-domain option of the VNA the 'low-pass' or the 'band-pass' modes are available. The low-pass mode can only be used for equidistant sampling in the frequency domain (equidistant with respect to DC), since the Fourier transform of a repetitive sequence of pulses has a line spectrum with equidistant spacing of the lines including the frequency zero. This implies that for a given frequency range and number of data points the instrument must first work out the exact frequencies for the low-pass mode (by using the soft key: set frequency Low-pass). Once these frequencies are defined, calibration can be applied. For a linear time-invariant system, frequency- and time-domain measurements are basically completely equivalent (excepting signal-to-noise ratio issues) and may be translated mutually via the Fourier transform. Note that the Fourier transform of a spectrum with constant density over a given frequency range (rectangular spectrum) has a sin(f)/f characteristic in the time domain. This characteristic shows undesired 'side-lobes' and thus an (amplitude) weighting function (= window) is applied in the frequency domain before entering the FFT. This weighting function is typically $\sin^2$ or Gaussian and helps to strongly suppress side-lobes in the time domain. Within the low-pass mode we can use the pulse and step function. The step function is nothing other than the integral over the pulse response. When using the gating function, keep in mind that gating is a non-linear operation and thus may artificially generate frequency components that were not present before gating. In the band-pass mode the spectral lines (frequency-domain data points) need no longer be equidistant with respect to DC but just within the frequency range of interest. The corresponding time-domain response for the same bandwidth is twice as long as in the low-pass mode and we get, in general, complex signals in the time domain. These complex signals are equivalent to the I and Q signals (I being in phase and Q quadrature) often found in complex mixer terminology. They can be directly displayed using soft keys 'real' or 'imaginary' in the format menu. The real part is equivalent to what one would see on a fast scope, i.e. an RF signal with a Gaussian envelope. The meaning of the time-domain band-pass mode response in linear magnitude format is the 'modulus of the complex envelope [SQRT{(re^2(t)+$imag^2$(t)}] of a carrier modulated signal'. Note that the time-domain mode can also be applied for CW excitation from the VNA, but to analyse a slowly time-variant response of the DUT (up to the IF bandwidth of 3 kHz).

To start with the time-domain option, follow these instructions.

– Preset the instrument, dial a frequency range of 300 kHz to 3 GHz (801 data points) and go into the time-domain, low-pass mode, step function. By this operation the VNA sets a frequency range, which is required for the low-pass mode Fourier-transform calculation. Check the frequency reading.

– Now you have to calibrate S_{11} as you did before, with the only difference being that you have to use the 'OPEN' from the calibration kit, as you are now measuring up to about 3 GHz. (Refer to the above descriptions of calibrating S_{11} if you do not remember.)

– Read out the pulse amplitude with the end connector of the cable open.

– Now connect a SHORT and read the signal.

– Discuss the meaning of the sign from the read-out for the reflected wave.

– Connect the 10 Ω test box with a SHORT at the end to the THRU.

  Can you calculate the resistance from the read-out?

  Remember the definition of S_{11} and which simple formula you have to apply: $\rho = (Z - Z_C)/(Z + Z_C)$; this is the reflection coefficient as seen in the reference plane of the DUT. Z_C stands for the characteristic impedance of the cable and usually amounts to 50 Ω. *Note that Γ and r can also be used to denote the reflection coefficient.*

– Repeat the experiment with the 100 Ω blue box using a SHORT at the end.

  Look at the 12(18) pF and 100 pF capacitor boxes (end = open). Discuss the traces and remember that it is a single-step response. The two numbers (12 and 18) for the capacitance indicate that a 12 pF capacitor mounted inside the test box returns a total capacity of 18 pF because of the connector feed-throughs and other parasitic capacitances.

– Now apply a 25 Ω DUT in the calibrated reference plane instead of the capacitor, using two 50 Ω loads connected in parallel via a coaxial T-piece.

– Use an appropriate vertical scale factor to obtain a good resolution on the screen.

– Store the result in the memory and display memory and data.

– Put a SHORT on the end of the calibrated cable (instead of the previously used 25 Ω) and compare the data trace with the memory trace.

– Plot the results (if a plotter is available) and discuss them.

– You may also try building a simple notch filter by attaching the T-piece to your calibrated reference plane and a 50 Ω load to one connector of the T-piece and an open (and later shorted) stub (say about 1 m of cable) to the other connector of the T.

– If you are interested in going further, repeat all the time-domain measurements mentioned above in the pulse mode (low-pass) instead of the step mode.

There is a wide range of applications for this synthetic pulse time-domain technique. A VNA in the time-domain, low-pass step mode has a very similar range of applications to a sampling scope (Figs. 1 and 2). However, it must always be kept in mind that carrying out a measurement in the frequency domain and then going via FFT or similar into the time domain implies strict linearity of the DUT. Thus a transient on a non-linear system, such as the onset of oscillations on some microwave oscillator with active elements after turn-on of the supply voltage, would not return meaningful results when using the technique mentioned above. The dynamic range of a typical sampling scope is limited at about 60 dB with a maximum input signal of 1 V and a noise floor around 1 mV. The VNA can easily go beyond 100 dB for the same maximum level of the input signal of about +10 dBm. Both instruments are using basically the same kind of detector, either a balanced mixer (four diodes) or the sampling head (two or four diodes), but the essential difference is the noise floor and the average power arriving at the receiver. In the case of the VNA we have a CW signal with bandwidth of a few hertz and thus can obtain with appropriate filtering a very good signal-to-noise ratio, since the thermal noise floor is –174 dBm/Hz. For the sampling scope we get a short pulse with a rather low repetition rate (typically around 100 kHz) and all the energy is spread over the full frequency range (typically 20 GHz bandwidth). With this low average power (around 1 μW) the spectral density is orders of

magnitude lower than in the case of the VNA and this ultimately makes the large difference in dynamic range (even without gain switching). Furthermore, the VNA permits a wide range of band-limited RF pulses to be tailored in the band pass, which would be very tedious with a sampling scope.

2.5 Amplifier measurements (second VNA lesson)

The description here is for the HP8753C. If you use another instrument you may have to adapt certain measurement parameters.

Whenever you measure medium- or high-power amplifiers, be sure that the power level cannot destroy the input of the VNA. For example, even measuring the input impedance of an amplifier may destroy the VNA, if the amplifier produces parasitic (self) oscillations.

– Preset the VNA. To protect the VNA against overload from the amplifier output, start with the following set-ups:

  Output power: –10 dBm

  Attenuator port 1: 20 dB; this leads to an input power of –30 dBm for the amplifier.

  Never remove the fixed 30 dB attenuator from the output of the amplifier. Assume that its attenuation is exactly 30.0 dB, constant over the complete frequency range.

– Carry out a response calibration in S_{21}.

– Measure the transmission coefficient from port 1 to port 2 ==> S_{21} (B/R) not S_{12}!

– Display the response using the auto-scale function; select trace averaging with an averaging factor of 10; measure the response (in this case the isolation) at 1 GHz; produce a hard copy (plotter) of the screen.

– Reduce the IF bandwidth (IF-BW) to 100 Hz; compare the result with the above hard copy.

– Go back to IF-BW = 3000 Hz and connect or turn on the DC voltage of 15 V to the amplifier. Use the full-screen display for a single channel; determine the small signal gain of the amplifier at 1 GHz; with marker1. Produce a hard copy.

– Measure the 3 dB bandwidth (and also 1 dB BW) of the amplifier (determine the frequencies where the gain is 1 dB lower than at 1 GHz) using marker functions. Use two markers and the statistic function. Produce a hard copy!

  Test other possibilities to determine the 1 dB and 3 dB bandwidth of this amplifier:

– Determine the 1 dB compression point of the amplifier at 1 GHz. Go via MENU to CW frequency and power sweep. Use marker functions. Produce a hard copy!

– Measure the frequency range over which the gain compression of 1 dB occurs, first in the specified frequency range of the amplifier using the CW mode at the lower, mid-band and upper frequency of the amplifier. Apply the display features of the 8753 for this measurement and the marker search function. Produce a hard copy!

– Now return to the frequency sweep display, set a suitable range (response calibration) to cover the 3 dB bandwidth of the amplifier and store the response trace for a small input signal (small enough that the output power is well below the 1 dB compression point). Select a vertical resolution of 1 dB/div and an adequate reference level to position the trace approximately in the middle of the screen.

– Now increase the input power until the read-out is approximately 1 dB below the previous trace. Note that the 1 dB compression level is frequency dependent. Copy the result.

What we see on a network analyzer

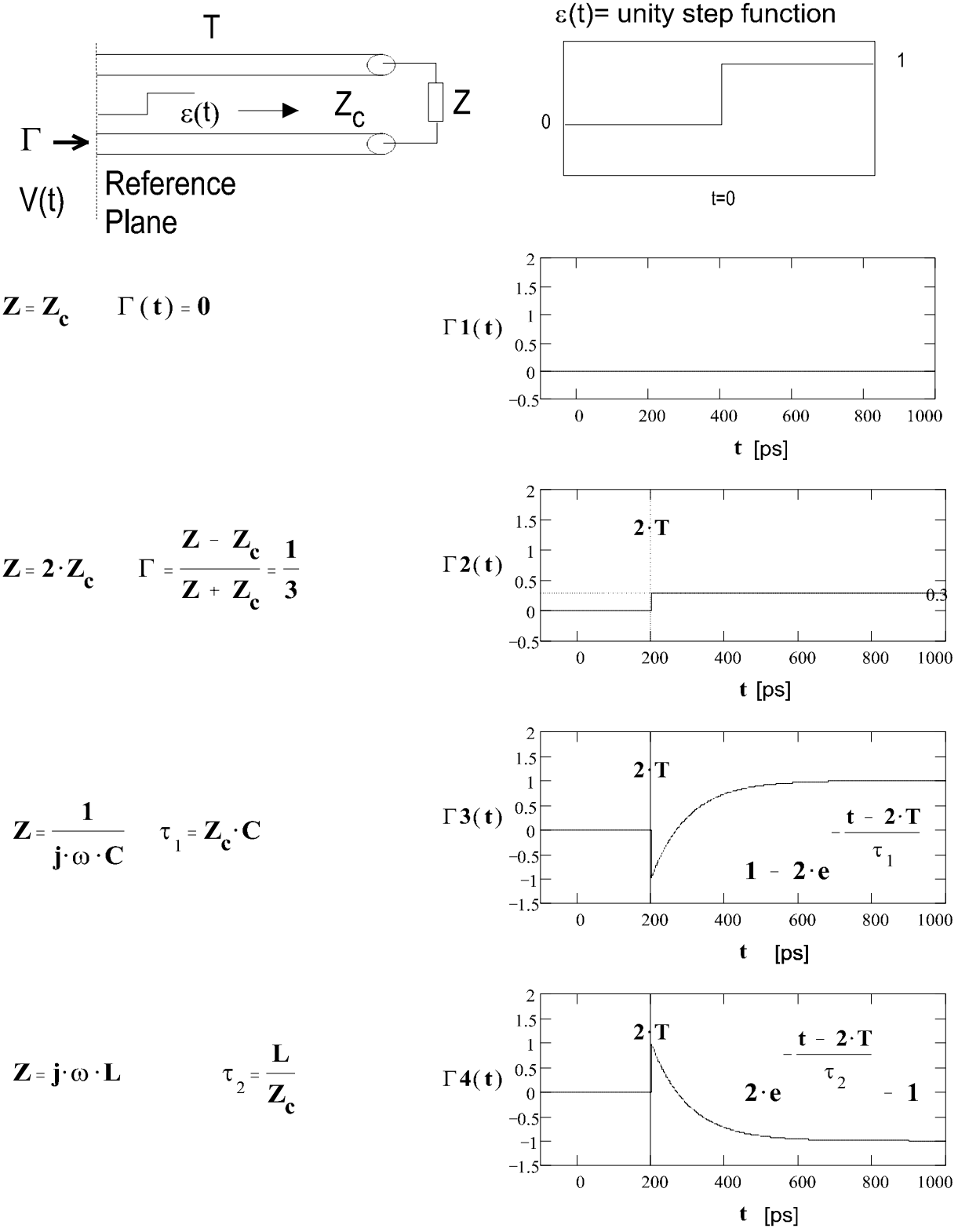


Fig. 1: Step-function response for different terminations on a VNA

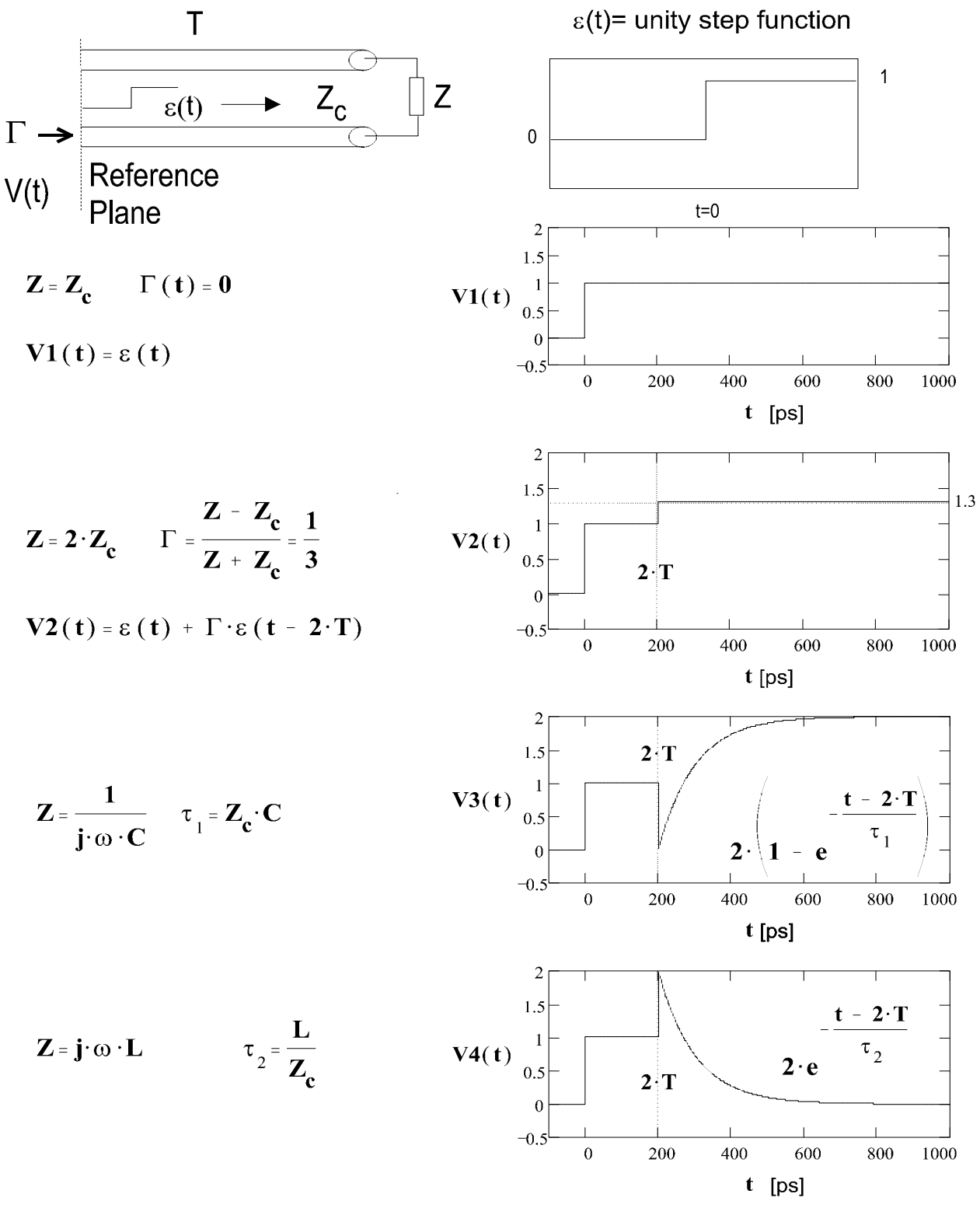


Fig. 2: Step-function response for different terminations on a sampling scope (through sampler)

Preset the VNA, set the power and attenuator of port 1 as at the beginning, set the frequency range and carry out an S_{11} one PORT calibration. Measure the Standing Wave Ratio (SWR) of the input of the amplifier in the specified frequency range. Determine the maximum SWR or Voltage Standing Wave ratio (VSWR).

Measure the deviation from linear phase and the group delay in transmission.

2.6 Directional couplers and cavities (third VNA lesson)

Now you have the choice of carrying out measurements on directional couplers, on cavities, or on a transmission line over a ground-plane (with an EMC probe).

In the directional-coupler section you will become familiar with the working principle of loop couplers, how to perform a precision calibration and how to measure the relevant parameters of different couplers.

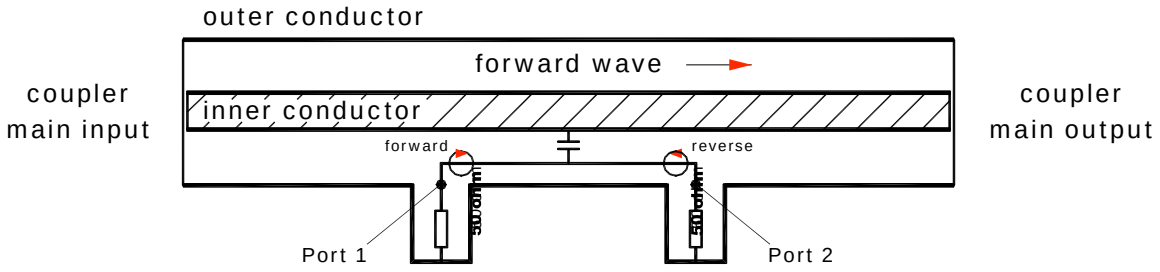
Cavity measurements include different ways to determine the loaded and unloaded Q of a cavity; tuning the coupling loop to critical coupling; measurements of the resonance frequencies of a pillbox for different modes in comparison with MAFIA calculations, and perturbation measurements.

With the EMC probe you can measure the effectiveness of RF shielding and the standing waves on a wire over a ground-plane. Decide yourself whether you will do a little of each or focus on your particular points of interest.

2.6.1 Directional couplers

The principle of a directional loop coupler is very simple. The capacitive coupling and the inductive coupling of the loop should be tuned to have exactly the same magnitude at the outputs of the coupling loop, terminated with 50 Ω.

From the capacitive part of the coupling we get equal output voltage at either end of the loop. The induced current, however, flows through the loop and leads to a positive voltage (with respect to ground at this location) at one end and a negative voltage at the other (Fig. 3).



Port 1		Port 2	
forward:	+ Ucap	forward:	+ Ucap
	+ Uind		- Uind
	= 2 Ucoupl. forw.		= 0 Ucoupl. forw.
reverse:	+ Ucap	reverse:	+ Ucap
	- Uind		+ Uind
	= 0 Ucoupl. rev.		= 2 Ucoupl. rev.

Fig. 3: Construction principle of a directional coupler

At the positive end (port 1), capacitively coupled and induced voltages combine, while on port 2 they cancel each other. This is valid for an incident (from the coupler main line input) wave in the forward direction.

Let us say that port 1 of the coupler then measures a proportional part of a wave in the forward direction, while port 2 does not see it. For a wave in the reverse direction it is just the opposite, as the

induced current in the loop flows in the opposite direction to the capacitively coupled part. Thus a backward-moving wave is proportionally measured on port 2, while port 1 does not see it and the directional coupler is ready.

The directivity is a measure of how close capacitive and inductive coupling match in magnitude and also in phase (or 180 degrees offset). For perfect balance the directivity is infinite. In this case you measure only the forward wave on port 1 and only the reflected wave on port 2.

In practice, you will always have a finite directivity, which implies that you measure a (small) part of the forward wave on port 2 and vice versa. The directivity of a directional coupler in your amplifiers should be better than 30 dB at the centre frequency.

Note that the 50 Ω terminations are also required to have a matching better than –30 dB as they directly influence the effective directivity!

The measurement error of the coupling coefficient should be below ±0.1 dB.

As directional couplers of high-power amplifiers have coupling coefficients in the range from –30 to –60 dB, it is not easy to calibrate these levels to 0.1 dB to absolute accuracy, as the strongly different input levels of the VNA have an impact on the uncertainty. Therefore it is necessary to calibrate the response with a known attenuator in the vicinity of the coupling coefficient.

Thus, for example, if you do not have attenuators calibrated in accordance with the Physikalisch Technische Bundesanstalt (PTB) in Germany or the National Institute of Standards and Technology (NIST) in the US, with a certificated attenuation at the desired frequency, you can help yourself by measuring attenuators up to 10 dB directly and add them until you get close to the coupling coefficient.

2.6.1.1 Precision calibration of a 108.406 MHz coupler with –30 dB coupling coefficient

(This procedure also applies, of course, for other adjustable, narrow-band couplers at different frequencies.)

– Preset the VNA. For all precision measurements the VNA should have had a warm-up time of at least 10 minutes, but two hours is better.

– Set the CW frequency to the centre frequency of the coupler.

– Disconnect the short cables from ports A and B of the analyser (not from the S-parameter test set).

This procedure allows simultaneous reading of the coupling AND the directivity when changing the orientation of the coupling loop. It is applicable to VNAs where the connections between the S-parameter test and the VNA unit are accessible. However, there also exist instruments that have two receiver input ports and thus allow true three-port measurements.

– Select input ports A/R for CH1 (channel 1) and B/R for CH2 (channel 2). (You have to monitor coupling and directivity at the same time.)

– Display: Dual channel on, split screen off.

The following section on the calibration of attenuators may be skipped or postponed if time is short.

– Do a response calibration for CH1 (A/R), Cal Type N 50 Ω, with port 1 of the S-parameter test set connected directly to input A of the VNA unit.

– Now insert (only one at a time) the three 10 dB attenuators into this signal path. Measure the attenuation of the three 10 dB attenuators and add the measured values numerically.

– Measure the attenuation of all three attenuators mounted together in the same signal path.

– Measure the attenuation of all three attenuators together with 0 dBm and +10 dBm output power. You will notice minor changes in the readings as a function of the generator power. Try to explain that effect.

– Repeat the last four steps, but now using channel CH2 (B/R).

– Average the result of the two channels (i.e. find the mean value between both channels). The purpose of this procedure is to get the maximum precision possible.

– Measure the loss between coupler input and output (main line) and note the value.

– Connect all three attenuators to the output of the coupler.

– Menu power: +20 dBm.

– Calibrate CH1 and CH2 (response) with the coupler and the three attenuators inserted. You may wish to use the average function for highest precision.

– Which reading from the VNA do you expect to give exactly –30.0 dB on your DUT?

Now we are ready for the actual tuning of the coupler.

– Connect the VNA input A to port 1 of one of the two coupling loops of the coupler.

– Connect the VNA input B to port 2 of the same coupling loop.

– Remove the three 10 dB attenuators at the output (main line of the coupler) and terminate this output with the 50 Ω load from the calibration kit.

– Turn the average off in both channels and set the REFERENCE POSITION to 9 vertical units.

– Now rotate slowly the coupling loop until you get the desired reading of _____dB at CH1.

– The reading on channel CH2 gives the directivity. (It is not worth correcting the small difference of the reference to –30.0 dB.)

– Do the same for the other coupling loop.

– Reconnect input A to the S-parameter test set.

– Calibrate S_{11} in the frequency range 20 MHz to 220 MHz (801 points).

– Measure the input matching of the coupler at 108.4 MHz using the termination resistor of the calibration kit at the output. Write down the matching in dB. How is the matching at other frequencies?

– Measure the coupling from 20 MHz to 220 MHz; note that the coupling is increasing by approximately 6 dB per octave. Can you explain why? Measure the remaining couplers (if time allows).

2.6.2 Cavities

We have three cavities available in this course: two pillbox cavities (cylindrical resonators) and one coaxial cavity.

The first pillbox cavity has a diameter of 30 cm; the second has a diameter of 31.5 cm, and its length is variable. The coaxial resonator is short-circuited at one end, while the capacitively loaded open end is variable at the location of the capacitor plates. This resonator was built to determine the dielectric losses of isolators.

2.6.1.2 Pillbox measurements

– PRESET the instrument. Set the frequency to between 500 MHz and 1.2 GHz. (There is no resonance below 500 MHz and there is no time available here to look for higher modes above 1.2 GHz.)

– Set the number of points to 1601 (to have enough data points at a high Q-resonance).

– Choose two small probes with N-connectors, look at how they are constructed.

– Calibrate (N) S_{11} on CH1 and S_{21} on CH2.

– Set the reference position of both channels to 9 vertical units.

– Connect the cables to the two installed inductive or capacitive probes and determine the frequency of all resonances you can find at S_{21} in the selected frequency range. (Display only CH2.)

– Calculate the length (actual length for pillbox 2) using the given mode patterns in your handouts and determine the type of the modes. Compare with results obtained using Figs. 4 and 5.

– Set the marker on the peak of the E_{010} mode (TM_{010}) and adjust it to the centre of the screen (MARKER => CENTER). For pillbox 2 (with variable length), the E_{010} mode is that which does not change its frequency if the length of the cavity is altered.

– Reduce the SCALE/DIV stepwise to 0.5 dB/DIV and the SPAN to fill the screen with the resonance curve. You may use the marker functions MARKER SEARCH MAX, MARKER => CENTER and MARKER => REFERENCE to do this.

– Then press MKR; MKR ZERO to define the maximum of the resonance curve with, as reference, the marker read-out about zero.

  Now you can use the width function to automatically get the –3 dB points. (Marker search, width on, width value –3 dB. You measure the nearly unloaded Q if both of the used probes are so small [noise level] that you get the result only by setting the following values: output power to 20 dBm, IF bandwidth to about 300 Hz, and the averaging on. It is also useful to insert a low-noise amplifier in front of port 2.)

– Write down the centre frequency, 3 dB bandwidth and the measured Q.

– Now try to reach a critical coupling with a bigger loop or a capacitive pick-up. Use channel 1 with increased bandwidth (as the new loop may shift the resonance frequency). Go to 10 dB/DIV and display S_{11} in the formats LOG MAG and Smith Chart. Adjust a circle that goes through point 1 (50 Ω). Turn the circle with the function PHASE OFFSET to the left side of the screen until it is symmetric to the locus of real impedance (de-tuned short position). Measure the loaded and unloaded Q using the markers. The points where the real and imaginary parts are equal give the bandwidth for the unloaded Q. (See *RF cavity higher order mode measurements* by J. Byrd and R. Rimmer in the appendix given to students participating in this tutorial and *Basic concepts I and II* by Heino Henke in this Report). You can find these points in the de-tuned short position looking at the real and imaginary parts of the marker.

  The loaded Q can be found at the crossing point of the circle with ±45 degree lines starting at the zero point. This can be easily done on paper but not on the analyser screen. It helps to know that the loaded points are always (not only at critical coupling) the highest and the lowest points of the circle, if this is brought to the de-tuned short position. For critical coupling you will find the points 10 Ω (or 0.2 when normalized to 50 Ω) for the real part and ±20 Ω for the imaginary part.

  But there is an even easier method for reading out the loaded and unloaded Q, if you turn the circle with the phase offset (do not use the electrical delay as it would deform the circle) to the de-tuned open position. For critical coupling the circle lies directly on the circle of the Smith Chart, where the (normalized) real part is 1. Therefore you find the points of the unloaded Q at the crossing to the lines where the (normalized) imaginary part is 1 too (X = R) and the loaded Q at the crossing to the lines where the (normalized) imaginary part is 2 (X = R + 1).

  Make sure that the de-tuned short and de-tuned open positions are well adjusted by checking for symmetry of the maxima of the imaginary parts using appropriate marker functions.

– Determine the loaded Q also by the –3 dB points of S_{11} for the critical coupling in the format LOG MAG.

– Determine the loaded Q in transmission by the 3 dB points of S_{21} with one probe in critical coupling and the other very small.

– Move your coupling loop to over-critical and under-critical coupling and calculate the coupling coefficient from the formula k = 1/(2/D – 1). D is the diameter of the circle; its unit is the radius of the Smith Chart. If D = 1, then the coupling is 1 and you have critical coupling with the centre frequency point at the centre of the Smith Chart. Weakly coupled resonators have small Q-circles, strongly coupled have large ones.

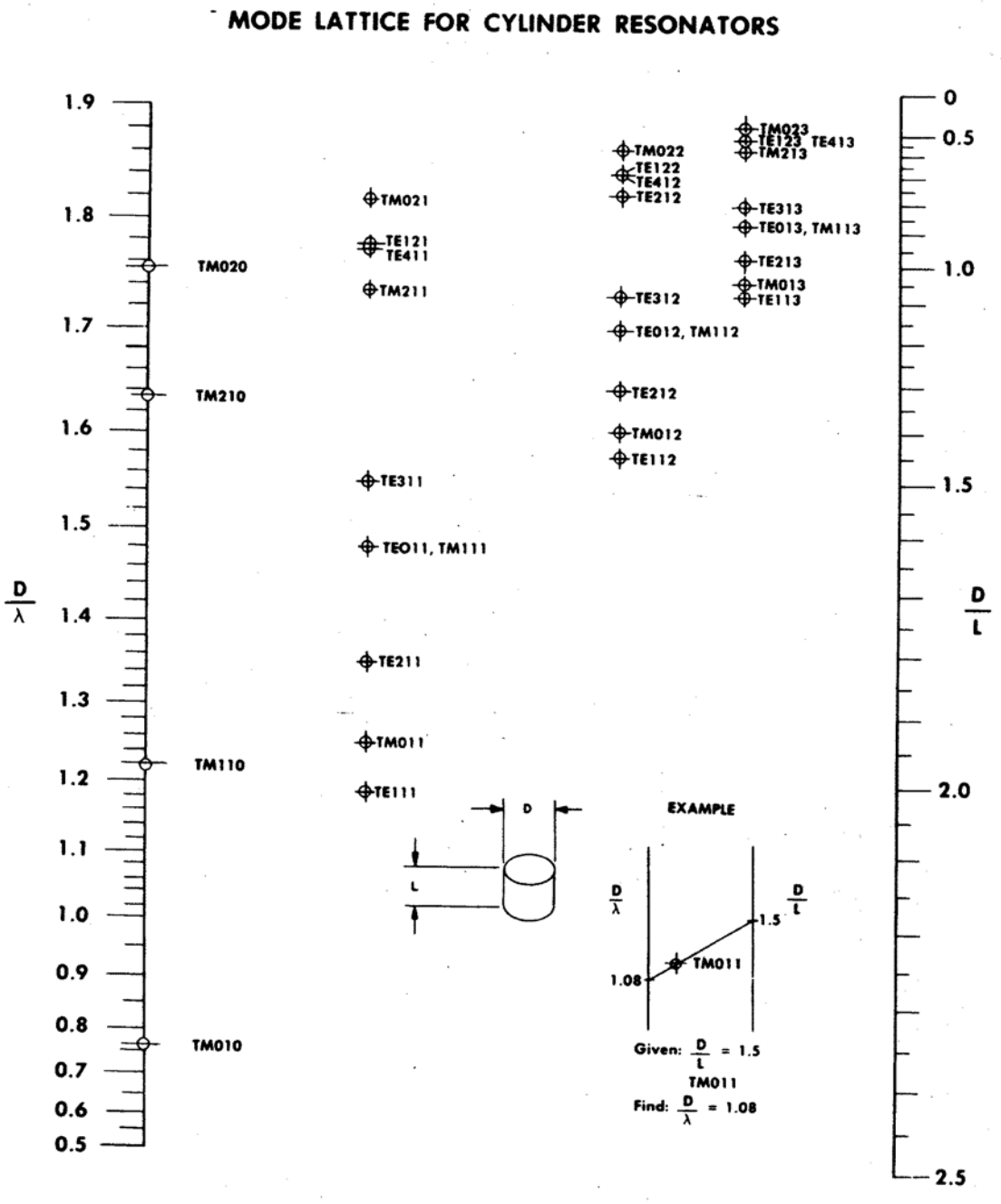


Fig. 4: Mode lattice for cylindrical resonators (reprinted from R.N. Bracewell, 'Charts for resonant frequencies of cavities', Proc. IRE, August 1947).

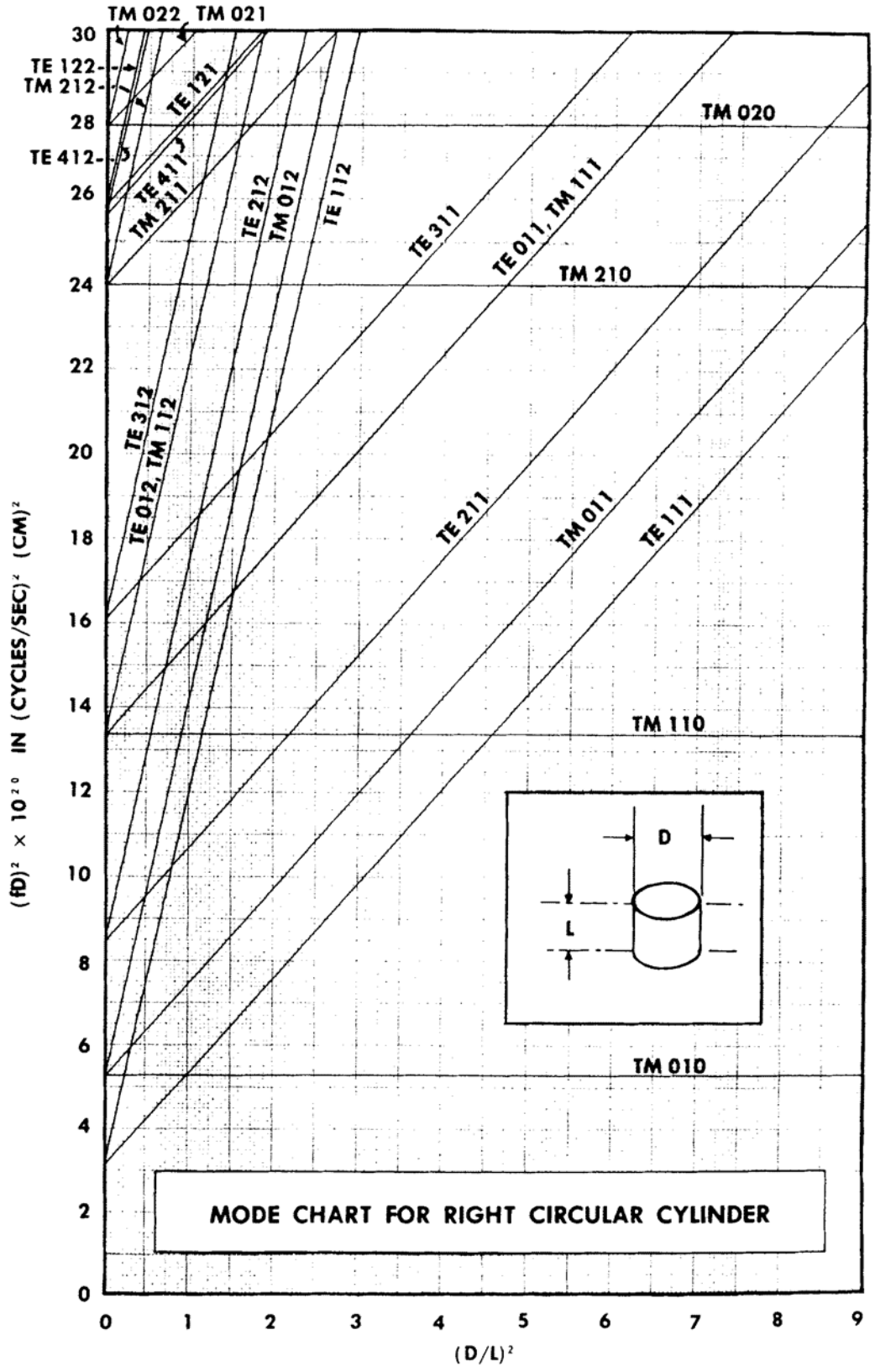


Fig. 5: Mode chart for a pillbox-type cavity (reprinted from *Microwave engineers handbook, Vol. 1*).

3. EXPERIMENTS WITH THE SPECTRUM ANALYSER

3.1 Becoming familiar with the spectrum analyser

– Display the spectrum of the signal coming from the 'CAL' output of the spectrum analyser (SPA).

– Display the spectrum of RF signals present in the classroom (using a short wire as an antenna).

– Measure the spectrum of an output signal from a signal generator (CW mode, no modulation) and look for second and third harmonics. How can you discriminate against SPA input mixer-related harmonics? (Do not exceed 0 dBm generator output power.)

– Measure the frequency response of an amplifier (scalar network analyser mode); if no tracking generator is available, measure 10 frequency points (100 MHz – 1 GHz) manually.

– Measure the 1 dB compression point (i.e. small signal gain reduction by 1 dB due to the beginning of saturation effects of the amplifier under test) at three different frequencies (low, mid-band, high).

– Measure the second-order intercept point (non-linear products at sum and difference frequency of the two input signals; both input signals (= tones) should have equal amplitude; select suitable frequencies for input signals in order to be able to display the sum and difference frequencies). Measure at three different amplitude levels; watch out for second and third order generator harmonics; you may use low-pass filters at the input.

– Measure the Third-Order Intercept (TOI) point (use two frequencies about 50 MHz apart). The IM3 products appear separated by the frequency difference from each tone. Use the automatic function (meas/user) button and TOI on (if available).

3.2 Noise and noise-figure measurements

– Make sure you are familiar with the most important functions of the SPA (frequency setting, Resolution Bandwidth, RBW, Video Bandwidth, VBW, amplitude scale). Note that the spectrum analyser should be used in the sample mode and not the usual peak detector mode.

– With resolution BW = 1 MHz, Start = 10 MHz, Stop = 1000 MHz, Video BW = 100 Hz, input attenuator = 0 dB display the baseline and read the power. How many dB is it above the thermal noise floor (thermal noise at 290 K = –174 dBm/Hz)?

– For the same settings, now connect the solid-state noise source (to be powered with +28 V DC via rear BNC connector) to the SPA input. The Excess Noise Ratio (ENR) of this device is close to 16 dB or a factor of 40 in spectral power density greater than the thermal noise of a common 50 Ω load. Use the table below to correct for absolute power reading. Note that for absolute power measurements with the spectrum analyser close to the noise floor the reading is too high by the amount indicated in the right column. The analyser should be set for this measurement in 'sample' mode and not in 'peak hold' mode, which may be a default setting.

– Load the noise measurement option software to the SPA (if available). Otherwise skip the next five points.

– Calibrate the SPA with the preamplifier. Record the measured noise figure of the system (SPA + preamplifier) from the reading on the CRT after calibration.

– Measure the gain and noise figure of some amplifiers. Convert noise figure into noise measure.

– Measure the gain and noise figure of some attenuators.

– Measure the noise figure of two amplifiers in cascade by the method already described. Measure the noise figure of an attenuator in the same way.

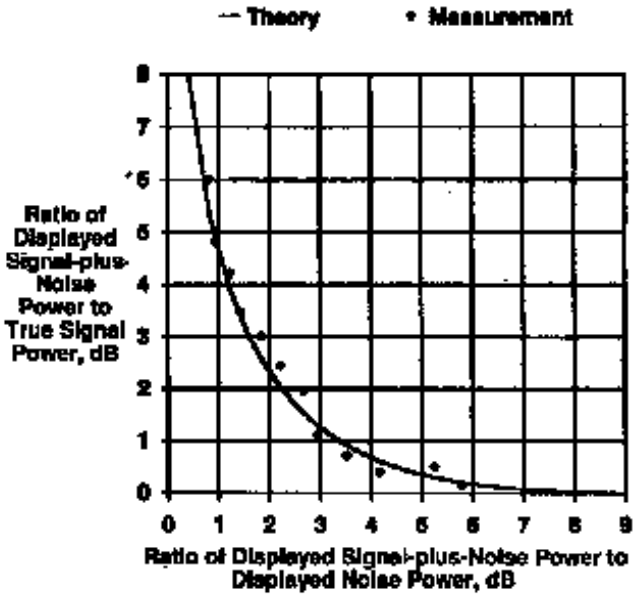


Fig. 6: Corrections for absolute noise power measurements on a spectrum analyser close to the instrument noise floor (from: A. Moulthrop, M. Muha, 'Accurate measurements of signals close to the noise floor on a spectrum analyser', IEEE Transactions MTT, Vol. 39, No. 11, 1991, pp. 1882–1884).

– Connect to the input of the preamplifier with a coaxial cable of 1–2 m length terminated by a short or open. Discuss the results observed (frequency range 10 MHz – 1 GHz). Also use a 50 Ω load or a triple stub tuner. Try to tune the 50 Ω input termination into an optimum noise source match using the triple stub tuner.

– If the noise measurement option is not available, connect a preamplifier to the SPA (input attenuator = 0 dB) and record the two traces 'noise source on' and 'noise source off'. Calculate from those traces the noise figure of the DUT. Convert noise figure into noise measure.

3.2.1 Some useful equations for noise-figure evaluation

There is frequently confusion over how to handle the dB (deci-Bel). The dB is used to describe a power ratio and thus is a dimensionless unit. As the power dissipated in a resistor is proportional to the square of the voltage or the square of the current, one may also take the ratio of these quantities into account. The dB is also used to describe absolute signal levels, but then there must be an additional letter to indicate which reference one refers to, e.g. +10 dBm (= 10 mW) is a power level of 10 dB above 1 mW (+20 dBm = 100 mW).

$$\alpha[dB] = 10 \cdot \log\left(\frac{P_1}{P_2}\right) = 20 \cdot \log\left(\frac{V_1}{V_2}\right)$$

$$10^{\frac{\alpha[dB]}{10}} = \frac{P_1}{P_2} \qquad 10^{\frac{\alpha[dB]}{20}} = \frac{V_1}{V_2}$$

The terms 'noise figure' and 'noise factor' are used to describe the poise properties of amplifiers. The term F is defined as signal-to-noise (power) ratio at the input of the DUT versus signal-to-noise power ratio at the output. F is always >1 for linear networks, i.e. the signal-to-noise ratio at the output of a two-port or four-pole is always more or less degraded. In other words, the DUT (which may be also an amplifier with a gain smaller than unity, i.e. an attenuator) always adds some of its own noise to the signal.

F [dB] is called the noise figure,

F [linear units of power ratio] is sometimes called noise factor,

F [dB] = 10 log F [linear units].

$$F[\text{linearunit}] = \frac{\text{ENR}[\text{linearunit}]}{Y[\text{linearunit}] - 1} = \frac{T_{ex}}{T_o \cdot (Y-1)} \quad \text{with} \quad T_{ex} = T_H - T_0 \ .$$

ENR is the excess noise ratio delivered by the noise diode and tells us how much 'warmer' than room temperature the noise diode appears. For an ENR of 16 dB this amounts to roughly a factor of 40 in power, or 40 × 300 = 12 000 K.

The quantity Y is the ratio of noise power densities measured on the SPA between the settings: 'noise source on' and 'noise source off'.

As shown in the equations below, the gain of the DUT can also be found from the two readings on the SPA. Thus one can simultaneously measure gain and noise figure.

The technique for noise-figure measurement described above is commonly used for noise-figure evaluation of an amplifier in the RF and microwave range for frequencies higher than about 10 MHz. For lower frequencies the characterization of noise properties is normally done in terms of specifying the (input) noise voltage and the (input) noise current of some amplifier, by placing a short between the input terminals or leaving the input open, respectively. Obviously one can define an optimum generator impedance (noise match) for which the combined effect of voltage and current noise is at minimum. This noise voltage and current may also be converted into an equivalent noise temperature. The noise temperature of an electronic amplifier at room temperature may be surprisingly low (depending on the technology applied) in the range 1 kHz to about 10 MHz: values below 1 K have been reported. This is not a contradiction to basic thermodynamic concepts, since a device like an amplifier (or also a forward-biased diode) connected to a power supply is no longer in thermo-dynamical equilibrium and may show equivalent noise temperatures well below its physical temperature. For the frequency range 100 MHz to about 10 GHz noise figures of 1 dB (= 70 K) and better are available over an octave bandwidth for non-cooled amplifiers (e.g. the low noise module of 12 GHz satellite receivers has a noise figure of around 1 dB and the antenna, looking into cold (3 K) space, has a noise temperature of about 30 K due to the losses of the atmosphere).

$$Y = \frac{measured\ \ DUT\ \ output\ \ power\ (density)\ \ with\ \ noise\ \ source = hot}{measured\ \ DUT\ \ output\ \ power\ (density)\ \ with\ \ noise\ \ source = cold}$$

$$ENR\ [linear\ \ unit] = \frac{(T_H - T_0)}{T_0}$$

$$ENR\ [dB] = 10 \cdot \log\left(\frac{(T_H - T_0)}{T_0}\right)$$

$$G_{(DUT)}\ [lin] = \frac{N(SPA + DUT,\ Diode\ \ on)\ [lin] - N(SPA + DUT,\ Diode\ \ off)\ [lin]}{N(SPA,\ Diode\ \ on)\ [lin] - N(SPA,\ Diode\ \ off)\ [lin]}$$

N = noise power measured on the SPA for, e.g., 1 MHz resolution bandwidth

$$F_{total}\ [linear\ \ units] = F_1[linear\ \ units] + \frac{F_2[linear\ \ units] - 1}{G_1[linear\ \ units]} + \cdots\cdots$$

ACKNOWLEDGEMENTS

The authors would like to thank in particular Agilent (Frankfurt and Meyrin) for generously providing a large number of test instruments as well as detailed documentation for the participants of the course. The smooth co-operation with both the CERN and GSI management permitted the authors to present a considerable number of test and demonstration objects.

BIBLIOGRAPHY

There is a large amount of very useful (and free!) information available on the Web, such as application notes by instrument manufacturers (Anritsu, Agilent (Hewlett-Packard), Rhode & Schwarz, Marconi, IFR, Tektronix, and many others). These application notes are usually very easy to read and give a lot of practical hints and examples after a brief theoretical introduction.

Apart from these application notes a few books are listed below that may also be helpful to gain a deeper insight to all questions of RF spectrum, network and noise measurements. Since they are not quoted in the text, no reference numbers are added (references for the figures are cited in full in the figure caption).

R.A. Witte, *Spectrum and Network Measurements* (Prentice Hall PTR, Eaglewood Cliffs, 1993, ISBN 0-13-030800-5).

T.S. Laverghetta, *Modern Microwave Measurements and Techniques* (Artech House, Norwood, MA, 1988, ISBN 0-89006-307-9).

G.H. Bryant, *Principles of Microwave Measurement* (Peter Peregrinus, 1988, ISBN 0-86341-296-3).

W.D. Schleifer, *Hochfrequenz- und Mikrowellen-Messtechnik in der Praxis* (Hüthig Verlag Heidelberg, 1981, ISBN 3-7785-0675-7).

P.C.L. Yip, *High frequency circuit design and measurements* (Chapman and Hall, London, 1990, ISBN 0-412-34160-3).

B. Schiek, *Mess-Systeme der Hochfrequenztechnik* (Hüthig Verlag, Heidelberg, 1984, ISBN 3-7785-1045-2).

J.M. Byrd, F. Caspers, 'Spectrum and network analysers', CERN-PS-99-003-RF, CERN Geneva, 1999. Also in S. Kurokawa, S.Y. Lee, E. Perevedentsev, and S. Turner (eds.), Joint US–CERN–Japan–Russia Particle Accelerators School on Beam Measurement, Montreux, 1998 (World Scientific, Singapore, 1999), pp. 703–722.

S. Turner, (ed.) 'CERN Accelerator School, RF engineering for particle accelerators', Oxford, 1991, CERN 92-03, CERN, Geneva, 1992.

SUPERCONDUCTING ELECTRON LINEAR ACCELERATORS AND RECIRCULATING LINACS

H.-D. Gräf and A. Richter
Darmstadt Technical University, Darmstadt, Germany

Abstract
As LEP reached the high-energy limit for electron–positron storage rings it was realized that linear accelerators, so-called linear colliders, would be the next generation of accelerators in high-energy physics. In nuclear physics the scheme of recirculating linacs dominates the scenario of accelerators producing continuous wave electron beams for coincidence experiments, while superconducting electron–positron accelerators in a certain energy regime offer a possibly superior alternative to normal conducting accelerators. Superconducting accelerator technology is presently the only choice for recirculating electron linacs at energies above 2 GeV. Both types, superconducting linacs and recirculating linacs, are analysed with respect to their special characteristics, and existing accelerators of both kinds are discussed.

1. INTRODUCTION

The properties of particle accelerators are always dictated by the special requirements of the experiments using their beam. Experiments in elementary particle physics usually require beam energies as high as possible, pushing the accelerators to their technical limits. Therefore elementary particle physics has for decades worked in the domain of circular accelerators and storage rings usually providing two beams of different particles that hit each other head-on in colliding beam experiments.

The obvious advantage of these accelerators is that non-interacting particles are not 'wasted', they are circulated again and only the energy lost due to synchrotron radiation has to be supplied by radio frequency (RF) cavities in the storage ring. Accelerators of this type provide fairly high beam currents (in the range of several mA) and thus a sufficient rate of events. However, there is a fundamental limit for storage rings set by the laws of electrodynamics: the energy loss per turn due to the emission of synchrotron radiation is inversely proportional to the radius of curvature of the particle trajectory, but proportional to the fourth power of γ, the relativistic Lorentz factor (the ratio of total energy to energy at rest). For electrons and positrons the Large Electron–Positron collider LEP II has reached this limit despite its impressive circumference of 27 km. Therefore, at least for light particles, the successors of the high-energy storage rings will be linear colliders: two linear accelerators (linacs) pointing at each other with their beams hitting head-on at almost unbelievably small cross sections, generated in sophisticated final focus systems.

Pioneering work on linear colliders was performed at SLAC, where the existing 50 GeV electron linac was modified to accelerate electron and positron beams of very much improved quality. Two arcs were added to enable head-on collisions between the electrons and positrons in a final focus arrangement.

To date, competitive design studies for electron–positron colliders in the energy regime up to 1 TeV have been completed for normal conducting linacs such as the Next Linear Collider (NLC, a joint American–Japanese effort), and for superconducting linacs such as the TeV Energy Superconducting Linear Accelerator (TESLA, a joint European–American effort). For superconducting accelerating cavities, the critical magnetic field of the material from which they are fabricated sets an upper limit to the achievable accelerating gradient. For niobium (still the most common material for superconducting cavities in electron accelerators) the gradient will be limited to some 50 MV/m. Therefore, for linear colliders considerably exceeding an energy of 1 TeV, normal

conducting cavities at very high frequencies may be chosen. A study of such a collider, the Compact Linear Collider (CLIC), using the two-beam accelerator technique at frequencies of 30 GHz and extremely high accelerating fields, is presently under way at CERN. It seems therefore that high-energy physics will need linacs in the near future.

In Section 2 superconducting electron linacs and their basic properties are discussed. Section 3 is devoted to the schemes and related properties of recirculating linacs. A variety of examples for both superconducting electron linacs and recirculating linacs is presented and discussed in Section 4 and followed by some concluding remarks.

2. SUPERCONDUCTING ELECTRON LINACS

Basically, electron linacs are simple devices since, due to their small rest mass ($mc^2 = 0.511$ MeV), electrons move at a speed close to the speed of light (c) once they have gained a kinetic energy of a few MeV (e.g. $\beta = v/c \geq 0.99$ for $E_{kin} \geq 3.1$ MeV). Therefore, in an electron linac (with the exception of its very front end where the electrons enter from the source) the accelerating structures (electromagnetic resonators oscillating at microwave frequencies) can all operate at the same phase velocity $v_{ph} = c$ in synchronicity with the speed of the electrons; they are therefore all identical in geometry. Since the topic of accelerating cavities is covered by several lectures (see the contributions by J. LeDuff and W. Wuensch in this report) we restrict ourselves here to considerations that apply to superconducting electron linacs and recirculating superconducting linacs in particular. For general information on superconducting accelerators we refer the reader to Ref. [1], and for the historical development of recirculating linacs to Ref. [2].

2.1 Why superconducting linacs?

The properties required of an accelerator designed for nuclear physics experiments are quite different than those of an accelerator for high-energy physics. In high-energy physics, it is most important to have as high an energy as possible available in the centre-of-mass system, thus colliding beam experiments are chosen. In addition the rate of events, dn/dt, has to be high enough to achieve significant results in a reasonable time. Since the event rate is equal to the product of the luminosity *L* (a property of the accelerator) and the cross section of the investigated reaction σ (a quantity given by nature), the accelerators have to produce high luminosities.

In nuclear physics (investigation of excitation and decay of nuclei, and properties and interaction of elementary particles in nuclear matter) it is commonly more important to have maximum luminosity rather than the highest possible energy. Therefore, the combination of a single beam hitting a fixed target is the usual choice in nuclear physics, since in this way luminosities can be reached that exceed those of colliding beams by three orders of magnitude. For modern experiments in nuclear physics there is one more important requirement: in order to optimize the ratio of truly coincident to accidentally coincident events, the time structure of the beam has to be as uniform as possible. Accelerators using electromagnetic fields at microwave (RF) frequencies always produce a beam consisting of bunches separated by at least one RF period.

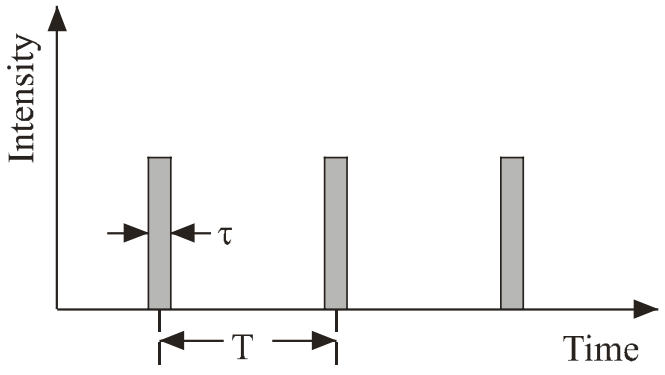


Fig. 1: Macroscopic time structure of a beam from a pulsed accelerator

Since the bunch repetition rate usually exceeds the maximum counting rate of the detectors by a long way, this so-called microstructure is neglected. If the accelerator is operated in a pulsed mode and produces (as indicated in Fig. 1) trains of bunches with a duration τ, separated by a time interval T, the macroscopic duty factor DF_{macro} of the beam is defined by $DF_{macro} = \tau/T$. If the accelerator produces a continuous train of bunches without any superimposed macrostructure, $DF_{macro} = 1$, the time structure is called Continuous Wave (CW). Continuous wave beams can either be achieved by expanding a pulsed beam in a stretcher ring, or naturally, by running the accelerator continuously. As will become clear in the following sections, the continuous operation of accelerating cavities at gradients of many MV/m is the natural domain of superconducting cavities.

2.2 Basic considerations

For an electric current oscillating at microwave frequencies, the surface resistivity of a superconductor does not vanish. According to BCS theory it has a finite value, depending on frequency and temperature. For temperatures below half the critical temperature of the superconductor ($T \leq T_c/2$) the surface resistivity R_s amounts to

$$R_s = A \cdot (f^{\kappa}/T) \cdot e^{-((\Delta/kTc)\cdot(Tc/T))} , \quad (1)$$

where f is the operating frequency (GHz), T the temperature (K), Δ the energy gap of the superconductor (eV), k the Boltzmann's constant, and T_c the critical temperature of the superconductor (K).

The quantities A and κ are usually determined experimentally. For niobium, which is still the only material in use for superconducting electron linacs (even in cavities with niobium sputtered on copper as is the case for most of the LEP cavities) values of $A = 9.0 \cdot 10^{-5}$ Ω K/(GHz)$^{\kappa}$, $\kappa = 1.9$, $\Delta/k\,T_c = 1.9$, and $T_c = 9.2$ K are found. Since the RF power dissipated in an accelerating cavity is proportional to R_s, Eq. (1) favours low operating frequencies (in contrast to normal conducting copper cavities), which unfortunately result in large cavities, cryostats, RF couplers, and other associated equipment. However, since the excitation of wakefields is very much reduced in cavities with large apertures, the final choice of the operating frequency will always be a trade-off between physics and financial considerations. In practice it has turned out that superconducting accelerating cavities for storage rings operate at frequencies below 1 GHz at temperatures of 4.2 K, whereas cavities for electron linacs operate in the range of 1.3–3 GHz at temperatures between 1.8–2 K. The fact that all superconducting accelerating structures operating in electron linacs are standing wave cavities is also a consequence of the low surface resistivity resulting from Eq. (1): the RF power dissipated in the cavity walls is only a small fraction of the power delivered to the beam and therefore, even in a pulsed application (as in a linear collider), the standing wave cavity has advantages over its travelling wave counterpart.

Beam dynamics in an electron linac is less complicated than in a circular accelerator. Longitudinally the electrons are synchronized to the phase velocity of the accelerating RF field and the electron bunches are usually positioned 'on crest' (at the maximum accelerating field) in order to minimize the effect of finite bunch length to energy spread (as discussed in Section 4.2, this can be different in recirculating linacs). The transverse focusing strength of the cavities (since they are standing wave cavities) is on the order of accelerating gradient/electron energy and is therefore only important at quite low energies. Quadrupoles, arranged in a FODO lattice, are the common choice for transverse beam optics, where in superconducting linacs they can either be positioned between or (in the case of very long cryomodules) inside cryogenic modules. In the latter case superconducting quadrupoles are used.

2.3 CW linacs: normal conducting vs superconducting

The following estimate of RF power, dissipated in a continuously running accelerating cavity, shows that it was the requirement for CW beams that led to the concept of recirculating linacs and in particular to superconducting recirculating linacs. The dissipated power P_{dis} per unit length (as sketched in Fig. 2) is given by

$$P_{dis} = E_{acc}^2/(R/Q)\cdot Q_0 , \quad (2)$$

where E_{acc} is the accelerating gradient, (R/Q) the normalized shunt impedance (depending on the geometry of cavity), and Q_0 the unloaded quality factor (depends on surface resistivity of material).

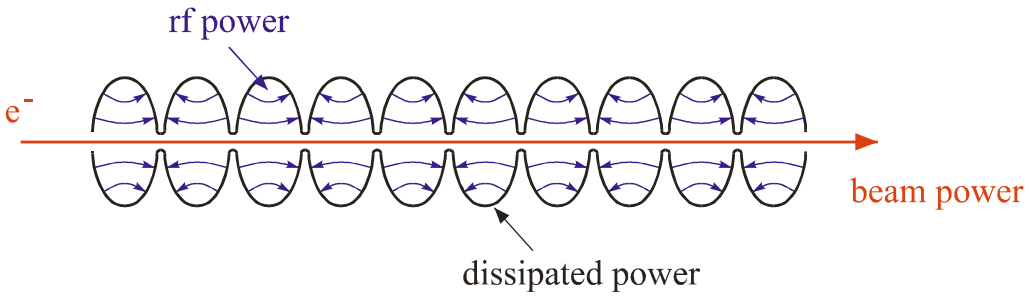


Fig. 2: Sketch of RF power distribution in CW linacs

Table 1 contains typical figures for well optimized normal conducting copper cavities and superconducting niobium cavities (both operating in S-band) as well as the dissipated power resulting from Eq. (2) for both cases.

Table 1: Comparison of typical S-band cavity parameters

Parameter	Normal conducting cavities	Superconducting cavities
(R / Q)	$4 \cdot 10^3$ Ω/m	$2 \cdot 10^3$ Ω/m
Q_0	$2.5 \cdot 10^4$	$3 \cdot 10^9$
E_{acc}	1 MV/m	5 MV/m
P_{dis}	10^4 W/m	4.2 W/m

Table 1 clearly indicates that the assumed gradient for normal conducting cavities is close to the technical limit and also that a linac, producing a beam of, for example, several hundred MeV would become unreasonably long.

For superconducting niobium cavities the small value of P_{dis} = 4.2 W/m indicates that gradients much higher than 5 MV/m are still reasonable. On the other hand, even at 10 MV/m a linac would still become very long. As a consequence, for both scenarios (normal conducting and superconducting linacs producing CW beams), a recirculating linac configured as described below is appropriate.

3. RECIRCULATING LINACS

The basic concept of a recirculating linac is that a beam of particles is accelerated several times in the same linac. Since the phase velocity of the accelerating field in the linac is fixed this scheme obviously only works for highly relativistic particles moving at velocities close to the speed of light, almost independent of their energy. Therefore, all the existing recirculating linacs are electron accelerators.

3.1 Basic schemes

The general layout of such a machine is indicated in Fig. 2. The accelerator consists of a linac placed between two ‘return devices’ (beam transport systems) that take the individual beams from the linac exit back to its entrance for the subsequent acceleration. (If the return devices are designed in such a way that the individual beams have a common return path between the two devices, a second linac can be placed there, essentially doubling the energy gain of the particles for each round trip in the accelerator; see for example the accelerator at the Thomas Jefferson National Accelerator Facility—TJNAF, or the fourth stage of the cascaded microtron MAMI C at Mainz, both discussed in Section 4.1). Two more devices are essential for a recirculating linac: one magnet system to inject the beam

from a source (usually an injector linac) into the recirculating linac and another magnet system for extraction of the beam after its final acceleration.

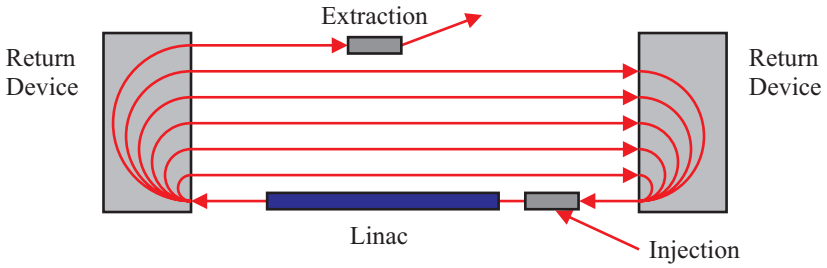


Fig. 2: Basic scheme of a recirculating linac

The final energy E_f of the recirculating linac is given by

$$E_f = N \cdot \Delta E_{linac} + E_{inj.} \,, \tag{3}$$

where N is the number of passes through the linac(s), ΔE_{linac} the energy gain in the linac(s), and E_{inj} the particle energy at injection.

Two basically different schemes exist for recirculating electron linacs: the polytron scheme and the scheme of 'independent orbit recirculation'. In a polytron a relatively small number of (large) dipole magnets forms the return devices. *All* of the recirculated beams pass through the same dipoles. The order of a polytron is usually indicated by an even number, p, which is equal to the number of dipoles. Each dipole bends the beams by an angle of $2\pi/p$. Figure 3 shows sketches of the lowest order polytrons, the racetrack microtron (p = 2), the double-sided racetrack microtron (p = 4), and the

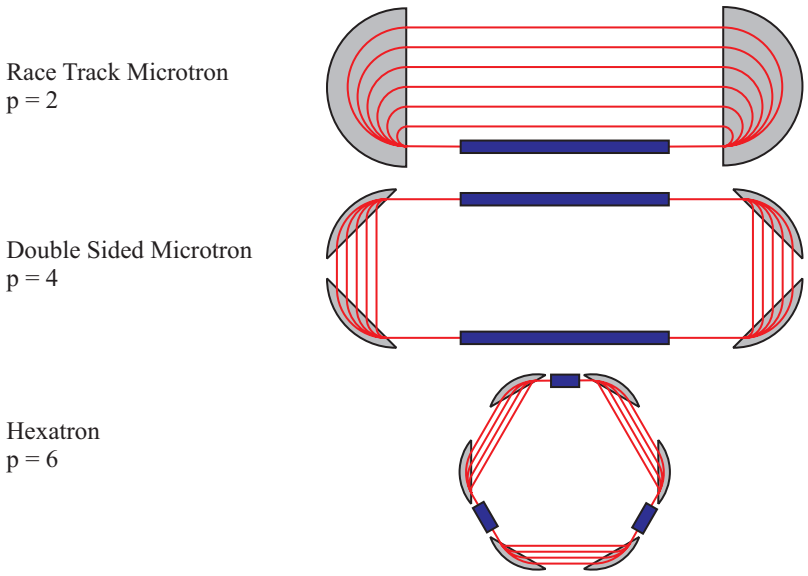


Fig. 3: Lowest order members of the 'polytron family'

hexatron (p = 6). The recirculating electron accelerator MAMI B at Mainz consists of three cascaded racetrack microtrons. A fourth stage is currently being developed and will be added to form MAMI C.

In a polytron, due to its high symmetry and the fact that the dipoles are common to all orbits, a relatively small number of parameters fixes the design of the accelerator. In a racetrack microtron (see Fig. 4 for a definition of quantities), two fundamental laws are obvious:

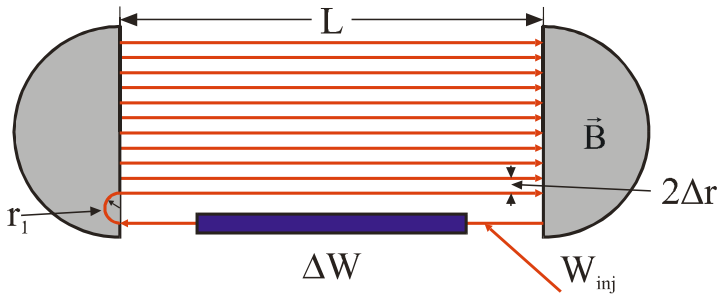


Fig. 4: Basic scheme of a racetrack microtron

i) the length of the first orbit must equal an integer number m of RF wavelengths λ $2\pi r_1 + 2L = m \cdot \lambda$, with r_1 being the radius of first orbit in dipole magnets and L the distance between dipole magnets;

ii) the length of each successive orbit must be longer than the previous orbit by an integer number ν of RF wavelengths λ
$2\Delta r = 2\pi\Delta W/ecB = \nu\lambda$, with 2Δr being the spacing of orbits, ΔW the energy gain in the linac, and B the magnetic field in dipoles.

As a consequence, once parameters ν and λ are chosen, the ratio of ΔW/B is fixed. Then selection of B fixes ΔW (or vice versa) and (for a given value of r_1) the injection energy. Finally, the number of recirculations, N, determines the size (weight and cost) of the dipole magnets.

The fundamentally different scheme of 'independent orbit recirculation' is sketched in Fig. 5. Here, many small dipole magnets are used in the return devices and, except for the splitters and recombiners (close to the entrance and exit of the linac), there are individual dipoles for each beam. The superconducting recirculating electron accelerators S-DALINAC at Darmstadt and the Continuous Electron Beam Accelerator Facility (CEBAF, the accelerator at the TJNAF) are examples of this scheme.

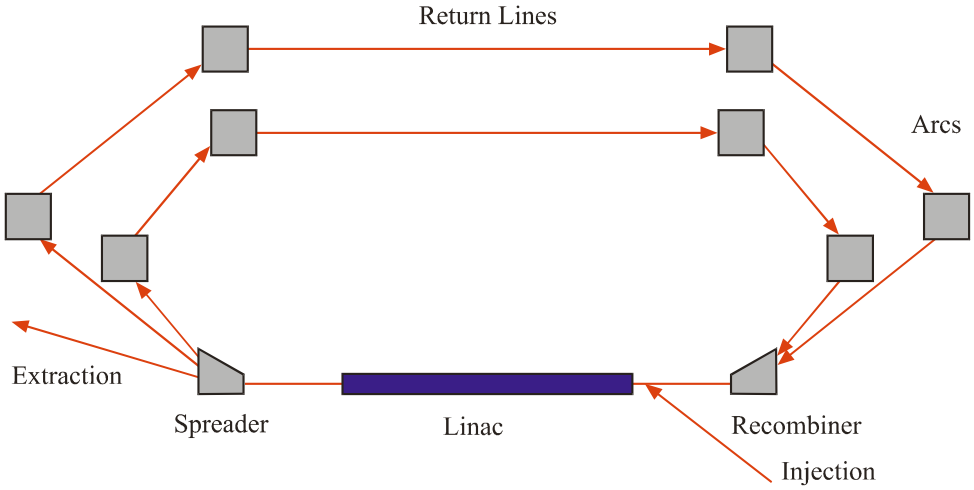


Fig. 5: Independent orbit recirculation

There is much more freedom in the design of the accelerator in the independent orbit recirculation scheme. Of course, the length of each orbit has to be equal to an integer number of RF wavelengths λ, but this number can be chosen independently for each orbit.

3.2 Efficiency of recirculating linacs

A brief analysis of the efficiency of recirculating linacs shows that different schemes are necessary for normal conducting and for superconducting linacs. Let η be the efficiency, defined as beam power divided by the electric power necessary to provide the RF power for acceleration. Then, for normal conducting linacs η is given by

$$\eta = (\eta_{rf} E_{acc} I_{beam} N)/[(E_{acc})^2/((R/Q)\, Q_0) + E_{acc} I_{beam} N] \,, \tag{4}$$

where η_{rf} is the power conversion efficiency (electric to RF), E_{acc} the accelerating gradient, I_{beam} the beam current, N the number of times the beam is accelerated in the linac(s), (R/Q) the normalized shunt impedance, and Q_0 the unloaded quality factor.

For superconducting linacs Eq. (4) has to be modified. The RF power dissipated in the accelerating cavities has to be taken away at a temperature of 1.8 or 2 K and thus must be divided by the efficiency η_{cryo} of the helium refrigerator, typically $\eta_{cryo} = 3.3 \times 10^{-3}$ for helium refrigerators of reasonable size. Then η is given by the expression

$$\eta = (\eta_{rf} E_{acc} I_{beam} N)/[(E_{acc})^2/((R/Q)Q_0)(1 + \eta_{rf}/\eta_{cryo}) + E_{acc} I_{beam} N]. \tag{5}$$

The result as shown in Fig. 6 is quite surprising, assuming $\eta_{rf} = 50\%$ and $\eta_{cryo} = 0.33\%$ for both types of accelerators, a beam current of 100 μA, and accelerating gradients of 1 MV/m and 5 MV/m for normal conducting and superconducting linacs, respectively. The normal conducting linac needs many recirculations in order to become effective (the upper limit of η is 50% since η_{rf} was chosen to be 50%!). This is the reason that the racetrack microtron configuration is very appropriate for normal conducting recirculating accelerators.

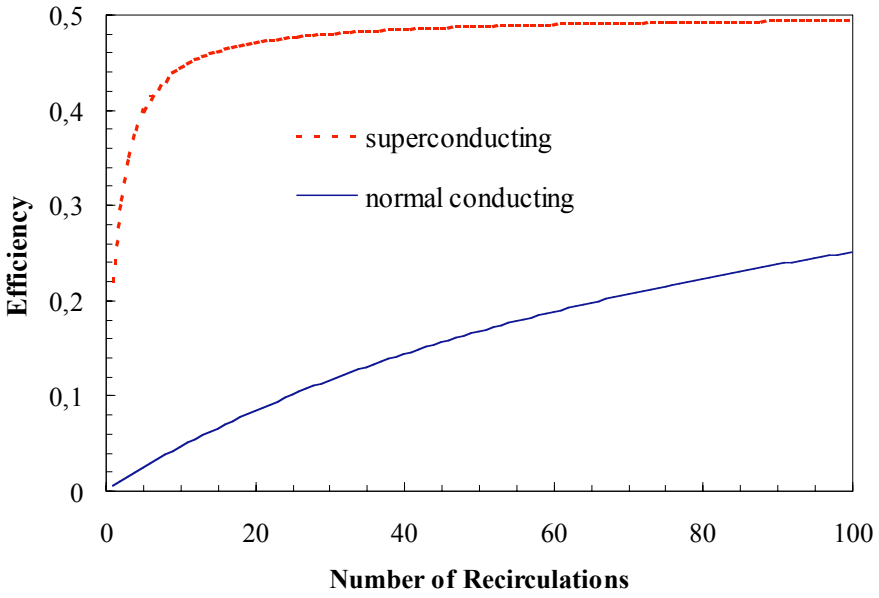


Fig. 6: Efficiency of recirculating linacs

The curve for superconducting accelerators (Fig. 6) shows that even a simple linac (N = 1) is effective and that a very small number of recirculations yields very good efficiencies. Superconducting recirculating electron linacs use recirculation schemes with independent optics on the individual recirculating beam lines.

3.3 Beam dynamics in recirculating linacs

In beam optics individual particles are described with respect to a reference particle that moves along the ideal trajectory with the correct longitudinal momentum p_0. For convenience a rectangular coordinate system x, y, z is used, whose origin moves along with the reference particle and whose z-axis is always tangential to the reference trajectory. Individual particles are then described in six-dimensional phase space by their spatial coordinates x, y, and z, and their corresponding momenta p_x, p_y, and p_z, or more commonly $\delta p_z = p_z - p_0$, the deviation from the correct longitudinal momentum. This characterization of individual particles is sketched in Fig. 7.

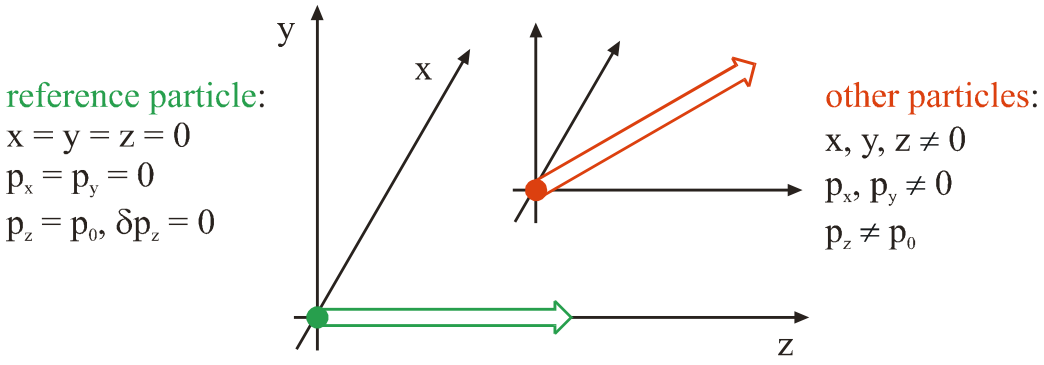


Fig. 7: Characterization of particles in phase space

In beam optics it is quite common for the inclination of the real trajectory with respect to the reference trajectory, given by $x' = p_x/p_0$ and $y' = p_y/p_0$, to be used instead of the transverse momenta. Whereas for the longitudinal coordinates the RF phase deviation ($\delta\phi = 2\pi z/\lambda$) and the relative momentum deviation ($\delta p_z/p_0$) are used. For particle ensembles suitable momenta of the whole ensemble have to be employed.

3.3.1 Longitudinal motion

The sketch in Fig. 8 illustrates the situation of a recirculating linac: particles are accelerated in the linac and after leaving it are returned to the entrance to be accelerated again. Even though the linac is always the same for each of the successive passes the energy gain of a particle (as indicated in the lower part of Fig. 8) depends on its phase with respect to the accelerating RF field, which in general will be different for each pass.

If φ_R denotes the correct phase of the reference particle, $\Delta W_R = eV_0 \cos(\varphi_R)$ is the correct energy gain for it. In general (for $\varphi \neq \varphi_R$) the energy gain ΔW will be $\Delta W = eV_0 \cdot \cos(\varphi_R + \delta\varphi) \approx eV_0 \cdot \cos(\varphi_R) - eV_0 \sin(\varphi_R)\delta\varphi = \Delta W_R + \delta W$. Since particle velocity and phase velocity of the RF field are assumed to be equal (very close to c) the phase deviation of the particle $\delta\varphi$ with respect to the reference particle remains unchanged. Thus, in a first order approximation, phase and energy deviation of a particle that enters the linac for the n^{th} time, $\delta\varphi_n$ and δW_n, transform into $\delta\varphi_{n+1}$ and δW_{n+1}, respectively, at the exit of the linac:

$$\begin{pmatrix} \delta\varphi_{n+1} \\ \delta W_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\Delta W_R \cdot \tan(\varphi_R) & 1 \end{pmatrix} \cdot \begin{pmatrix} \delta\varphi_n \\ \delta W_n \end{pmatrix}. \tag{6}$$

In the case of the racetrack microtron the longitudinal properties of the returns are those of the two 180° dipole magnets (see Section 3.1). Therefore (again to a first order approximation) for the m^{th} passage of a particle through the return, phase and energy deviation transform in the following way:

$$\begin{pmatrix} \delta\varphi_{m+1} \\ \delta W_{m+1} \end{pmatrix} = \begin{pmatrix} 1 & 2\pi \cdot \nu / \Delta W \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \delta\varphi_m \\ \delta W_m \end{pmatrix}. \tag{7}$$

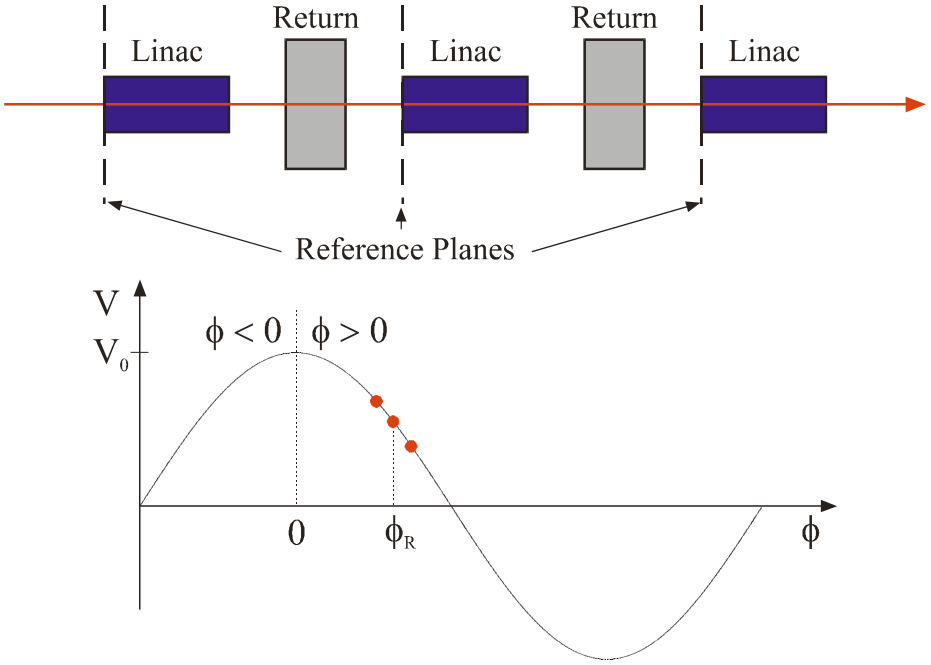


Fig. 8: Longitudinal motion

For a complete orbit (transport from one to the next reference plane in Fig. 8) first order transformation yields

$$\begin{pmatrix} \delta\varphi_{n+1} \\ \delta W_{n+1} \end{pmatrix} = \begin{pmatrix} 1-2\pi\cdot\nu\cdot\tan(\varphi_R) & 2\pi\cdot\nu/\Delta W \\ -\Delta W_R\cdot\tan(\varphi_R) & 1 \end{pmatrix}\cdot\begin{pmatrix} \delta\varphi_n \\ \delta W_n \end{pmatrix} \equiv \mathbf{M}\cdot\begin{pmatrix} \delta\varphi_n \\ \delta W_n \end{pmatrix}. \quad (8)$$

If the matrix **M** is the same for all orbits and Det(**M**) = 1 (conservation of phase space), the longitudinal motion of the particles is stable if Tr(**M**) < 2, i.e. the particles perform synchrotron oscillations in longitudinal phase space. The phase advance μ of these oscillations for one round trip in the racetrack microtron is defined by Tr(**M**) = 2 cos(μ). Using this definition of the phase advance μ, matrix **M** takes the general form

$$\mathbf{M} = \begin{pmatrix} \cos(\mu)+\alpha\cdot\sin(\mu) & \beta\cdot\sin(\mu) \\ -((1+\alpha^2)/\beta)\cdot\sin(\mu) & \cos(\mu)-\alpha\cdot\sin(\mu) \end{pmatrix}, \quad (9)$$

with α and β being constants.

Comparison of Eq. (9) with Eq. (8) gives the relation between phase advance μ and reference phase φ_R:

$$\cos(\mu) = 1 - \pi\nu\tan(\varphi_R) \ . \quad (10)$$

Considering that cos(μ) has to be less than one finally yields the condition for choosing φ_R such that the synchrotron motion of the particles in the recirculating linac is stable:

$$0 < \pi\nu\tan(\varphi_R) \ . \quad (11)$$

Independent orbit recirculation schemes are more flexible and allow for so-called isochronous designs, where the path length of the particles does not depend on the particle energy in the returns to first order approximation. In this case the recirculating linac is longitudinally equivalent to a straight single linac; the returns have no influence on the phase deviation of the particles. This situation is illustrated in Fig. 9 below.

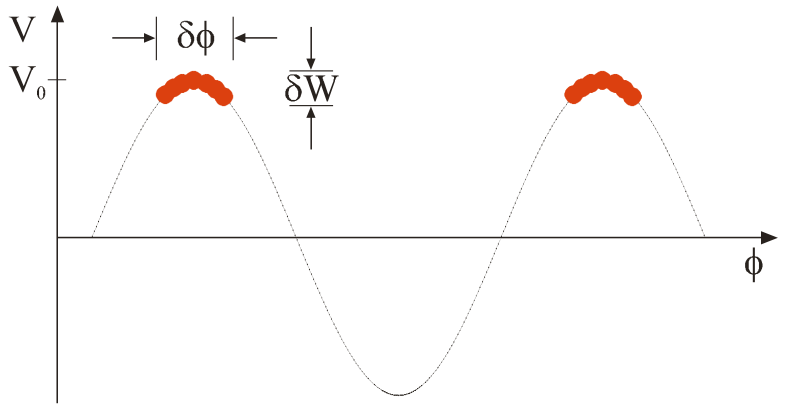


Fig. 9: Isochronous recirculation

In this scheme (neglecting all instabilities) the energy spread δW of the particles is determined by the bunch length δφ (see Fig. 9). Particles are usually accelerated 'on crest', using the accelerating field most efficiently. In this case the resulting energy spread δW is given by

$$\delta W = eV_0 \cdot (1 - \cos(\delta\varphi/2)) \approx eV_0 \delta\varphi^2/8 \; . \qquad (12)$$

and depends only to second order on the phase deviation δφ. However, since this scheme has no longitudinal stability the influence of any amplitude jitter or phase jitter of the accelerating field is not suppressed but fully reflected in the energy spread of the accelerated particle beam.

3.3.2 Transverse motion

In recirculating accelerators of the polytron type there are two possible schemes for achieving transverse stability. The first scheme, as sketched in Fig. 10, uses individual focusing on each return. The obvious advantage is that one can provide the same focusing strength for each orbit despite the increasing beam energy. However, there are two major disadvantages: the first is the large number of quadrupoles necessary and the second (possibly even more severe) is the fact that all quadrupoles are located in dispersive sections. This scheme of focusing on the individual returns has not been realized to date.

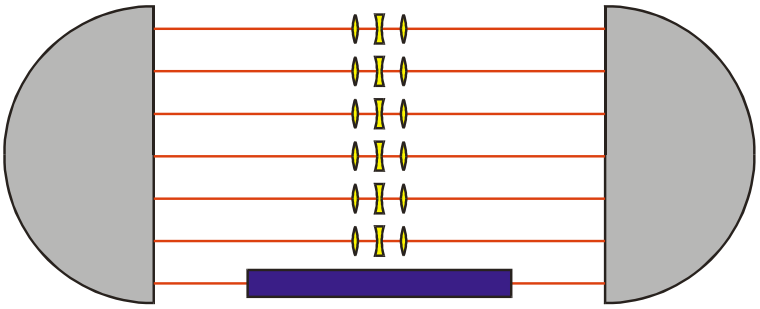


Fig. 10: First focusing scheme in a racetrack microtron

The second scheme, as indicated in Fig. 11, uses focusing elements only on the linac axis that is common to all orbits. Advantages of this method are first the very small number of quadrupoles needed, and second the fact that there is no dispersion on the linac axis, and therefore the quadrupoles do not influence the dispersive trajectory of the lattice. A disadvantage is the fact that the focusing strength decreases from orbit to orbit because of the steadily increasing beam energy. This finally leads to a limitation in the possible ratio of input to output energy ('Herminghaus condition': $W_{out}/W_{in} < 10$) for the racetrack microtron. This limitation in turn results in the configuration of multi-staged cascaded racetrack microtrons, where each stage can increase the beam energy by a factor of ten. The most successful examples of this type of recirculating linac are the cascaded racetrack microtrons of MAMI at Mainz, discussed in more detail in Section 4.

Fig. 11: Second focusing scheme in a racetrack microtron

The situation is completely different in machines that use independent orbit recirculation. There is (as indicated in Fig. 12 below) great flexibility in the layout of the lattice. Constraints are that the lattice has to be free of dispersion (or even isochronous) on the axis of the linac(s) and that it should also have no dispersion on the straight returns. One advantage of this scheme is the fact that magnets can be small (even the dipoles) since (except for spreaders and recombiners) they belong to individual orbits. Therefore, it is possible to place quradupoles in the arcs in locations with or without dispersion in order to tailor the optical properties of the lattice. On the straight return lines quadrupoles can be used either in a telescope-like imaging scheme or in a FODO arrangement.

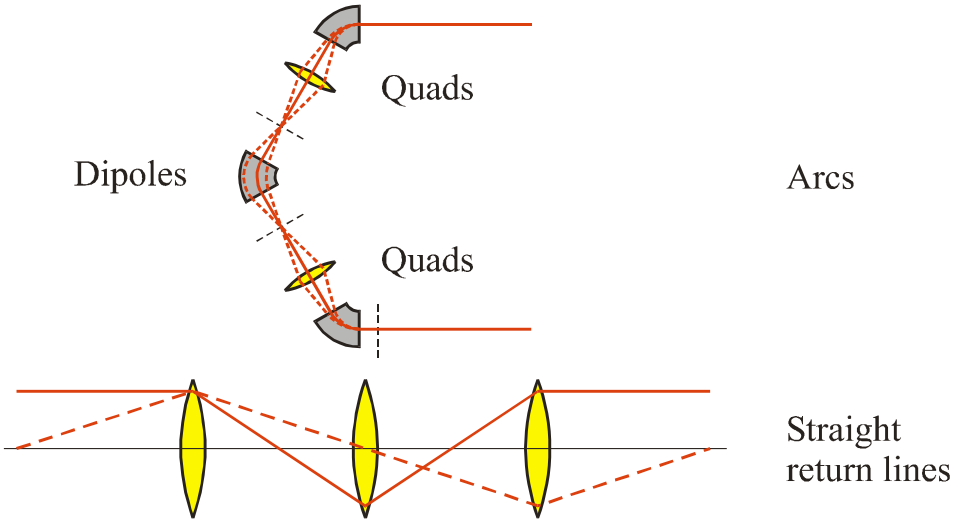


Fig. 12: Lattice schemes used in independent orbit recirculation

4. EXAMPLES

We now present in some detail examples of the different accelerator categories. The cascaded racetrack microtrons MAMI at Mainz have been chosen as a normal conducting recirculating electron linac since they display in a most beautiful way the features of this accelerator type (as explained in Section 3). Three examples of superconducting recirculating electron linacs are presented: the 130 MeV machine S-DALINAC at Darmstadt, the 5 GeV accelerator CEBAF at the TJNAF, and the proposed 25 GeV ELFE installation at CERN. Being a small installation, the S-DALINAC is best suited to looking into details of special features of this particular type of superconducting electron accelerator. CEBAF is the 'workhorse' operating at the high-energy end of CW electron accelerators in nuclear physics. It represents the biggest installation world-wide as far as the number of superconducting accelerator cavities is concerned. ELFE proposes to use the superconducting cavities of LEP II after its shutdown in a next-generation high-energy nuclear physics accelerator. Regarding superconducting electron linacs without recirculation, the installation of the TESLA Test Facility (TTF) at DESY and of course TESLA itself, the proposed 500 GeV electron–positron linear collider, are chosen. As examples of drivers of Free-Electron Lasers (FELs) the superconducting electron linacs of the S-DALINAC, the TTF, and the Infra Red Demonstration Free-Electron Laser (IR DEMO FEL)

at TJNAF have been selected. For detailed information on the individual accelerators we refer the reader to the Web pages of the respective institutes [3].

4.1 Recirculating linacs

4.1.1 MAMI at Mainz

Figure 13 shows the layout of the present accelerator installation MAMI B at Mainz (without experimental halls). It consists of two electron sources, a conventional (unpolarized) thermionic gun (labelled EKAN in Fig. 13) and a polarized source using a laser-driven strained GaAs photocathode (labelled PKA1). Both sources produce a 100 keV electron beam. A short injector linac (located between PKA1 and RTM1 in Fig. 13) with three RF structures increases the energy to 3.97 MeV. RTM1 is the first of the three cascaded racetrack microtrons of MAMI B. It uses one RF structure and 18 turns to increase the beam energy to 14.86 MeV. The second stage, RTM2, has two RF structures and in 51 turns the beam energy is increased to 180 MeV. The final stage, RTM3, has a linac consisting of five RF structures and the beam is recirculated as many as 89 times before it reaches its final energy of 855 MeV and is extracted to the experimental halls.

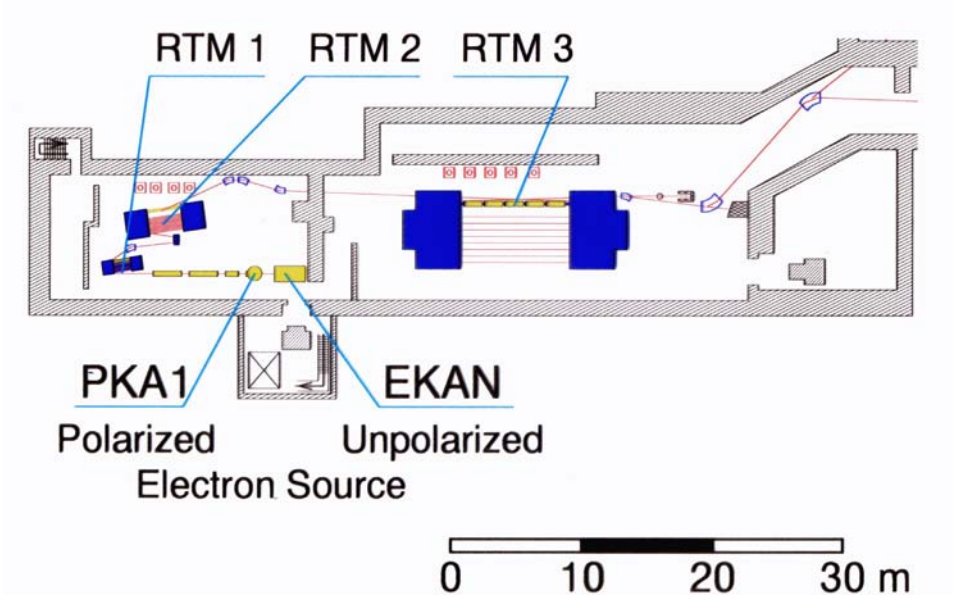


Fig. 13: Layout of MAMI B

Some of the most important parameters of the accelerators at Mainz are collated in Table 2. The data for the racetrack microtrons RTM1, RTM2, and RTM3 clearly display some of the particular features of this type of accelerator as discussed in Section 3. All three microtrons use the second focusing scheme (common focusing for all orbits on the linac axis) and Herminghaus' rule is only slightly violated in RTM2, where the ratio of output to input energy amounts to 12, rather than 10. One can derive the energy gain per RF structure from the number of turns and the number of RF structures contained in each linac. It amounts to 1.5–1.6 MeV (except for RTM1, where it is lower) and corresponds to $E_{acc} \approx 1$ MV/m, in agreement with the discussion of normal conducting accelerator cavities in Section 2.3. The exceptionally small energy spread (penultimate row of Table 2) is due to the longitudinal stability discussed in Section 3.3, inherent to the racetrack microtron scheme. Another salient feature of recirculating electron accelerators is also apparent in Table 2. The normalized emittances presented in the bottom row show a much stronger increase for the horizontal emittance than for the vertical emittance, due to the stochastic nature of synchrotron radiation. This effect requires that the number of recirculations should be kept low, particularly at energies approaching or exceeding 1 GeV, or in other words, for a given output energy the energy gain per turn should be high. This fact finally sets a limit on CW recirculating electron accelerators using normal conducting cavities.

Table 2: Collection of MAMI C parameters

Parameter	(Unit)	Injector	RTM1	RTM2	RTM3	HDSM
Energies	(MeV)	3.97	14.86	180	855	1500
Turns			18	51	90	43
Dipoles	(# / t)		2 / 2	2 / 45	2 / 450	4 / 250
Field	(T)		0.1026	0.555	1.284	1.53–0.95
Frequency	(GHz)	2.4495	2.4495	2.4495	2.4495	2.4495 / 4.899
RF structures		3	1	2	5	5 / 8
ΔE^a	(keV)	1.2	1.2	2.8	13	110
$\varepsilon_{x,n}/\varepsilon_{y,n}{}^a$	($\pi \cdot 10^{-6}$m)	0.05 / 0.04	0.07 / 0.07	0.25 / 0.13	13 / 0.84	27 / 1.2

[a] 1σ values

A fourth microtron is currently under construction at Mainz as an energy upgrade from 855 to 1500 MeV (conversion from MAMI B to MAMI C). It is not a racetrack microtron but rather a double-sided microtron (DSM). This is for two main reasons: first, as stages RTM1 to RTM3 show, the magnet weight for 180° dipole magnets would become excessive at 1500 MeV; second, since a DSM has a common axis for the beam return lines a second linac can be placed there and double the energy gain per turn. Finally, careful design studies revealed that a significant increase in longitudinal stability can be achieved if one of the two linacs operates at the first harmonic of the fundamental RF frequency of 2.4495 GHz. The resulting configuration therefore is called a Harmonic Double-Sided Microtron (HDSM).

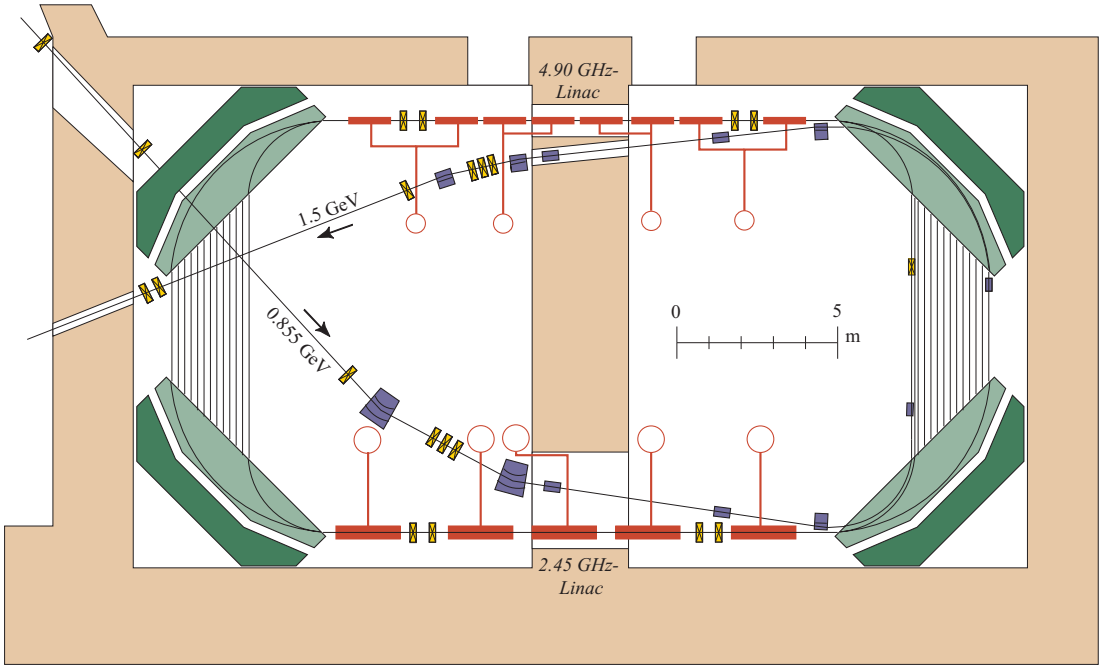


Fig. 13: Layout of the harmonic double-sided microtron (HDSM) at Mainz

The last column of Table 2 displays some of the notable data for the HDSM, and its layout is shown in Fig. 13. A look at the area covered by the four dipole magnets shows that, compared with a racetrack microtron, some 60% of magnet weight could be saved in the HDSM. The obvious drawback of the configuration, the strong focusing of the inclined magnet edges, is compensated by a

carefully tailored non-uniform field in the dipoles. Since the ratio of output to input energy is less than two, the focusing of the 43 beams is realized on both linac axes (the second focusing scheme). The harmonic 4.899 GHz linac consists of eight, and the fundamental 2.4495 GHz linac of five, RF structures; thus, both linacs have about the same physical length. Due to its compactness the HDSM fits into one of the existing experimental halls.

4.1.2 S-DALINAC at Darmstadt

Historically the S-DALINAC was the third superconducting recirculating electron accelerator after the racetrack microtron at the University of Illinois and the recirculating linac at Stanford University. Because of the advances in the design and fabrication of superconducting accelerating cavities that had been achieved at the time of its design and construction the S-DALINAC could be considered as a pilot project for its big brother CEBAF (discussed in the following subsection). The S-DALINAC's principle of operation is illustrated by its layout, shown in Fig. 15.

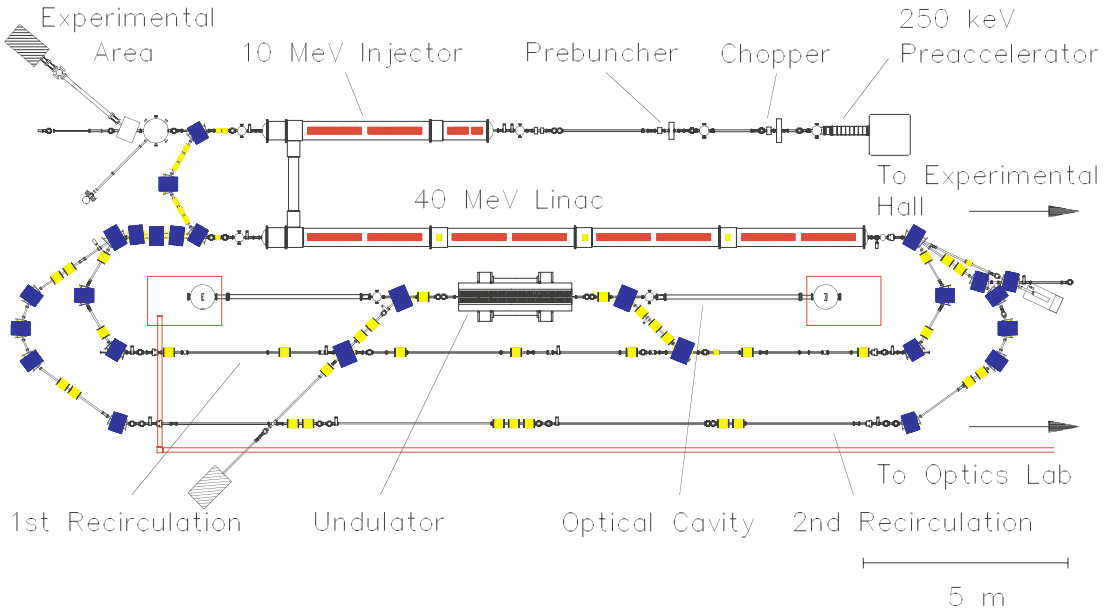


Fig. 15: Layout of the S-DALINAC

The electron source is located on a high-voltage terminal (top right) at 250 kV. The electrostatically preaccelerated beam receives its 3 GHz time structure, which is necessary for successive acceleration in the superconducting cavities to the chopper–prebuncher section at room temperature. The DC current from the source is first chopped into 30 ps long packages, which are then bunched to a length of 5 ps when they enter the superconducting injector linac. Acceleration is achieved by a 2-cell capture cavity ($\beta = 0.85$) followed by a 5-cell capture cavity ($\beta = 1.0$) and two 20-cell cavities, all fabricated from RRR = 280 niobium and operated in liquid helium at 2 K. When leaving the injector, the beam has an energy of up to 10 MeV and can either be used for low-energy experiments in an experimental area straight ahead or can be bent by 180° for injection into the main linac. There, eight 20-cell cavities (originally designed at the University of Wuppertal) installed in four identical cryomodules increase the beam energy by up to 40 MeV. When leaving the main linac the beam can either be extracted to the experimental hall or recirculated and reinjected one or two times by the appropriate beam transport systems (lower part of Fig. 16). The maximum beam energy after three passes through the main linac therefore amounts to 130 MeV. The central part of Fig. 15 shows the arrangement of the FEL. There, the beam from the straight section of the first recirculating beam line is bent over and passed through the undulator and then is either dumped or reinjected into the first recirculation for energy recovery experiments. Two mirrors to the left and right of the undulator form the 15 m long optical cavity. The evacuated transfer tube of some 50 m in length, for the radiation generated in the optical cavity, is indicated in the lower part of Fig. 15. Table 3 summarizes the most important parameters of the S-DALINAC.

Table 3: Design parameters of the S-DALINAC

Beam energy	130 MeV
Energy spread	$\pm 10^{-4}$
Beam current	20 μA
Duty cycle	CW
Cavity material	Niobium (RRR = 280)
Frequency	2.9975 GHz
Temperature	2 K
Quality factor	$3 \cdot 10^9$
Accelerating gradient	5 MV/m
RF losses at 5 MV/m	4.2 W/m

Fig. 16 shows a photograph of a 3 GHz superconducting 20-cell cavity. The cavity is a very slim object and, since high purity niobium is a rather soft material, cleaning, handling, and assembly of such cavities are delicate operations.

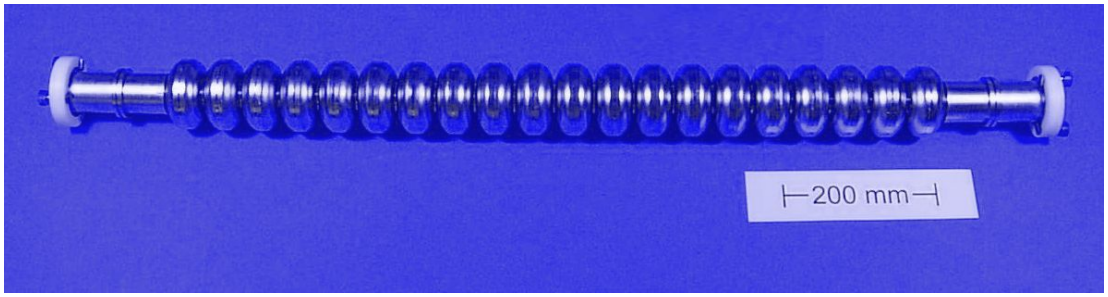


Fig. 16: Superconducting 3 GHz 20-cell cavity

All the cavities installed in the S-DALINAC clearly exceed the design gradient of 5 MV/m, averaging 6.7 MV/m. The only remaining drawback is the fact that the unloaded quality factors Q_0 are lower than the expected value of 3 x 10^9. The present average amounts to 7.7 x 10^8, resulting in increased dissipated RF power at 2 K. Since the helium refrigerator of the entire accelerator has only a capacity of slightly less than 120 W at 2 K it is the dissipated RF power that has up to now limited the maximum beam energy to 120 MeV. Therefore work continues on the improvement of Q_0 by, for example, applying new surface preparation techniques and the development of improved magnetic shielding.

The RF input couplers are very important parts of a superconducting accelerator and usually delicate. They have to transfer to the cavity the RF power necessary to achieve the correct gradient and accelerate the beam current. Since in superconducting cavities the wall losses are usually much smaller than the RF power transferred to the beam, the cavity represents a load that very much depends on the beam current, i.e. the transition from the RF input coupler to the cavity can change from being matched to being almost fully reflective. Thus one tries to optimize the input coupler in such a way that for maximum beam current (minimum loaded Q of the cavity) there is a minimum of reflected RF power (maximum power transfer efficiency). In this case the external quality factor (Q_{ext}) as presented by the input coupler, RF transfer line, and transmitter, equals the loaded Q of the cavity. Usually one defines the input coupling factor, or input coupling strength β_1, as the ratio Q_0 / Q_{ext}. Then the reflected power P_{refl} in terms of the forward power P_{forw} is given by by the following expression:

$$P_{refl} / P_{forw} = (1 - \beta_1)^2 / (1 + \beta_1)^2 . \qquad (13)$$

Figure 17 indicates how much forward RF power is needed for the S-DALINAC 20-cell cavities to accelerate different beam currents at a gradient of 5 MV/m as a function of the input coupling strength.

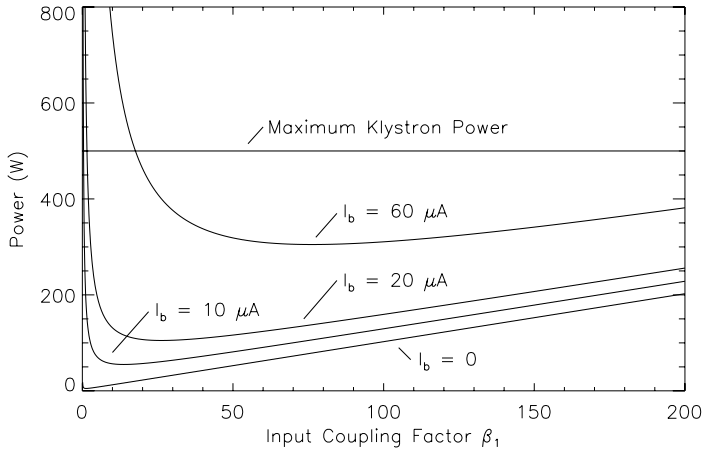


Fig. 17: Input coupling characteristics for the S-DALINAC 20-cell cavities at 5 MV/m

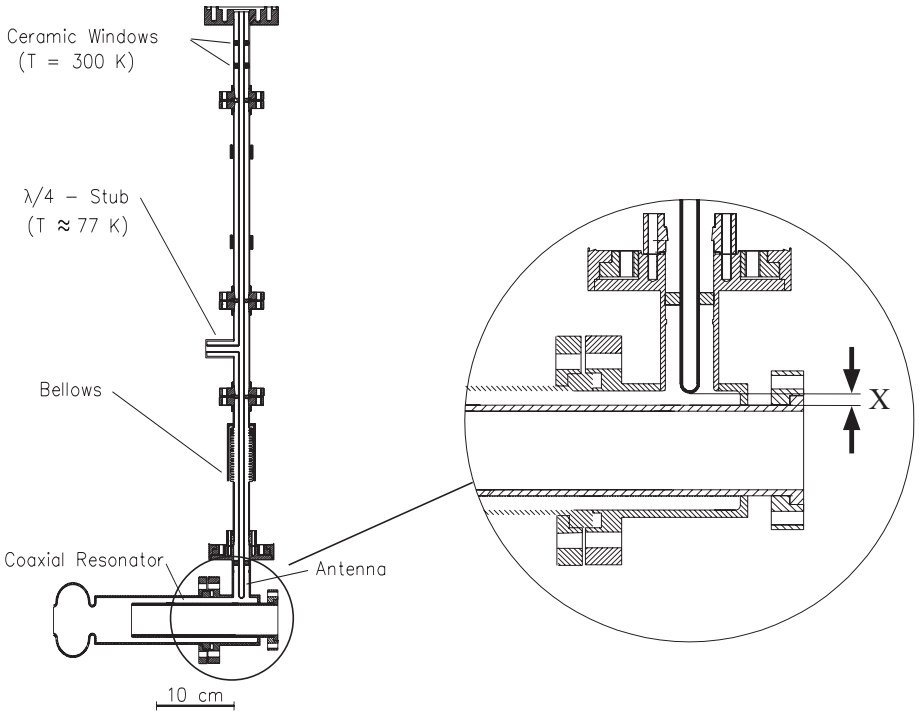


Fig. 18: Variable superconducting RF input coupler

For each beam current the corresponding curve has a minimum for a certain value of β_1 (matched condition), which increases with increasing beam current. Since the slope of all curves towards stronger coupling is less steep than for the undercoupled case it is common to use a little overcoupling. For diagnostic purposes, however, it is quite often desirable to match to the unloaded Q of the cavity (without beam) in order to obtain a resonance of minimum width and thus maximum diagnostic sensitivity.

The RF input couplers in the S-DALINAC allow for this option: they provide variable coupling strength over a wide range. The left part of Fig. 18 shows a cross section of the coupler with its coaxial RF input line. The interior of this line is common to the beam vacuum and is therefore sealed with two ceramic windows at room temperature (the volume between the two windows is common to

the insulating vacuum of the cryostat). Heat flow into the helium bath is minimized by the λ/4-stub, which is thermally connected to a liquid nitrogen cooled reservoir at 77 K. Transfer of RF power from the input line to the cavity is accomplished in two steps: the coaxial line couples to the coaxial resonator (lower part of Fig. 18), which in turn couples to the cavity (indicated by its first cell). Coupling between the coaxial line and resonator is determined by the distance x (right part of Fig. 18), which can be varied by shifting the top part of the coaxial line vertically while keeping the position of the coaxial resonator fixed. This is possible because of the bellows incorporated in the outer conductor of the coaxial line. The resulting variation in the overall coupling strength, expressed in terms of the external quality factor, is impressive and is shown in Fig. 19, which displays a measurement at 2 K (left part). It is clear that in situ diagnostic measurements at $Q_{ext} \approx 10^9$ (right part of Fig. 19) are possible with this coupler. Of course optimum matching to different beam loading conditions at $Q_{ext} \approx$ a few 10^7 is also possible by adjusting the input coupling strength accordingly.

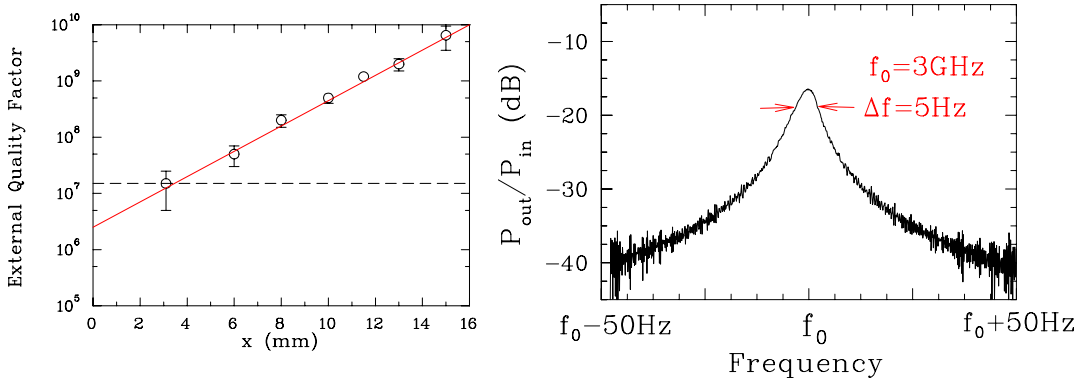


Fig. 19: Variable coupling strength and matched coupling without beam

In a low-energy electron accelerator like the S-DALINAC the beam is not yet completely relativistic and therefore the time needed for one recirculation still slightly depends on energy. On the other hand at the entrance of the main linac the injection phase has to be the same for all three beams. Thus it is necessary (since the accelerator has to provide a wide range of beam energies) to match the lengths of the recirculating beam transport systems for each beam energy. Figure 20 shows how this is done at the S-DALINAC. The encircled parts of the layout (blown up in the insets) indicate where the path length adjustment is performed.

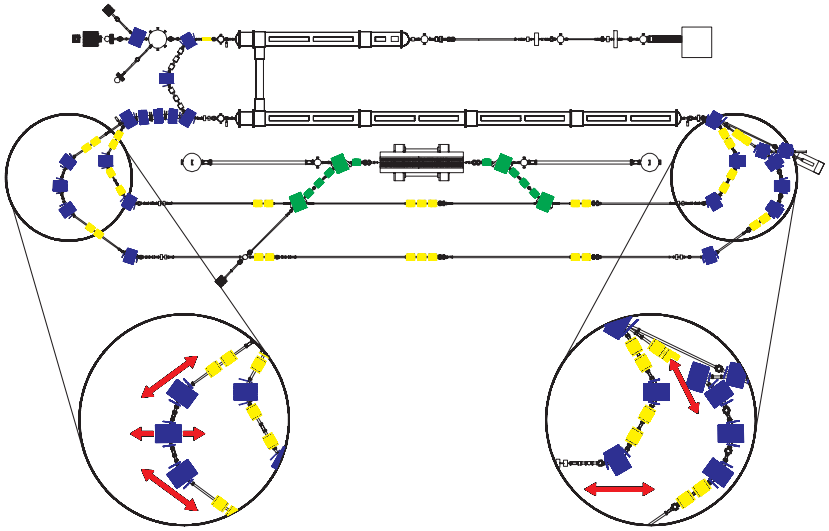


Fig. 20: Path length adjustment in the recirculating beam lines

In the first recirculation two dipole magnets and two quadrupoles (due to modified beam optics, there is now only one quadrupole between the two dipoles) are shifted on linear bearings in the direction of the arrows, allowing for a path length variation of 50 mm, corresponding to 180° of the

RF phase. For the second recirculation (left inset in Fig. 20) a length variation of ΔL = 62 mm (220° of RF phase) is achieved by shifting three dipoles in the indicated directions.

Since good diagnostics is important for the operation of any accelerator a few examples are discussed below to show how transverse and longitudinal properties of the electron beam are determined at the S-DALINAC. Measurements of the transverse intensity distribution of the beam have to be performed for a quantitative analysis of transverse beam parameters. For this purpose either wire scanners, giving projections of the intensity distribution, or CCD cameras, taking an image of the beam via Optical Transition Radiation (OTR) are used. A typical OTR diagnostics set-up is shown in Fig. 21.

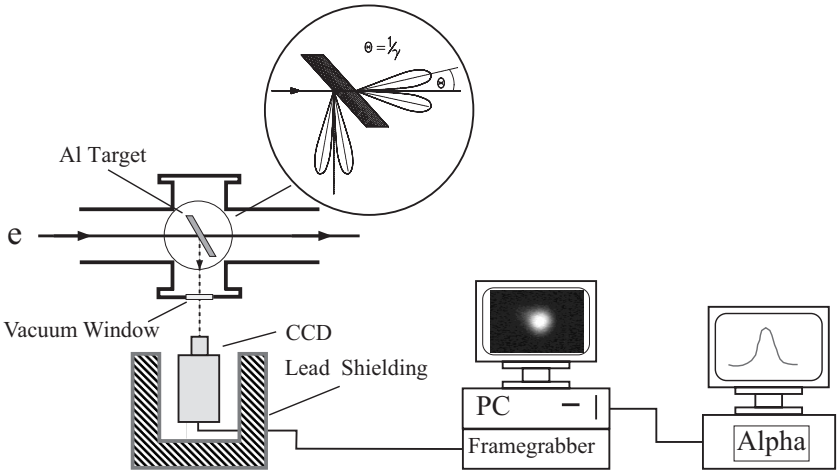


Fig. 21: Schematic set-up of an OTR diagnostics station

The inset in Fig. 21 shows the typical emission characteristics of OTR. The radiation produced by the electron beam hitting a 25 μm thick aluminum target is observed through a standard vacuum window by a well-shielded CCD camera, whose output is digitized in a PC equipped with a framegrabber board. Final analysis is performed on a workstation using IDL [4] as dedicated software. OTR is used for diagnostic purposes at energies ranging from as low as 250 keV up to 120 MeV requiring a minimum beam current of 0.5 μA down to 10 nA, respectively. Results from a typical measurement of transverse beam parameters (injector beam at 5.4 MeV) are displayed in Fig. 22, where the left part shows the two-dimensional intensity distribution of the beam on the aluminum foil (the original is a false colour presentation). The right part of Fig. 22 shows the horizontal half-widths of different intensity distributions obtained by variation of the focusing strength of a quadrupole, located upstream of the diagnostics station. A three parameter fit to the data yields the beam parameters and thus as the usually quoted global quantity the horizontal emittance given in Fig. 22.

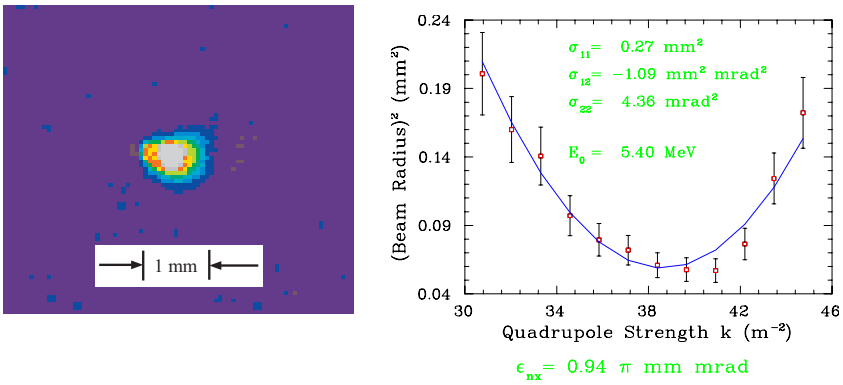


Fig. 22: Determination of transverse beam parameters via OTR

Determination of the transverse intensity distribution of the electron beam in a dispersive section of the lattice contains information on the energy spread of the beam. Results from such a measurement, taken from the beam after leaving the main linac, are displayed in Fig. 23. The two

intensity distributions shown in the left part of the figure correspond to an optimized (left picture) and a non-optimized (right picture) setting of the RF phases in the main linac. The right part of Fig. 23 shows a quantitative on-line display of the measured data: the intensity distribution is projected onto an energy scale and a Gaussian fit with its corresponding half-width are displayed to help the operator in an optimization process.

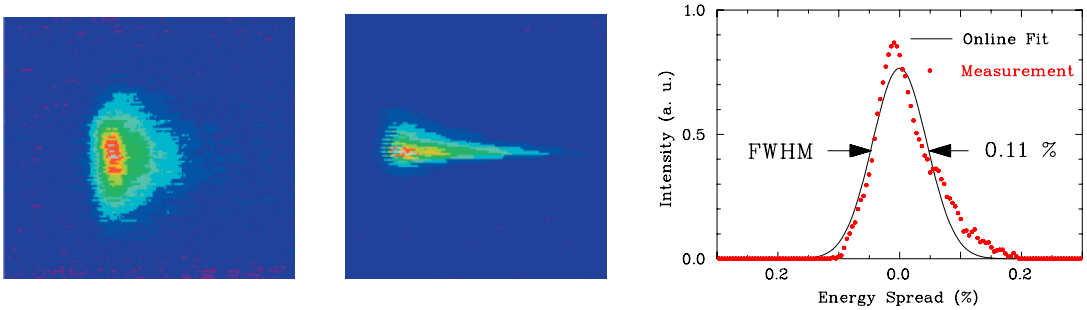


Fig. 23: Determination of energy spread via OTR

Knowledge of the bunch length and thus the peak current is essential for operation of the FEL. The apparatus shown in Fig. 24 (known as the Martin Puplett Interferometer [5]) makes use of the coherent Far InfraRed (FIR) part of transition radiation (the radiation is coherent for wavelengths comparable to or longer than the electron bunch length). A parabolic metal mirror deflects the radiation emitted from the aluminum foil onto a wire grid, which acts as a polarizer. The part of FIR radiation that is polarized perpendicularly to the plane of the drawing is deflected to the right onto a second wire grid, oriented under 45° with respect to the direction of polarization, thus acting as a beam splitter. The split parts of the radiation hit a fixed (top) and a moveable (right) roof mirror that rotate the direction of polarization by 90°. The FIR radiation is then recombined by the beam splitter and has a direction of polarization that depends on the position of the moveable mirror. A second parabolic mirror directs the radiation onto a third wire grid (analyser), from where the part to be detected is reflected into the feed horn of a pyroelectric detector.

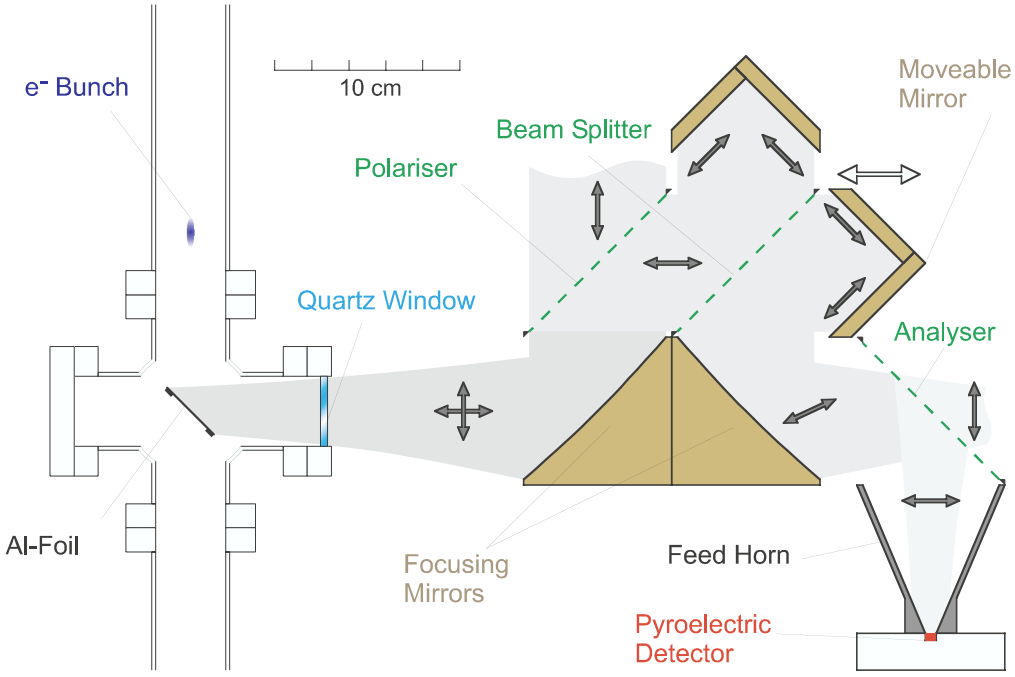


Fig. 24: Interferometer for the FIR part of transition radiation

The response of the detector (autocorrelation) as a function of the mirror position, expressed as optical path difference and converted into picoseconds, is drawn in the left part of Fig. 25. It has to be deconvoluted with respect to the very structured spectral response of the apparatus (including the

detector). The final result, the longitudinal intensity distribution of the electron bunch is shown in the right part of Fig. 25 (electron beam parameters are given in the inset).

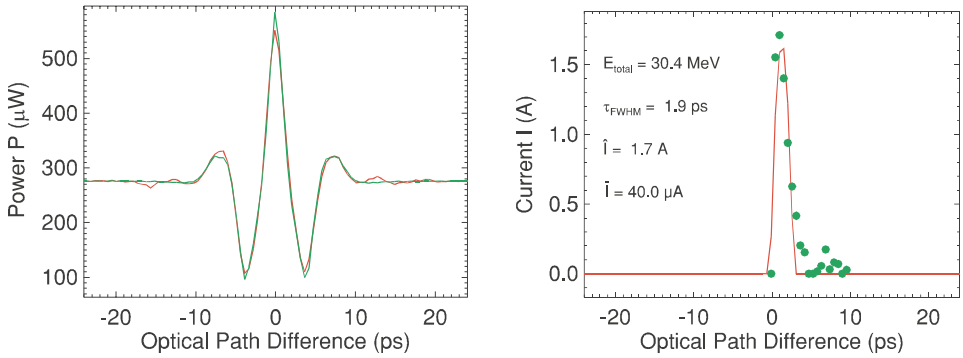


Fig. 25: Measured autocorrelation and corresponding bunch shape

For non-interceptive diagnosis of beam intensity and position RF cavity monitors in the TM_{010}- and TM_{110}-like modes are used. The cavities (a pair is shown in Fig. 26 with the associated electronics) are fabricated from stainless steel. They consist of a common centrepiece and two lids, sealed with copper gaskets. On the outside, the lids allow for a direct connection to standard CF 38 vacuum flanges. The antennas, four on the position monitor and two on the intensity monitor, are connected to 3.5 mm RF connectors through ceramic 50 Ω vacuum feedthroughs welded onto CF 16 flanges.

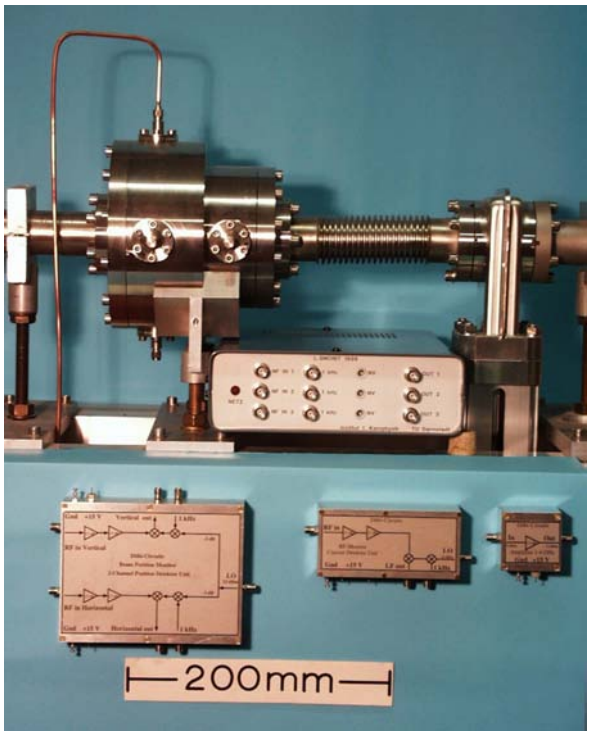


Fig. 26: Beam position and intensity monitors with associated electronics

The cavities have low quality factors ($Q \leq 10^3$), therefore only initial frequency tuning (after fabrication) is necessary and no special temperature stabilization is required. The resulting rather low RF sensitivities (15 nW/(μA)^2 for the intensity monitor and 15 pW/(mm·µA)^2 for the position monitor) are compensated for by the lock in detection technique, resulting in typical resolutions of 10 nA for the beam current and 0.1 mm for the beam position at a current of 1 μA.

4.1.3 CEBAF at TJNAF

CEBAF at TJNAF represents the high-energy end of CW electron accelerators for nuclear physics. A layout of the facility is sketched in Fig. 27 with design specifications for energy, current, emittance, and energy spread of the electron beam.

The electron source of the accelerator (left part of Fig. 24 on top of experimental hall A) produces a polarized 100 keV beam from a strained GaAs photocathode. After a special preparation of the bunch sequence, described below, the beam is accelerated to 500 keV in a 1.5 GHz CW normal conducting accelerating structure before it enters the superconducting injector linac. The superconducting accelerator cavities of CEBAF, originally developed at Cornell University, are 5-cell cavities fabricated from solid niobium, operating at 1.5 GHz. A pair of cavities, assembled head to head (with their RF input couplers in the centre) form a cryounit, which has its own helium vessel. Four such cryounits assembled together form a cryomodule (one cryostat). The injector linac consists of one cryounit (with a special cryostat) and two standard cryomodules. It accelerates the beam to 45 MeV. The main accelerator consists of two linacs that each contain 20 cryomodules (160 cavities). Between successive cryomodules beam diagnostics stations, quadrupoles, and vacuum pumps are installed. The nominal energy gain of each linac (which is exceeded in the meantime) is 400 MeV. After leaving the first linac the beam is deflected vertically in a so-called spreader section. The first pass (lowest energy) goes to the top level of the following 180° arc, is then deflected down to the linac level in a so-called recombiner section, and enters the second linac to gain another 400 MeV. In the same manner, in a second set of 180° arcs the beam is brought back to the entrance of the first linac for its second pass acceleration. Transition to the second linac is accomplished through the second arc from the top one in the set of five 180° arcs (right side of Fig. 27). Since there are four arcs contained in the left set the beam can pass up to five times through each linac. Each one of the nine 180° arcs is composed of many compact dipoles, quadrupoles, and sextupoles. Longitudinally, the beam transport is isochronous. The accelerator serves three experimental halls, A, B, and C (left part of Fig. 27) and it is the very special property of CEBAF that beam can be delivered simultaneously to all three halls, even at different energies and intensities.

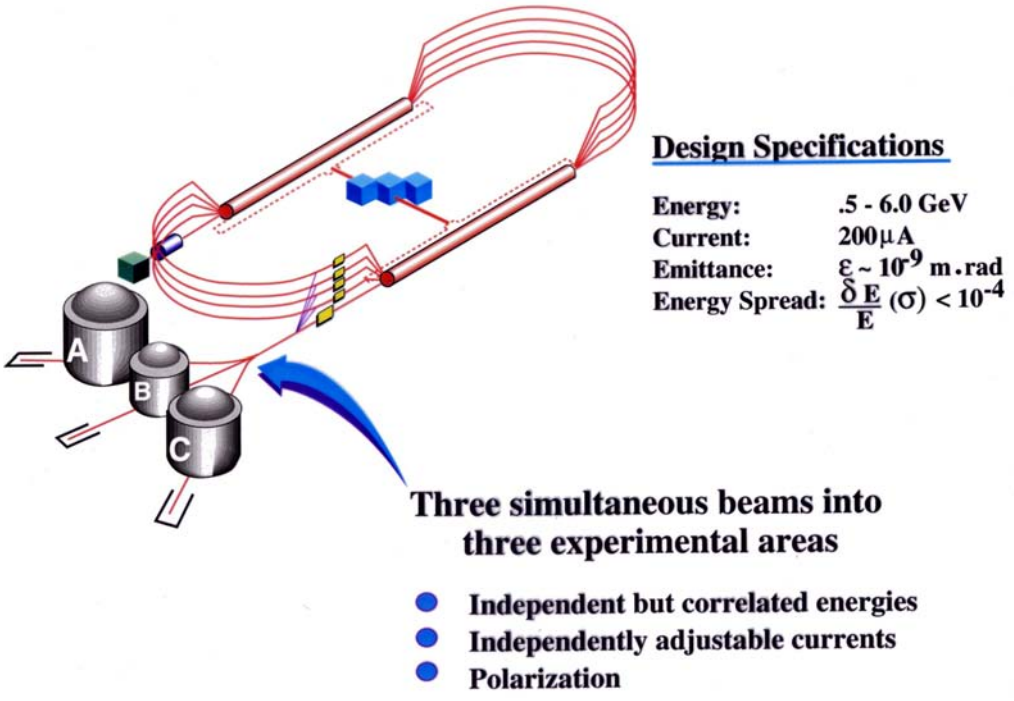


Fig. 27: Layout of CEBAF at TJNAF

The way this is accomplished is illustrated in the following two figures. The principle of basically accelerating three individual beams in the same accelerator is sketched in Fig. 28. It requires, to some extent, individual preparation of successive electron bunches. For example, in the case of CEBAF only every third bunch is identical. This is indicated by the empty, dotted, and full small circles in Fig. 28.

The second requirement is that one must find a way to direct only every third bunch to the same experimental hall and to extract the bunches in between after a different—or the same—number of passes through the accelerators, and to direct them to the other two experimental halls.

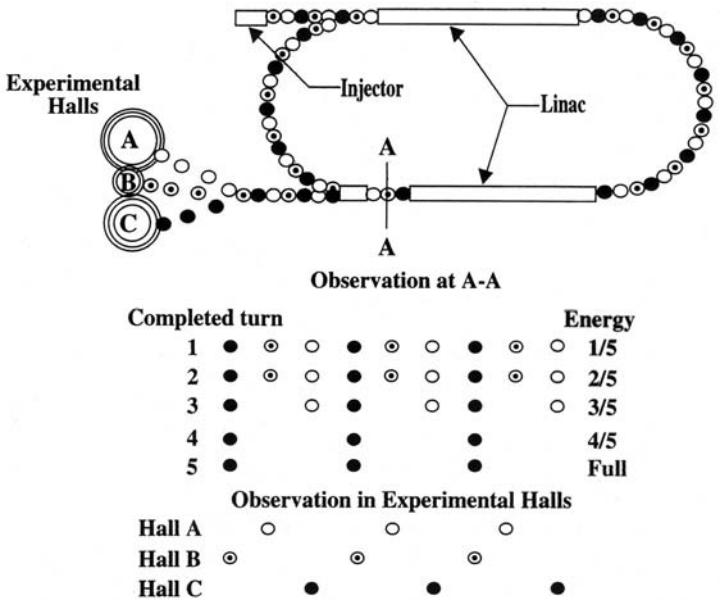


Fig. 28: Scheme of simultaneous beam delivery

Provided this is possible, a situation as sketched in the lower part of Fig. 28 can be achieved: the injector periodically prepares a sequence of three bunches with different charge. At the exit of the second linac (location A-A) the bunches pass in the same periodic order as they leave the injector. After the second pass the dotted bunches, for example, are extracted to hall B and thus are missing at A-A. If the bunches indicated by empty circles are then extracted to hall A the last two passes contain only the full-circle bunches, to be directed finally to hall C at maximum energy. In this case each hall gets its own beam of individual intensity (determined by the preparation in the injector): hall B at 2/5 of the full energy, hall A at 3/5, and hall C at full energy.

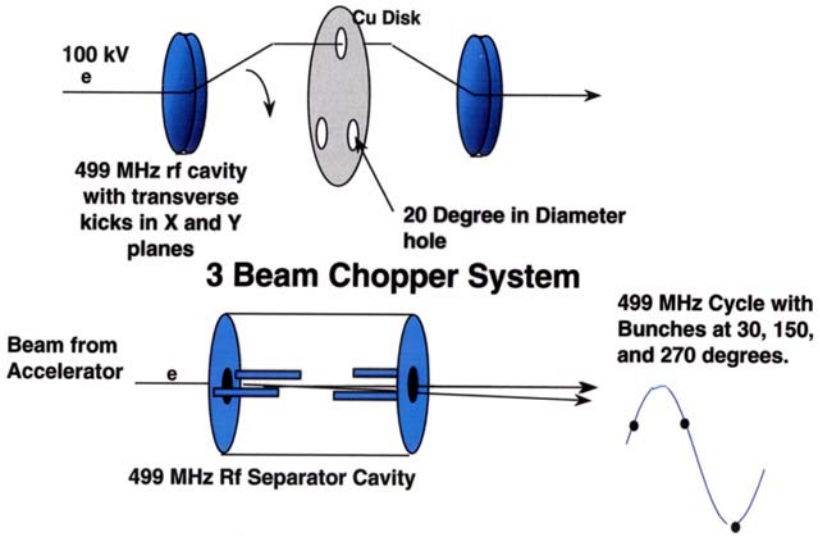


Fig. 29: Generation and separation of three independent beams

The upper part of Fig. 29 shows how (periodically) the individual bunches are prepared in the injector. The 100 keV beam from the electron gun is deflected transversely by an RF cavity which operates at 1/3 of the accelerator frequency. Deflections in the x and y direction are out of phase by

90°, i.e. the beam is deflected tangentially to a cone. If this beam hits the Cu disc with three holes, indicated in Fig. 29, three packets of electrons will pass through the disc while the beam performs one round trip on the cone. The Cu disc is surrounded by a solenoidal lens, which bends the electron packets back towards the original beam axis, where a second RF cavity brings the packets back on axis. Each packet fits into one RF bucket of the accelerator, since the deflection was performed at one-third of the accelerator frequency. Different intensities in the three packets can be achieved either by modulating the laser that illuminates the photocathode properly (in this case the diameter of the holes in the Cu disc are just limiters for the duration of the electron packets) or by replacing the three holes with adjustable slits (the width of the slits determines the charge contained in the electron packets). Extraction of the individual bunches after (partial) acceleration is more difficult because of the increased electron energy, but the first device is again an RF cavity operating at one-third of the fundamental frequency. This cavity is a true separator cavity, consuming a significant amount of RF power and deflecting the beam in one direction only. The right part of Fig. 29 shows that for a certain phasing of the separator two successive bunches are deflected to one side by the same amount, while the third bunch is deflected more strongly to the opposite side. The RF separator is followed by a series of septum magnets to finally direct the individual bunches into the different extraction beam lines.

Despite producing data at an impressive rate, the TJNAF has already studied upgrade possibilities very carefully. The version currently preferred is shown in Fig. 30. The important modifications necessary for the upgrade are listed in the top left of the figure and are indicated in the sketch of the accelerator: five additional cryomodules with newly developed 7-cell cavities must be installed behind each linac. Six of the installed cryomodules must be replaced or reworked. The capacity of the helium refrigerator has to be increased to 10 kW. Magnets and power supplies have to be replaced in the 180° arcs and in the spreader and recombiner sections.

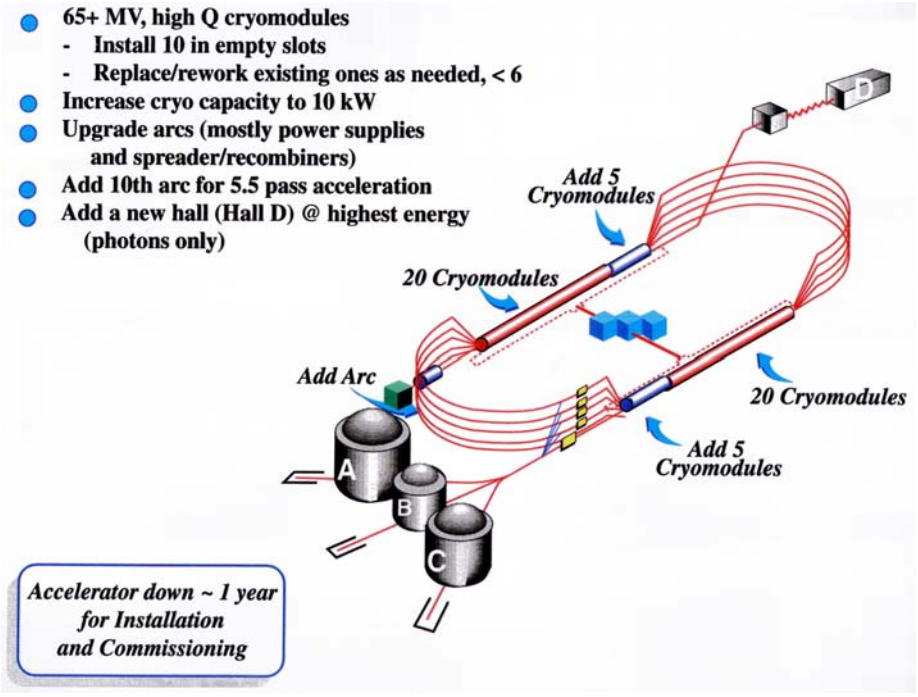


Fig. 30: Proposed upgrade to CEBAF at TJNAF

These modifications are supposed to increase the beam energy for the existing three experimental halls to 10 GeV. For photon experiments at energies up to 12.5 GeV a fifth arc in the left arc section, allowing for a sixth pass through the first linac, and a fourth end station (top right) are proposed.

4.1.4 ELFE at CERN

An Electron Laboratory For Europe (ELFE) significantly exceeding even the energy of the CEBAF upgrade has been under discussion for several years. Following a first study at CERN [6], CERN and NuPECC set up a joint study for ELFE at CERN, a superconducting recirculating 25 GeV electron accelerator that would use the superconducting accelerator cavities of LEP II and other components not required for LHC. In parallel an international study group at CERN looked into the design aspects of such an accelerator and in December 1999 a conceptual design report [7] was published.

A schematic view of ELFE at CERN is shown in Fig. 31 and a brief summary of its main parameters is given in Table 4.

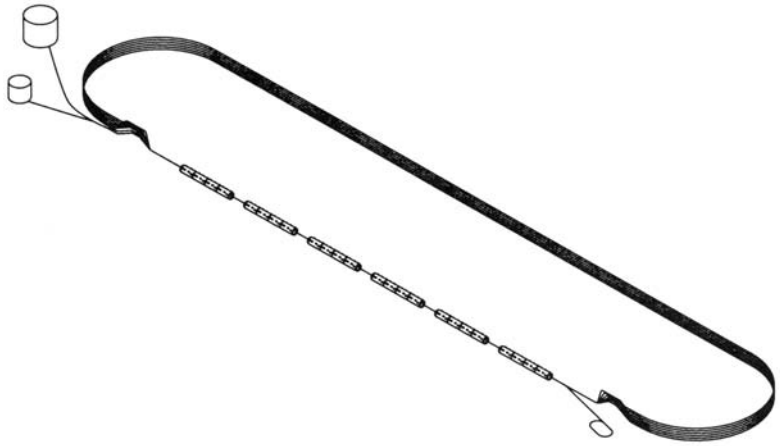


Fig. 31: Schematic view of the ELFE accelerator

ELFE will use a polarized electron source (semiconductor photocathodes as in MAMI and CEBAF). Two scenarios have been investigated for the injector: a two-stage, normal conducting racetrack microtron (very similar to MAMI B, except for two stages instead of the three at MAMI B), or a superconducting recirculating accelerator with two recirculations (quite similar to the scheme of the S-DALINAC). Both injectors have to provide 800 MeV of beam energy. Different from CEBAF, ELFE will use one linac only in its recirculation scheme. Its energy gain will be 3.5 GeV and thus seven passes of the electron beam through the linac are needed to achieve the final energy of 25 GeV. The arcs of the return lines for the beam are arranged in a similar scheme as in CEBAF (on top of each other, with the same radius) but there is only one spreader section (at the exit of the main linac) and one recombiner section (in front of the main linac). Therefore, six straight return lines (parallel to the main linac and on top of each other) are necessary in this recirculation scheme.

Table 4: The most important ELFE design parameters

Top energy	25 GeV
Beam current on target	100 μA
Bunch separation on target	2.8 ns
Injection energy	0.8 GeV
Number of passes	7
Energy gain per pass	3.5 GeV
Relative r.m.s. momentum spread	$\leq 10^{-3}$
Emittance	$\leq$ 30 nm

The first two figures in Table 4 show the enormous amount of beam power (2.5 MW in the case of ELFE) that hits the target in such a high energy nuclear physics accelerator. The fairly large bunch separation of 2.8 ns is of course due to the low operating frequency of the LEP cavities (350 MHz) typical of storage rings. The estimated capital expenditure for the construction of ELFE at CERN amounts to some 400 MCHF, split equally between accelerator components and conventional construction.

4.2 Superconducting electron linacs

The next generation of accelerators providing colliding beams of electrons and positrons at energies in the TeV regime will be linear colliders, two linacs directed from opposite directions at a common interaction point. In order to reach the necessary luminosity (required by the small cross sections at TeV energies) bunch charges of several nC and thus extremely high beam power are required. As a consequence linear colliders are pulsed accelerators with a duty factor of the order of one per cent or less. Traditionally this has been the domain of normal conducting linacs and (as discussed in Section 2) the advantages of superconducting linacs are most obvious for CW accelerators. Detailed studies, however, have shown that superconducting cavities, which provide good power efficiency and can be operated at comparably low frequencies, are very likely the superior choice for a linear collider. They are preferable because of the strongly reduced excitation of wakefields. Since about a decade the feasibility of TESLA has been investigated and, centred at DESY, an international collaboration has worked on the layout of TESLA [8] and established the TESLA Test Facility to demonstrate that such an accelerator can be built, to push superconducting accelerator technology to its limits, and to find the most reliable and cost-efficient way for the construction of a superconducting linear collider.

A schematic layout of the TESLA linear collider is shown in Fig. 32. A train of low-emittance electron or positron bunches is extracted from a source system, compressed longitudinally, and injected into two opposing linear accelerators at an energy of 10 GeV. The bunches are carefully aligned along the accelerator axis as defined by the RF cavities and the quadrupole magnets and are accelerated to 250 GeV in each linac. After acceleration the beam halo is removed by a set of precision collimators, and the bunches are transported and focused down to a few hundred nanometres in the horizontal direction, and to the order of 10 nm in the vertical plane, and are collided head-on. The large bunch spacing (compared with normal conducting linear colliders) allows the use of a bunch-to-bunch feedback system, necessary to ensure that opposing bunches collide head-on at the interaction point. The spent beams are extracted from the interaction region and are either dumped (positrons) or used (electrons) to produce the next batch of positrons.

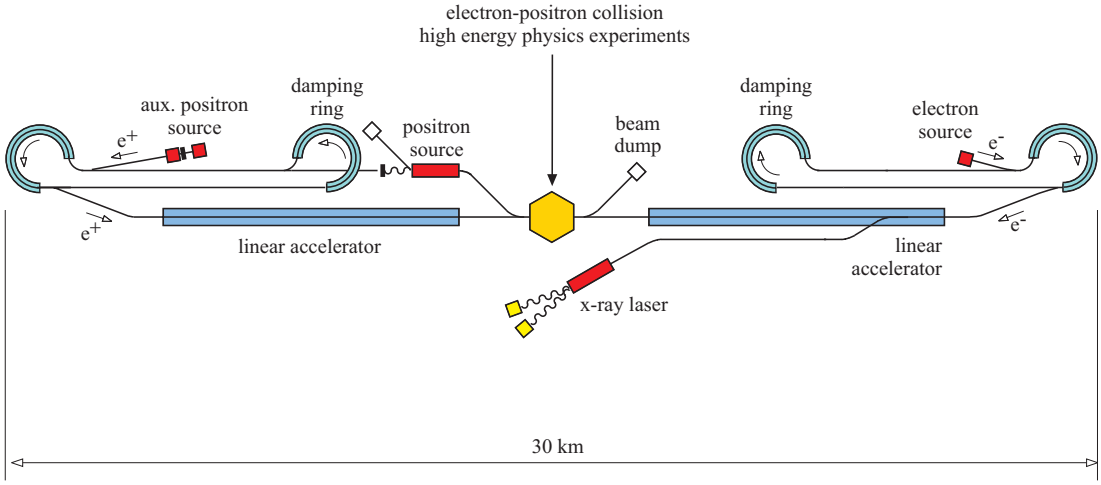


Fig. 32: Schematic layout of TESLA

Interleaved with the bunches for particle physics, bunches of electrons originating from a low-emittance RF photocathode gun are longitudinally compressed and accelerated to energies of between 10 and 50 GeV. They are then extracted from the linac and passed through long, high-precision undulators. The low-emittance electron bunches traversing the undulators will yield a very bright, very short burst of transversely coherent light with tunable wavelengths in the sub-nanometre region. Using

a transverse mode cavity to deflect the beam it is possible to feed several undulators during the same RF pulse, and consequently such a facility can serve a large number of users without interfering with the particle physics programme.

Some of the most important parameters of the 500 GeV linear collider TESLA 500 and of the (original) TTF Linac (TTFL) are summarized in Table 5. A comparison of the figures for the TTFL and for TESLA 500 shows that the most important criteria for the big machine have yet to be demonstrated in the test facility.

Of course, the beam energy is much lower and the number of cryomodules is much smaller in the TTFL, but the 9-cell superconducting cavities developed for TESLA (which in the meantime have become the standard for superconducting cavities) were expected to operate at 15 MV/m and to demonstrate that 25 MV/m for TESLA were not unrealistic. Due to numerous improvements in quality control, fabrication procedures, treatment, handling, and assembly of the cavities a gradient of 25 MV/m is presently already regarded as a standard. Also an unloaded Q-value of 5×10^9 is quite usual for the TESLA cavities. The first injector in TTF used a thermionic source with a bunch charge of a few 10^8 electrons per bunch with a much higher bunch repetition rate. Since then, a photocathode in an RF gun has been used as the source. This provides the same bunch charge and repetition rate as will be used in TESLA 500.

Table 5: Comparison of parameters for TTFL and TESLA 500

Parameter		**TTFL**	**TESLA 500**
Main linac			
Energy	(GeV)	0.500	250
RF frequency	(GHz)	1.300	1.300
Accel. gradient	(MV/m)	15/25	25
Q_0		3×10^9	5×10^9
Cryomodules		4	2500
Injector			
Energy	(GeV)	0.020	10
Particles/bunch		5×10^{10}	5×10^{10}
Bunch separation	(μs)	1.0	1.0
Macro pulse length	(μs)	800	800

In the course of its development and construction the layout of the TTFL was modified with respect to its original design [9]. The present configuration is sketched in Fig. 33.

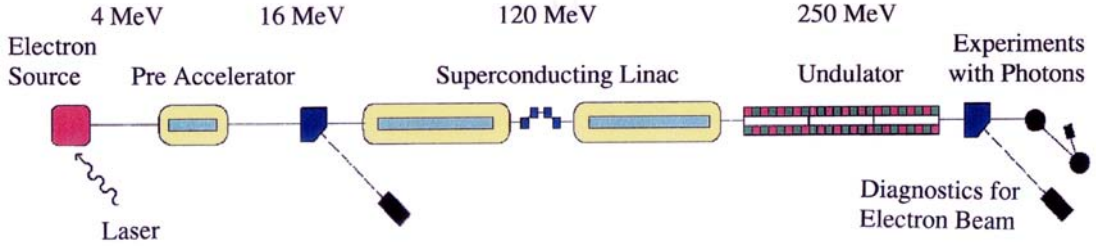


Fig. 33: Layout of the TESLA Test Facility Linac

The tunable X-ray laser envisaged for TESLA is based on the principle of Self-Amplified Spontaneous Emission (SASE) as there are no mirrors available for an optical cavity in the sub-nanometre wavelength range. In order to prove that the SASE principle works at very short wavelengths the number of cryomodules for the TTFL was reduced and a very precise 15 m long undulator was installed downstream of the linac with the necessary diagnostics for the electron beam and the generated laser radiation. First experiments with the configuration (shown in Fig. 33) were successful in the spring of this year; the expected development of the installation is discussed in Section 4.3.

4.3 Superconducting linacs driving an FEL

Free-Electron Lasers require fairly high peak currents, usually provided by short bunches and high bunch charges. Electrostatic accelerators as well as normal conducting and superconducting linacs are used as drivers. For FELs operating in a CW mode superconducting linacs are the only choice (except at low energies where electrostatic accelerators can be used). The basic principle of a conventional FEL is illustrated in Fig. 34. The electron beam from the accelerator is directed through an undulator, which provides a periodically alternating magnetic field (the length of the magnetic period being λ_u). In Fig. 34 the undulator field is oriented horizontally and the electrons perform vertical oscillations about their reference trajectory while passing through the undulator. At the same time they emit synchrotron radiation (spontaneous radiation) that, for small amplitudes of the oscillatory motion, consists of a line spectrum rather than the usual continuous synchrotron radiation spectrum.

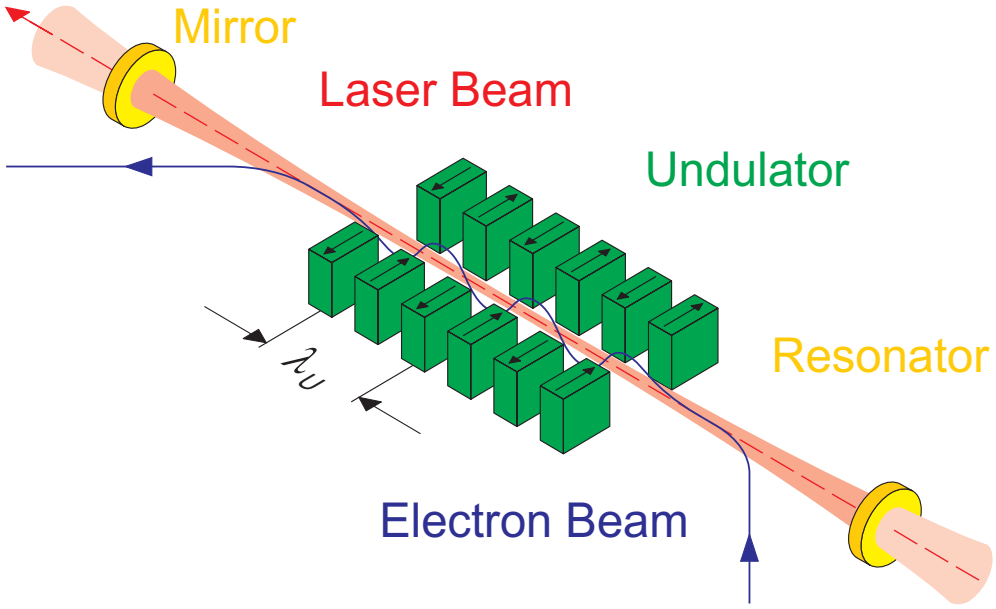


Fig. 34: Principle of conventional free-electron lasers

The fundamental wavelength λ of the spontaneous radiation is given by $\lambda = (\lambda_u/2\gamma^2)(1 + K^2/2)$ with K, the undulator parameter, containing the magnetic properties of the undulator ($K \approx 1$). The pulse of spontaneous radiation generated by an electron bunch while passing through the undulator travels down the optical cavity, is reflected by the downstream (left) mirror, and hits the upstream (right) mirror where it is reflected again. If the length of the optical cavity (distance between the two mirrors) is adjusted correctly the pulse of spontaneous radiation when passing through the undulator from right to left will overlap with the successive electron bunch. Due to their oscillatory motion in the undulator region the electrons can transfer energy into the electromagnetic field of the spontaneous radiation. This amplification for one round trip of the spontaneous radiation is called small signal gain and depends among other parameters on the peak current of the electron beam. If the small signal gain exceeds the losses inside the optical cavity the intensity of the spontaneous radiation increases with each round trip until the laser finally (due to limiting processes) reaches saturation. The laser radiation is coupled out through one of the mirrors; in the case of dielectric mirrors by using one mirror with the

appropriate transmission, in the case of metallic mirrors by a hole of appropriate size in the centre of one mirror.

The rather high peak current necessary to drive an FEL population of each successive RF bucket in a superconducting linac with an electron bunch would result in excessive beam power and demand for RF power. The situation at the S-DALINAC (Fig. 35) illustrates this fact in a most impressive way.

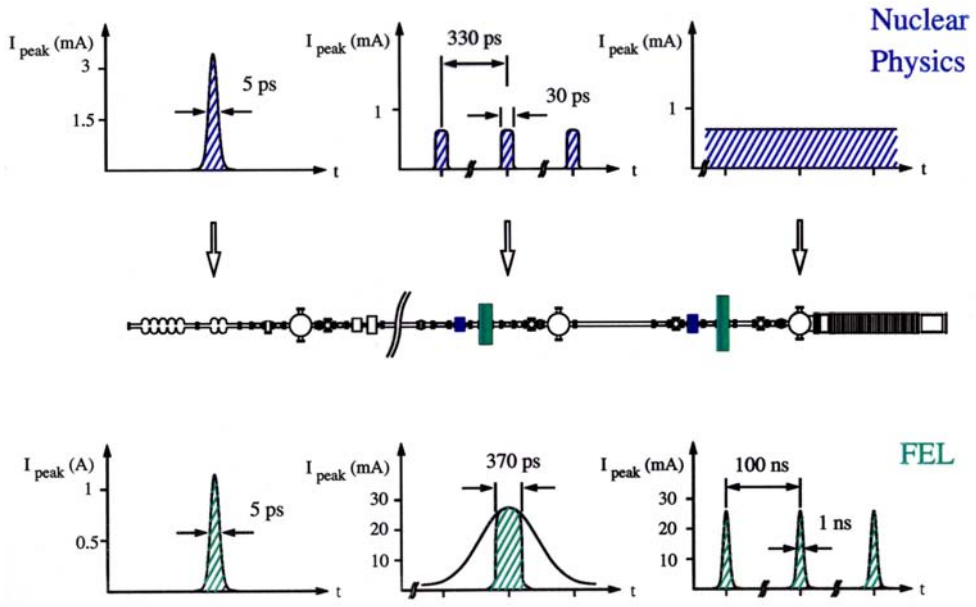


Fig. 35: Time structure for nuclear physics and for FEL operation

Each accelerating cavity has its own small RF transmitter and can provide 300 W for acceleration of the beam. At an energy gain of 5 MeV in the 1 m long cavity this corresponds to a maximum average beam current of 60 μA. Considering a bunch length equal to two degrees of RF phase, the maximum peak current amounts to 10 mA if every RF bucket is populated, which is the case when the accelerator is used for nuclear or radiation physics experiments. The FEL, however, in order to provide enough small signal gain for lasing operation, requires a peak current of some 3 A. This can only be achieved by increasing the bunch charge by a factor of 300 and consequently populating only every 300th RF bucket (thus staying below the RF power limit).

To allow for this type of operation a pulsing option for the electron gun, operating at a repetition rate of 10 MHz (300th subharmonic of the accelerator frequency) and a 600 MHz (5th subharmonic) chopper/prebuncher system had to be incorporated into the 250 keV injection of the accelerator. The central part of Fig. 36 displays a schematic layout of the modified injection, consisting (from right to left) of a 10 kV thermionic gun, a 250 kV electrostatic preaccelerator tube, a 3 GHz chopper/ prebuncher system for nuclear physics operation, a 600 MHz chopper/prebuncher system for FEL operation, chopping aperture, and the superconducting 2-cell and 5-cell capture sections.

Figure 36 shows the time structure of the electron beam at the locations indicated by the arrows for the two modes of operation. For nuclear physics experiments the chopper cavity and the water-cooled chopper aperture chops the DC current from the electron gun into pulses of 30 ps. The prebuncher cavity compresses the width of these pulses to 5 ps at the entrance of the capture section of the superconducting injector linac. Since the beam is both accelerated and bunched in the capture sections, the bunch length is further reduced to 2 ps at the end of the injector.

For FEL operation the 1 ns wide pulses from the electron gun are chopped to a width of 370 ps by the 600 MHz subharmonic chopper and the chopper aperture. The subharmonic prebuncher then compresses the pulse width again to 5 ps at the entrance of the capture section. Similar to the 3 GHz CW operation, the capture sections reduce the bunch length to 2 ps while increasing the peak current. Thus, the peak current corresponding to the emitted electron current of 27 mA from the gun amounts to 2.7 A.

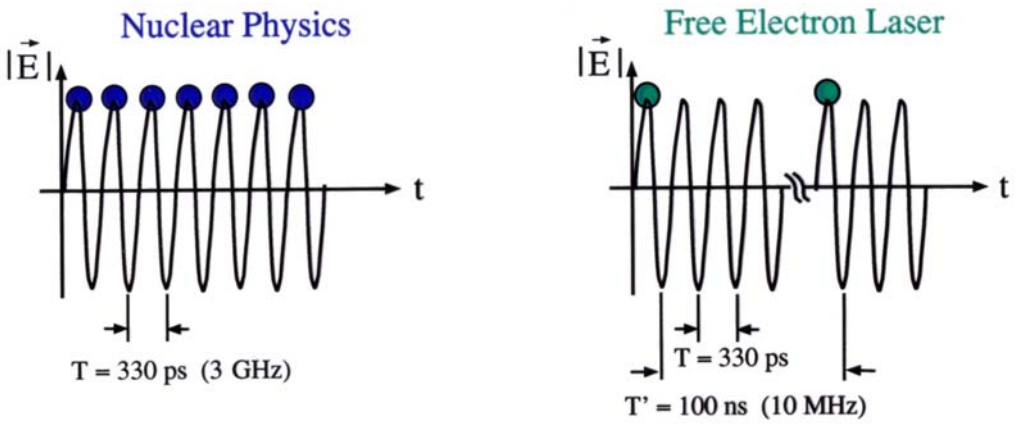


Fig. 36: Time structure of the beam for nuclear physics (3 GHz CW) and FEL (10 MHz CW) mode of operation

The difference between spontaneous radiation at the very beginning of the FEL start-up and the laser radiation during saturated operation is shown in Fig. 37, which displays the spectra for the two conditions.

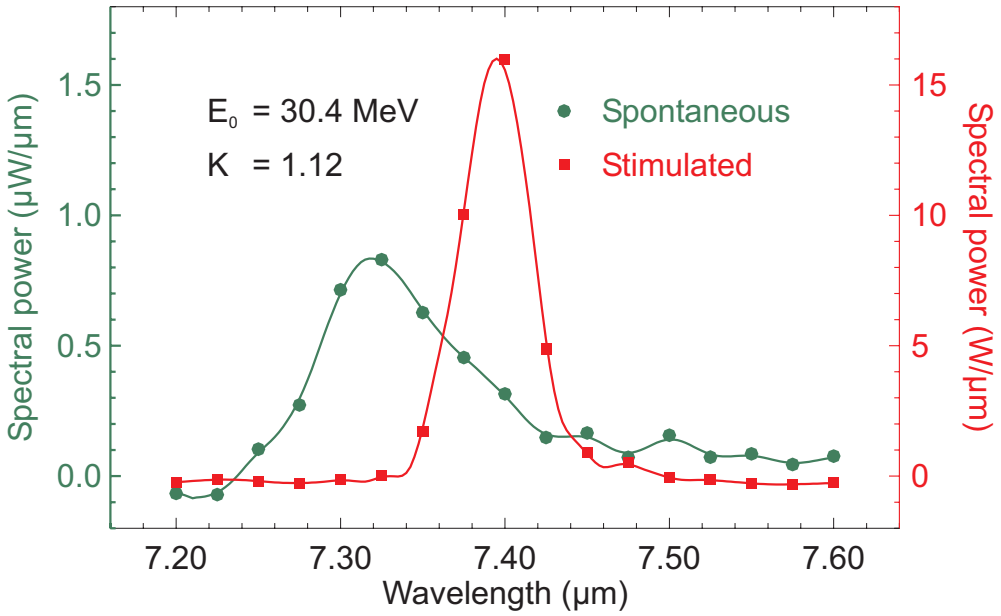


Fig. 37: Spectra of spontaneous (left ordinate) and laser radiation (right ordinate) from the S-DALINAC FEL

The spectral power displayed on the ordinate axes (left axis for spontaneous radiation, right axis for stimulated radiation) increases by more than seven orders of magnitude, in agreement with the fact that some 10^7 electrons in a bunch emit coherently in stimulated emission. Another fact is also beautifully displayed by Fig. 37: according to FEL theory (Madey's Theorem) the spectral width of the spontaneous radiation is reduced by a factor of two in the case of stimulated emission.

Presently the most powerful (average power) FEL is the IR DEMO at TJNAF. The installation (roughly sketched in Fig. 38) uses a 350 kV electron gun with a laser-driven photocathode. The superconducting injector linac consists of a short cryomodule, housing two CEBAF 5-cell cavities. The injector produces a 10 MeV beam with an r.m.s. energy spread of 15 keV, the bunch length amounts to 2 ps and bunch repetition frequencies ranging from 4.7 to 74.8 MHz can be provided. The linac currently consists of a standard CEBAF cryomodule equipped with eight 5-cell cavities. It provides an energy gain of 38 MeV. The beam after leaving the linac is passed around the upstream mirror of the optical cavity and the bunch length is magnetically compressed to 0.4 ps before the beam is directed through the undulator (wiggler).

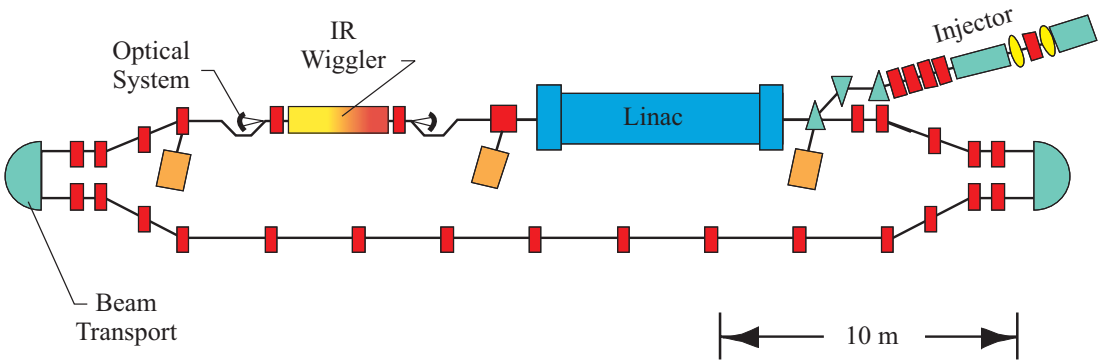


Fig. 38: Layout of the IR DEMO FEL at TJNAF

After passing the undulator and the downstream mirror of the optical cavity the electron beam can be recirculated by means of a very carefully designed beam transport system. In the case of the IR DEMO FEL the purpose for recirculating the beam is *not* to accelerate it one more time, but to decelerate the beam to 10 MeV and return the energy it draws from the linac cavities during its acceleration directly back to the RF field in those cavities. The beam is dumped at an energy of 10 MeV after its second passage through the linac following separation from the first pass beam by a magnetic chicane. Because, after passing the undulator, the electron beam has an r.m.s. energy spread of 2 MeV, it covers large transverse dimensions in dispersive sections of the recirculation system. An isochronous recirculation would leave the energy spread of the beam unaltered, even after deceleration, i.e. the 10 MeV beam would still have an r.m.s. energy spread of 2 MeV. Therefore transverse and longitudinal properties of the recirculation system have been carefully designed, including higher order corrections. As a result the decelerated beam leaves the linac with an r.m.s. energy spread of 0.5 MeV and can be dumped easily. The IR DEMO FEL has beautifully shown that this energy recovery scheme works almost perfectly. A beam current of up to 5 mA has been accelerated, producing an infra red beam with 1.7 kW of average power. At the same time it was demonstrated that in the energy recovery mode of operation the amount of RF power necessary to excite the linac cavities to their respective gradients did not depend on the beam current, whereas without energy recovery acceleration of 1 mA of beam current doubled the linac cavities' RF power consumption. The injector linac of course has to provide the full power necessary to accelerate the 5 mA beam to 10 MeV.

The SASE FEL of the TTF at DESY in its present configuration (displayed in Fig. 33) has successfully demonstrated that an FEL based on the SASE principle works in the UV wavelength region. Laser operation with a single pass gain of $G \approx 1000$ was achieved in the 80–180 nm wavelength range. This shows that long undulators with the extreme precision required by a SASE FEL can be built and that the transverse position of the electron beam can be controlled to within a few tens of μm.

Figure 39 indicates how the development of SASE FELs at DESY will continue. For phase one a third TESLA cryomodule with eight cavities will be inserted to increase the beam energy to 390 MeV. Two stages of bunch compression will reduce the bunch length to 250 μm. With the present undulator a laser beam of 44 nm wavelength, corresponding to a photon energy of 28 eV, will be produced. Then in a major upgrade of the facility (phase two) a third stage of bunch compression will be added at an energy of 390 MeV to reduce the bunch length to 50 μm. Five more TESLA cryomodules will then produce an electron beam of 1 GeV. The length of the present undulator will be doubled (necessary for the SASE principle to work at 1 GeV) and the laser beam generated by this facility will have a wavelength of 6 nm (200 eV photon energy). This source of continuously tunable coherent light in the VUV wavelength range with extreme brightness will of course become a true users' facility.

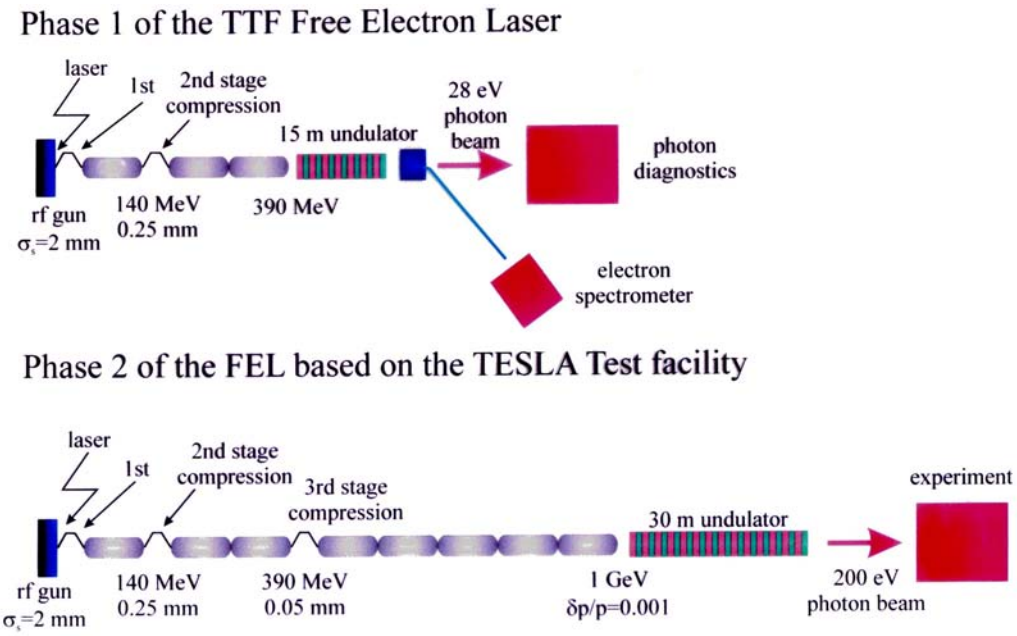


Fig. 39: Steps toward a 6 nm SASE FEL at DESY

5. CONCLUSION

First attempts to use superconducting RF structures for particle acceleration were undertaken at Stanford University and the Kernforschungszentrum Karlsruhe in the 1960s. The optimistic expectations from the new technology were inevitably followed by a few experimental drawbacks. It was only after some years of development that the first electron accelerators were assembled at the University of Illinois and at Stanford University, and the first superconducting booster for heavy ions became operative at Argonne National Laboratory. The S-DALINAC is historically the third superconducting electron linac. In the meantime the technology has matured. Superconducting RF structures have for many years been proven to operate not only in linacs but also in storage rings, even with extremely high beam currents, e.g. the Cornell B-factory. Large installations involving many cavities, like LEP II (some 200 cavities) or CEBAF (more than 330 cavities), have demonstrated the reliability of the technology. Therefore, old prejudices have disappeared; designs using superconducting technology are carefully studied and quite often appear to be the most effective and attractive solution for almost all big future accelerator projects, be they linear colliders, high-intensity proton accelerators, or accelerators for radioactive beams.

ACKNOWLEDGEMENTS

We are grateful to L.S. Cardman, S. Döbert, K.H. Kaiser, G.R. Neil, and H. Weise for generously providing us with information and material on the accelerators at their respective home institutions. In particular we are indebted to S. Kostial for his great help in the course of the preparation of the tutorial and this manuscript.

This work has been supported by the Deutsche Forschungsgemeinschaft under contract FOR 272/2-1 and the Graduiertenkolleg GRK 410.

REFERENCES

[1] H. Padamsee, J. Knobloch, and T. Hays, *RF Superconductivity for Accelerato* (John Wiley & Sons, New York, 1998)

[2] R. E. Rand, *Recirculating Electron Accelerators* (Harwood Academic Publishers, London, 1984)

[3] IKP Mainz (MAMI) Web page, http://www.kph.uni-mainz.de at the time of publication.

IKP Darmstadt (S-DALINAC) Web page, http://www.s-dalinac.de at the time of publication.

TJNAF (CEBAF, IR DEMO FEL) Web page, http://www.cebaf.gov/accel/index.html at the time of publication.

CERN (ELFE) Web page, http://hbu.home.cern.ch/hbu/Elfe/Elfe.html at the time of publication.

DESY (TTF, TESLA) Web page, http://tesla.desy.de at the time of publication.

[4] Interactive Data Language, Version 3.0, Research Systems (1993) 4.

[5] D.H. Martin, E. Puplett, *Inf. Phys.* **10** (1970) 105.

[6] G. Geschonke and E. Keil, 'A recirculating electron accelerator (ELFE) using the LEP superconducting RF cavities', CERN–SL–98–060–RF, CERN, Geneva (1998).

[7] H. Burkhardt (ed.), 'ELFE AT CERN, Conceptual Design Report', CERN 99–10, CERN, Geneva (1999).

[8] R. Brinkmann, G. Materlik, J. Rossbach and A. Wagner (eds.), 'Conceptual design of a 500 GeV e^+e^- linear collider with integrated X-ray laser facility', DESY Report 1997–048, ECFA Report 1997–182 (1997).

[9] D.A. Edwards (ed.), 'TESLA Test Facility linac—Design report', TESLA 95–01 (1995).

PARTICIPANTS

AGHAJAFARI, R.	NRCAM, Karaj, Iran
AHRENS, J.	GSI, Darmstadt, Germany
ALESINI, D.	INFN-LNF, Frascati, Italy
APPELBEE, C.W.	RAL, Didcot, UK
ARNAUD, E.	LURE, Orsay, France
AYVAZYAN, V.	DESY, Hamburg, Germany
BALDINI, P.	INFN-LNF, Frascati, Italy
BALLANTINI, R.	INFN Genova, Genoa, Italy
BERNAUDIN, P-E.	CEN-Saclay, Paris, France
BLAS, A.	CERN, Geneva, Switzerland
BLEDNYKH, A.	Moscow Engineering Physics Inst., Moscow, Russia
BLELL, U.	GSI, Darmstadt, Germany
BROERE, J.	CERN, Geneva, Switzerland
BUSSE-GRAWITZ, M.E.	PSI, Villigen, Switzerland
BYLINSKI, Y.	CERN, Geneva, Switzerland
CARUSO, A.	INFN-LNS, Catania, Italy
CHAI, J-S.	KCCH/KAERI, Seoul, Korea
CHUN, M-H.	PAL, POSTECH, Kyungbuk, Korea
CRAIEVICH, P.	Sincrotrone Trieste S.c.p.A., Trieste, Italy
DAMERAU, H.	Univ. of Dortmund, Germany
DAVINO, D.	Università Federico II di Napoli, Naples, Italy
EICHHORN, R.	GSI, Darmstadt, Germany
EILERS, G.	FU Berlin, Germany
EMMERLING, M.	GSI, Darmstadt, Germany
GOPYCH, M.	TU Darmstadt, Germany
GORBATCHEV, E.	JINR, Dubna, Russia
GRAWER, G.	CERN, Geneva, Switzerland
GUTOWSKI, W.	GSI, Darmstadt, Germany
HALE, P.	RAL, Didcot, UK
HELLGREN, H.	CERN, Geneva, Switzerland
HOUK, Y.	Bogolyubov Inst. Theoretical Physics, Kiev, Ukraine
IVAN, G.	Nat. Inst. of Lasers, Bucharest, Romania
KASPAR, K.	GSI, Darmstadt, Germany
KESTER, O.	LMU München, Munich, Germany

KOENIG, H.G.	GSI, Darmstadt, Germany
KOSTIAL, S.	TU Darmstadt, Germany
LAD, M.	FZK, Karlsruhe, Germany
LAIER, U.	TU Darmstadt, Germany
LAMANNA, G.	INFN Bari, Italy
LOSITO, R.	CERN, Geneva, Switzerland
LU, J.	TRIUMF, Vancouver, BC, Canada
LU, H-L.	Shanghai University, China
MARQVERSEN, O.	CERN, Geneva, Switzerland
MASULLO, M.R.	INFN Complesso Univ. M.S.A., Naples, Italy
MCMONAGLE, G.	CERN, Geneva, Switzerland
MIKKELSEN, F.	University of Aarhus, Denmark
MORVILLO, M.	CERN, Geneva, Switzerland
MOTOS-LOPEZ, T.	CERN, Geneva, Switzerland
OANE, M.	Nat. Inst. Lasers, Bucharest, Romania
PALMIERI, A.	INFN LNL, Legnaro, Italy
PANDE, S.A.	BESSY GmbH, Berlin, Germany
PEDROZZI, M.	PSI, Villigen, Switzerland
PEPLOV, V.	INR of RAS, Moscow, Russia
PIVI, M.	CERN, Geneva, Switzerland
PLATZ, M.	TU Darmstadt, Germany
PRAESTEGAARD, L.	ISA, Aarhus, Denmark
RIPPON, C.	LURE, Orsay, France
RUNKEL, S.	J.W. Goethe-University, Frankfurt, Germany
SCHNEGG, A.	FU Berlin, Germany
SHULYAKOVSKY, R.	Institute of Physics, Minsk, Belarus
SØBY, L.	CERN, Geneva, Switzerland
SOULIMOV, A.	Moscow Engineering Physics Inst., Moscow, Russia
STASSEN, R.	Forschungszentrum Jülich, Germany
STEVENS, A.	RAL, Didcot, UK
STINGELIN, L.	PSI, Villigen, Switzerland
STIRBET, M.	CERN, Geneva, Switzerland
URUSOVA, E.	Institute of Nuclear Physics, Tashkent, Uzbekistan
VARDANYAN, A.	Yerevan Physics Institute, Yerevan, Armenia
WATZLAWIK, S.	TU Darmstadt, Germany
ZENNARO, R.	CERN, Geneva, Switzerland