

Plasma Confinement

Allen H. Boozer
Columbia University

I. Introduction
II. Basic Plasma Physics
III. Toroidal Plasmas
IV. Mirror Confined Plasmas
V. Transport

GLOSSARY

Ambipolar Equal loss rate of ions and electrons. If there is more than one particle-loss process, the individual processes need not be ambipolar. However, quasi-neutrality forces the sum of all the loss processes to be ambipolar.

Anomalous transport The transport of plasmas across magnetic field lines is generally much more rapid than would be expected if the magnetic and electric fields were smooth on a spatial scale small compared to that of the overall plasma equilibrium. The enhanced transport due to small-scale electric and magnetic fields is said to be anomalous.

Debye length The Debye length λ_d is the thermal velocity of the electrons divided by the plasma frequency. It is approximately the distance that the electrons and ions can be separated before the electrostatic energy equals the thermal energy.

Hamiltonian mechanics The Newtonian equations of motion $d\mathbf{p}/dt = -\boldsymbol{\nabla} V(\mathbf{x}, t)$ are obviously equivalent to Hamilton's equations, $d\mathbf{p}/dt = -\partial H/\partial \mathbf{x}$ and $d\mathbf{x}/dt = \partial H/\partial \mathbf{p}$, with the Hamiltonian $\mathrm{H}(\mathbf{p}, \mathbf{x}, t) = \frac{1}{2} m v^2 + V(\mathbf{x}, t)$ and the canonical momentum $\mathbf{p} = m\mathbf{v}$. Any set of differential equations that can be written in the form of Hamilton's equations are said to be Hamiltonian with $\mathbf{p}$ the canonical momentum and $\mathbf{x}$ the canonical position, regardless of the form of H. A particle of charge q in an electromagnetic field with vector potential $\mathbf{A}$ and electric potential V has a canonical momentum $\mathbf{p} = m\mathbf{v} + q\mathbf{A}$. The potential in the Hamiltonian is $V = q\Phi$.

Kinetic theory Theory that uses the distribution of electrons and ions in both position and velocity is called kinetic theory. The basic differential equation of the theory is the Fokker–Planck, or Boltzmann, equation, which is derived under the assumption that the number of electrons in a sphere with a radius of a Debye length is very large compared to unity.

Magnetohydrodynamic (MHD) theory The MHD theory of plasmas makes the simplifying assumption that a plasma can be viewed as a conducting fluid with well-defined conductivities for heat and electrical currents. Individual equations based on MHD theory can often be justified under much more general assumptions than those required to obtain the complete MHD theory.

Plasma frequency ω_p The plasma frequency is the natural frequency of the electron oscillation that would

result from a separation of the electrons and the ions that form the plasma.

Quasi-neutrality approximation The density of positive and negative charges is almost equal in most plasmas. The ratio of the net charge density to the density of electron charge is approximately the square of the ratio of the Debye length to the size of the plasma. The electric field in plasma physics is normally determined by equating the electron and the ion charge densities, which is called the quasi-neutrality approximation.

A PLASMA is a near ideal gas that consists of ions and electrons. All substances become plasmas at a sufficiently high temperature—temperatures above about $10^4\,^\circ$C. The confinement of a laboratory plasma for a time long compared to the transit time of a thermal ion requires the use of magnetic fields to balance the plasma pressure. Astrophysical plasmas may be confined by gravitational forces. Gravitationally confined plasmas are discussed as a part of other subjects such as stellar structure. The equilibrium, stability, and transport properties of plasmas that are confined by an embedded magnetic field are major topics of research and are the subject of this article. The physics of plasma confinement is closely related to a number of other areas such as fluid mechanics, kinetic theory, and Hamiltonian mechanics.

I. INTRODUCTION

Matter at temperatures sufficiently high compared to atomic ionization potentials, about $10^4\,^\circ$C, forms a plasma over a broad range of density. The confinement of plasma by a magnetic field was originally studied to gain understanding of space and astrophysical phenomena, particularly phenomena of the solar corona and the magnetosphere of the earth. Although space and astrophysical plasmas remain an active area of study, plasma physics research has been dominated since the 1950s by the effort to achieve adequate plasma confinement for thermonuclear fusion. The thermonuclear fusion application places requirements on the plasma temperature, density, and energy confinement time as well as an upper limit on the magnetic field strength. These requirements have tended to define the plasma regimes of research interest. This is especially true since astrophysical plasmas are frequently in similar dimensionless parameter regimes.

The basic concept of magnetic confinement is simple. A current of electric charge, with density $\mathbf{j}$, interacts with a magnetic field $\mathbf{B}$ to produce a force $\mathbf{j}\times\mathbf{B}$ per unit volume. This force can balance the force per unit volume ∇p that is produced by a gradient in the plasma pressure, or $\nabla p = \mathbf{j}\times\mathbf{B}$. Ampere's law, $\nabla\times\mathbf{B} = \mu_0\mathbf{j}$, implies the confining current modifies the magnetic field. Force balance, Ampere's law, and the condition that the magnetic field be divergence free, $\nabla\cdot\mathbf{B} = 0$, define a large fraction of the physics of plasma confinement.

A simple result of plasma confinement theory is that a plasma cannot be confined purely by a self-produced magnetic field. Either there must be an external current-carrying circuit, which produces some part of the magnetic field, as in laboratory plasmas or in the magnetosphere of the earth, or there must be an additional force, such as gravitation in astrophysical plasmas. To prove that an externally produced magnetic field is needed to confine a plasma, write the magnetic field as the sum of a field produced by plasma currents and an externally produced field, $\mathbf{B} = \mathbf{B}_{\rm pl} + \mathbf{B}_{\rm ex}$. The equilibrium equation is then $\nabla p = \mathbf{j}_{\rm pl}\times(\mathbf{B}_{\rm pl}+\mathbf{B}_{\rm ex})$ with $\mu_0\mathbf{j}_{\rm pl} = \nabla\times\mathbf{B}_{\rm pl}$. Multiplying this form of the equilibrium equation by the position vector $\mathbf{x}$ and integrating over all of space, one finds (using vector identities, such as $\nabla\cdot\mathbf{x} = 3$) that

$$\int\left(3p + \frac{B_{\rm pl}^2}{2\mu_0}\right)d^3x = -\int\mathbf{x}\cdot(\mathbf{j}_{\rm pl}\times\mathbf{B}_{\rm ex})\,d^3x. \quad (1)$$

The left-hand side of this equation is greater than zero, so the external magnetic field must be nonzero.

Suppose a plasma is to be confined in a bounded region of space by a magnetic field. A number of questions should be answered. First, what properties must the magnetic field have in order to confine the plasma? Second, how high a plasma pressure can be confined by the magnetic field? The relevant dimensionless parameter is called beta, $\beta \equiv 2\mu_0 p/B^2$, and is a measure of the efficiency of the utilization of the magnetic field. Third, how long can the plasma be confined or, equivalently, what sources of energy and particles are required to maintain a given plasma configuration? A plasma in thermodynamic equilibrium cannot be confined by a magnetic field, because plasma particles would then have a Maxwellian distribution that carries no current $\mathbf{j}$. Plasma confinement by a magnetic field implies entropy production, which is balanced either by the decay of the configuration or by external sources of energy and particles. Although these questions identify the types of issues that must be addressed, they represent too simple a view of plasma confinement to be answered definitively.

Important but subtle points constantly arise in the theory of plasma confinement. Consider a trivial consequence of the force balance equation, $\mathbf{B}\cdot\nabla p = 0$. The magnetic field vector must lie in a surface of constant pressure. What are the implications? First, the constant pressure surfaces must be toroidal. It is a mathematical theorem that the only surface in three dimensions that can have a

finite vector field everywhere tangent to it is a topological torus. Second, the magnetic field line trajectories, which satisfy the differential equation $d\mathbf{x}/d\tau = \mathbf{B}(\mathbf{x})$, must lie in the toroidal surfaces wherever the pressure gradient is nonzero. Toroidal surfaces formed by magnetic field lines are called magnetic surfaces. (The parameter τ is just a label for points along a field line trajectory with $d\tau = dl/B$ and dl the differential distance along a field line.) As shown in Section III.A, the differential equation for the field lines of a divergence-free field is equivalent to a one degree of freedom, time-dependent, Hamiltonian mechanics problem. The condition that the magnetic field lines lie in nested toroidal surfaces, the surfaces of constant pressure, is only satisfied by a magnetic field in special cases (corresponding to integrable Hamiltonians). An arbitrary magnetic field configuration will not confine a plasma with a scalar pressure. The presence of a continuous symmetry guarantees the existence of magnetic surfaces. However, the only continuous symmetry that is consistent with a confined plasma is toroidal. In theoretical studies, plasmas that have cylindrical and helical symmetry are considered, but in both cases the plasma is not confined in a symmetry direction.

Even if the externally imposed magnetic field has perfect toroidal symmetry, the plasma configuration need not be symmetric. The free energy contained in the plasma pressure gradient and current can spontaneously create asymmetries. These spontaneously produced asymmetries, called plasma instabilities, will be discussed in Section III.C. Plasma instabilities place the practical constraint on the plasma beta, $\beta \equiv 2\mu_0 p/B^2$.

This discussion has implicitly assumed that the plasma is confined a long time compared to the ion and electron collision frequencies, which is a requirement for efficient fusion energy production. A long confinement time compared to the collision frequencies implies that the distribution functions of the ions and the electrons are nearly Maxwellian and the force exerted by the plasma is given by the gradient of a scalar pressure p. If the confinement time is comparable to a collision frequency, the force exerted by the plasma must be represented using a tensor rather than a scalar pressure. A tensor pressure permits plasma confinement along the magnetic field lines to regions of low magnetic field strength. This is called magnetic mirror confinement. See Section IV.

The conditions required for a magnetic field to confine a near-Maxwellian plasma are greater than just the existence of magnetic surfaces. For plasmas of thermonuclear fusion interest, the ions and electrons move a distance far longer (tens of kilometers) than the size of the plasma (meters) between collisions. Consequently, thermonuclear plasmas are often described as collisionless. Nonetheless, the confinement time for energy must be about 100 ion collision times and 10^4 electron collision times. Consequently, thermonuclear plasmas are sufficiently collisional to have near-Maxwellian electrons and ions. The paradoxical collisionality regime of plasmas of interest for thermonuclear fusion places a constraint on the properties of the particle trajectories. The particle trajectories in the self-consistent magnetic and electric fields must remain close to the constant-pressure surfaces. This constraint is not easily satisfied, except in plasmas with toroidal symmetry, and will be discussed further in Section V.

The electric field has a major role in determining the particle trajectories, but essentially no direct role in the overall plasma force balance. Why is this? The electric field $\mathbf{E}$ prevents the particle trajectories from reaching points in space when the associated electric potential is comparable to the thermal energy. This is equivalent to $eEL \sim T$ with L the gradient scale length, e the magnitude of the electron charge, and T the plasma temperature using energy units. (If conventional temperature units are used, T is replaced by $k_B T$ with k_B the Boltzmann constant.) The net charge density, $e\tilde{n}$, is related to the electric field by Gauss's law, $\nabla \cdot \mathbf{E} = e\tilde{n}/\varepsilon_0$, or $E/L \sim e\tilde{n}/\varepsilon_0$. Consequently the electric field becomes sufficiently strong to control the location of the plasma particles when $\tilde{n}/n \sim (\lambda_d/L)^2$ with $\lambda_d \equiv \varepsilon_0 T/ne^2$ the Debye length and n the number density of electrons. Typically, $\lambda_d/L < 10^{-3}$, so the fractional charge imbalance is minute—the plasma is quasi-neutral. The quasi-neutrality of the plasma means the electric field exerts an almost equal and opposite force on the electrons and the ions and almost no force on the overall plasma. In plasma physics, the electric field is generally calculated using the quasi-neutrality constraint, $\tilde{n}/n \to 0$, and is called the ambipolar or self-consistent electric field. The large scale (greater than λ_d) electric field normally has a magnitude that approximately balances the pressure force of one of the species, the ions or the electrons.

The self-consistent electric field can have the beneficial effect of holding the particle trajectories closer to the constant-pressure surfaces. However, plasmas frequently establish a complicated, asymmetric electric potential even when the magnetic field is symmetric. A small asymmetric potential, $e\tilde{\Phi}/T \sim 0.01$, can significantly alter the particle trajectories and degrade confinement. This degradation in confinement is called anomalous transport. Anomalous ion and electron transport is frequently observed in laboratory plasmas and will be discussed in Section V.

II. BASIC PLASMA PHYSICS

This section covers the principles of basic plasma physics that are required to understand the remainder of the article.

The topics that will be considered are the motion of charged particles in given electric and magnetic fields, the effects of collisions between plasma particles, the plasma kinetic equation, and the two fluid equations.

A. Particle Drift Motion

In many plasmas of interest, the particles move a much greater distance than the size of the plasma between collisions. Plasma confinement in these cases is equivalent to the confinement of particle trajectories in the self-consistent magnetic and electric fields. Unfortunately, particle trajectory calculations are generally quite difficult, even using numerical methods with given fields. Considerable insight and simplification result from the use of the asymptotic guiding center, or drift, equations for the particle motion. A vector form for these equations will be given in this section. The topic will be considered further in Section V after the structure of the magnetic field is discussed.

The equation of motion of a particle in a given magnetic and electric field appears simple

$$m\frac{d\mathbf{V}}{dt} = e\mathbf{V} \times B + e\mathbf{E}, \tag{2}$$

where m is the particle mass and e the charge. Indeed, if $\mathbf{B}$ and $\mathbf{E}$ are constant in space and time the motion is simple. When the magnetic and electric fields are constant, Eq. (2) can be simplified by letting $\mathbf{V} = \mathbf{v} + \mathbf{E} \times \mathbf{B}/B^2$. Equation (2) can then be rewritten as $md\mathbf{v}_\perp/dt = e\mathbf{v}_\perp \times \mathbf{B}$ and $md\mathbf{v}_\parallel/dt = e\mathbf{E}_\parallel$. The perpendicular and parallel signs mean relative to the magnetic field direction. Along the magnetic field, the particle is subject to a simple one-dimensional acceleration. Across the magnetic field, the particle moves in a circle at the cyclotron frequency, or gyrofrequency, $\omega_c = eB/m$ with a radius, the cyclotron radius or gyroradius, $\rho = v_\perp/\omega_c$. The center of the circle, which is called the gyrocenter, moves with a velocity $\mathbf{v}_\parallel$ along the magnetic field and drifts with a velocity $\mathbf{v}_E \equiv \mathbf{E} \times \mathbf{B}/B^2$ across the magnetic field.

The particle trajectories are far more complicated if the magnetic field depends on position $\mathbf{x}$. However, if the gyroradius ρ is small enough compared to the spatial variation of the fields, an approximate constant of the motion exists, the adiabatic invariant or magnetic moment

$$\mu \simeq \frac{mV_\perp^2}{2B}. \tag{3}$$

The existence of this invariant follows from a very general result of Hamiltonian mechanics. Suppose a Hamiltonian system has periodic motion in a canonical coordinate x and its conjugate momentum p. If the Hamiltonian is perturbed so the parameters of the oscillatory motion change only slightly over a period, then the integral over a period $\oint p(dx/dt)\,dt$ is an adiabatic invariant. It is easy to see that the magnetic moment μ is just a constant times the standard Hamiltonian adiabatic invariant. The actual conservation properties of adiabatic invariants is a complex subject in Hamiltonian mechanics. Here it will suffice to say that if the fields are slowly varying analytic functions, then there is an invariant μ, which is accurately approximated by $mV_\perp^2/2B$, that is either conserved or of bounded variation except for exponentially small terms due to so-called Arnold diffusion. By a slowly varying field, we mean the gyroradius is small compared to the spatial scale along the magnetic field and the gyrofrequency is large compared to the rate of the time variation of the magnetic field. A rapid spatial variation perpendicular to the magnetic field is irrelevant to the existence of an adiabatic invariant.

The constancy of the magnetic moment μ simplifies the evaluation of particle trajectories. Formally this occurs through a reduction in the number of degrees of freedom of the Hamiltonian from three, for a general particle trajectory in three dimensions, to two. The canonical variables that are associated with the two degree of freedom Hamiltonian are a function of the three components of position as well as the average of the parallel velocity over a gyro-orbit. More intuitively, the spatial variation of the magnetic field causes a nonclosure of the gyro-orbits (see Fig. 1) and produces a slow drift, $\sim\rho V/L$, across the lines. The lowest order expression for this drift velocity across the field was derived by Alfvén and is

$$\mathbf{v}_d = \frac{\mathbf{B}}{eB^2} \times \left(\mu\nabla B + mv_\parallel^2 \hat{b} \cdot \nabla\hat{b}\right) + \frac{\mathbf{E} \times \mathbf{B}}{B^2}, \tag{4}$$

with $B = |\mathbf{B}|$ and $\hat{b} \equiv \mathbf{B}/|\mathbf{B}|$. Alfvén's expression for the drift velocity can be obtained heuristically. If the force due to the electric field in Eq. (2), $e\mathbf{E}$, is replaced by an arbitrary

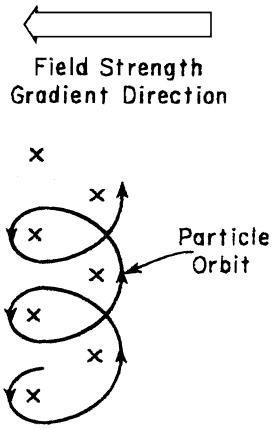


FIGURE 1 Particle orbits. The variation in the particle gyroradius with field strength causes the particle gyrocenter to drift and gives the particle drift orbit.

force **F**, then one finds the gyrocenter drifts across the field lines at a velocity $\mathbf{v}_\perp = \mathbf{B}\times\mathbf{F}/(eB^2)$. The force **F** in a spatially varying field can be approximated as $\mathbf{F} = -\boldsymbol{\nabla}(\mu B) - mv_\parallel^2\,\hat{b}\cdot\nabla\hat{b}$, which yields Eq. (4). The first term in **F** is the force due to the equivalent potential energy of the gyromotion μB. The second term in **F** is the centrifugal force associated with particle moving along curved field lines. The magnitude of the curvature vector, $\hat{b}\cdot\nabla\hat{b}$, is one over the radius of curvature of the magnetic field lines, and the direction of the curvature vector is toward the center of curvature.

The drift Hamiltonian, or particle energy, is to lowest order

$$\mathrm{H} = \frac{1}{2}mv_\parallel^2 + \mu B + e\Phi, \tag{5}$$

as one would expect. However, the canonical coordinates for the drift motion have a nontrivial relation to the ordinary spatial coordinates: a canonical description will not be given until Section V. If the drift velocity is derived from Hamiltonian equations, one would expect the qualitative features of the trajectories to be given correctly, even on a long timescale using the lowest order drifts, due to the so-called KAM theorem of Hamiltonian mechanics.

B. Collisions and the Kinetic Equation

The motion of the ions and the electrons, which form the plasma, is determined not only by large-scale electric and magnetic fields, but also by collisions. Collisions in a plasma have a different character than those in an ordinary gas. In an ordinary gas a collision leads to a sudden, large change in the constants of the motion of a collisionless trajectory. In a plasma, constants of the motion, like the energy and the magnetic moment, are constantly diffusing. The mathematical operator that represents the collisional effects will be discussed in this section as well as the kinetic equation that is used to find the distribution of plasma particles in both position and velocity.

The standard collision operator for plasma problems is the Fokker–Planck operator, which is an integrodifferential operator in the velocity space of the electron and the ion distribution functions. The operator conserves the total energy and the total momentum of the plasma at each spatial point as well as the number of electrons and ions. A simple model operator, the Lorentz operator, which unfortunately does not conserve momentum, helps clarify what is meant by a diffusive collision operator. Let $f(\mathbf{x}, \mathbf{V}, t)$ be the distribution function, which is the number of particles with given position and velocity per unit volume in position **x** and velocity **V** space. Said differently, the total number of particles in a region of position and velocity space is $\int f\, d^3x\, d^3V$. The Lorentz operator scatters the pitch λ of the velocity vector. Let $V_\parallel$ be the component of the velocity along the magnetic field then $\lambda \equiv V_\parallel/V$. The Lorentz operator is

$$C_\mathrm{L}(f) = \frac{\nu_\Omega}{2}\frac{\partial}{\partial\lambda}(1-\lambda^2)\frac{\partial f}{\partial\lambda}, \tag{6}$$

with ν_Ω the pitch angle collision frequency. To understand the effect of this operator, consider its eigenfunctions and eigenvalues, $C_\mathrm{L}(f_l) = -\nu_l f_l$. The eigenfunctions are the Legendre polynomials $f_l \propto P_l(\lambda)$, and the eigenvalues are $\nu_l = l(l+1)\nu_\Omega/2$. A distribution function $f(\lambda)$ can be expanded in the Legendre polynomials—the more localized the region in which f varies, the higher the Legendre polynomials that are required and the more rapid the relaxation of f to a function that is independent of λ. The analogous model collision operator for energy scattering is

$$C_\mathrm{E}(f) = \frac{1}{V^2}\frac{\partial}{\partial V}\left\{V^2\nu_\mathrm{E}\left(Vf + \frac{T}{m}\frac{\partial f}{\partial V}\right)\right\}, \tag{7}$$

with T the plasma temperature and m the mass of the particles being scattered.

The largest collision frequencies in a plasma are the electron pitch angle scattering on both the electrons and the ions and the electron energy scattering on the other electrons. The approximate value for the electron scattering is $\nu_\mathrm{e} \approx (5\times10^{-11}/\mathrm{sec})\, n/T_\mathrm{e}^{3/2}$ with n the number of electrons per cubic meter and T_e the electron temperature measured in electron volts. The next largest collision frequencies are the ion pitch angle scattering and the ion energy scattering on ions, which are smaller than the electron scattering by roughly the square root of the electron to ion mass ratio. That is, $\nu_\mathrm{i} \approx (10^{-12}/\mathrm{sec})\, nZ^3/(A^{1/2}T_\mathrm{i}^{3/2})$ with Z the charge, A the atomic number, and T_i the temperature in electron volts of the ions. Equilibration of the ion and electron temperatures is very slow. This rate is smaller than the electron scattering rate by approximately the electron to ion mass ratio. Another peculiarity of plasma collisions is that if the speed V of the particle being scattered is greater than the thermal speed $V_\mathrm{T} \equiv \sqrt{T/m}$ of the species on which it is scattering, then the collision frequencies are reduced by approximately $(V_\mathrm{T}/V)^3$. The slowness of the electron/ion equilibration implies a plasma can have near-Maxwellian electrons and ions but with two distinct temperatures and two distinct entropies.

The charge density of a plasma species comes not from a smooth function of space but from a collection of point charges. This causes spatial fluctuations in the electric field and plasma collisions. The natural scale of these fluctuations is between the Debye length λ_d, which was discussed in Section I, and the distance b at which the electrostatic energy equals the kinetic energy of the themal motion, $T \approx e^2/\varepsilon_0 b$. To calculate the fluctuating electric field, we

note that the electric field at a location $\mathbf{x}$ due to a charge q at position $\mathbf{x}_1$ is

$$\mathbf{E}(\mathbf{x},\mathbf{x}_1) = \frac{q}{4\pi\varepsilon_0}\frac{\mathbf{x}-\mathbf{x}_1}{|\mathbf{x}-\mathbf{x}_1|^3} = -\frac{q}{\varepsilon_0}\int \frac{i\mathbf{k}}{k^2} e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}_1)}\frac{d^3k}{(2\pi)^3} \tag{8}$$

as long as the distance between $\mathbf{x}$ and $\mathbf{x}_1$ is less than a Debye length, so negligible shielding occurs. [The validity of the Fourier transform can be demonstrated using $\boldsymbol{\nabla}\cdot\mathbf{E} = (e/\varepsilon_0)\,\delta(\mathbf{x}-\mathbf{x}_1)$]. The spatially fluctuating electric field is

$$\tilde{E}^2 \equiv \int n d^3x_1\, \mathbf{E}(\mathbf{x},\mathbf{x}_1)\cdot\mathbf{E}(\mathbf{x},\mathbf{x}_1) = n\left(\frac{q}{\varepsilon_0}\right)^2 \int \frac{1}{k^2}\frac{d^3k}{(2\pi)^3}, \tag{9}$$

with n the number of electrons per cubic meter. The wavenumber k must satisfy $k\lambda_\mathrm{d} \gg 1$ to justify the approximation that the distance between the relevant points is much less than a Debye length. Let $\tilde{\mathcal{E}} \equiv \sqrt{n}(q/\varepsilon_0)/k$ be the electric field fluctuation at wavenumber k. The fluctuation in the velocity of a superthermal thermal particle due to that fluctuation is $\tilde{V} \approx (q/m)\,\tilde{\mathcal{E}}\Delta t$ with $\Delta t = 1/(kV)$ the time it takes the particle to cross the fluctuation. The scattering of the particles is a random walk, so the time, $1/\nu_k$, for fluctuations of wavenumber k to scatter the particle is $1/\nu_k \approx (V/\tilde{V})^2\Delta t$. Adding the scattering from all wavenumbers,

$$\nu \approx \left(\frac{q}{m}\right)^2 \frac{1}{V^3}\int_{1/\lambda_\mathrm{d}}^{1/b} \frac{\tilde{\mathcal{E}}^2}{k}\frac{d^3k}{(2\pi)^3}, \tag{10}$$

which implies $\nu \approx \{\omega_\mathrm{p}/(n\lambda_\mathrm{d}^3)\}\{V_\mathrm{T}/(2\pi V)\}^3\,\ln(\lambda_\mathrm{d}/b)$ for super thermal particles, $V > V_\mathrm{T}$. The quantity $\ln(\lambda_\mathrm{d}/b)$ is called the Coulomb logarithm and is roughly 15 in laboratory plasmas.

The distribution of particles in a plasma as a function of position and velocity is determined both by trajectory and collisional effects. The partial differential equation that describes the evolution of the distribution function is

$$\frac{\partial f}{\partial t} + \mathbf{V}\cdot\boldsymbol{\nabla} f + \frac{e}{m}(\mathbf{E}+\mathbf{V}\times\mathbf{B})\cdot\frac{\partial f}{\partial \mathbf{V}} = C(f), \tag{11}$$

which is called the Fokker–Planck equation or kinetic equation. The right-hand side of this equation is the Fokker–Planck collision operator. The left-hand side is a first-order, hyperbolic differential operator, which has a single characteristic curve. This characteristic is the trajectory of a particle in the electric and magnetic fields. If the drift velocity $\mathbf{v}_\mathrm{d}$, Eq. (4), is used to obtain the particle trajectories instead of exact velocity $\mathbf{V}$, the kinetic equation is called the drift kinetic equation.

C. Two Fluid Equations

Although the kinetic equations for the ions and the electrons provide an essentially complete description of plasma confinement, the complexity of this description obscures the basic properties of plasmas in which the gyroradii are small compared to the size of the plasma. These properties can be studied by considering the first velocity moment of each kinetic equation with the additional assumption that the distribution functions $f(\mathbf{x},\mathbf{V},t)$ are independent of the phase of the particles in their nearly circular gyro-orbits. The resulting pair of equations, one for the electrons and one for the ions, is called the two fluid equations.

The first moment of the kinetic equation is determined by multiplying Eq. (11) by $m\mathbf{V}$ and integrating over velocity space. Denoting the various species by a subscript α,

$$\frac{\partial n_\alpha m_\alpha \mathbf{u}_\alpha}{\partial t} + \boldsymbol{\nabla}\cdot(n_\alpha m_\alpha \mathbf{u}_\alpha\mathbf{u}_\alpha) + \boldsymbol{\nabla}\cdot\overleftrightarrow{\mathbf{p}}_\alpha - e_\alpha n_\alpha(\mathbf{E}+\mathbf{u}_\alpha\times\mathbf{B}) = -e_\alpha n_\alpha \mathbf{R}_\alpha, \tag{12}$$

with

$$n_\alpha \equiv \int f_\alpha\, d^3V, \qquad \mathbf{u}_\alpha \equiv \frac{1}{n_\alpha}\int \mathbf{V} f_\alpha\, d^3V, \tag{13}$$

$$\overleftrightarrow{\mathbf{p}}_\alpha \equiv \int m_\alpha(\mathbf{V}-\mathbf{u}_\alpha)(\mathbf{V}-\mathbf{u}_\alpha) f_\alpha\, d^3V, \tag{14}$$

$$\overleftrightarrow{\mathbf{R}}_\alpha \equiv -\frac{1}{e_\alpha n_\alpha}\int m_\alpha V C_\alpha(f_\alpha)\, d^3V. \tag{15}$$

The physical interpretation of these definitions is n_α is the number of particles of species α per cubic meter, $\mathbf{u}_\alpha$ is the mean velocity, $\overleftrightarrow{\mathbf{p}}_\alpha$ is the pressure (or with a change of sign, the stress tensor), and $e_\alpha n_\alpha \overleftrightarrow{\mathbf{R}}_\alpha$ is the force between species due to collisions.

The assumption that the distribution function is independent of gyrophase, which follows from the smallness of the gyroradius compared to the system size, restricts the form of the pressure tensor. Let

$$p_\parallel^{(\alpha)} \equiv \int m_\alpha \hat{b}\cdot(\mathbf{V}-\mathbf{u}_\alpha)\hat{b}\cdot(\mathbf{V}-\mathbf{u}_\alpha) f_\alpha\, d^3V, \tag{16}$$

$$p_\perp^{(\alpha)} \equiv \frac{1}{2}\int m_\alpha\{\hat{b}\times(\mathbf{V}-\mathbf{u}_\alpha)\cdot\hat{b}\times(\mathbf{V}-\mathbf{u}_\alpha)\} f_\alpha\, d^3V, \tag{17}$$

with the parallel and perpendicular signs meaning parallel or perpendicular to the magnetic field and $\hat{b} \equiv \mathbf{B}/|\mathbf{B}|$ the unit vector along the magnetic field. In the limit of a small gyroradius to system size the pressure tensor has the form

$$\overleftrightarrow{\mathbf{p}}_\alpha = p_\parallel^{(\alpha)}\hat{b}\hat{b} + p_\perp^{(\alpha)}(\overleftrightarrow{1} - \hat{b}\hat{b}) \tag{18}$$

with $\overleftrightarrow{1}$ the unit tensor. Note that the validity of Eq. (18) does not depend on the distribution function being close to Maxwellian. The only requirement is that the distribution function be independent of gyrophase.

In many plasma problems of interest, the flow velocities are much smaller than the thermal velocity, $m_\alpha u_\alpha^2 \ll p_\parallel^{(\alpha)}$. In this case, the inertial terms can be ignored, so Eq. (12) implies

$$\nabla \cdot \overleftrightarrow{\mathbf{p}}_\alpha - e_\alpha n_\alpha (\mathbf{E} + \mathbf{u}_\alpha \times \mathbf{B}) = -e_\alpha n_\alpha \mathbf{R}_\alpha \tag{19}$$

for each species in the plasma. One equation of the form of Eq. (19) for the ions and another for the electrons are the two fluid equations of an equilibrium plasma.

The equilibrium equation for the plasma is obtained by summing over all the plasma species:

$$\nabla \cdot \overleftrightarrow{\mathbf{p}} = \mathbf{j} \times \mathbf{B}, \tag{20}$$

with the current density given by

$$\mathbf{j} = \sum_\alpha e_\alpha n_\alpha \mathbf{u}_\alpha. \tag{21}$$

In deriving the equilibrium equation, the quasi-neutrality condition $\sum e_\alpha n_\alpha = 0$ was used as well as the condition that interparticle collisions exert no net force on the plasma, $\sum e_\alpha n_\alpha \mathbf{R}_\alpha = 0$.

The generalized Ohm's law, which is the electron equilibrium equation, frequently arises in discussions of plasma confinement. Electron collisions are sufficiently frequent that the electron fluid is often isotropic, $p_\parallel^{(e)} = p_\perp^{(e)} = p_e$, so

$$\mathbf{E} + \mathbf{u}_e \times \mathbf{B} + \frac{\nabla p_e}{en} = \mathbf{R}_e, \tag{22}$$

with $n \equiv n_e$.

The generalized Ohm's law that is commonly used in plasma studies is a simplified form of Eq. (22). The term $\nabla p/en$ is often ignored because in toroidal plasmas it usually has no curl, which means this term in Ohm's law does not affect the evolution of the magnetic field. (However, a strong magnetic field can be generated if $\nabla \times (\nabla p/en)$ is nonzero as is often the case in laser produced plasmas.) The electron and the plasma flow velocities $\mathbf{u}$ are identified and the dissipative force is written as $\mathbf{R}_e = \eta \mathbf{j}$ with η the plasma resistivity. The standard form for the generalized Ohm's law is

$$\mathbf{E} + \mathbf{u} \times \mathbf{B} = \eta \mathbf{j}. \tag{23}$$

The difference between the electron and the ion flow velocities goes to zero as the gyroradius to system size, so the identification of the plasma velocity $\mathbf{u}$ with the electron flow velocity is accurate for flows rapid compared to the diamagnetic speed $|\nabla p/(enB)|$. The equation $\mathbf{R}_e = \eta \mathbf{j}$ is only approximate for two reasons: First, the resistivity is a diagonal tensor with distinct components parallel and perpendicular to the magnetic field, which differ by approximately a factor of two. Second, the heat flux contributes to the collisional force through thermodynamic cross terms.

III. TOROIDAL PLASMAS

A noted in Section I, the torus is the only plasma shape that can confine an isotropic plasma, $p_\parallel = p_\perp$. To simplify the discussion of toroidal plasmas, the pressure will be assumed isotropic. The effects of pressure anisotropy are discussed in Section IV.

Toroidal devices have always had a central position in the magnetic fusion program. The simplest toroidal device would have only a symmetric toroidal field, but as we will find, this is not consistent with a plasma equilibrium. A symmetric toroidal equilibrium, called a *tokamak* (Fig. 2), can be produced by inducing a plasma current along the externally produced toroidal magnetic field. The toroidal current produces an outward force that must be balanced [Eq. (1)] by an externally produced vertical magnetic field, which means parallel to the axis of symmetry of the torus. There are probably more tokamaks worldwide than all other magnetic fusion devices combined. The largest tokamak is the JET (Joint European Torus) at Culham, England. The major and minor radii are 3.1 and 1.25 m, respectively. The magnetic field is 3.6 T and a typical current is 4.2 MA. This device has produced 14 MW of DT fusion power with a central ion temperature of 28 keV, an electron temperature of 14 keV, an electron number density of $4.1 \times 10^{19}/\mathrm{m}^3$, an energy confinement time of 0.9 sec, and a power input of 25 MW.

By adding a helical twist to the coils that produce the toroidal magnetic field, the magnetic field lines can be given a twist, similar to that of a tokamak, but without the

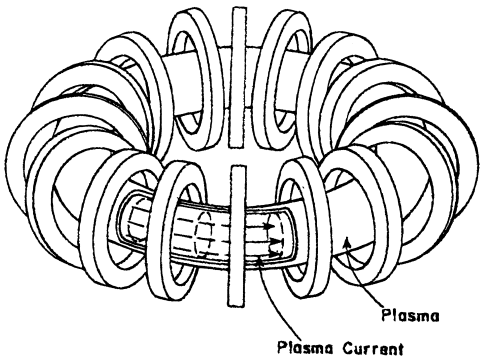


FIGURE 2 Tokamak. The toroidal field coils and the plasma current, which produces the poloidal field, are illustrated. In addition, a vertical magnetic field is required to achieve an equilibrium.

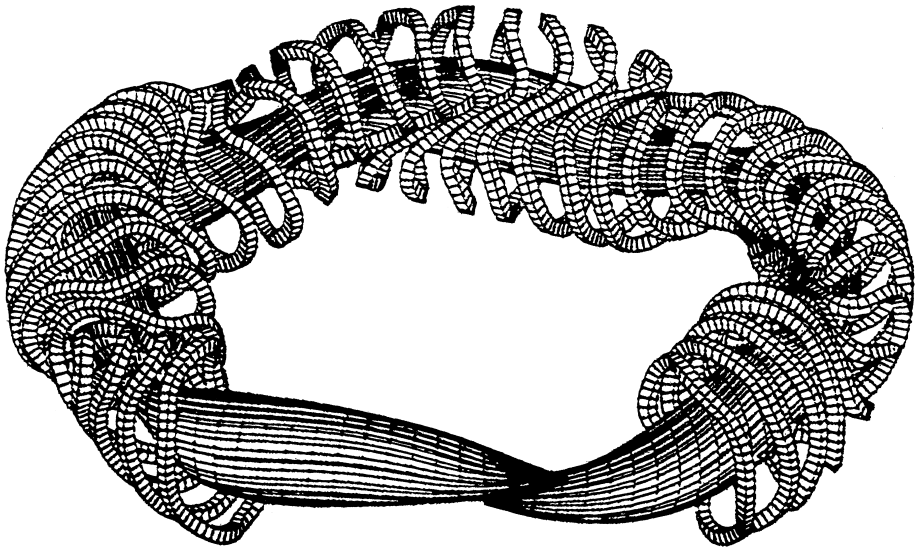


FIGURE 3 Stellarator. Field coils with a helical twist can produce both a toroidal and a poloidal magnetic field. The W7-X stellarator, which is being built in Germany, is illustrated.

need for a net toroidal plasma current. Such configurations are called *stellarators* (Fig. 3). Existing stellarators are second only to tokamaks in their confinement properties. The largest stellarator is the LHD (Large Helical Device) located at Toki, Japan. The LHD can operate at a field of 3 T and has a major radius of 3.9 m and an average minor radius of 0.6 m. In early operations ion and electron temperatures of a few kilovolts were achieved, and confinement times were several hundred milliseconds.

Three topics will be discussed in this section: the properties of magnetic field lines, the constraints of plasma equilibria, and the stability of equilibria. The properties of magnetic field lines underlie the entire theory of toroidal plasmas because equilibrium, $\nabla p = \mathbf{j} \times \mathbf{B}$, implies that pressure is constant along the magnetic field lines, $\mathbf{B} \cdot \nabla p = 0$. The magnetic field lines will be found to have identical properties to those of particle trajectories in one spatial dimension with time-dependent forces. Such systems are called one-and-a-half degree of freedom Hamiltonian systems. The action-angle coordinates of Hamiltonian mechanics, which are called magnetic coordinates in toroidal plasma physics, are the basis for many calculations. The Hamiltonian picture of the magnetic field separates the properties of the field that can change arbitrarily rapidly from those that are conserved except for slow plasma dissipation.

The constraints of equilibrium are largely determined by force balance, $\nabla p = \mathbf{j} \times \mathbf{B}$, and Ampere's law, $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$. Remarkably, the current density, $\mathbf{j}$, given by these equations has a singular mathematical form. The singularities are demonstrated explicitly by the use of magnetic coordinates. The physical resolution of these singularities determines much of toroidal equilibrium, stability, and transport theory. For example, toroidal equilibria are frequently not uniquely defined by the boundary conditions. Nonuniqueness can cause the tokamak to spontaneously break the toroidal symmetry. This breaking of the symmetry provides critical limitations, called stability limits, on the obtainable plasma parameters.

Plasma equilibria and their stability can be efficiently calculated using the energy principle. The energy associated with the plasma is the sum of the plasma thermal energy and the magnetic energy. Plasma equilibria are extrema of the energy and stable equilibria are minima of the energy.

A. Properties of Magnetic Field Lines

The field lines of divergence-free fields, such as the magnetic field, have special mathematical properties which are those of particle trajectories in a one-and-a-half degree of freedom Hamiltonian mechanics. The fundamental properties of the magnetic field that can be derived from the divergence-free condition will be given in this section.

Three coordinates are required to describe a magnetic field. They are conventionally the three Cartesian coordinates of the position vector $\mathbf{x}$. However, many properties of the magnetic field become apparent only when a more subtle canonical coordinate system $(\psi_t, \theta, \varphi)$ is employed. The coordinates θ and φ can be any poloidal and toroidal angles (see Fig. 4), and $\psi_t = \int \mathbf{B} \cdot d\mathbf{a}_\varphi$ is the toroidal magnetic flux enclosed by a constant ψ_t surface. (To simplify formulas, the angles θ and φ are chosen to have a period of unity rather than the conventional

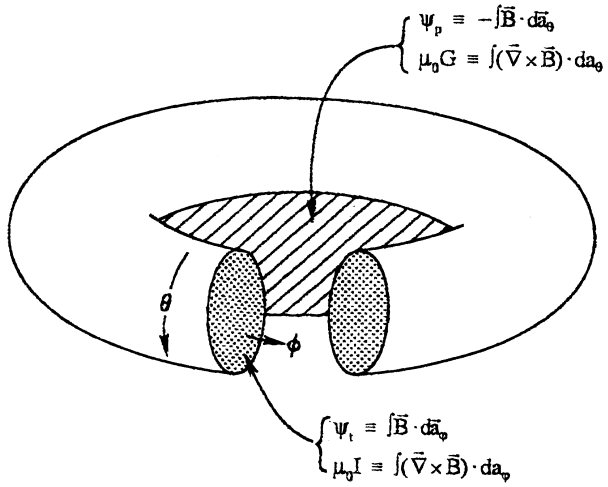


FIGURE 4 Magnetic coordinate systems. The poloidal angle is θ and the toroidal angle is φ. The poloidal magnetic flux is $-\psi_p$ and the toroidal flux is ψ_t. The poloidal and toroidal currents are G and I.

2π. A conventional toroidal angle would be $\varphi_c = 2\pi\varphi$.) Any divergence-free vector can be written in the so-called canonical form,

$$\mathbf{B} = \nabla\psi_t \times \nabla\theta + \nabla\varphi \times \nabla\psi_p, \tag{24}$$

with $\psi_p = -\int \mathbf{B}\cdot d\mathbf{a}_\theta$ the poloidal magnetic flux enclosed in the hole of the torus (Fig. 4). The canonical form can be derived by noting that an arbitrary vector can be written as $\mathbf{A} = a_r\nabla r + a_\theta\nabla\theta + a_\varphi\nabla\varphi$ using any well-behaved coordinates (r, θ, φ). Let $\partial G/\partial r \equiv a_r$, $\psi_t \equiv a_\theta - \partial G/\partial\theta$, and $\psi_p \equiv -a_\varphi + \partial G/\partial\varphi$. Then, $\mathbf{A} = \psi_t\nabla\theta - \psi_p\nabla\varphi + \nabla G$, and Eq. (24) is obtained from $\mathbf{B} = \nabla\times\mathbf{A}$ since any globally divergence-free field is the curl of a vector potential.

The importance of the canonical form for the magnetic field [Eq. (24)] is that it leads immediately to Hamilton's equations of motion for the field lines. Along a field line, $d\mathbf{x}/d\tau = \mathbf{B}(\mathbf{x})$, any function $f\mathbf{x}$ varies as $df/d\tau = (\partial f/\partial\mathbf{x})\cdot d\mathbf{x}/d\tau = \mathbf{B}\cdot\nabla f$. Writing the poloidal flux ψ_p as a function of the toroidal flux ψ_t and the two angles, the field line equations $d\psi_t/d\varphi = \mathbf{B}\cdot\nabla\psi_t/\mathbf{B}\cdot\nabla\varphi$ and $d\theta/d\varphi = \mathbf{B}\cdot\nabla\theta/\mathbf{B}\cdot\nabla\varphi$ imply

$$\frac{d\psi_t}{d\varphi} = -\frac{\partial\psi_p}{d\theta} \quad \text{and} \quad \frac{d\theta}{d\varphi} = \frac{\partial\psi_p}{\partial\psi_t}. \tag{25}$$

The poloidal flux $\psi_p(\psi_t, \theta, \varphi)$ is the Hamiltonian for the magnetic field lines, the toroidal flux ψ_t is the canonical momentum, the poloidal angle is the canonical position, and the toroidal angle is the canonical time. The canonical coordinates $(\psi_t, \theta, \varphi)$ are well behaved in any region of space in which the toroidal magnetic field, $\mathbf{B}\cdot\nabla\varphi = (\nabla\psi_t\times\nabla\theta)\cdot\nabla\varphi$, is nonzero, for the Jacobian of $(\psi_t, \theta, \varphi)$ coordinates is $\mathcal{J} = 1/\mathbf{B}\cdot\nabla\varphi$.

Topological properties are independent of the coordinate system; the topological properties are the same in canonical $(\psi_t, \theta, \varphi)$ coordinates as in Cartesian coordinates. Therefore, full topological information about the magnetic field lines is given by the field line Hamiltonian, $\psi_p(\psi_t, \theta, \varphi)$. Examples of topological properties are whether the magnetic field lines form nested toroidal surfaces and the average number of poloidal circuits a field line makes while making one toroidal circuit, which is called the rotational transform and denoted by the Greek letter iota, ι.

Concepts from Hamiltonian mechanics play a major role in the theory of toroidal plasmas—though often with a different name than in mechanics texts. These concepts include Poincaré plot, integrability, rotational transform, magnetic coordinates, magnetic islands, and stochastic regions. A Poincaré plot is illustrated in Fig. 5. Each time a field line goes the long (toroidal) way around the torus, a point is marked on a plane that is transverse to the torus [Fig. 5(a)]. The Poincaré plot is, ideally, the plot of the points from an infinite number of traversals. If the points that form the Poincaré plot of each field line in a region of space lie on a smooth curve, then the field is said to be integrable in that region of space. If the magnetic field were associated with a plasma equilibrium, each smooth curve would be the intersection of a constant-pressure surface with the Poincaré plane. As the Poincaré plot of an integrable field is constructed, there is an average rate at which the poloidal angle advances for each toroidal transversal. The average advance is the rotational transform, which is denoted by ι.

A magnetic field is integrable if and only if canonical coordinates $(\psi_t, \theta, \varphi)$ exist such that the poloidal flux ψ_p is a function of the toroidal flux ψ_t alone. In the Hamiltonian mechanics literature such canonical coordinates are called action-angle coordinates. However, in the plasma physics literature they are called magnetic coordinates. In magnetic coordinates, the equation of a field line is ψ_t equal to a constant and $\theta - \iota\varphi$ equal to a constant. The rotational transform is $\iota \equiv d\psi_p/d\psi_t$.

A magnetic field $\mathbf{B}$ $(\mathbf{x})$ that has no continuous symmetries is rarely integrable just as few Hamiltonians of the form $H(p, x, t)$ are. The Poicaré plot may show magnetic islands [Fig. 5(b)] or stochastic regions in which a single field line covers (comes arbitrarily close to every point in) a finite volume. Clearly, regions occupied by stochastic field lines must be limited in extent to confine a plasma with nonzero pressure.

When a magnetic field is time dependent, $\mathbf{B}(\mathbf{x}, t)$, the Hamiltonian $\psi_p(\psi_t, \theta, \varphi, t)$ evolves slowly, on a timescale set by the resistivity of the plasma, while the transformation equations, which give the ordinary spatial coordinates

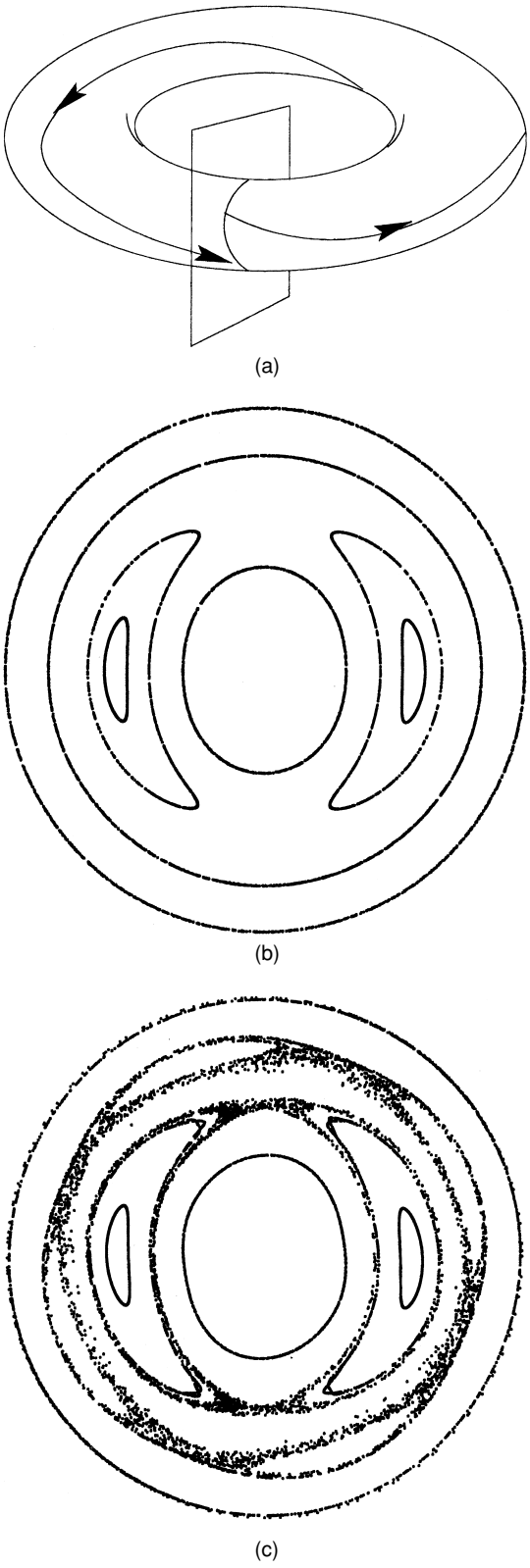


FIGURE 5 Poincaré plot (a) Beginning of the construction of a Poincaré plot. (b) Plot of a field that has a magnetic island. (c) A field with stochasticity.

in terms of the canonical coordinates $\mathbf{x}(\psi_t, \theta, \varphi, t)$, evolve as rapidly as required to maintain force balance. The two timescales may differ by more that six orders of magnitude in laboratory experiments.

The time development of a magnetic field is most easily examined using the vector potential $\mathbf{A}$, which is defined by $\mathbf{B} = \nabla \times \mathbf{A}$. As noted earlier, the canonical form for the magnetic field, Eq. (24), is obtained from the vector potential $\mathbf{A} = \psi_t \nabla\theta - \psi_p \nabla\varphi + \nabla G$ with G an arbitrary gauge function. Coordinate transformation theory can be used to prove the identity

$$\left(\frac{\partial \mathbf{A}(\mathbf{x}, t)}{\partial t}\right)_{\mathbf{x}} = -\left(\frac{\partial \psi_p}{\partial t}\right)_c \nabla\varphi + \mathbf{v}_c \times \mathbf{B} + \nabla s, \quad (26)$$

with the subscript c implying that the canonical coordinates $(\psi_t, \theta, \varphi)$ are held constant and with $\mathbf{v}_c \equiv (\partial \mathbf{x}/\partial t)_c$ the velocity of the canonical coordinates. The function $s(\psi_t, \theta, \varphi, t)$, or $s(\mathbf{x}, t)$, can be considered arbitrary due to the arbitrariness of the gauge function G. The arbitrary function s is generating function for infinitesimal canonical transformations of Hamiltonian mechanics.

By far, the most important component of the equation for $\partial \mathbf{A}/\partial t$ is the component parallel to the magnetic field,

$$\mathbf{B} \cdot \left(\frac{\partial \mathbf{A}}{\partial t}\right)_{\mathbf{x}} = -\left(\frac{\partial \psi_p}{\partial t}\right)_c \mathbf{B} \cdot \nabla\varphi + \mathbf{B} \cdot \nabla s, \quad (27)$$

If a function $s(\mathbf{x}, t)$ exists such that $\mathbf{B} \cdot \nabla s = \mathbf{B} \cdot (\partial \mathbf{A}/\partial t)_{\mathbf{x}}$, then a Hamiltonian exists that is independent of time. If a Hamiltonian exists that is independent of time, then the topology of the magnetic field lines cannot change.

Faraday's law implies the electric field has the form $\mathbf{E} = -\partial \mathbf{A}/\partial t - \nabla\Phi$ with Φ the electric potential. The generalized Ohm's law, $\mathbf{E} + \mathbf{u} \times \mathbf{B} = \mathbf{R}_e$ with $\mathbf{u}$ the plasma velocity, relates the electric field to the dissipation $\mathbf{R}_e = \eta \mathbf{j}$. Letting $\Phi_c \equiv \Phi + s$, which is the electric potential transformed into canonical coordinate space, one finds

$$\mathbf{R}_e = \left(\frac{\partial \psi_p}{\partial t}\right)_c \nabla\varphi + (\mathbf{u} - \mathbf{v}_c) \times \mathbf{B} + \nabla\Phi_c. \quad (28)$$

This equation demonstrates the separation of timescales between the Hamiltonian ψ_p and the velocity of the canonical coordinates $\mathbf{v}_c$. The component of Eq. (28) parallel to the magnetic field determines the evolution of ψ_p. The perpendicular components determine the evolution of $\mathbf{u} - \mathbf{v}_c$ the difference between the plasma velocity and the velocity of the canonical coordinates. The plasma velocity, and hence the velocity of the canonical coordinates, can only be found by bringing in additional physics, namely, force balance.

In the case of an integrable field, the time evolution of the Hamiltonian, or poloidal flux $\psi_p(\psi_t, t)$, is given by

$$\frac{\partial \psi_p(\psi_t, t)}{\partial t} = V(\psi_t, t), \quad (29)$$

with the loop voltage defined by

$$V(\psi_t, t) \equiv \frac{\partial}{\partial \psi_t} \int \mathbf{E} \cdot \mathbf{B}\, d^3x. \tag{30}$$

This relation follows from Eq. (27) and $\mathbf{E} = -\partial \mathbf{A}/\partial t - \nabla \Phi$ using the fact that $d^3x = d\psi_t\, d\theta\, d\varphi / \mathbf{B} \cdot \nabla \varphi$. In a tokamak a loop voltage is provided through the transformer action of changing the poloidal flux in the central hole of the torus. If the loop voltage is held constant for a sufficiently long time, the magnetic field in the plasma can become time independent with $\psi_p(\psi_t, t) = \Psi_p(\psi_t) + Vt$. Equation (29) implies the loop voltage is constant everywhere in the plasma when the magnetic field in the plasma is independent of time. In other words, if the timescale for evolution is short compared to the resistive timescale, the magnetic field line Hamiltonian ψ_p is conserved, but if the timescale is long, the loop voltage V becomes constant across the plasma with the value of V set by the change in the magnetic field outside of the plasma.

The existence of a Hamiltonian structure for the magnetic field comes from the field being divergence free. Other divergence-free fields, such as the vorticity field $\boldsymbol{\omega} = \nabla \times \mathbf{u}$ of fluid mechanics, have a similar structure. For the vorticity field there is an analogue of Eq. (26), but this is not true for all divergence-free fields.

B. Toroidal Equilibrium

To specify a plasma equilibrium, two functions of the toroidal flux must be given as well as the shape of the plasma surface. Many choices of the two functions are possible. One choice is the plasma pressure $p(\psi_t)$ and the rotational transform $\iota(\psi_t)$. Stable plasma equilibria are minima of the sum of the plasma and the magnetic energy within the region bounded by the plasma surface. As demonstrated by Eq. (1), an external magnetic field is required to support a plasma equilibrium. In the case of the axisymmetric tokamak, this field strengthens the poloidal component of the magnetic field (the component that encircles the torus the short way around) on the outboard side of the torus and weakens it on the inboard side. The stronger poloidal field on the outboard side implies the distance between poloidal flux contours is smaller there. Since the pressure is a function of the poloidal flux alone, $p(\psi_p)$, the center of the plasma is shifted toward the outboard side of the torus. This outward shift limits the plasma beta ($\beta \equiv 2\mu_0 p/B^2$) to approximately $\beta < \varepsilon \iota^2$ with $\varepsilon \equiv a/R$ the ratio of the minor, a, to the major, R, radius of the torus.

Some of the basic properties of equilibria can be derived using the property that an equilibrium is an extremum of the energy

$$W = \int_{\text{plasma}} \left(\frac{p}{\gamma - 1} + \frac{B^2}{2\mu_0} \right) d^3x, \tag{31}$$

with γ the adiabatic index, which is 5/3 for an ideal gas. While extremizing the energy, the plasma boundary and the entropy per particle, $s = \ln(p/n^\gamma)^{1/\gamma - 1}$ should be held fixed, and the magnetic field should be evolved using the ideal Ohm's law, $\mathbf{E} + \mathbf{u} \times \mathbf{B} = 0$. Particle conservation, $\partial n/\partial t + \nabla \cdot (n\mathbf{u}) = 0$ and fixed entropy $\partial s/\partial t + \mathbf{u} \cdot \nabla s = 0$ imply the pressure evolution is $\partial p/\partial t + \mathbf{u} \cdot \nabla p + \gamma p \nabla \cdot \mathbf{u} = 0$. The ideal Ohm's law implies $\partial \mathbf{B}/\partial t = \nabla \times (\mathbf{u} \times \mathbf{B})$. Because neither the plasma flow nor the magnetic field can cross the plasma boundary, one finds

$$\frac{dW}{dt} = \int_{\text{plasma}} \mathbf{u} \cdot (\nabla p - \mathbf{j} \times \mathbf{B})\, d^3x, \tag{32}$$

which satisfies the equilibrium equation, $\nabla p = \mathbf{j} \times \mathbf{B}$, at extrema. Since the equilibrium equation, $\nabla p = \mathbf{j} \times \mathbf{B}$, is independent of the adiabatic index γ, one can choose the index to have the nonphysical value of zero while calculating equilibria. If this is done, the pressure remains a fixed function of the toroidal flux throughout the relaxation, $p(\psi_t)$. The ideal Ohm's law conserves the poloidal flux as a function of the toroidal flux and, therefore, the rotational transform $\iota(\psi_t) \equiv d\psi_p/d\psi_t$. An equilibrium solver based on the relaxation of the energy uses the functions pressure $p(\psi_t)$ and the rotational transform $\iota(\psi_t)$ to specify the equilibrium.

As noted in the introduction an equilibrium plasma must be supported in part by an external magnetic field. Equation (1) can be used to estimate this field for a large aspect ratio, axisymmetric torus. If the external magnetic field is almost constant across the plasma, the right-hand side of Eq. (1) is $-2\mathbf{m} \times \mathbf{B}_{ex}$ with $\mathbf{m} \equiv \frac{1}{2} \int (\mathbf{x} \times \mathbf{j}_{pl})\, d^3x$, the dipole moment of the plasma. If the aspect ratio R/a is large with (R, φ, Z) cylindrical coordinates and a the minor radius of the plasma, $\mathbf{m} \approx \hat{z} I \pi R^2 + \hat{\varphi} \int \{B_\varphi^2 - (B_\varphi^{(ex)})^2\}\, d^3x/(4B_\varphi^{(ex)})$. Let $B_p \equiv \mu_0 I/(2\pi a)$, then

$$\frac{\mathbf{B}_{ex} \cdot \hat{z}}{B_p} = -\frac{a}{4R} \frac{\int \left(3p + \frac{B_{pl}^2}{2\mu_0} + \frac{B_\varphi^2 - (B_\varphi^{(ex)})^2}{2\mu_0} \right) d^3x}{\frac{B_p^2}{2\mu_0} \int d^3x}, \tag{33}$$

with $\int d^3x = (\pi a^2)(2\pi R)$ the volume of the plasma. If the externally produced vertical field, $\mathbf{B}_{ex} \cdot \hat{z}$, becomes larger in magnitude than B_p, the poloidal magnetic field has a point of zero magnitude on the inboard side of the plasma. This point is an X point on a separatrix between field lines that encircle the plasma and field lines that intercept the chamber in which the plasma lies. The formation of a

separatrix sets the maximum pressure limit for which one can have an equilibrium plasma.

A magnetic field that is consistent with a plasma equilibrium, $\nabla p = \mathbf{j} \times \mathbf{B}$, has important properties in addition to those of a general magnetic field. These properties lead to formulas for the plasma current $\mathbf{j}$, which have a singular mathematical from. Much of toroidal equilibrium, stability, and transport theory is associated with these singularities.

In any region of an equilibrium plasma in which the pressure gradient is nonzero, the equation $\mathbf{B} \cdot \nabla p = 0$ implies the magnetic field is locally integrable and can be written in magnetic coordinate form,

$$\mathbf{B} = \nabla\psi_t \times \nabla\theta + \nabla\varphi \times \nabla\psi_p(\psi_t). \tag{34}$$

The equation $\mathbf{j} \cdot \nabla p = 0$, which also follows from plasma equilibrium, implies the current obeys an equation related to that of Eq. (34). Using Ampere's law, $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$, this implies the magnetic field has a second representation,

$$\mathbf{B} = \mu_0 G(\psi_t)\nabla\varphi + \mu_0 I(\psi_t)\nabla\theta + \beta_*(\psi_t, \theta, \varphi)\nabla\psi_t. \tag{35}$$

In the language of the theory of general coordinates, Eq. (34) is the contravariant representation of the magnetic field and Eq. (35) is the covariant representation. The simultaneous existence of simple contravariant and covariant representations of the magnetic field permits a great simplification in the theory of equilibrium plasmas. The quantities G and I in Eq. (35) have a simple physical interpretation: $G(\psi_t)$ is the poloidal current outside a constant ψ_t surface, and $I(\psi_t)$ is the toroidal current enclosed by a constant ψ_t surface (see Fig. 4). The quantity β_* is closely related to the Pfirsch–Schlüter current, which is defined below. Given a $\mathbf{B}(\mathbf{x})$ that is consistent with equilibrium, codes exist to find the transformation, $\mathbf{x}(\psi_t, \theta, \varphi)$ and $\psi_p(\psi_t)$, to magnetic coordinates. The Jacobian of $(\psi_t, \theta, \varphi)$ coordinates, $\mathcal{J}$, is defined so $d^3x = \mathcal{J}\, d\psi_t\, d\theta\, d\varphi$, so $1/\mathcal{J} = (\nabla\psi_t \times \nabla\theta) \cdot \nabla\varphi$. By dotting together the two forms for the magnetic field, Eqs. (34) and (35), one finds $\mathcal{J} = \mu_0(G + \iota I)/B^2$.

The equations obeyed by an equilibrium current, $\nabla p = \mathbf{j} \times \mathbf{B}$ and $\nabla \cdot \mathbf{j} = 0$, are equivalent to $\mathbf{j}_\perp = (\mathbf{B} \times \nabla p)/B^2$ and

$$\mathbf{B} \cdot \nabla\left(\frac{j_\parallel}{B}\right) = -\nabla \cdot \mathbf{j}_\perp, \tag{36}$$

with the current written as $\mathbf{j} = (j_\parallel/B)\mathbf{B} + \mathbf{j}_\perp$. The most interesting component of the current, the parallel current $j_\parallel$, is the sum of two terms, $j_\parallel = j_{ps} + j_n$. The first term, which is known as the Pfirsch–Schlüter current j_{ps}, is the special solution of the inhomogeneous differential equation for $j_\parallel$ [Eq. (36)]. The second term, which is known as the net current j_n, is the solution of the associated homogeneous equation to Eq. (36). The net current is made unique by letting $k(\psi_t) \equiv \mu_0 j_n/B$ and $(G + \iota I)k(\psi_t) = d(\int \mathbf{j} \cdot \mathbf{B}\, d^3x)/d\psi_t$. With this choice the simple Ohm's law $\mathbf{E} \cdot \mathbf{B} = \eta \mathbf{j} \cdot \mathbf{B}$ implies the loop voltage, and Eq. (30) is given by $V = \eta(G + \iota I)k$.

Using the contravariant Eq. (34) and covariant Eq. (35) representations, it is easily shown that $\mathbf{j}_\perp$, which is called the diamagnetic current, is given by

$$\mathbf{j}_\perp = \frac{G(\psi_t)\nabla\varphi \times \nabla\psi_t - I(\psi_t)\nabla\psi_t \times \nabla\theta}{B^2(\psi_t, \theta, \varphi)} \frac{dp(\psi_t)}{d\psi_t}. \tag{37}$$

The Pfirsch–Schlüter current, the part of the parallel current that varies on a constant-pressure surface, can be given explicitly if $1/B^2$ is written as a Fourier series,

$$\frac{1}{B^2(\psi_t, \theta, \varphi)} = \frac{1}{B_0^2(\psi_t)}\left\{1 - 2{\sum_{m,n}}' \varepsilon_{mn}\, e^{2\pi i(n\varphi - m\theta)}\right\}, \tag{38}$$

with the prime on the sum implying the $(m = 0,\ n = 0)$ term is omitted. The contravariant representation of the magnetic field [Eq. (34)] trivializes the inversion of the differential operator $\mathbf{B} \cdot \nabla(j_\parallel/B)$. If $j_\parallel/B$ is written as a Fourier series, then $\mathbf{B} \cdot \nabla$ multiplies each term in that series by $2\pi(n - \iota m)$ with $\iota \equiv d\psi_p/d\psi_t$ the rotational transform. By this technique, the Pfirsch–Schlüter current, the special solution to Eq. (36), is shown to be

$$\frac{j_{ps}}{B} = -\frac{1}{\pi B_0^2}\frac{dp}{d\psi_t}{\sum_{m,n}}' \frac{mG + nI}{n - \iota m}\varepsilon_{mn}\, e^{2\pi i(n\varphi - m\theta)}. \tag{39}$$

This equation has apparent singularities on every ψ_t surface on which the rotational transform is rational, $\iota = N/M$ the ratio of two integers. To avoid these singularities, either the pressure gradient must vanish, $dp/d\psi_t = 0$, or the resonant ε_{mn} values must vanish. On a rational surface, which means ι rational, each magnetic field lines closes on itself and, therefore, does not cover the surface. One can show that the requirement that the resonant ε_{mn} values vanish on a rational surface with a pressure gradient is equivalent to $\oint dl/B$ being the same for every field line; dl is he differential distance along a line. The constancy of $\oint dl/B$ on a constant-pressure surface made up of closed magnetic field lines, which is called the Hamada condition, is not fulfilled by a symmetric toroidal magnetic field. Toroidally symmetric equilibria require a nonzero rotational transform and a net toroidal current $I(\psi_t)$.

Equations (38) and (39) for the various parts of the plasma current can be used to obtain expressions for variation of the poloidal, G, and toroidal, I, current:

$$\frac{dI}{d\psi_t} + \frac{\mu_0}{B_0^2}\frac{dp}{d\psi_t} I = \mu_0 k(\psi_t), \tag{40}$$

$$\frac{dG}{d\psi_t} + \iota\frac{dI}{d\psi_t} + \frac{\mu_0(G+\iota I)}{B_0^2}\frac{dp}{d\psi_t} = 0, \tag{41}$$

$$\beta_* = \frac{G+\iota I}{\pi B_0^2}\frac{dp}{d\psi_t}\sum_{m,n}{}' \frac{i\varepsilon_{mn}}{n-\iota m} e^{2\pi i(n\varphi - m\theta)}. \tag{42}$$

The first of these three equations, Eq. (40), relates the toroidal current to the net parallel current. The maintenance of toroidal current requires either a loop voltage or terms in the parallel Ohm's law, $\mathbf{E}\cdot\mathbf{B}$, in addition to $\eta\mathbf{j}\cdot\mathbf{B}$. Equation (41) is called the Kruskal and Kulsrud average equilibrium equation. The third relates the coefficient β_* in the covariant expression for the magnetic field [Eq. (35)] to the pressure gradient and the variation of the magnetic field strength within a pressure surface.

The equations that have been given, particularly Eq. (39) for the Pfirsch–Schlüter current, allow an iterative solution for equilibria. For example, it is the Pfirsch–Schlüter current that produces a vertical magnetic field in a tokamak that leads to an outward shift of the inner magnetic surfaces relative to the outer surfaces as the plasma pressure is increased.

If the magnetic field is toroidally symmetric, it can be written in a hybrid covariant-contravariant form using (R, φ, Z) cylindrical coordinates,

$$\mathbf{B} = \mu_0 G(\psi_p)\nabla\varphi + \nabla\varphi\times\nabla\psi_p. \tag{43}$$

The curl of this expression gives the current $\mathbf{j} = -\nabla\varphi\times\nabla G + (\Delta_*\psi_p)\nabla\varphi/\mu_0$ with the Grad–Shafranov operator defined by

$$\Delta_*\psi_p \equiv R\frac{\partial}{\partial R}\left(\frac{1}{R}\frac{\partial\psi_p}{\partial R}\right) + \frac{\partial^2\psi_p}{\partial Z^2}. \tag{44}$$

The equilibrium equation, $\nabla p = \mathbf{j}\times\mathbf{B}$, then yields the Grad–Shafranov equation for toroidally symmetric equilibria,

$$\nabla_*\psi_p = -\mu_0\left(\mu_0 G\frac{dG}{d\psi_p} + (2\pi R)^2\frac{dp}{d\psi_p}\right). \tag{45}$$

The pressure and the poloidal current outside at constant-ψ_p surface are the natural functions for defining the equilibrium in a Grad–Shafranov solver.

C. Stability of Equilibria

The stability of plasma equilibria is determined by whether perturbations increase or decrease the energy of the plasma plus the magnetic field [Eq. (31)]. Suppose an equilibrium plasma is displaced a small distance $\boldsymbol{\xi}(x,t)$. Using Eq. (32) with $\mathbf{u} = \partial\boldsymbol{\xi}/\partial t$, one finds that the change in the energy is given by an integral over the plasma volume, $\delta W = -\frac{1}{2}\int\boldsymbol{\xi}\cdot\mathbf{F}(\boldsymbol{\xi})\,d^3x$ with the force associated with the perturbation $\mathbf{F}(\boldsymbol{\xi})$ linearly dependent on the displacement $\boldsymbol{\xi}$. The force is $\mathbf{F} = \mathbf{j}_p\times\mathbf{B} + \mathbf{j}\times\mathbf{b} - \nabla\tilde{p}$ with $\mathbf{j}_p$ the perturbed current and $\mathbf{b}$ the perturbed field. The perturbed pressure is $\tilde{p} = -\boldsymbol{\xi}\cdot\nabla p - \gamma\nabla\cdot\boldsymbol{\xi}$, the perturbed vector potential is $\mathbf{a} = \boldsymbol{\xi}\times\mathbf{B}$, the perturbed magnetic field is $\mathbf{b} = \nabla\times\mathbf{a}$, and the perturbed current is $\mathbf{j}_p = \nabla\times\mathbf{b}/\mu_0$. Now $\boldsymbol{\xi}\cdot(\mu_0\mathbf{j}_p\times\mathbf{B}) = \nabla\cdot(\mathbf{a}\times\mathbf{b}) - b^2$. Since the integral of $\nabla\cdot(\mathbf{a}\times\mathbf{b})$ over all of space is zero, the integral of $\nabla\cdot(\mathbf{a}\times\mathbf{b})$ over the plasma volume can be replaced by the integral of $-b^2/\mu_0$ over the region outside of the plasma. Consequently,

$$\delta W = \delta W_p + \int_{\text{exterior}}\frac{b^2}{2\mu_0}\,d^3x, \tag{46}$$

with δW_p an integral over the plasma volume, which can be written in many ways. One of the most useful forms is

$$\delta W_p = \frac{1}{2}\int_{\text{plasma}}\left\{\frac{b_\perp^2}{\mu_0} + \frac{B^2}{\mu_0}(\nabla\cdot\boldsymbol{\xi}_\perp + 2\boldsymbol{\xi}_\perp\cdot\boldsymbol{\kappa})^2 - \frac{j_\parallel}{B}\mathbf{b}\cdot\mathbf{a} - 2(\boldsymbol{\xi}_\perp\cdot\boldsymbol{\kappa})(\boldsymbol{\xi}_\perp\cdot\nabla p) + \gamma(\nabla\cdot\boldsymbol{\xi})^2\right\}d^3x. \tag{47}$$

The field line curvature $\boldsymbol{\kappa}\equiv\hat{b}\cdot\nabla\hat{b}$, with $\hat{b}\equiv\mathbf{B}/|\mathbf{B}|$, points toward the center of curvature and has a magnitude of one over the radius of curvature. The perpendicular and parallel signs are relative to the unperturbed magnetic field.

A plasma is unstable if a displacement $\boldsymbol{\xi}(\mathbf{x})$ exists for which δW is negative. Only two terms in the entire expression for δW can be negative, the term involving $j_\parallel/B$ and the term involving ∇p. The pressure gradient term is also positive, and therefore stabilizing, if the pressure gradient is in the opposite direction to the curvature vector. This sense of curvature is called good curvature. Unfortunately, the curvature of the magnetic field lines cannot be good everywhere on a toroidal magnetic surface.

Unless the magnetic field lines close on themselves, the displacement with the most negative δW is divergence free. The reason is that the parallel displacement arises only in a positive definite term $\gamma(\nabla\cdot\boldsymbol{\xi})^2$. If the field lines do not close on themselves, the equation $\mathbf{B}\cdot\nabla(\boldsymbol{\xi}_\parallel/B) = -\nabla\cdot\boldsymbol{\xi}_\perp$ can be solved to make $\nabla\cdot\boldsymbol{\xi} = 0$. When this is done the minimum δW is independent of the adiabatic index γ.

The response of a plasma to perturbations of its boundary can be examined not only by δW techniques but also using the equilibrium equations. For simplicity consider a force-free equilibrium, which means $\nabla p = 0$. The equilibrium equation is $\nabla\times\mathbf{B} = k(\psi_t)\mathbf{B}$ with net current $j_n = kB/\mu_0$. The net current distribution k must be constant on irrational magnetic surfaces since the divergence

of the equilibrium equation implies $\mathbf{B}\cdot\nabla k=0$. This is true both for the perturbed, and the unperturbed, plasma equilibrium. To linear order in the perturbation the distribution of net current $k(\psi_t)$ is unchanged by the perturbation except on rational surfaces. On a rational surface the net current can have a term proportional to the Dirac delta function $\delta(n-\iota m)$ and still satisfy $\mathbf{B}\cdot\nabla k=0$. The singularities represented by these delta functions can occur if the perturbation satisfies the ideal Ohm's law, $\mathbf{E}+\mathbf{u}\times\mathbf{B}=0$ so $\iota(\psi_t)$ is conserved. However, the delta function singularities cannot arise if the perturbation satisfies the condition of being a long-lived resistive state, which means the loop voltage is a fixed function of the toroidal flux $V(\psi_t)$. It is the existence of singularities in the net current at the rational surfaces that is the distinguishing feature between ideal and resistive instabilities.

The equation for the magnetic field perturbation in a perturbed force-free equilibrium is

$$\nabla\times\mathbf{b}=\{k(\psi_t)-k(\psi_0)\}\mathbf{B}_0+k(\psi_t)\mathbf{b}, \qquad (48)$$

with $\psi_t(\mathbf{x})$ the toroidal flux enclosed by the perturbed magnetic surfaces and $\psi_0(\mathbf{x})$ the toroidal flux enclosed by the unperturbed magnetic surfaces. Because $\mathbf{B}\cdot\nabla k=0$, the perturbation to the current distribution, $\tilde{k}\equiv k(\psi_t)-k(\psi_0)$, is given by $\mathbf{B}_0\cdot\nabla\tilde{k}\approx-\mathbf{b}\cdot\nabla k(\psi_0)$. Using magnetic coordinates, Eq. (34), $\mathbf{B}_0\cdot\nabla\tilde{k}=(\partial\tilde{k}/\partial\varphi+\iota\partial\tilde{k}/\partial\theta)\mathbf{B}_0\cdot\nabla\varphi$. The magnetic perturbation can be written as

$$\frac{\mathbf{b}\cdot\nabla\psi_0}{\mathbf{B}\cdot\nabla\varphi}=\sum_{m,n}b_{mn}e^{2\pi i(n\varphi-m\theta)}, \qquad (49)$$

so

$$\tilde{k}=\frac{dk}{d\psi_t}\sum_{m,n}\frac{ib_{mn}}{2\pi(n-\iota m)}e^{2\pi i(n\varphi-m\theta)}. \qquad (50)$$

The right-hand side of the equation for the perturbed magnetic field, $\nabla\times\mathbf{b}=\tilde{k}\mathbf{B}_0+k\mathbf{b}$, has two terms. The term $\tilde{k}\mathbf{B}_0$ is by far the more important due to its amplification by the $n-\iota m$ singularity in toroidal devices like the tokamak. The equation for perturbed equilibria is simplest for cylindrically symmetric equilibria, which are a model for toroidal equilibria of very large aspect ratio. Typical shapes for the radial component of the magnetic perturbation using three different magnitudes of the plasma current gradient $dk/d\psi_t$ are shown in Fig. 6. All three have the resonant surface, $n=\iota m$, for the perturbation, which is proportional to $\cos[2\pi(n\varphi-m\theta)]$, at the same radial location. In the absence of any currents in the plasma region, the radial magnetic field would have a radial dependence r^{m-1}. The radial perturbation is assumed smooth at the rational surface for all cases shown, so these equilibria define the resistive stability of the plasma. In case 1, which has the smallest $dk/d\psi_t$, a radial perturbation defined at the plasma boundary is significantly larger at the rational surface that it would be in the absence of a plasma. In case 2, which has a larger $dk/d\psi_t$, even a perturbation of zero amplitude at the plasma boundary causes a nonzero perturbation at the rational surface. In case 3, which has the largest $dk/d\psi_t$, the perturbation at the plasma boundary and at the rational surface have opposite phase. By calculating the energy required to drive a current at the plasma surface that would produce the perturbation, one can show that case 1 is stable and case 3 is unstable (energy is removed from the system, for example, by resistivity, as the perturbation grows). One can show that if $dk/d\psi_t$ is negative and k is positive, which is the normal sense of the current profile in toroidal devices, that the region of the plasma in which $\iota>n/m$ is destabilizing and the region $\iota<n/m$ is stabilizing. In a tokamak with a normal current profile, the transform ι is a maximum at the center of the plasma, so the most unstable perturbations tend to arise when a rational surface is just outside of the plasma. These modes are called external kinks. The case in which $dk/d\psi_t=0$ is clearly particularly stable and is called a Taylor state.

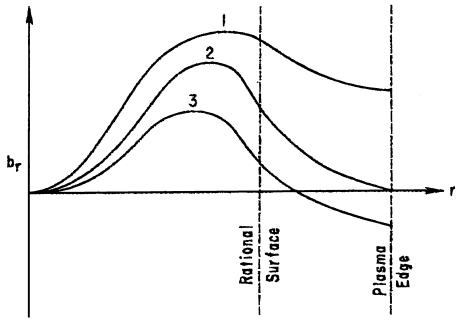


FIGURE 6 Magnetic perturbations. The radial dependence of a magnetic perturbation $b_r=f(r)\cos\{2\pi(n\varphi-m\theta)\}$ is given for three values of the gradient of the force-free current, dk/dr.

Pressure-driven modes can also be studied as perturbed equilibria. The perturbed current for a pressure-driven mode contains a quadratic singularity at a rational surface for the mode $1/(n-\iota m)^2$. One factor of $1/(n-\iota m)$ arises from the singularity of the Pfirsch–Schlüter current, and the other from the $1/(n-\iota m)$ factor in the distortion of the surfaces, which gave the singularity in $\tilde{k}$. The quadratic singularity is sufficiently strong that pressure-driven perturbations can be unstable when they are of arbitrarily short wavelength across the magnetic surface. This is in distinction to perturbations driven by $dk/d\psi_t$, which can only be unstable if their radial wavelength is comparable to the plasma radius.

IV. MIRROR CONFINED PLASMAS

In general, a stationary plasma in which the gyroradius is small compared to the system size has distinct

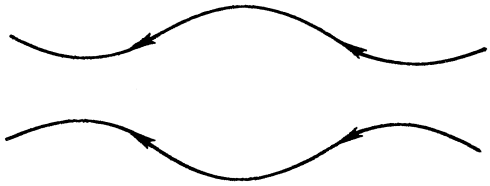


FIGURE 7 Simple mirror. The field of a mirror is strongest where the field lines are closest together.

parallel and perpendicular components of the pressure tensor, $\overleftrightarrow{\mathbf{p}} = p_\parallel \hat{b}\hat{b} + p_\perp(\overleftrightarrow{1} - \hat{b}\hat{b})$, [Eq. (18)]. Collisions tend to reduce the pressure anisotropy, so the maintenance of an anisotropy implies a large amount of power is flowing through the plasma. One can show that $(p_\parallel - p_\perp)/(p_\parallel + p_\perp) < 1/\sqrt{\tau_E \nu}$ where τ_E is the energy confinement time and ν is a collision frequency. Toroidal plasmas can have an isotropic pressure, but the focus of this section is on the confinement of anisotropic plasmas in nontoroidal, open-ended configurations.

A simple mirror (Fig. 7) confines the plasma ions by the variation in the magnetic field strength coupled with the conservation of the magnetic moment μ. The energy of a small gyroradius particle was given as $H = \frac{1}{2}mv_\parallel^2 + \mu B + e\Phi$ [Eq. (5)]. The electrons are often isotropic due to their higher collision frequency and are confined by an ambipolar electric potential Φ. For the fusion application a simple mirror has two problems. First, it is unstable to pressure-driven instabilities, but this problem can be solved by a more complicated field configuration in which the magnetic field line curvature is everywhere in the good direction. Second, the energy confinement time of a simple mirror, $\tau_E \approx 1/\nu_i$ with ν_i the ion collision frequency, is too short to be consistent with an ignited fusion plasma. However space plasmas, such as the plasma in the magnetosphere of the earth, are often mirror confined.

A. Mirror Equilibria

The equilibria of open confinement systems are calculated using $\nabla \cdot \overleftrightarrow{\mathbf{p}} = \mathbf{j} \times \mathbf{B}$. This equation can be written in a more useful form by evaluating the divergence of the pressure tensor to obtain

$$\nabla p_\parallel + \frac{p_\perp - p_\parallel}{B^2} \nabla\left(\frac{1}{2}B^2\right) = \mathbf{j} \times \mathbf{B}, \tag{51}$$

with

$$\mathbf{J} \equiv \frac{1}{\mu_0} \nabla \times (\sigma \mathbf{B}) \tag{52}$$

and

$$\sigma \equiv 1 + \mu_0 \frac{p_\perp - p_\parallel}{B^2}. \tag{53}$$

This is the same equilibrium equation as that of a fluid with magnetic permeability μ_0/σ and pressure $p_\parallel$.

As in toroidal systems, magnetic coordinates greatly simplify the study of equilibria. However, the field lines never circle back on themselves inside an open confinement system; so magnetic coordinates always exist such that the field line Hamiltonian ψ_p is zero in the plasma region. In other words, the magnetic field always has the contravariant representation $\mathbf{B} = \nabla\psi \times \nabla\theta$, which is called the Clebsch representation.

Equilibria of open confinement systems are specified by the parallel pressure in Clebsch coordinates, $p_\parallel(\psi, \theta, B)$. In these variables, the parallel component of Eq. (51) is

$$\left(\frac{\partial p_\parallel}{\partial B}\right)_{\psi,\theta} = -\frac{p_\perp - p_\parallel}{B}. \tag{54}$$

The components perpendicular to the magnetic field are given by

$$\left(\frac{\partial p_\parallel}{\partial \psi}\right)_{\theta,B} \nabla\psi + \left(\frac{\partial p_\parallel}{\partial \theta}\right)_{\psi,B} \nabla\theta = \mathbf{J} \times \mathbf{B}. \tag{55}$$

Sometimes the magnetic flux ψ and the angle θ can be chosen so $\partial p_\parallel/\partial\theta = 0$. This is possible if the drift orbits of the particles deviate little from the ψ surfaces. Although this property of the drift orbits is generally desirable, it is not required unless the θ coordinate is in a symmetry direction. When one can assume $\partial p_\parallel/\partial\theta = 0$, the anisotropic equilibrium equations have many similarities to the isotropic equations. For example, since $\mathbf{J} \cdot \nabla\psi = 0$ one can find a coordinate w such that the magnetic field has a simple covariant representation, $\sigma\mathbf{B} = \nabla w + \beta_*(\psi, \theta, w)\nabla\psi$. Using the fact that the current must vanish at the two ends of the equilibrium, one finds that

$$\int \left(\frac{\partial p_\parallel}{\partial \psi}\right)_B \frac{dl}{B} \tag{56}$$

must be the same on each field line of a ψ surface.

V. TRANSPORT

To maintain a steady-state plasma, sources of particles, energy, and sometimes magnetic flux are required. The determination of these requirements is the subject of transport.

If the mean free path of the particles is shorter than the scale of the plasma, the transport properties for particles and magnetic flux can be studied using the generalized Ohm's law [Eq. (23)]. Energy transport is studied using an equation for the heat flux,

$$\mathbf{q} = -K\left\{\nabla T - \frac{2e}{5p}\mathbf{q} \times \mathbf{B}\right\}, \tag{57}$$

which is analogous to the generalized Ohm's law and is the $\frac{1}{2}mV^2\mathbf{V}$ moment of the kinetic equation [Eq. (11)]. The thermal conductivity $K \approx (5p/2m)/\nu$ is analogous to the electrical conductivity $1/\eta$.

For most plasmas of interest, the mean free path of the particles is much longer than the scale of the plasma. For such plasmas the transport properties must be studied by a direct solution of the kinetic equation, Eq. (11). The central part of such a study is determining the collisionless trajectories of particles in the magnetic electric fields in which the plasma is embedded. This is the subject of Section V.A.

The observed transport of magnetic flux in plasmas is accurately predicted in stable plasmas by a relatively simple application of the kinetic equation. Indeed, the transport of flux is not badly approximated by the generalized Ohm's law [Eq. (23)]. However, the transport of particles and heat is generally much faster than is predicted by a simple application of the kinetic equation. The cause is the presence a low-level fluctuating electric potential. The enhanced transport associated with these fluctuations is called anomalous transport and is discussed in Section V.B.

A. Low Collisionality Transport

When the collisionality of a plasma is high, the effect of the particle drifts (Section II.A), is approximated by the perturbation the drifts make to the Maxwellian distribution that is enforced by collisions. If collisional effects are weak, as they would be in a fusion power plant, the distribution functions must be approximately constant along the drift trajectories of the particles. This means the distribution functions must be accurately approximated by functions of the constants of the drift motion. Nevertheless, the distribution function of each species must be close to Maxwellian if the confinement time is to be long compared to the collision time of that species. If the distribution function is written as $f = (1 + \tilde{f})f_M$ with f_M a local Maxwellian, entropy production arguments imply $|\int \tilde{f} C(\tilde{f} f_M)\, d^3V| \approx n/\tau_E$ with τ_E the energy confinement time and $C(f)$ the collision operator. This implies $\tilde{f} \sim 1/\sqrt{\nu\tau_E}$. The closeness with which a local Maxwellian can be approximated by constants of the drift motion determines the confinement time in the low collisionality limit. If the magnetic field strength has no continuous symmetries in magnetic coordinates, the drift trajectories are generally complicated, and there may be no constants of the motion to keep the particle trajectories near a pressure surface. When this occurs, the thermal conductivity across the pressure surfaces scales inversely with the collision frequency, ν, and the energy confinement time is proportional to the collision frequency, $\tau_E \propto \nu$. This form of transport is generally too rapid to be consistent with the requirements of a fusion power plant.

The essential element in the calculation of transport at low collisionality is the determination of the particle drift motion. If every drift trajectory that was started on a pressure surface remained on that surface, then the drift motion would cause no transport. Transport is largely determined by the extent to which the drift trajectories deviate from the pressure surfaces. Therefore, the drift trajectories are most usefully calculated in magnetic coordinates in which the magnetic surfaces have a simple form ψ_t equal to a constant.

The Hamiltonian form of the drift equations is especially simple in the magnetic coordinates that are defined by Eqs. (34) and (35). The drift Hamiltonian $H = \frac{1}{2}mv_\parallel^2 + \mu B + e\Phi$ has only two degrees of freedom. The two canonical coordinates of the drift Hamiltonian are the angles θ and φ. The momentum conjugate to a coordinate θ in Hamiltonian mechanics is $p_\theta = \mathbf{p} \cdot \partial\mathbf{x}/\partial\theta$ with $\mathbf{p} = m\mathbf{v} + e\mathbf{A}$ the momentum conjugate to the position $\mathbf{x}$. To zeroth order in the gyroradius, the gyro-averaged velocity is $\mathbf{v}_g = (\mathbf{V} \cdot \mathbf{B})/B$. A simple calculation yields the conjugate momenta:

$$p_\theta = (\mu_0 I/B) m v_\parallel + e\psi_t$$

and

$$p_\varphi = (\mu_0 G/B) m v_\parallel - e\psi_p. \tag{58}$$

Hamilton's equations, $dp_\varphi/dt = -\partial H/\partial\varphi$, etc., give the drift trajectories.

In a plasma with perfect magnetic surfaces, the properties of the drift trajectories are determined by $B(\psi_t, \theta, \varphi)$ alone. The trajectories are of two types: trapped or passing. If the electric potential is ignored for simplicity and $B_{max}(\psi_t)$ is the maximum of the field strength on a surface, a particle is trapped if $H < \mu B_{max}$ and passing if $H > \mu B_{max}$. The trapped particles, as the name implies, can only move a limited distance along the field lines. They bounce back and forth between the points where $B = H/\mu$. The passing particles can cover an entire irrational magnetic surface by moving along the field lines. The rapid bounce motion of the trapped particles leads to a second adiabatic invariant, $J = \oint m v_\parallel\, dl$, which is used in studies of particle drift motion.

If the magnetic field strength is toroidally symmetric, $\partial B/\partial\varphi = 0$, the canonical momentum p_φ is conserved. If the field is time independent, the Hamiltonian H is also conserved and the shape of the drift trajectories can be analytically determined. For trapped particles these trajectories have a characteristic banana shape in magnetic coordinates.

If the field strength has no symmetries, the trapped particles are very sensitive to the exact form of the field

strength. Even a small breaking of the toroidal symmetry of a tokamak, by even less than 1%, can greatly enhance the low collisionality transport. The passing particles are insensitive to the exact form of the field strength, but they are very sensitive to any stochasticity in the magnetic field, which can be represented in the drift Hamiltonian by taking ψ_p to be of the form $\psi_p(\psi_t, \theta, \varphi, t)$. The trapped particles are insensitive the magnetic field stochasticity due to their relatively short excursions along the magnetic field lines.

B. Anomalous Transport

As noted earlier, the energy and particle transport that is observed in toroidal experiments is generally much larger than would be expected from calculations based on fluid or low collisionality theory. This enhanced, or anomalous, transport is generally thought to arise from a small variation in electric potential in the magnetic surfaces, $\tilde{\Phi}(\psi_t, \theta, \varphi, t)$. A breakup of the magnetic surfaces, $\psi_p(\psi_t, \theta, \varphi, t)$, could also cause enhanced transport, but the experimental evidence appears consistent with the electric potential being the dominant cause for anomalous transport.

As implied by the name, anomalous transport is not in accordance with theories based on smooth electric and magnetic fields. However, the general magnitude is in agreement with that expected from instabilities associated with drift waves. The phase velocity of a wave is the angular frequency ω divided by the wavenumber k. For drift waves the phase velocity along the magnetic field lines, $\omega/k_\parallel$, is much faster than the ion thermal velocity but much slower than the electron thermal velocity. If $\omega/k_\parallel$ were of the order of the thermal velocity of either species, the particles that move with the velocity $\omega/k_\parallel$ would interact strongly with the waves and damp them. This damping is called Landau damping.

The properties of drift waves can be calculated using the kinetic equation, Eq. (11), but the basic features can be demonstrated much more simply. The electrons can respond to the variations in electric potential by moving along the field lines, so they remain in thermodynamic equilibrium along the magnetic field, $n_e \propto \exp(e\tilde{\Phi}/T)$, or

$$\frac{\tilde{n}_e}{n} \approx \frac{e\tilde{\Phi}}{T}, \tag{59}$$

with $\tilde{n}_e$ the perturbation in the electron density and $\tilde{\Phi}$ the perturbation in the electric potential. The ions move so slowly compared to the wave that the variation in the potential along the magnetic field has no direct effect on their motion. Variations in the ion density $\tilde{n}_i$ are given by $\partial\tilde{n}_i/\partial t + \nabla\cdot(n\tilde{v}_\mathbf{E}) = 0$ with $\tilde{v}_\mathbf{E} \equiv (-\nabla\tilde{\Phi})\times\mathbf{B}/B^2$. A simple calculation yields the result that the ion perturbation is given by $\tilde{n}_i/n \approx (e\tilde{\Phi}/T)(\omega_*/\omega)$ with

$$\omega_* \equiv -k_s \frac{T}{eB}\frac{1}{n}\frac{dn}{dr}, \tag{60}$$

where k_s is the wavenumber perpendicular to the field within the magnetic surface. These surfaces are labelled by a radial distance r. Quasi-neutrality, $\tilde{n}_i = \tilde{n}_e$, then implies that the frequency of drift modes is ω_*, which is called omega star. Ions do not $\mathbf{E}\times\mathbf{B}$ drift at the velocity $\tilde{v}_\mathbf{E}$ if $k_s\rho_i$ is larger than one with ρ_i the ion gyroradius, and $k_s < 1/\rho_i$ for drift waves. The condition that $\omega/k_\parallel > v_i$ then implies that $k_s\rho_i > Lk_\parallel$ with L the scale of the density gradient, $1/L \equiv d\ln(n)/dr$. Since $\rho_i/L \approx 10^{-3}$ in a modern tokamak, the drift modes are very extended along the magnetic field lines and have short wavelengths across the lines. The direction of the magnetic field lines is not the same on all of the magnetic surfaces of a toroidal device. As the field lines change their direction, so must the drift modes in order to keep $k_\parallel$ small. This shear in the direction of the modes can force the radial wavenumber to be large, $k_r \approx k_s$. That is, the wavenumber can have a comparable magnitude, $k_\perp$, in the two directions that are orthogonal to the magnetic field lines.

Drift waves can be made unstable by the trapped electrons in a toroidal device. Trapped electrons have too little parallel energy to move over the full length of a field line, $H < \mu B_{max}$, so they cannot move along the field lines in response to electric potential variations of long wavelength. If the electron mean free path is long, the trapped electrons that are drifting in the magnetic surface at the same velocity as the phase velocity as the wave interact strongly with the wave and can make it unstable. A careful checking of signs demonstrates that this interaction only occurs if the trapped electrons are on average in a region in which the magnetic field strength decreases in the same direction as the density. Unfortunately, the field strength always has this unfavorable form for some trapped electrons and usually for most.

The characteristic diffusion coefficient of drift modes is $D_d \approx \omega_*/k_\perp{}^2$ with $k_\perp\rho_i \approx 1$. Written differently $D_d \approx \rho_i^2 v_i/L$. This transport coefficient is of the correct order of magnitude in most experiments on tokamaks and stellarators. However, many features of transport are not explained by the theory of drift waves. In particular the radial dependence of the plasma temperature is not in agreement with drift wave theory. However, numerical simulations of the kinetic equation—usually based on the drift kinetic equation—are achieving closer agreement with the experimentally observed transport.

Another method of studying plasma transport uses dimensionless parameters. Only three dimensionless plasma parameters are thought to be of primary importance: the

ratio of the ion gyroradius to the radius of curvature of the field lines, ρ_i/R; the ratio of the distance it takes a field line to encircle the plasma to the mean free path of the ions $(R/\iota)/(V_i/\nu_i)$; and the ratio of the plasma pressure to the magnetic field pressure, β. Other dimensionless parameters, such as the rotational transform ι, and the inverse aspect ratio of the torus $\varepsilon = L_n/R$, are more geometric in character but may also enter expressions for transport. The geometric parameters are within an order of magnitude of unity while the three plasma parameters are all small in a fusion reactor. Although dimensionless parameters can be used to scale the confinement times in tokamaks, there is no generally agreed form for this scaling. If there were no critical values for these dimensionless parameters, the confinement time times the cyclotron frequency would have to be given by a power law, $\tau_E\omega_c \propto (\rho_i/R)^{\alpha_1}[(R/\iota)(V_i/\nu_i)]^{\alpha_2}\beta^{\alpha_3}\varepsilon^{\alpha_4}\ldots$. Such scaling laws are used to design experiments, but the uncertainties intrinsic to such extrapolations make it difficult to define the minimal cost experiment for demonstrating a self-sustained fusion burn.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC AND MOLECULAR COLLISIONS • HEAT TRANSFER • ION KINETICS AND ENERGETICS • NUCLEAR FUSION POWER • NUCLEAR PHYSICS • PLASMA SCIENCE AND ENGINEERING • SPACE PLASMA PHYSICS

BIBLIOGRAPHY

Baumjohan, W., and Treumann, R. A. (1996). "Basic Space Plasma Physics," Imperial College Press, London.

Goldston, R. J., and Rutherford, P. H. (1995). "Introduction to Plasma Physics," Institute of Physics Publishing, Bristol.

Priest, E. R. (1982). "Solar Magnetohydrodynamics," Reidel, Dordrecht.

Wesson, J. (1997). "Tokamaks," 2nd ed., Clarendon Press, Oxford.



Plasma Diagnostic Techniques

Clifford W. Mendel, Jr.
Sandia National Laboratories (Retired)

Edl Schamiloglu
University of New Mexico

I. Introduction
II. Magnetic Diagnostics
III. Particle Flux Measurements
IV. Refractive Index Measurements
V. Scattering of Radiation by Plasma Particles
VI. Optical Diagnostics

GLOSSARY

Bremsstrahlung Radiation resulting from free-electron transitions which occur when one particle suffers a deceleration on encountering another particle. It is an important component of continuum radiation.

Čerenkov radiation Radiation that results when a charged particle loses energy while moving through a medium at a speed greater than the wave velocity in the medium.

Cyclotron radiation Radiation emitted by a charged particle moving in a circular orbit in a magnetic field. It occurs at the cyclotron frequency and its harmonics. Also called synchrotron radiation.

Doppler shift A shifting of a spectral line due to the relative velocities of radiating or light-scattering particles.

Electromagnetic scattering Process of radiation emission due to the acceleration of charged particles by electromagnetic waves.

Gridded analyzer Device that utilizes biased grids to measure the flux of ions as a function of their energy.

Interferometer Device that is arranged to compare the relative phases of two paths of coherent radiation, one of which passes through the plasma.

Ion spectrometer Device that can resolve different ion species and record their currents. Usually these involve deflection of the ions by electric and/or magnetic fields.

Langmuir probe Relatively small electrode exposed to a plasma where it collects electrons or ions when a potential is applied.

Laser-induced fluorescence Excitation of atoms or ions by a laser tuned to a selected transition of the atomic species being detected. The resultant radiation is called fluorescence.

Line broadening Increase in the width of a spectral line due to some physical process in the plasma.

Rogowski coil Multiturn toroidal inductor that encircles a distributed current which is to be measured.

Stark shift Shifting of a spectral line due to electric fields in the vicinity of the radiating particle.

Thompson scattering Classical limit of light scattering by free charges.

Zeeman effect Splitting of a spectral line in the presence of a magnetic field.

PLASMA DIAGNOSTIC TECHNIQUES are used to observe physical processes that reveal parameters that characterize a plasma. These parameters include spatial and temporal distributions of constituent particle densities and temperatures and localized magnitudes of electric and magnetic fields. The techniques used include those that have applications in other areas of science and those that have been developed for their unique applications to plasmas.

I. INTRODUCTION

Plasma diagnostics is the observation of physical processes that allows one to infer parameters that characterize a plasma. It had its inception in the late nineteenth century with the observation of colored glows from gas-filled discharge tubes. These plasmas were low in temperature and weakly ionized. By weakly ionized we mean that only a small fraction of the neutral gas is ionized—yet the plasma is quasi-neutral. The field grew rapidly in the first part of the twentieth century after the discovery of electrons, ions, and the ionizing effects of X-rays. The methods used for diagnosing the plasmas relied on electrostatic probes and basic current and voltage characteristics of the discharge tubes. The advent of quantum mechanics led to the use of spectroscopy.

Interest in plasma diagnostics waned until the mid-1950s, when the emerging field of controlled thermonuclear fusion spawned new interest in the subject. The high temperatures and densities of fusion plasmas required the development of new nonintrusive diagnostics. The accompanying development of lasers and sensitive solid-state detectors made possible new diagnostic techniques and the application of known techniques to the new parameter regimes.

The requirements for manufacturing microelectronic and optoelectronic devices and the applications to materials processing in general have caused a resurgence of interest in glow discharge plasmas. These plasmas are low in temperature and weakly ionized, in contrast with their fusion counterparts. The materials processing community is now utilizing many of the plasma diagnostic techniques that were advanced by research in controlled fusion.

Recently, the plasma diagnostics community has been challenged by the need to characterize the strongly coupled plasmas produced in laser-driven inertial confinement fusion experiments and pulsed power-driven Z-pinch radiation sources. In this regime the predominant diagnostic utilizes X-ray imaging techniques coupled with neutron and proton diagnostics to determine the parameters of the target plasma. Advances in X-ray instrumentation and components are leading to greater measurement sensitivity and finer spatial resolution.

TABLE I Plasma Diagnostic Techniques

Diagnostic	Parameter	M or F
Magnetic	T_e, E, B	F
Particle flux	f_e, f_i, n_e, n_i, T_e, T_i, E, v_i	M
Refractive index	n_e, B	M, F
Scattering	f_e, n_e, n_i, T_e, T_i	F
Optical emission	T_e, n_e, n_i	M, F
Line radiation	n_e, n_i, n_0, T_e, T_i, v_i, B	M, F

The two most common plasmas found in the laboratory can be categorized as (1) high-temperature highly ionized plasmas generated in magnetic confinement fusion research experiments, and (2) weakly ionized low-temperature plasmas utilized in the materials processing and gaseous electronics communities. Although we are restricting our discussion of diagnostic techniques to fundamental techniques applicable to laboratory plasmas, many of these methods apply to space and astrophysical plasmas as well.

The diagnostic techniques that we will be describing are listed in Table I. The parameters indicated are f_α (particle distribution function), n_α (particle density), T_α (particle temperature), v_α (particle velocity), E (electric field), and B (magnetic field). Subscript α refers to electrons (e), ions (i), or neutrals (0). The diagnostics are categorized as being most useful in materials processing plasmas (M) or fusion plasmas (F). Many of the diagnostics are equally useful in both types of plasmas.

The diagnostic technique, or combination of techniques, chosen for a particular application represents a compromise among the precision to which the measurement needs to be made, the spatial and temporal resolution required, and the time and expense available to perform the measurement. Plasmas are very complex, often being turbulent and perhaps impure, and the theory of the diagnostic can be equally complicated. In many instances, diagnostic access and extreme plasma parameters limit the available choices to a meager few. For this reason, plasma diagnostics is often referred to as an "art" as well as a science.

II. MAGNETIC DIAGNOSTICS

A. External Measurements

Much can be learned from measuring the magnetic fields outside of a plasma. Figure 1 shows a toroidal plasma being heated ohmically, as is common in magnetically confined fusion experiments. In these devices the plasma is the secondary circuit of a transformer. The primary is a coil like those in most transformers. The toroidal electric field in the plasma is generated by the flow of magnetic flux into the core of the transformer. The resulting voltage about the

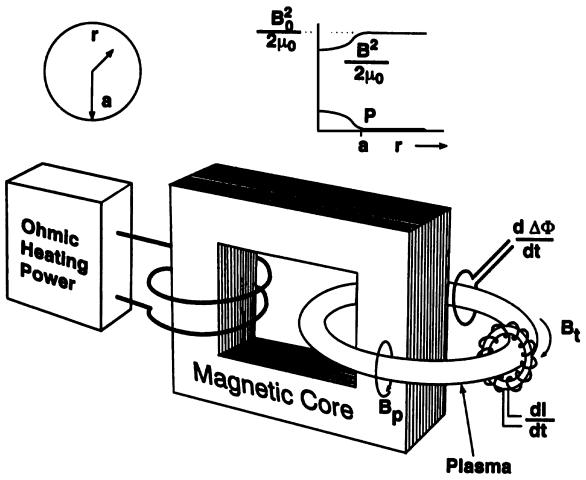


FIGURE 1 Schematic of a toroidal plasma being heated ohmically. The plasma acts as the secondary circuit of a transformer. Also shown is a plot of the magnetic pressure and plasma pressure across the minor radius of the toroid.

loop of plasma V_{loop} can be measured by a conducting loop parallel to the plasma loop.

The current in the plasma loop is also easily measured. From Ampere's law the loop current is given by

$$I = \frac{1}{\mu_0} \oint \mathbf{B} \cdot d\mathbf{l}, \tag{1}$$

where the path of the integral encloses the current to be measured. By recording the integrated voltage from a Rogowski coil (Fig. 1) about the plasma, the integral of the poloidal magnetic field around a loop enclosing the plasma is measured, and therefore the current is measured. From these measurements the power Π into the plasma and the resistance R of the plasma are given by

$$\Pi = IV_{\text{loop}}, \quad R = \frac{V_{\text{loop}}}{I}. \tag{2}$$

In some cases the power needs to be corrected for change in the magnetic field energy.

The resistance of the loop of plasma can, in many cases, be related to the characteristics of the plasma. In the simplest case the electron temperature, which is generally different from the ion temperature, is measured. The resistance of a fully ionized plasma is proportional to the square of the number of electrons removed from each ion divided by the electron temperature to the 3/2 power. It is almost independent of the plasma electron density because the number of scattering centers (i.e., ions) is proportional to the number of charge carriers (i.e., electrons).

Measuring electron temperature by measuring the plasma resistance applies when currents are small enough that the electron drift velocity is smaller than the velocities of the many waves a plasma can support. In cases where this is not true the electrons leave a wake of waves not unlike Čerenkov light from particles moving faster than the speed of light in a material. These waves make local concentrations of charge which increase the resistance of the plasma. This is somewhat like raising the number of times each ion is ionized. In these cases extensive and sophisticated theory is needed to interpret the data, but generally the conductivity is a function of plasma density.

The toroidal magnetic field B_t necessary to confine and stabilize a toroidal plasma is shown in Fig. 1. It is generated by field coils not shown. In an actual experiment this can be taken into account if required. The plasma pressure P is contained by the magnetic field, primarily by the toroidal component. The magnetic field has a pressure

$$P_{\text{magnetic}} = \frac{B^2}{2\mu_0} \tag{3}$$

normal to the field lines. There is also a "tension" of $B^2/2\mu_0$ along field lines, but this can be neglected if the radius of curvature of field lines is large. If the curvature is negligible then

$$P + P_{\text{magnetic}} = P + \frac{B^2}{2\mu_0} = \text{constant}, \tag{4}$$

and the magnetic pressure and plasma pressure look like the plot inset in Fig. 1. Since this is a static situation, the total pressure is constant. Because the magnetic pressure is usually large compared to the plasma pressure for reasons of stability, it can be shown that the flux excluded from the diagnostic loop about the plasma is given by

$$\Delta\Phi = \frac{\mu_0}{B_0} \int_0^{r_{\text{loop}}} 2\pi r P \, dr, \tag{5}$$

where r_{loop} is the radius of the diagnostic loop. Since the magnetic field is B_0 everywhere outside the plasma, the change in flux is actually independent of r_{loop}.

This is therefore a measure of the average plasma pressure or, more correctly, a measure of the pressure perpendicular to the magnetic field, which can be different from the pressure along the field lines. Since this pressure is simply the kinetic energy density, the total perpendicular energy is given by

$$W_\perp = \int_0^{r_{\text{loop}}} 2\pi r P \, dr = \frac{B_0 \Delta\Phi}{\mu_0}. \tag{6}$$

This energy includes both the ion and electron perpendicular kinetic energies. If the density profile is known from other diagnostics the total perpendicular energy reveals the sum of the electron and ion temperatures averaged over the plasma.

B. Internal Measurements

If the axial current in the plasma (I in Fig. 1) is large enough, the poloidal magnetic field B_p must be taken into account. The measurement of current with the Rogowski coil is a measure of the poloidal magnetic field outside the plasma, and if the current profile is known the pressure can still be calculated. If the current profile is not known, it can be measured with magnetic probes. These devices are simple, small coils of wire, with output voltage equal to the rate of magnetic flux change through the coils. The number of turns and the size of each turn depend on the application. Usually the turns have a radius of one to a few millimeters. For magnetically confined fusion plasmas the coils need a large number of turns to get sufficient voltage from the slow changes. For faster fields like those seen in inertially confined fusion experiments, the coils usually have only one or a few turns. The magnetic field is proportional to the integral of the coil voltage.

There are two difficulties with these probes. Like any physical probe placed inside a plasma, these probes disturb the plasma and can lead to rapid plasma loss. In addition, plasmas are much hotter than the melting points of materials. Probe envelopes are usually glass or ceramic, but a probe still must either be in a plasma for a short time or have a method of removing heat at a sufficient rate. A less instrusive method of measuring internal fields is with ion probes. These diagnostics look at the deflection of energetic ions by the magnetic fields. These diagnostics are generally difficult to field and interpret. The orbits of the probing ions in magnetic fusion plasmas are usually complex and are also affected by electric fields at the plasma edge.

III. PARTICLE FLUX MEASUREMENTS

Particle flux measurements generally involve measurements of particle energy distributions or measurements of ion species fractions. One of the first plasma diagnostics was the Langmuir probe. This device consists of an electrode placed inside the plasma. The probe is a perturbative diagnostic since its dimension is usually much larger than the electrostatic shielding distance, that is, the Debye length

$$\lambda_D = \left(\frac{\varepsilon_0 k_B T_e}{e^2 n_\infty}\right)^{1/2}, \tag{7}$$

where $k_B = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant, T_e is the electron temperature and n_∞ is the electron density far from the probe. The voltage on the probe is swept slowly in time and the probe current recorded. Figure 2(a) shows such a probe and the current to the probe plotted

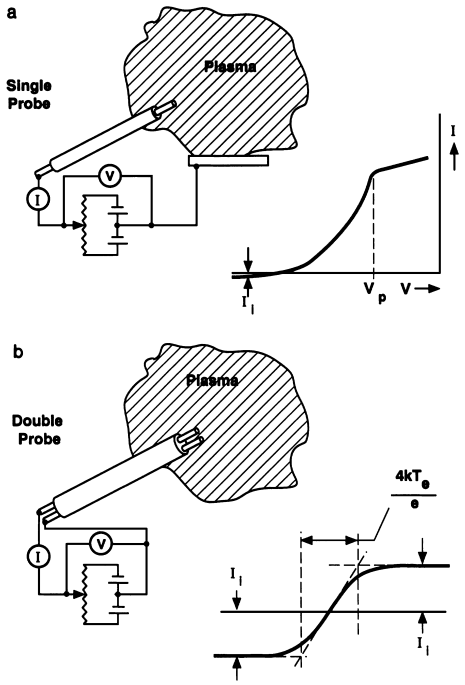


FIGURE 2 (a) Schematic of a Langmuir probe measuring electron and ion current in a plasma. (b) A double probe configuration and the analysis of a double probe $I - V$ characteristic.

versus voltage. The current consists of both ion and electron parts. The ion current density to the probe is given by

$$J_i = \frac{1}{2} n_i q_i v_a, \quad v_a = \left[\frac{q_i k_B (T_e + T_i)}{e m_i}\right]^{1/2}, \tag{8}$$

where n_i is the ion density, q_i is the ion charge, and v_a is the acoustic velocity in the plasma with $T_e \gg T_i$ in most cases. This expression is strictly true for a planar probe and the current is approximately independent of the probe voltage. A more accurate expression can be obtained by solving for the potential distribution in the vicinity of the probe, taking into account the actual probe geometry.

The electron current density is similarly

$$J_e = \frac{1}{4} n_e e v_e = n_e e \left[\frac{k_B T_e}{2\pi m_e}\right]^{1/2}, \tag{9}$$

where v_e is the mean velocity for an electron distribution which is Maxwellian in the direction toward the probe and zero in the direction away from the probe. The factor of one-quarter comes from a factor of one-half which accounts for the fact that particles come from all directions so that the probe area appears one-half its actual area when averaged over all angles, and another factor of one-half because the density near the probe is one-half the value far from the probe. The latter is because there are electrons moving towards the probe but none coming back from it at

the energies that are being collected. The electron density at the surface of the probe is given by

$$n_e = n_\infty e^{eV/k_B T_e} \tag{10}$$

if the electrons are described by a Maxwellian velocity distribution. The total probe current density is thus

$$J_{\text{probe}} \simeq \frac{1}{4} n_\infty e\left[2v_a - v_e e^{eV/k_B T_e}\right]. \tag{11}$$

Clearly, the electron temperature can be obtained from a current plot like that in Fig. 2(a). More typically the ion current is substracted off and the data plotted on semilogarithmic scale, in which case the data lie on a straight line if the distribution is Maxwellian. The reciprocal of the slope of the line is proportional to electron temperature. The product of plasma density and acoustic velocity can be obtained from the ion current (i.e., the current at large negative probe voltage), and the plasma density can be calculated. The probe current at moderate negative voltage is often above the current expected for a Maxwellian in the case of active discharges such as those used in semiconductor processing plasmas. This excess current is due to high-energy electrons that are involved in ionizing background gases.

By measuring the current between two probes held at a voltage difference, but with no net current to the pair, the electron temperature and density can again be found. The above equation for the probe current is applied to both probes at their differing voltages. In addition, the sum of the two currents is set to zero. This method, shown in Fig. 2(b), disturbs the plasma less. The diagnostic is called a double probe. Figure 2(b) also shows a way the data can be analyzed.

Electric probes are very easy to field and, therefore, are often one of the first diagnostics used. It is difficult to get very accurate density data with these probes. The effective area of the probe changes with surface conditions of the probe and nearby insulators. The theory is in general quite complex, particularly if there is a strong magnetic field or a significant plasma drift. Electron temperature measurement generally works well, even in magnetic fields. Nevertheless, Langmuir probes are quite useful when quick or easy measurements are required.

A slightly more complicated electric particle diagnostic is the gridded analyzer shown in Fig. 3(a). This device is external to the plasma. In the case of pulsed plasmas of short output duration, it can be used to measure the fractions of ion charge in the plasma. To do this the analyzer is placed at a distance from the plasma so that the time of flight of the fastest ions is long compared to the length of the plasma pulse at the source. In this situation the velocity of the ions arriving at time t is given by d/t if d

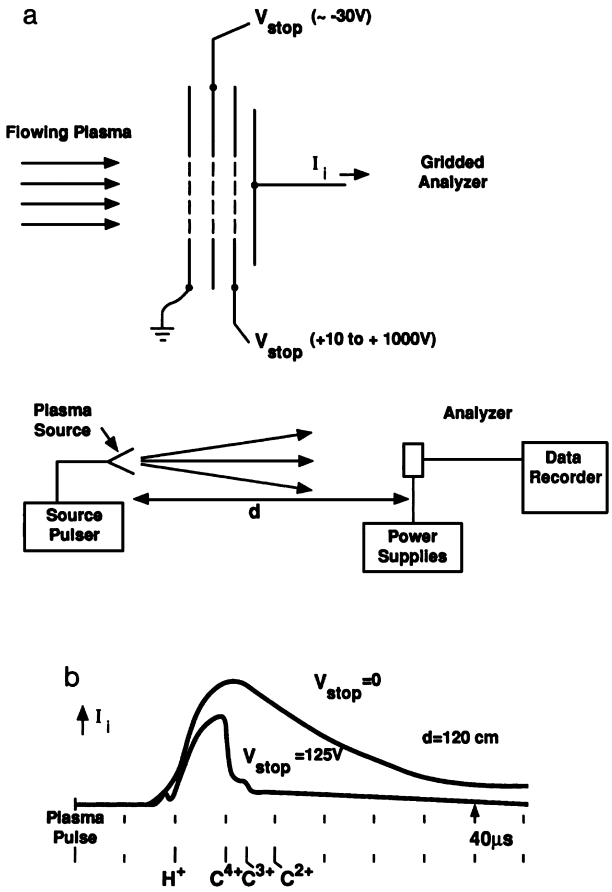


FIGURE 3 (a) Schematic of a gridded analyzer. (b) Data from a gridded analyzer taken 1.2 m downstream from a carbon plasma source.

is the distance from source to analyzer. The analyzer consists of three grids. The first is at the same potential as the plasma source. The second is at a negative voltage that is large compared to the electron temperature of the plasma, which is rarely higher than a few electron volts. This second grid removes the electrons from the plasma but allows the ions to pass. The third grid is at a high positive voltage, V_{stop}. Only ions of kinetic energy greater than their charge times V_{stop} can pass through the third grid to the collector. Thus, the collector measures the current of those ions that meet the criterion

$$\frac{m_i d^2}{2t^2} > q_i V_{\text{stop}} \Rightarrow V_{\text{stop}} t^2 < \frac{m_i d^2}{2q_i}. \tag{12}$$

Figure 3(b) shows data taken 1.2 m from a small carbon plasma source. There are two analyzers, one with 125 V, and the other with 0 V on the stopping grid. The collector current steps down as H^+ (protons), C^{4+}, and C^{3+} ions are stopped. The ratio of steps in the collector current of the 125-V analyzer to the current of the 0-V analyzer is the charge fraction of the particular ion species at that

time. By taking data at a number of stopping voltages, the experimenter can find the charge fractions at various times or kinetic energies.

Another way these devices are used is with steady-state plasmas such as those found in processing plasmas or space plasmas. In these applications, the plasma ions flow through a small aperture in a containment wall and into the analyzer. The plasma is not at the same potential as the wall, since the electron current to the wall would then be too large. For this reason, the ions arrive with kinetic energy equal to $q_i V_{\text{plasma}}$ plus the thermal energy. The plasma potential can be obtained from the data, and so can the ion temperature if it is sufficiently high. The plasma potential can generally be related to the electron temperature.

Devices like the gridded analyzer, but without the ion collector, are often used to preselect ions for an ion spectrometer. The ion spectrometer then measures the concentration of various ion species in the plasma. Ion spectrometers generally use electric and/or magnetic fields to deflect the ions.

IV. REFRACTIVE INDEX MEASUREMENTS

Plasmas support a great many waves, and the phase velocities of these waves are related to important plasma parameters. Generally the waves are measured by comparing the number of wavelengths in a known path with and without the plasma in place. Figure 4 presents a schematic of an interferometer used to make such a measurement. With no plasma, the number of wavelengths in the reference leg is adjusted to be one-half wavelength more or less than the measurement leg. The plasma in this case acts like a refractive medium, and the plasma density can be related to the index of refraction.

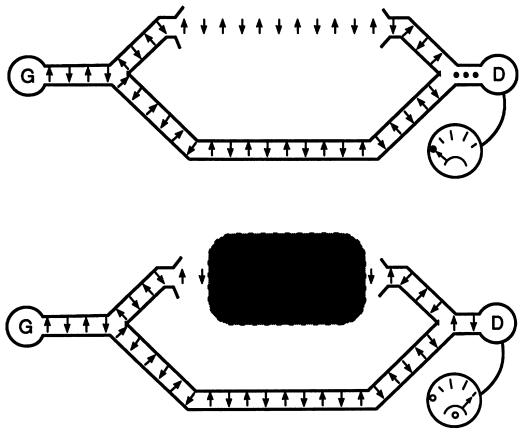


FIGURE 4 Schematic of an interferometer measurement of phase shift in a plasma. G, generator; D, detector.

When plasma is placed in the path, the number of wavelengths in the measurement leg changes. This results in a detector output that passes through successive maxima and minima as the number of wavelengths in the measuring leg goes through odd and even multiples of half wavelengths. By counting the number of these "fringes," the experimenter can ascertain the number of wavelengths in the path. In the example shown, the number decreases. An example of this is the case of an unmagnetized plasma and microwave propagation. The initial number of wavelengths is given by the measurement length times the source frequency divided by the speed of light c. When plasma is placed in the path, the wavelength of the waves is

$$\lambda_{\text{plasma}} = \frac{c}{\left[f^2 - f_p^2\right]^{1/2}},$$

where

$$f_p = \frac{1}{2\pi}\left[\frac{n_e e^2}{\varepsilon_0 m_e}\right]^{1/2} \tag{13}$$

so that the number of wavelengths is

$$N_{\text{plasma}} = N_0 \frac{\left[f^2 - f_p^2\right]^{1/2}}{f}, \tag{14}$$

where N_0 is the number of wavelengths in the path in the absence of plasma. The plasma frequency $f_p = \omega_p/2\pi$ is therefore given by

$$\frac{f_p^2}{f^2} = 1 - \frac{N_p^2}{N_0^2}, \tag{15}$$

where N_p is the number of fringes counted from the time when the plasma density is zero. The plasma electron density can then be calculated from the definition of the plasma frequency. When the plasma density reaches a value such that $f_p = f$, the number of wavelengths goes to zero, and the waves are cut off, so no measurement can be made at higher densities.

When there is a magnetic field in the plasma, the wavelength depends on the plasma frequency shown above and also on the electron cyclotron frequency given by

$$f_c = \frac{\omega_c}{2\pi} = \frac{eB}{2\pi m_e}. \tag{16}$$

The wavelength also depends on the direction of propagation and the direction of the electric vector of the (transverse) waves. If the direction of propagation is normal to the magnetic field, and the electric vector is parallel to the magnetic field, the wave is called an ordinary wave because the wavelength is given by the zero magnetic field expression above. If the propagation is normal to the magnetic field and the electric vector is normal to the field and to the direction of propagation, the wave is called an

extraordinary wave, and the wavelength depends on the plasma frequency and the cyclotron frequency.

Waves propagating along magnetic field lines are broken into left and right polarized circular waves, and their wavelength depends on which of these polarizations they have. If a wave propagating along the magnetic field is polarized in a plane, it will consist of equal amplitudes of left and right polarized waves, and its plane of polarization will rotate along the field.

Generally, these waves have cutoffs and resonances. Cutoffs are regions where the combination of wave frequency, plasma density, and magnetic field strength results in evanescent waves that do not propagate, whereas resonances are regions where the combination of the above parameters results in zero wavelength. By making measurements at very high frequencies, the wavelengths become close to the vacuum wavelengths, and averages of the plasma parameters tend to be measured.

Although the interpretation of these measurements can be complicated, the data are quite accurate when the measurement can be understood. By taking measurements along many chords through the plasma, profiles of plasma density usually can be measured.

The source of the waves depends on the source of the plasma. For plasmas in the 10^{11} to 10^{15} cm^{-3} density range, microwaves are generally the choice. At very high densities like those found in laser or particle beam generated plasmas, laser light can be used. For astrophysical plasmas, radio-frequency waves are used. For example, the dispersion of radio waves from pulsars (rapidly rotating neutron stars) allows the measurement of the number of electrons per unit area between earth and the pulsar.

V. SCATTERING OF RADIATION BY PLASMA PARTICLES

The scattering of electromagnetic radiation from plasmas is a nonperturbing method of obtaining detailed information about the distribution function of electrons and, in certain instances, ions as well. It is generally a difficult diagnostic to implement technically. This diagnostic method is most widely applied in fusion plasma experiments.

A classical view of scattering consists of an electromagnetic wave impinging on a charged particle. The wave electric fields accelerate the particle and cause it to emit electromagnetic radiation in all directions. This radiation is termed the scattered wave. This description is valid, provided the incident wave photon mass is much smaller than the charged particle rest mass, that is, $h\nu \ll m_0 c^2$, where $h = 6.63 \times 10^{-34}$ J · sec is Planck's constant, ν is the incident wave frequency, m_0 is the charged particle rest mass, and c is the speed of light. For plasma diagnostic applications, the incident photons are typically on the order of 1 eV, while the rest mass of an electron is 0.5 MeV. This classical limit of scattering by free charges is called Thomson scattering.

The theory of scattering of electromagnetic waves by plasmas was first applied to studies of the interaction of microwaves with the earth's ionosphere. Initial laboratory applications utilized microwaves as the incident radiation source as well. The cross section for scattering of electromagnetic radiation by an electron is given by the Thomson cross section,

$$\sigma = \frac{8\pi}{3} r_e^2, \tag{17}$$

where $r_e = e^2/4\pi\varepsilon_0 m_e c^2 = 2.82 \times 10^{-15}$ m is the classical electron radius. Since the cross section is of the order of the electron size, which is to say very small, early laboratory experiments utilized microwaves scattered from groups of electrons. As a result the experiments provided information concerning the collective motion of the plasma wherein the motions of electrons were correlated.

The development of pulsed high-power lasers in the visible to far-infrared frequencies has allowed scattering from random motions of individual electrons to be measured. The total scattering from a plasma is a result of the superposition of the electric field contributions from each individual electron. Both amplitude and phase of the scattered electric fields are required to calculate the net scattered power. If the phases of the scattered waves are totally uncorrelated, then the powers of the scattered fields, obtained by taking the modulus squared of the electric fields, are simply summed. This is termed incoherent scattering. If there is a significant correlation between the electrons, then a coherent sum of the electric fields from the individual scatterers must be performed. This is termed coherent scattering. Incoherent scattering can be distinguished from coherent scattering by comparing the scattering wavenumber k with the Debye shielding length $\lambda_D = v_e/\omega_p$. Incoherent scattering occurs when $k\lambda_D \gg 1$. By measuring the scattered spectrum, that is, scattered power as a function of frequency, the one-dimensional velocity distribution function is measured and the plasma electron temperature and density are obtained.

An example of an incoherent Thomson scattering experiment is shown in Fig. 5. Consider a beam of electromagnetic waves propagating a characteristic length L across a plasma with electron density n_e. A fraction of the incident photons given by $\sigma n_e L$ will be incoherently scattered, and only a fraction of those scattered photons will be actually collected. Generally, in a fusion plasma experiment where typically $n_e \geq 10^{14}$ cm^{-3} and $L = 10^2$ cm, only 10^{-13} or so of the incident photons will be detected. This points out the practical difficulty of implementing

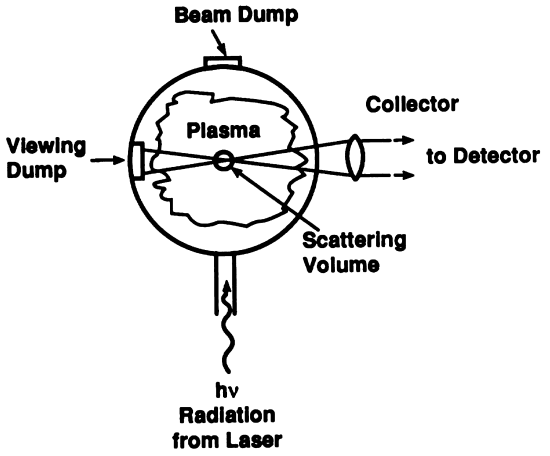


FIGURE 5 Schematic of a Thomson scattering experiment.

this diagnostic. One must usually resort to an energetic pulsed laser in such an experiment to ensure a measurable scattered signal. Stray light, plasma emission, and other sources of parasitic radiation need to be minimized to achieve an acceptable ratio of scattered to background light.

In the case where there is significant correlation between the electrons in the plasma, that is, $k\lambda_D \ll 1$, one must perform a coherent sum of the electric fields from the individual scatterers. One is interested in obtaining specific information concerning macroscopic fluctuations in the plasma. The collective effect of ions may be important in this case. This regime of scattering requires detailed calculation of a scattering form factor in order to interpret the scattered power distribution.

The CO_2 laser has been extensively used for the purpose of plasma density fluctuation measurement. Heterodyne or homodyne detection techniques are commonly employed with continuous radiation sources. Homodyne detection involves using a portion of the incident source to mix with the scattered signal, whereas heterodyne detection involves using a separate frequency source. Interferometry is functionally the equivalent of homodyne detection of scattering at zero angle.

VI. OPTICAL DIAGNOSTICS

Optical diagnostics refers to the observation of electromagnetic radiation that is emitted from a plasma. The emission may occur naturally because of physical processes that are taking place among the plasma constituents or between the constituents and electromagnetic fields, or the emission may be induced by the introduction of an external source of electromagnetic radiation or, possibly, by the introduction of atoms or charged particles. This is a nonperturbing method of characterizing plasma properties that is well established in fusion plasma experiments and is gaining importance in low-temperature, weakly ionized plasmas utilized by the materials processing community. Optical diagnostics are generally difficult to implement because they require careful alignment of optical components, and the interpretation of measurements generally requires the use of appropriate physical models describing the particular process being employed. Our discussion of optical diagnostics will be divided into radiation emission from free electrons and line radiation from atoms and ions that are not completely ionized.

A. Radiation from Free Electrons

Charged particles emit electromagnetic radiation when they are accelerated. Electrons have low mass and experience greater accelerations than heavier particles. As a result, electrons emit radiation more abundantly than other particles. Accelerations occur because of the influence of electric or magnetic fields. In many instances, plasmas are imbedded in an external magnetic field, and the resultant gyration of the electrons causes the emission of radiation, which occurs at frequencies that are multiples of $\omega_c = eB/m_e$. This is called cyclotron radiation.

Steady electric fields usually result in negligible radiation, but rapidly time varying electric fields can cause significant emission of radiation. Indeed, this is the process of electromagnetic scattering. Electrons moving in a plasma can be significantly slowed down because of undergoing Coulomb collisions with the nearly stationary ions. This process is called bremsstrahlung radiation from the German "braking radiation." Bremsstrahlung radiation is typically observed over a very broad frequency spectrum and is an important component of continuum radiation, that is, background emission.

Once the measurement of radiation from free electrons is attributed to one or more of the above processes, an appropriate model for the emission can be used to obtain spatial and/or temporal distribution of electron temperature in the plasma. Electron density and sometimes ion density information can be obtained as well.

B. Radiation from Bound Electrons

Atoms, ions, and impurity components of a plasma emit radiation when the bound electrons undergo atomic transitions from one energy level to another. The transition is characterized by the emission of radiation in a narrow spectral line. Different information about plasma properties can be obtained from the intensity of the lines observed and the shape of the lines. Each case will be discussed separately.

1. Spectral Line Intensity

Measurement of the absolute intensity of a spectral line provides an estimate of the density of an atomic species. Since the line intensity is a measurement of the density of the excited state, an appropriate equilibrium model relating the excited state density to the ground state density allows the determination of the density of the species. This is used to obtain information about plasma impurities as well as ratios of the abundance of the constituent atoms.

Impurities in hot plasmas can be used to provide a rough estimate of the electron temperature. This is because the ionization state of an impurity atom is a strong function of electron temperature. One needs to be careful when performing this measurement since knowledge of the relative abundance of the impurity atom is not precise and the impurities may not be in ionization equilibrium. In addition, there may be gross variations in electron temperature throughout the plasma. As a result, one may incorrectly infer the electron temperature by assuming the radiation is uniformly emitted from a plasma volume.

A good method for estimating electron temperature is to measure the relative ratio of line intensities from the same species of the plasma, or to measure the ratio of line-to-continuum intensity. This measurement is important in astrophysical plasmas as well as laboratory plasmas. An example of a time-dependent electron temperature measurement in a beam-plasma experiment is shown in Fig. 6. Light from a quiescent magnetically confined hydrogen plasma is collected by a lens and fed into a grating spectrometer, which disperses the photons according to their wavelength. The time-dependent intensity of the H_γ line of the Balmer series is measured simultaneously with the intensity of continuum radiation. Fiber optic cables are used to transmit the spectrum from the spectrometer to be measured with photomultiplier tubes in this case. The line-to-continuum ratio is observed to decrease when an energetic proton beam is injected into the plasma (Fig. 6). An increase in electron temperature from 5 to 6 eV is measured, consistent with plasma heating due to Coulomb collisions between the beam protons and plasma electrons in the experiment.

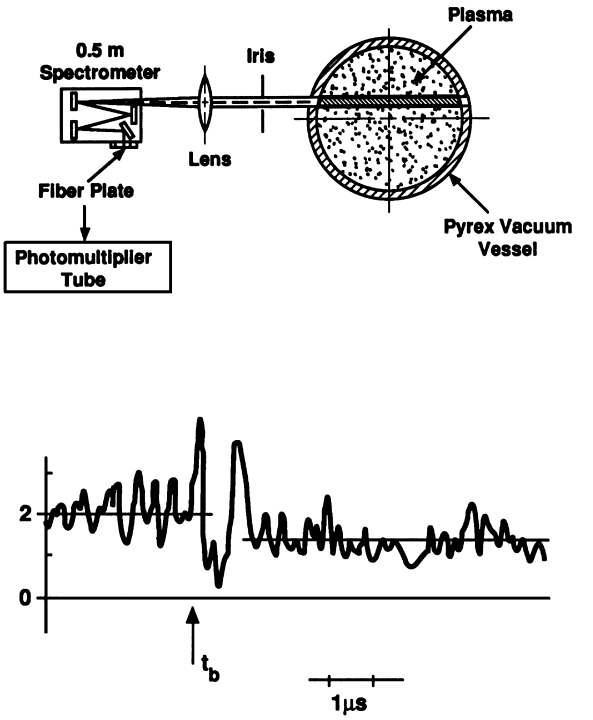


FIGURE 6 Schematic of a line-to-continuum intensity measurement of electron temperature in a beam-plasma experiment (top). The time-dependent measurement indicates that the intensity ratio decreases when the plasma is heated by a proton beam at time t_b (bottom).

2. Spectral Line Shape

The shape of an emitted spectral line changes when various physical processes occur. These changes can be used to ascertain information concerning plasma parameters and electric and magnetic field values in the vicinity of the measurement. A spectral line shape is typically characterized by a Lorentzian distribution. This distribution takes into account natural line broadening that occurs because of the spread in energy of the quantum states of an atom. The full width at half maximum (FWHM) is the common method of describing a line width and is related to the lifetime of an excited state of an atom.

The thermal motion of atoms in a plasma results in a Doppler shift of the emitted radiation frequency. This appears as a broadening of the spectral line distribution. A measurement of the Doppler half-width allows one to ascertain the ion temperature for the species being measured.

Another process that results in the broadening of the spectral line distribution is pressure broadening. This is also known as collisional, or Stark, broadening and is most prevalent in high-density, low-temperature ($n > 10^{15}$ cm^{-3}, $T_e \leq 4$ eV) plasmas. This effect arises from the influence of the electric field of particles in the vicinity of the radiating atom. The frequency of radiation is shifted because of the perturbing effect of the random electric fields, called the Stark shift. If the plasma parameters are such that the Stark shift is the dominant broadening mechanism, then a measurement of the electron density can be ascertained from the line width. In high-temperature fusion plasmas, Stark broadening is a negligible effect when compared with Doppler broadening.

A diagnostic that is gaining importance, especially in radio-frequency (RF) plasma reactors used for materials processing, is fluorescence. This diagnostic method is based on the active perturbation of the excited state populations to obtain additional information from the emitted line radiation. Laser-induced fluorescence (LIF) occurs when a small volume of the plasma is irradiated with an electromagnetic source that is tuned to a resonant line of a constituent atom of the plasma (Fig. 7). The atomic species that are accessible by laser fluorescence techniques are limited by the wavelength tunability and the power of available sources. Neutral metal atoms are of interest to many plasma processing systems. These species have many allowable transitions between the near ultraviolet and near infrared wavelength regions. A tunable pulsed dye laser is generally used as the radiation source for these investigations. Laser-induced fluorescence is a very sensitive technique for obtaining temporal and three-dimensional spatial resolution of number densities.

Tunable infrared lasers are gaining importance as sensitive probes of species in plasma processing. The most commonly used continuously tunable source is the diode laser. These lasers are used to identify a number of stable molecules, free radicals, and ions from the laser absorption spectra. It is possible to obtain translational temperatures of atoms from the width of the absorption line and the number density from a measurement of the absolute infrared absorption cross sections. A disadvantage of this method as a general plasma diagnostic is that it is very cumbersome to tune over a wide spectral range. This limits the species that can be detected. In addition, since the sources used are continuous rather than pulsed, time-dependent information cannot be obtained in the important time regime below 100 nsec.

Finally, the presence of a magnetic field imbedded in a plasma causes the splitting of emission lines. The magnitude of the field may be deduced from the magnitude of the splitting due to the Zeeman effect. Splitting from even natural line radiation can be observed, in principle. This diagnostic is primarily relevant to magnetically confined plasma experiments in tokamaks and to astrophysics. The main difficulty is ensuring that the splitting of the line dominates over Doppler broadening. Resonance fluorescence techniques may be used to enhance the observed emission intensity.

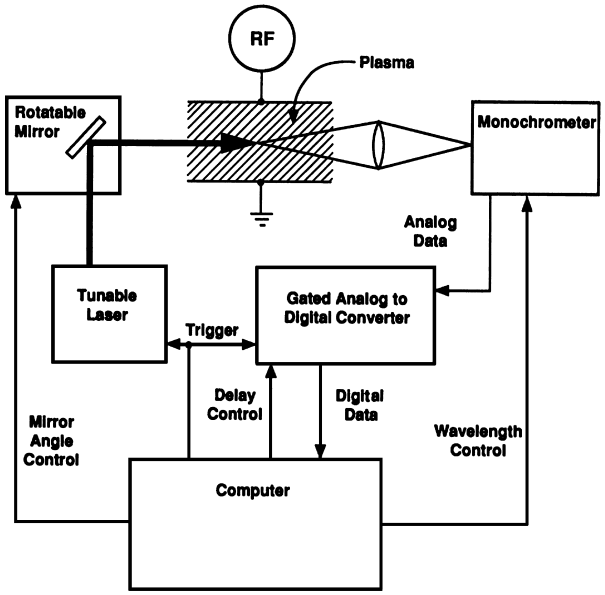


FIGURE 7 Schematic of a laser-induced-fluorescence experiment in an RF plasma reactor.

SEE ALSO THE FOLLOWING ARTICLES

PLASMA CONFINEMENT • PLASMA SCIENCE AND ENGINEERING

BIBLIOGRAPHY

Auciello, O., and Flamm, D. L., eds. (1989). "Plasma Diagnostics, Vol. 1, Discharge Parameters and Chemistry," Academic Press, San Diego.
Auciello, O., and Flamm, D. L., eds. (1989). "Plasma Diagnostics, Vol. 2, Surface Analysis and Interactions," Academic Press, San Diego.
Bekefi, G. (1966). "Radiation Processes in Plasmas," Wiley, New York.
Griem, H. R. (1997). "Principles of Plasma Spectroscopy," Cambridge University Press, Cambridge, UK.
Heald, M. A., and Wharton, C. B. (1978). "Plasma Diagnostics with Microwaves," 2nd ed., Krieger, Malabar, FL.
Huddlestone, R. H., and Leonard, S. L., eds. (1965). "Plasma Diagnostic Techniques," Academic Press, New York.
Hutchinson, I. H. (1987). "Principles of Plasma Diagnostics," Cambridge, New York.
Lochte-Holtgreven, W. (1995). "Plasma Diagnostics," American Institute of Physics, Woodbury, NY.
Manos, D. M., and Flamm, D. L., eds. (1989). "Plasma Etching: An Introduction," Academic Press, San Diego.
Oks, E. (1995). "Plasma Spectroscopy: The Influence of Microwave and Laser Fields," Springer-Verlag, Berlin.
Salzmann, D. (1998). "Atomic Physics in Hot Plasmas," Oxford University Press, New York.
Sheffield, J. (1975). "Plasma Scattering of Electromagnetic Radiation," Academic Press, New York.
Silver, E. H., and Kahn, S. M., eds. (1994). "UV and X-Ray Spectroscopy of Astrophysical and Laboratory Plasmas," Cambridge University Press, Cambridge, UK.



Relativistic Ion Physics

William A. Zajc
Columbia University

I. Introduction
II. The Relativistic Heavy Ion Collider
III. RHIC Experiments
IV. RHIC Physics
V. Future Directions

GLOSSARY

Baryon A particle such as a neutron or proton consisting of three quarks and assigned a baryon number equal to one. The baryon number is a conserved quantity. Protons and neutrons are the lightest variety of baryons; heavier versions either are excited states built on the same quark basis as the proton and neutron or contain heavier quarks than the up and down quarks found in protons and neutrons.

Central collision Collisions between nuclei that are approximately head-on. Experimentally, this selection is typically made by selecting a fraction $\sim$5 to 10% of the most energetic events.

Color The analog to charge in quantum chromodynamics; describes the strength of the interaction between quarks and gluons.

Gluon The massless, spin-one boson that mediates the strong force of quantum chromodynamics. Gluons carry one of eight different color "charges."

Hadron Any particle that interacts via the strong interaction. All hadrons are composed of quarks: Baryons contain three quarks, while mesons consist of a quark–antiquark pair.

Heavy ion In this context, the nucleus of an atom used as a colliding species.

Meson Any strongly interacting particle that is not a baryon. All known mesons consist of a quark and antiquark pair.

Minimum bias Refers to a selection of collisions with the least possible trigger bias, ideally representing the full interaction cross section.

Nucleon A generic name for a proton or neutron—that is, the baryonic constituents of a nucleus.

Parton A generic name for the constituents of an elementary (hadronic) particle. Partons are composed of quarks and/or gluons but represent any portion of a hadron that may be treated as an independent object for some collision process.

Quantum chromodynamics (QCD) The theory that describes the interaction of quarks and gluons interacting via the exchange of color. QCD is based on analogy to quantum electrodynamics (QED), which describes the interaction of charged particles via the exchange of photons. Unlike the photon in QED, the force carrier in QCD itself carries a charge; that is, gluons have color charge.

Quark As elementary constituent of hadrons. Six types

of quarks, grouped into three "generations," are known to exist: up, down; charm, strange; and top, bottom.

Quark–gluon plasma (QGP) A state of matter in which quarks and gluons are not bound within their constituent particles but are deconfined over some much larger volume. This state is thought to exist only at temperatures comparable to a hadron mass, or at densities comparable to the mass densities of hadrons themselves.

Strong interaction The short-range force between hadrons responsible for binding nucleons into nuclei. More generally, the strong interaction refers to the interactions between quarks and gluons, so that the force binding nuclei is in fact a special case of this interaction.

RELATIVISTIC heavy ion physics studies collisions of atomic nuclei at very high energies. Such collisions produce the hottest, densest matter ever formed in the laboratory. The properties of nuclear matter at such extreme density and temperature are intrinsically interesting quantities which probe the boundaries of our understanding of matter itself. A particular emphasis of the field is the formation of a new state of matter, the quark–gluon plasma (QGP), in which quarks and gluons are no longer confined within nucleons but are "deconfined" over a volume comparable to or somewhat larger than the initial nuclear volume. The best prospects for the systematic production and measurement of QGP are provided by experiments at a dedicated facility such as Brookhaven National Laboratory's Relativistic Heavy Ion Collider (RHIC), which began operation in 2000.

I. INTRODUCTION

A fundamental goal of many branches of physics is the understanding of properties of matter. The range of sizes, densities, and temperatures over which it is appropriate to describe a "state of matter" is quite extraordinary. The universe itself is a system having a mean density of order 10^{-25} gm/cm^3 and a temperature of 2.7 K. Condensed matter physicists study samples of ordinary densities $\sim$1 gm/cm^3 yet with temperatures a trillion times lower, while plasma physicists seek to control diffuse atomic plasmas of density 10^{-9} gm/cm^3 at temperatures exceeding 10^4 K.

Far and away the hottest, densest conditions ever formed in the laboratory are found in the collisions of "relativistic heavy ions," for which densities surpass 10^{15} gm/cm^3 and temperatures are in excess of 10^{12} K. (In the remainder of this article, temperature will be given in energy units of kT, so that 10^{12} K $\sim 10^8$ eV $= 100$ MeV.) Heavy ions in this context are simply atomic nuclei of heavy elements (typically gold or lead), stripped of all their electrons. The collisions are relativistic in the sense that the collision energies are large in comparison with the rest mass of the nuclei. Under these conditions, between 10^3 and 10^4 particles emerge from the collision, giving rise to the possibility of a statistical description in which quantities such as energy density and temperature are appropriate variables. In addition, our current understanding of such hot, dense matter predicts that the system will pass through a *phase transition*, in which the individual elementary particles temporarily cease to exist, and the matter is characterized instead by a deconfined state of quarks and gluons. This transition is of basic interest as it represents the only experimentally accessible phase transition in a fundamental theory (quantum chromodynamics). It also has relevance to studies of the early universe since this deconfined state was the dominant form of matter in the first few microseconds after the Big Bang. Finally, study of hadronic systems under such extreme conditions is the logical extension of decades of effort to understand the equation of state of "ordinary" nuclear matter both in atomic nuclei and in neutron stars.

Extensive measurements of the charge distribution of atomic nuclei have demonstrated that the radius of a nucleus with atomic number A (when modeled as a uniform sphere) is approximated by:

$$R(A) \approx r_0 A^{1/3} \tag{1}$$

with $r_0 = 1.2 \times 10^{-15}$ m $\equiv 1.2$ fm. Therefore, nuclear matter in its ground state should have a characteristic density ρ of order:

$$\rho = \frac{A}{\frac{4\pi}{3}R(A)^3} = \frac{1}{\frac{4\pi}{3}r_0^3} = 0.14 \text{ nucleons/fm}^3 \tag{2}$$

A more accurate treatment that takes into account the nonuniform distribution near the surface of the nucleus gives a somewhat higher value for the central density of 0.17 nucleons/fm^3, which is thought to be the appropriate value for the ground state of nuclear matter in the limit of very large A (provided gravitational effects are neglected, which is valid for all laboratory experiments with nuclei). For "relativistic" heavy ion collisions, it is usually more convenient to consider the *energy* density rather than the spatial density of nuclear matter. Since each nucleon has an energy associated with its rest mass of 0.94 GeV (neglecting the very slight mass difference between protons and neutrons), the energy density ρ_0 of nuclear matter in its ground state is approximately 0.16 GeV/fm^3.

The detailed description of nuclear matter in terms of the known properties of protons, neutrons and the interactions between them was a major triumph of nuclear physics over a 40-year period, beginning in the early 1930s. During

that time, it sufficed to treat the nucleons as elementary particles, and in fact for much of that period it was assumed that the nucleon was in fact an indivisible object. However, it is now known that protons, neutrons, and all other hadrons are composed of quarks, bound together via the exchange of gluons. The interactions of quarks and gluons is described by quantum chromodynamics (QCD), which assumes that each type of quark may have one of three "color" charges, conventionally labeled as red, green, or blue. Gluons couple to this charge, just as photons couple to conventional electrical charge, but with the crucial distinction that gluons themselves carry one of eight color charges. The distinction is crucial in the sense that the color charges of gluons imply that gluons also couple to other gluons, which in turn is responsible for *color confinement*—that is, the property that quarks are never observed as isolated objects. While QCD does not predict the number of quark types, analysis of the rich spectrum of baryons and mesons has demonstrated that there are six "flavors" of quarks, grouped into three "generations": up and down, charm and strange, and top and bottom. The generations are ordered by mass, with up and down quarks being the lightest. While the masses of the heavy quarks can be estimated quite reliably to be $m_c = 1.15$–1.35 GeV, $m_B = 4.0$–4.4 GeV, and $m_t = 174$ GeV, the masses of the up, down, and (to some extent) the strange quark are much more difficult to determine. Current ranges are $m_u = 1$–5 MeV, $m_d = 3$–9 GeV, and $m_s = 75$–175 MeV.[1]

The striking range of quark masses remains one of the most interesting puzzles in our current understanding of the elementary consituents of matter. Of particular interest here are the very light values for the up (u) and down (d) quarks. In terms of the natural mass scales of hadronic physics (0.1 to 1.0 GeV) they are essentially massless. At the same time, massive particles such as protons and neutrons consist of (uud) and (udd) triads of quarks. The mass of such hadrons results from the confinement mechanism of the QCD vacuum, which is often modeled as producing a constant energy per unit volume B of order 0.2 GeV/fm^3. (This quantity is generally referred to as the *bag constant*, after models in which hadrons are modeled as a set of quarks contained within a bag of this energy per unit volume.) Our picture of normal nuclear matter is thus one in which protons and neutrons retain their identities, each consisting of a bag containing three quarks that remain isolated from the other nearby quarks in neighboring nucleons. It is then natural to ask, "What conditions of temperature and density are required for a *phase transition* from normal nuclear matter to a state characterized by quarks and gluons?"

A useful estimate of the required conditions may be obtained by a simple argument from statistical mechanics, based on the following set of assumptions:

- The confinement of quarks within hadrons results is modeled as the result of the bag constant B.
- Normal hadronic matter is approximated as a gas of noninteracting pions, which is further approximated by treating the pions as being massless.
- Quark–gluon matter is also treated as a noninteracting gas of quarks and gluons, also considered as approximately massless.

It should first be noted that these assumptions are not directly related to nuclear matter containing protons and neutrons, but instead to hadronic matter in which the dominant degrees of freedom are pions. This will turn out to be an appropriate approximation for the matter created in very relativistic heavy ion collisions, as will be the assumption that the resulting transition temperature is sufficiently low that the neglect of heavy species of mesons and baryons is justified. These assumptions, together with a standard result from statistical mechanics for the pressure of a massless gas of bosons at temperature T, $P(T) = \frac{\pi^2}{90} T^4$, allows one to calculate the pressure in the two phases:

- *Pion phase*: There are three species of pions (π^-, π^0, π^+), so the pressure in this phase is

$$P_\pi(T) = 3 \cdot \frac{\pi^2}{90} T^4 \tag{3}$$

- *QGP phase*: Both gluons and quarks contribute to the pressure. Quarks come in three colors, two spin states, and (by assumption) two "active" flavors. Both quarks and antiquarks contribute, leading to another factor of two, so the total number of contributing species is $3 \times 2 \times 2 \times 2 = 24$. These are fermionic degrees of freedom, which have a contribution to the pressure that is $\frac{7}{8}$ that for bosons which results from the difference in the integrals over the respective distribution functions; that is,

$$\int_0^\infty \frac{x^3\,dx}{e^x + 1} = \frac{7}{8} \int_0^\infty \frac{x^3\,dx}{e^x - 1}$$

There are eight varieties of gluons, each having two spin states, so that the total pressure in the QGP phase is

$$p_{QGP}(T) = \left\{ \frac{7}{8} \times 3 \times 2 \times 2 \times 2 + 8 \times 2 \right\} \cdot \frac{\pi^2}{90} T^4 - B \tag{4}$$

Nature will choose the phase with the larger pressure, which implies a transition at a temperature T_C given by $P_\pi(T_C) = P_{QGP}(T_C)$, which implies $T_C \simeq 0.72 \cdot B^{1/4}$. The corresponding energy densities ϵ_π and ϵ_{QGP} are

substantially different in the two phases. For the case of free pions, one has the standard result for a massless gas:

$$\epsilon_\pi = 3P_\pi \tag{5}$$

while for the quarks and gluons, the same kinetic relation applies, but proper treatment of the bag constant gives:

$$\epsilon_{QGP} = 3 \cdot \left\{ \frac{7}{8} \times 3 \times 2 \times 2 \times 2 + 8 \times 2 \right\} \cdot \frac{\pi^2}{90} T^4 + B \tag{6}$$

At the critical temperature, these results then give:

$$\epsilon_{QGP}(T_C) - \epsilon_\pi(T_C) = 4B \tag{7}$$

indicating that there is a large latent heat between the two phases. Numerically, $T_C \sim 0.14$ GeV and $\epsilon_{QGP}(T_C) \sim 0.85$ GeV/fm^3 (i.e., a few times the energy density of normal nuclear matter).

It must be emphasized that these results for the critical temperature and the nature of the phase transition are no more than qualitative guides. A complete treatment of the hadronic sector would both extend the calculation to all known hadrons and take into account the interactions between them. Similarly, the calculation in the QGP phase should allow for the interactions between quarks and gluons described by the QCD Lagrangian, and should in fact generate a confining term similar to the bag constant. To date, the only method of performing such a calculation is via *lattice QCD*, which is a numerical simulation of QCD on a discrete space–time lattice. Although calculations that consider only the gluon degrees of freedom are relatively straightforward, a full treatment that includes the quarks has become technically feasible only very recently, via massively parallel custom computers with approximately a teraflop of computing power. Even then, typical calculations remain limited to lattices of 16 or 24 points in each space dimension and a smaller number in the temporal dimension. Nonetheless, the systematic variation of the results with the number of lattice points and with the lattice spacing provides some confidence in the results. The world average of the most sophisticated calculations suggests $T_C \sim 0.15$–0.17 GeV and $\epsilon_{QGP} \sim 0.7 \pm 0.3$ GeV/fm^3, in surprising (qualitative) agreement with the simple estimated presented above.

A discussion of the phases of nuclear matter in the terms of temperature and energy density implicitly assumes an underlying thermal distribution. This is clearly not the case prior to the collision—each nucleus may be assumed to be in its ground state, with a large component of momentum directed along the collision axis. The collision transforms this *directed* energy into *randomized* energy. If (in the center of mass) the collision resulted in an isotropic distribution for the momenta of the produced particles, then it would be a relatively straightforward exercise to determine the energy density at the time these particles were produced: Measure the energy and momentum of all produced particles, trace them back to an assumed production volume (presumably the size of the original nucleus), then calculate the energy density by dividing the total energy observed by the production volume.

However, this approach fails for relativistic nuclear collisions. Both the dynamics of parton–parton scattering and the confinement mechanism strongly favor particle production with "limited" momentum transverse to the initial collision axis, so that the final-state distribution of particles is strongly peaked along the original direction of motion of the incident nuclei. It is precisely the relativistic nature of these collisions that leads to this distribution; in fact, an analysis using variables appropriate to this leads to the conclusion that these distributions are nonetheless *locally* approximately thermal.

If z is taken as the collision axis, it is useful to distinguish between *longitudinal* momentum $p_z = |\vec{p}| \cos\theta$ and the magnitude of *transverse* momentum $p_T = |\vec{p}| \sin\theta$ perpendicular to the collision axis, which is invariant under Lorentz boosts along the z axis. The *rapidity*, defined as:

$$y \equiv \frac{1}{2} \ln \left\{ \frac{E + p_Z}{E - p_Z} \right\} \tag{8}$$

is *additive* under Lorentz boosts along z, which implies that differences $y_1 - y_2$ in rapidity (and, consequently, differentials dy) are invariant between frames. As a result, a convenient method to describe particle production is in terms of *invariant cross-sections*, defined as:

$$E \frac{d^3\sigma}{d^3p} = \frac{d^3\sigma}{dy\, d^2p_T} \tag{9}$$

Closely related to this is the invariant number distribution:

$$\frac{d^3n}{dy\, d^2p_T} \equiv \frac{1}{\sigma} \cdot \frac{d^3\sigma}{dy\, d^2p_T} \tag{10}$$

which as defined is normalized to the average total multiplicity $\langle n \rangle$ produced in the collision

$$\int \frac{d^3n}{dy\, d^2p_T}\, dy\, d^2p_T = \langle n \rangle \tag{11}$$

The integral over just the transverse momentum gives the rapidity distribution:

$$\frac{dn}{dy} = \int \frac{d^3n}{dy\, d^2p_T}\, d^2p_T \tag{12}$$

Prior to the collision, this is of course localized to $\pm\, y_B$, where y_B is the rapidity of (one of) the beams. The value of dn/dy at $y = 0$ therefore measures the extent to which the initial momentum of the beams is converted into particles produced near the center-of-mass

of the collision. A related quantity is the *pseudo*-rapidity distribution $dn/d\eta$, where η is the pseudo-rapidity:

$$\eta \equiv \frac{1}{2} \ln \left\{ \frac{p + p_Z}{p - p_z} \right\} \tag{13}$$

While this variable has no simple Lorentz transformation properties, it has the virtue of (1) depending on only the polar angle of emission (the above definition reduces to $\eta = -\ln \tan \theta/2$) and (2) being approximately equal to rapidity for relativistic particles satisfying $E \gg p$. In particular, the exact result $dy = \beta d\eta$ implies that after integration over the transverse momentum spectrum,

$$\frac{dn}{dy} \approx \frac{1}{\langle \beta \rangle} \frac{dn}{d\eta} \tag{14}$$

where $\langle \beta \rangle$ is a suitable average velocity of the particles in the distribution. If it is known that $\langle \beta \rangle \approx 1$, or even if the value can be determined by other means, this has the useful experimental consequence that the angular distribution of particles can be related to a rapidity distribution, *without* determining the identities of each particle.

Two other quantities are of great utility in analyzing the properties of relativistic collisions. The first is the transverse mass, defined for a particle with mass m and transverse momentum p_T as $m_T \equiv \sqrt{p_T^2 + m^2}$. The transformation between (E, p_z) and (y, m_T) can then be written as:

$$E = m_T \cosh y, \qquad p_Z = m_T \sinh y \tag{15}$$

The transverse mass, p_T, y, etc. are all defined per particle. It is also useful to define E_T, the *total* transverse energy emitted into some region. Conceptually,

$$E_T^{(th)} \equiv \sum_i m_T^{(i)} \tag{16}$$

which is explicitly Lorentz invariant under boosts along the collision axis, while the experimentally measured form:

$$E_T^{(exp)} \equiv \sum_i E^{(i)} \sin \theta^{(i)} \tag{17}$$

is only approximately so. Nonetheless, the approximation is reasonably good for relativistic collisions, as most produced particles (dominantly pions) have mean transverse momenta significantly greater than their rest mass, so that $E^{(i)} \sin \theta^{(i)} \approx m_T^{(i)}$. The transverse energy density per unit rapdity dE_T/dy, therefore, is an invariant quantity which expresses the extent to which the initial directed energy of the colliding nuclei has been transformed to (presumably randomized) motion transverse to the collision axis.

With additional assumptions, the transverse energy distribution can also be related to the *initial energy density* ϵ immediately after the collision. In a nonrelativistic case, it would suffice to simply retrace each final-state particle to some assumed initial volume, and then compute the energy density by dividing the total energy of all particles by the initial volume. In the case of relativistic collisions, time-dilation effects and the anisotropy of final-state particles require a more sophisticated approach. The most commonly used method is due to Bjorken,[2] who modeled the collision geometry as a one-dimensional expansion of matter starting from a very dense initial state at $t = 0$, so that the location of any fluid element is given by $z = v_z t$. The expansion produces a spread in longitudinal velocities, so that a slab of width Δz contains a velocity spread satisfying $\Delta z = \Delta v_z t$. Recalling that the expansion is assumed to be one dimensional (along the collision axis) and that dimensions transverse to the collision axis are Lorentz invariant and noting that $\Delta v_z = \Delta y$, the volume contained in the slab of longitudinal width Δz and transverse radius R_T may be written as:

$$\Delta V = \pi R_T^2 \Delta z = \pi R_T^2 \Delta v_z \, t = \pi R_T^2 \Delta y \, \tau \tag{18}$$

where in the last step the time has also been expressed as the invariant quantity $\tau \equiv \sqrt{t^2 - z^2}$. The energy contained within this slab will ultimately produce a transverse energy ΔE_T, which provides the connection between an observable quantity and the energy density ϵ:

$$\Delta E_T = \epsilon \Delta V = \epsilon \pi R_T^2 \Delta y \, \tau \Rightarrow \epsilon = \frac{1}{\pi R_T^2} \frac{1}{\tau} \frac{dE_T}{dy} \tag{19}$$

With the assumption that R_T is the initial radius of the nucleus (plausible for "central" collisions in which the nuclei collide essentially head-on) and following measurement of the transverse energy distribution $\frac{dE_T}{dy}$, the determination of ϵ reduces to the selection of an appropriate time, τ. Typically a value $\tau \sim 1$ fm/c characteristic of the strong interactions is chosen, leading to values of the energy density in head-on nuclear collisions at the highest available energies of 3 to 5 GeV/fm^3, well in excess of that required for a phase transition to a new state of matter.

II. THE RELATIVISTIC HEAVY ION COLLIDER

Prior to the year 2000, all experiments with truly heavy ions were performed in *fixed-target* mode, in which a beam of ions is extracted from an accelerator and collided on a stationary target. Beams of gold ions with an energy per nucleon of 10.6 GeV were first produced by the Brookhaven Alternating Gradient Synchrotron (AGS) in 1992. A year later, operations began with lead beams at 158 GeV per nucleon at the CERN Super Proton Synchrotron (SPS). The corresponding center-of-mass

energies per nucleon $\sqrt{s_{NN}}$ obtained by the standard four-vector expression:

$$s = (p_{BEAM} + p_{TARGET})^2$$
$$= m_{BEAM}^2 + m_{TARGET}^2 + 2p_{BEAM} \cdot p_{TARGET} \quad (20)$$
$$\approx 2m_{TARGET}E_{BEAM} \quad (\text{for} \quad E_{BEAM} \gg m_{BEAM}) \quad (21)$$

correspond to $\sqrt{s_{NN}(AGS)} = 4.8$ GeV and $\sqrt{s_{NN}(SPS)} = 17.2$ GeV. It is also useful to characterize these energies in terms of the separation in rapidity between the target and the beam, which is 3.2 (AGS) and 5.8 (SPS). The empirical observation that single nucleon–nucleon collisions at these energies shift the rapidity distribution of the target and beam nucleons by 1 to 2 units of rapidity implies that the central region of rapidity at the AGS is baryon rich, and that a substantial baryon content could still be expected (and was, in fact, observed) at the CERN SPS.

In the summer of 2000, the Relativistic Heavy Ion Collider (RHIC) at Brookhaven began operations.[3] After a brief commissioning period at a lower energy, beams of 65 GeV per nucleon gold ions were brought into collision with counter-rotating beams of gold ions with the same energy. The collision energy of $\sqrt{s_{NN}} = 130$ GeV, corresponding to a rapidity interval of 11.3, greatly increases the energies and extends the domain over which heavy ion collisions are studied.

A. RHIC Parameters

RHIC is configured as two interlaced rings, each of circumference 3.83 km, which intersect each other at six equally spaced intervals (Fig. 1). The RHIC design permits injection, capture, acceleration, and storage of beams from protons to Au ions. This flexibility results from the injection complex, which consists of a tandem van de Graaff accelerator (for nuclear ions) and a linac (for protons), either of which injects particles first into a fast-ramping Booster synchrotron, followed by further acceleration in the Alternating Gradient Synchroton. Beams from tha AGS with energy of order 10 GeV per nucleon are injected into the RHIC ring, formed into bunches, accelerated to full energy, then stored for up to 10 hours. The design specification calls for a top energy of $\sqrt{s_{NN}} = 200$ GeV for Au ions, which implies a maximum energy for other ion species of $\sqrt{s_{NN}(Z, A)} = \frac{Z/79}{A/197} \cdot 200$ GeV. A particularly interesting special case is $Z = A = 1$, that is, protons for which RHIC's top energy is 500 GeV.

Particles are stored in bunches in each RHIC ring. The radio-frequency harmonic that shapes the bunches permits 360 "buckets" equally spaced around the ring. The initial operation of the machine filled every sixth such bucket to create 56 full bunches and a sequence of 4 successive empty bunches in each ring, which implies a time between bunch crossings of 212 ns. The sequence of 4 empty bunches is required for terminating the store

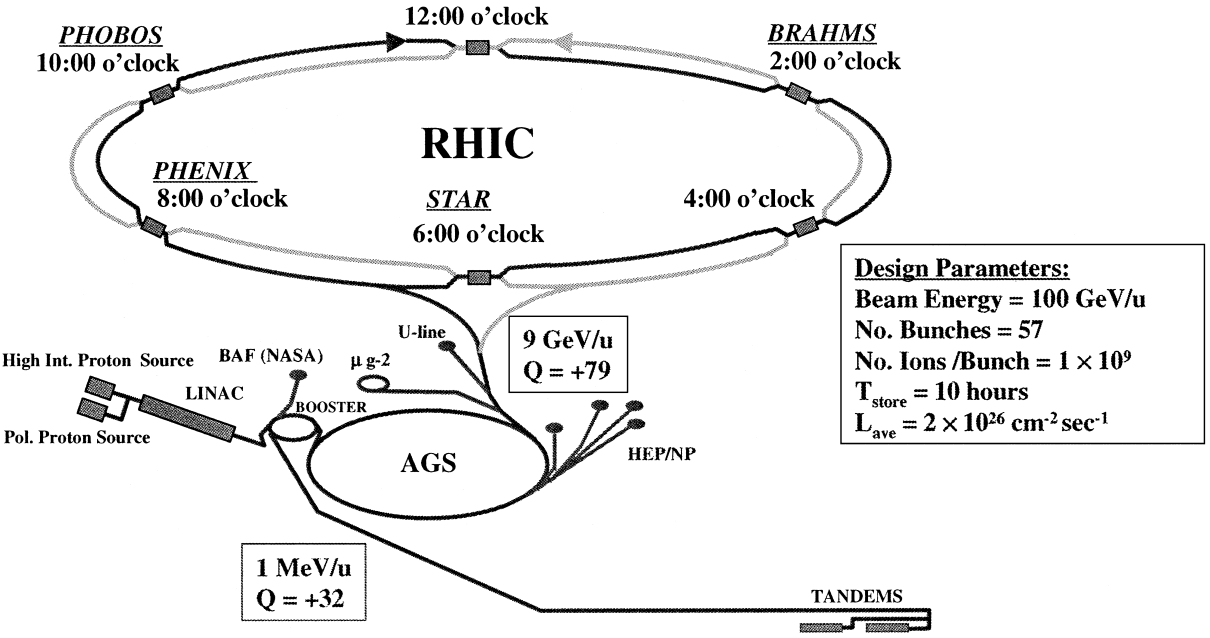


FIGURE 1 Configuration of Brookhaven National Laboratory's Relativistic Heavy Ion Collider (RHIC) and its injection complex and detectors.

in a controlled fashion. Future plans call for doubling the number of bunches and thus halving the crossing time. For gold ions, the design specification for RHIC calls for 10^9 ions per bunch, which produces a luminosity of $\mathcal{L}_{Au-Au} = 2 \times 10^{26}$ cm^{-2} s^{-1}. The total cross section (Coulomb + nuclear) for Au–Au interactions is 10.7 barns, so that the interaction rate in each intersection region is

$$\mathcal{R}_{Au-Au} = \mathcal{L}_{Au-Au} \cdot \sigma_{Au-Au} \sim 2 \text{ kHz} \qquad (22)$$

Four of the six intersection regions are instrumented with experiments, which are discussed in the following section.

An important feature of RHIC is the ability to collide a great variety of nuclear species, at a range of energies from (at least) the design value of 200 GeV per nucleon down to roughly 10% of this value. This flexibility provides a crucial ability to perform *in situ* variations of both the mass and energy of the colliding species. Previous experience with heavy ion collisions has shown that such systematic studies are essential for isolating truly new effects in nucleus–nucleus collisions from those that might reasonably be inferred from proton–proton or proton–nucleus collisions. By a happy accident of the properties of the injection complex, the increase in the number of ions per bunch for lower mass species compensates for the decreasing cross section, so that the equivalent proton-proton interaction rate,

$$\mathcal{R}_{\text{eq}} \equiv A^2 \mathcal{L}_{A-A} \cdot \sigma_{A-A} \qquad (23)$$

is roughly constant. This has the benefit that comparison studies with different species can be conducted with approximately equal amounts of running time, with the important caveat that the experiments must be designed to accomodate both a very broad range of interaction rates (∼1 kHz for Au–Au collisions to ∼10 MHz for p–p collisions) and a correspondingly large variation in the multiplicity of produced particles (∼10–100 for p–p up to ∼10^4 for Au–Au). For the case of proton–proton collisions, RHIC has the additional capability to maintain and manipulate the spin of the colliding protons, which will permit the study of the spin structure of the proton wave function, particularly the contributions of gluons and sea quarks to the proton spin. It is anticipated that most if not all proton–proton comparision running for RHIC will be conducted with polarized beams.

III. RHIC EXPERIMENTS

Each of the four experiments at RHIC has been designed to pursue specific signals emerging from RHIC collisions. Nonetheless, certain themes common to all four experiments include:

- *Event characterization*: A key parameter for each collision is the impact parameter—that is, the minimum distance between the centers of the nuclei. This can be determined on an event-by-event basis by relating the observed final state multiplicity to a model of the initial geometric overlap. Each RHIC experiment has both specific subsystems that provide this information as well as a common subsystem consisting of zero-degree calorimeters.
- *Sampling*: Each experiment samples a fraction of the collision products. No attempt is made to perform a total reconstruction of all particles emitted in a collision.
- *Hadron identification*: The great majority of particles produced in a relativistic heavy ion collision are hadrons, primarily pions, kaons, and nucleons. While there is a broad spectrum of transverse momentum, the average value of order 0.5 GeV/c is relatively modest, so particle identification may be achieved by standard techiques based on energy loss, time-of-flight, or Cerenkov radiation.
- *Multiple channels*: The complex task of relating observed particles back to fundamental properties of the quark–gluon plasma is best accomplished by correlated observations performed in several detection channels. Each RHIC detector, to varying extents, has been designed to have this capability.

As a result of these shared features, there is substantial overlap between the various experiments, which provides for a robust experimental program of complementary measurements.

A. BRAHMS

The BRAHMS experiment[4] is designed to provide quality identified hadron spectra measured over the broadest possible range in rapidity and transverse momentum. BRAHMS is configured as two small-aperture magnetic spectrometers, both using dipole magnetic fields together with charged particle tracking together with time-of-flight and Cerenkov particle identification techniques (see Fig. 2). A mid-rapidity spectrometer can be rotated to cover the region $-0.08 < \eta < 1.3$ in $\Delta\eta$ steps of roughly 0.3, while a forward spectrometer can be similarly deployed to sample a region spanning $1.3 < \eta < 3.9$. Tracking of charged particles is performed with a series of drift and time expansion chambers. Particle identification employs time-of-flight hodoscopes, segmented gas Cerenkov counters, and a ring-imaging Cherenkov (RICH) counter. Event characterization is achieved via a silicon multiplicity counter covering $|\eta| \leq 2.2$, while timing and trigger information is provided by beam–beam counters (BBCs) in the interval $3.2 \leq |\eta| \leq 4.3$.

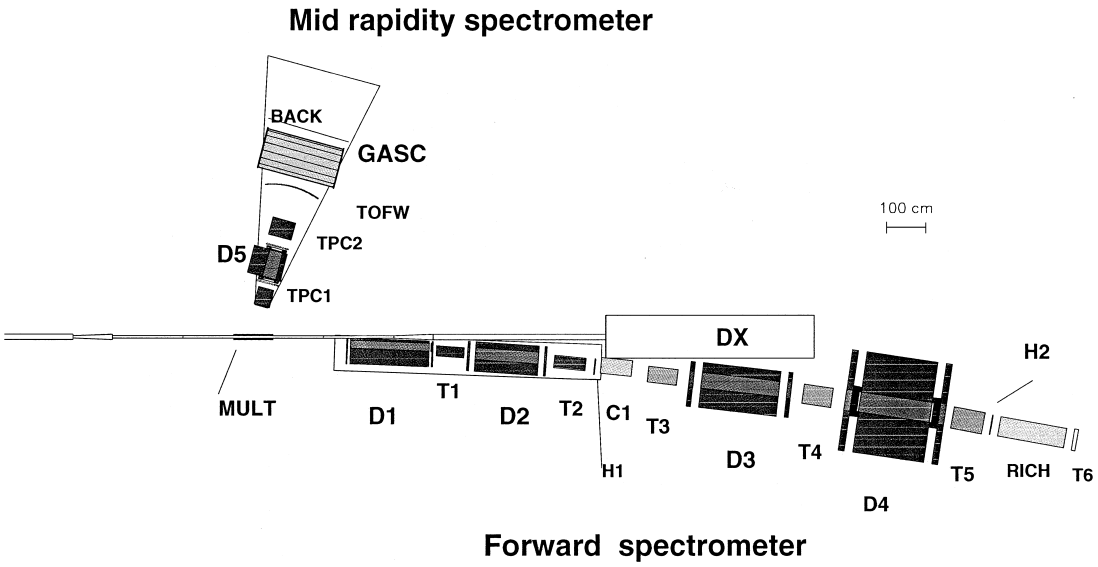


FIGURE 2 The BRAHMS experiment at RHIC.

A particular emphasis of the BRAHMS experiment is tracking the net baryon number in the collision. To do so, the results from the two small-aperture magnetic spectrometers are combined to provide inclusive spectra for $\pi^{\pm}$, $K^{\pm}$, and $p^{\pm}$ over the rapidity interval $0 < y < 3.8$ for $0.2 < p_T < 3$ GeV/c. The difference between the proton and antiproton rapidity distributions measured over this wide rapidity range allows BRAHMS to determine the net baryon number distribution over a very substantial interval. This coverage at RHIC is unique to the BRAHMS experiment, and information derived from the net baryon distribution measured in this interval will establish the mechanism by which ordered longitudinal momentum is transformed to disordered transverse energy ("stopping").

B. PHENIX

PHENIX[7] specializes in the detection of "penetrating probes," that is, real and virtual photons that emerge from the plasma without further interaction. Real photons are measured in two varieties of electromagnetic calorimeter (see Fig. 3). Because virtual photons of mass M decay in times of order $\hbar/M$ to "di-lepton pairs" consisting of e^+e^- or $\mu^+\mu^-$, they are detected in PHENIX via charged particle spectrometers which specialize in electron or muon measurements. The electrons and photons are measured in two central arms, each subtending $|\eta| < 0.35$ and $\Delta\Phi = 90^\circ$. Tracking of charged particles in the cental arms is accomplished via drift chambers, pixel-pad chambers, and a time expansion chamber (TEC), which together determine the momentum of charged particles via their curvature in an axial magnetic field. Electron identification is provided by a RICH counter, lead glass and lead-scintillator calorimeters, and energy loss in the TEC. As shown in Fig. 4, these systems taken together provide a rejection power in excess of 10^3 against the (much more numerous) pions over the momentum range 200 MeV/c $< p_T <$ 4 GeV/c. Charged hadrons are identified in the central arms via time-of-flight measurements in the two calorimeters and in a high-precision dedicated scintillator hodoscope covering a portion of one central arm. This provides unambiguous separation of pions, kaons, protons, and antiprotons over a broad momentum band. The start signal for these timing measurements is provided by BBCs consisting of arrays of 64 Cerenkov counters on either side of the interaction region.

Muons are detected in two dedicated spectrometers covering the full azimuth in the rapidity interval $1.1 < |y| < 2.4$. All particles from the collision vertex pass through ~50 cm of iron before entering the muon spectrometers. This greatly reduces the flux of hadrons (through their nuclear interactions) and, by destroying pions and kaons via such interaction, reduces the background of muons that result from the decay of these mesons. The primordial muons, which interact only electromagnetically and are hence much more penetrating, are momentum-analyzed via their curvature in an approximately radial magentic field, as measured in three stations of cathode-strip chambers. The muons are identified by their penetration into an iron absorber which is interwoven with five layers of Iarocci tubes.

Global event characterization is achieved by measurement of the total charged multiplicity measurements in the

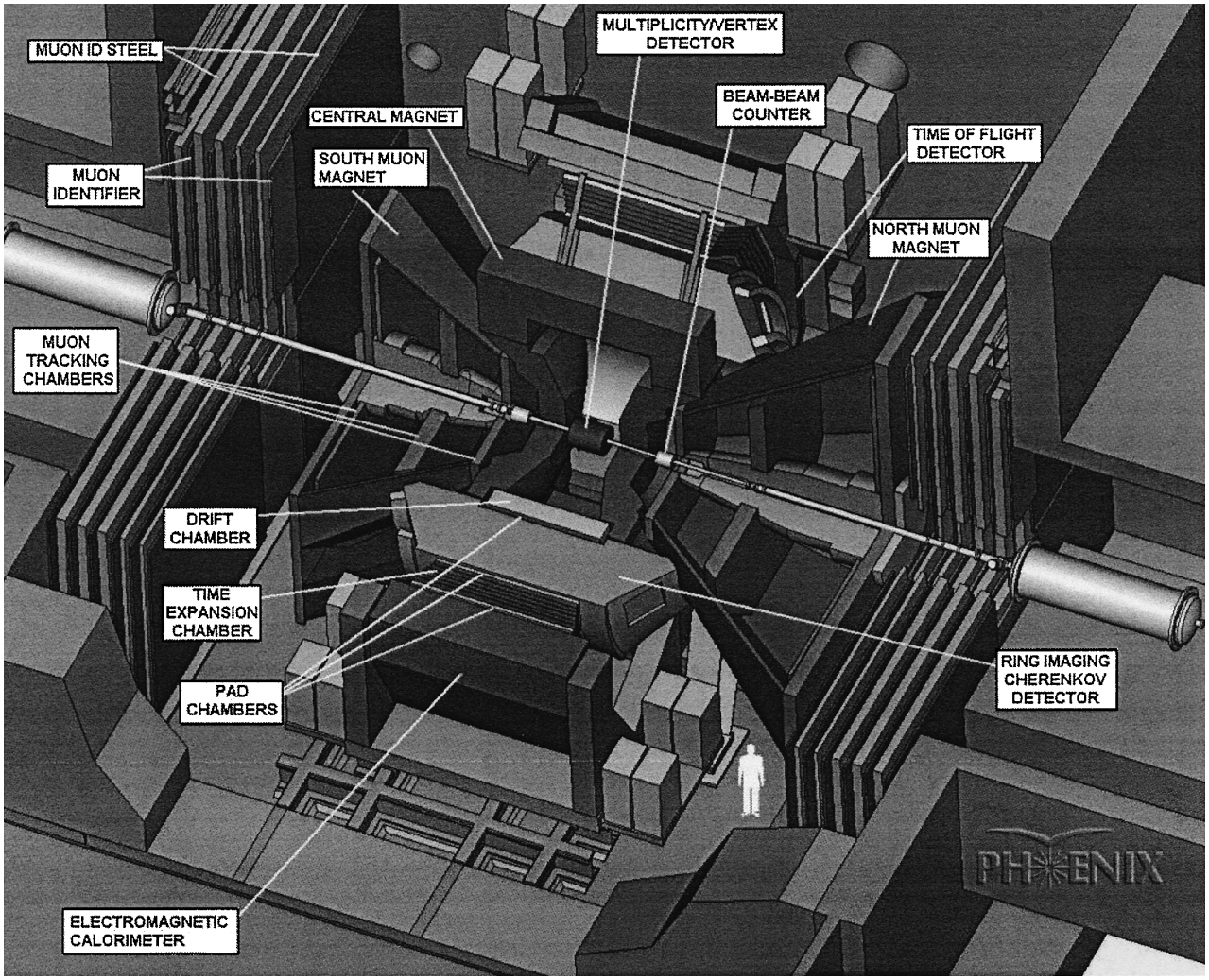


FIGURE 3 The PHENIX experiment at RHIC.

region $|\eta| < 2.5$ with silicon strips and pads. This is supplemented with Cerenkov-based BBCs covering the interval $3 < |\eta| < 4$, which also provide the interaction trigger and serve as precision start counters for time-of-flight measurements.

The unique geometry of PHENIX results from the conflicting design constraints from the electron arms, in which open geometry and low material in the active aperture are highly desirable, and the muon arms, in which removal of hadrons via heavy absorbers is the primary requirement. This optimization has resulted in a detector that is sensitive to nearly all particle species emitted from nuclear collisions. In particular, the measurement of electron and muon pairs is also the ideal method for detecting decays of vector mesons, ω, φ, and J/Ψ, in the central arms and J/Ψ, Ψ', and Υ, in the muon arms. The systematics of the production and subsequent suppression of these vector mesons are considered to be one of the principal discovery modes for detection of a QGP. The measurement of these rare objects requires in addition a high-bandwidth data-acquisition system in conjunction with a sophisticated trigger system, which in turn also provides for an extensive program of plasma diagnostics based on hard scattering of partons, detected via high-p_T single particles.

C. PHOBOS

The PHOBOS experiment[5] combines a broad coverage of the produced charged particles with a two-arm spectrometer specializing in the precision measurement of charged hadrons at low transverse momenta near midrapidity (see Fig. 5). The emphasis on low transverse momenta is a general one based on the uncertainty principle—spatial structure at large distances will manifest itself as

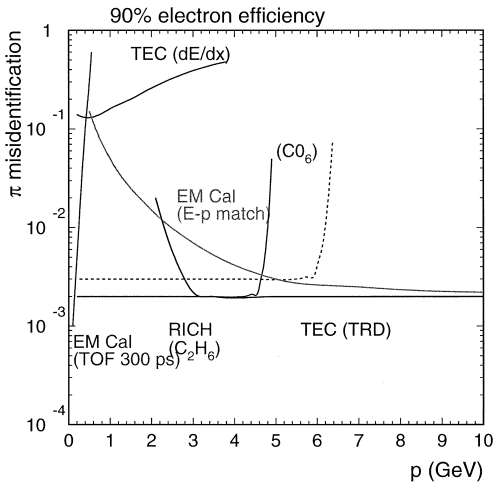


FIGURE 4 Pion rejection versus momentum in PHENIX.

enhancements at small momenta. The number of technologies in the PHOBOS subsystems has been minimized to essentially two: (1) silicon strips and pads for tracking, dE/dx measurements, and event characterization, and (2) a time-of-flight hodoscope for extended particle identification. The tracking detectors are arranged in two arms, each consisting of 15 planes of silicon strips or pads, with ∼1-mm × 1-mm pixels in the (high-density) front planes and ∼0.7-mm × 19-mm strips in the rear planes, leading to a total of 55K channels/arm. Event characterization is obtained via a silicon-strip array covering essentially the entire phase space of $|\eta| < 5.2$, which gives PHOBOS the broadest angular coverage of the RHIC experiments.

PHOBOS measures the spectra of low to moderate transverse momentum identified hadrons, with a special emphasis on two-particle correlations, ϕ physics, and the search for exotic fluctuations in the ratio between charged and neutral pions expected from the restoration of chiral symmetry. These signals will be correlated with the global event structure as measured by the very complete coverage of its multiplicity array, which also permits analysis of spatial anisotropies from flow and/or dynamical fluctuations. The data-acquisition system has been designed to give an unbiased look at a large number ($\sim 10^9$) of collisions in order to provide maximal sensitivity to unusual fluctuations due to new phenomena.

D. STAR

The STAR experiment[6] focuses on measurement of the final-state hadrons, measured over the largest possible aperture. The centerpiece of the experiment is a time projection chamber (TPC), designed to measure all charged hadrons emitted in the central two units of rapidity (see Fig. 6). The TPC consists of a cylindrical drift region (4-m length ×2-m radius) containing a mixture of 90% Ar

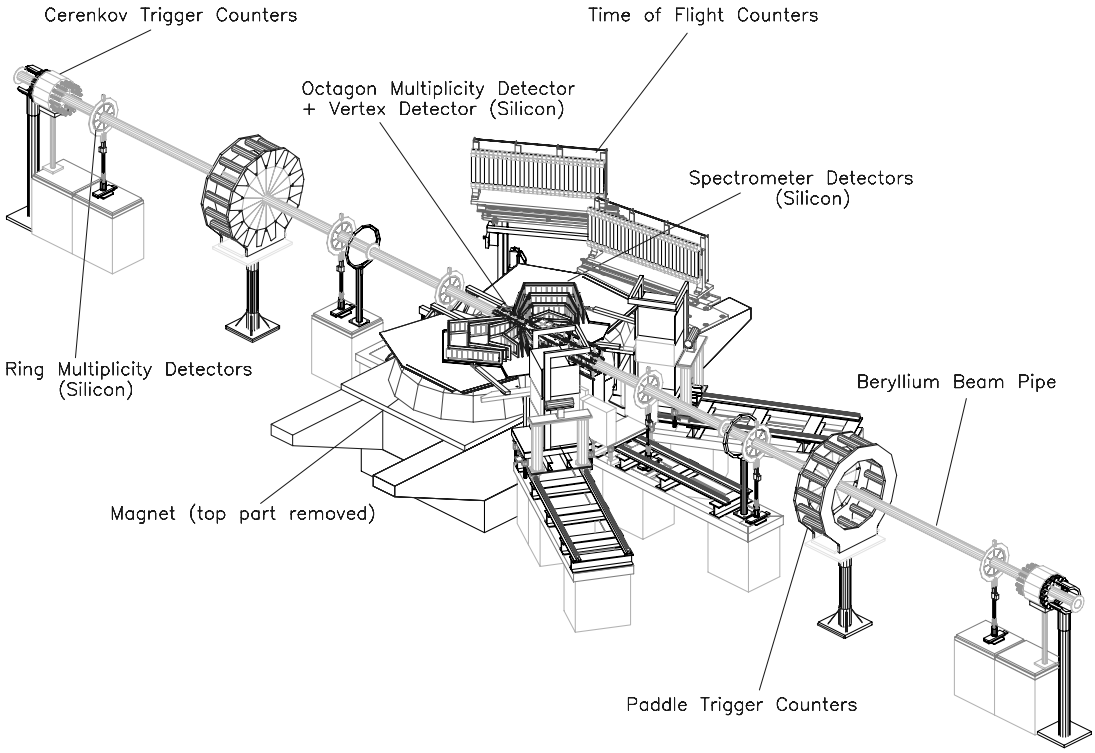


FIGURE 5 The PHOBOS experiment at RHIC.

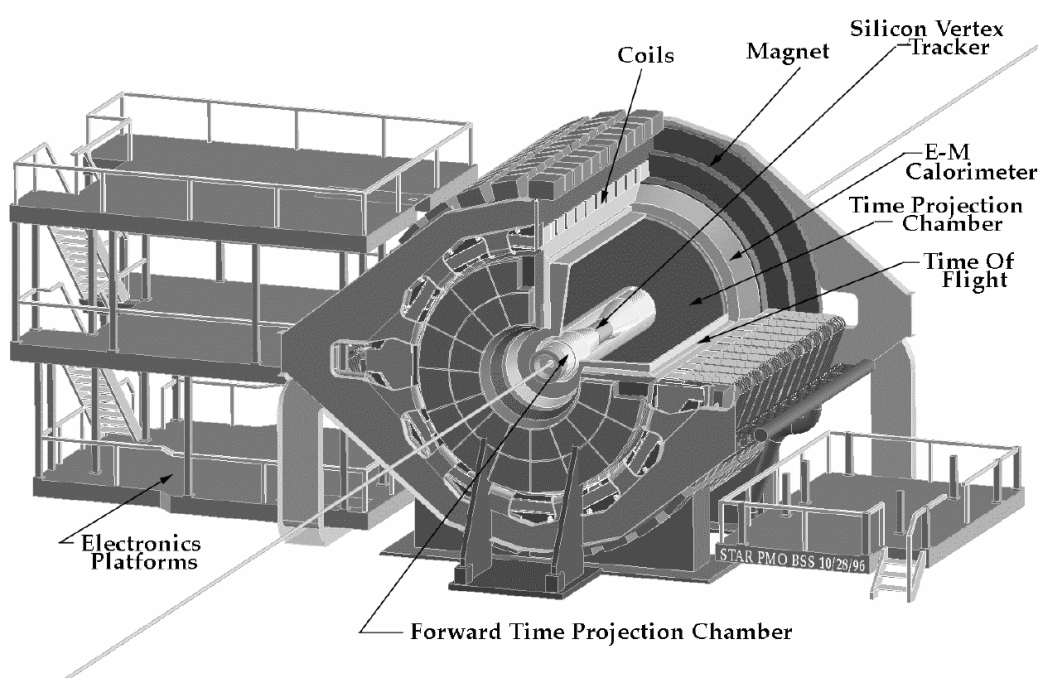


FIGURE 6 The STAR experiment at RHIC.

and 10% CH_4. Charged particles which traverse this region create ionization in the gas, which then drifts in a uniform electric field parallel to the beam axis to the endcaps of the cylinder, which are instrumented with 140,000 readout pads. Because the drift time is proportional to the distance along the beam axis, measurement of the drift time in addition to the deposited charge on each of these pads creates a three-dimensional digital image of the tracks in the event, as shown in Fig. 7.

The STAR TPC is located in a 0.5-Tesla solenoidal magnetic field. The curvature of each track in the magnetic field provides a measurement of its momentum. This information, when combined with the total ionization collected along the track, results in clear bands for each type of particle which permits identification of particles in the regions of nonoverlap evident in Fig. 8.

The TPC tracker in STAR is augmented by several other subsystems. Three layers of silicon drift and one layer of silicon strips provide an inner tracking system, significantly enhancing the range over which the decays of multiply strange hyperons may be identified. A trigger barrel surrounds the TPC, followed by an electromagnetic calorimeter of moderate $(\Delta\eta \times \Delta\Phi = .05 \times .05)$ segmentation which is located just inside the solenoid yoke. A small-aperture prototype RICH detector is also mounted on the inner yoke surface. Forward TPCs with radial drift extend the rapidity coverage for charged tracks to $|\eta| < 4.5$. An endcap calorimeter is currently under contstruction, with primary application to the spin program.

Correlations between the particles measured in STAR are used to determine both the size and shape of the emission region. This information, along with the inclusive momenta spectra and observed particle yields, provides a complete characterization of the final-state hadronic distribution. In addition, it is anticipated that the huge volume of information measured by STAR permits event-by-event determinations of quantities such as the temperature, chemical potential, and particle flow.

E. Zero Degree Calorimeters

The four RHIC experiments each have been instrumented with zero-degree calorimeters (ZDCs), located 19 m on either side of the collision region.[8] These devices are hadronic calorimeters located in the area where the two beam pipes begin to diverge after the collision region, so that all charged fragments are either transported in the beam pipe or swept further away from the ZDC aperture (see Fig. 9). As a result, the ZDCs measure the total energy of all neutrons emitted in a cone of half-angle 2 milliradians. By requiring a time coincidence between the two ZDCs on either side of the intersection region, experimenters can determine if a nuclear interaction occured in a given bunch crossing. The total cross section for Au–Au collisions to create such a coincidence is calculated to be

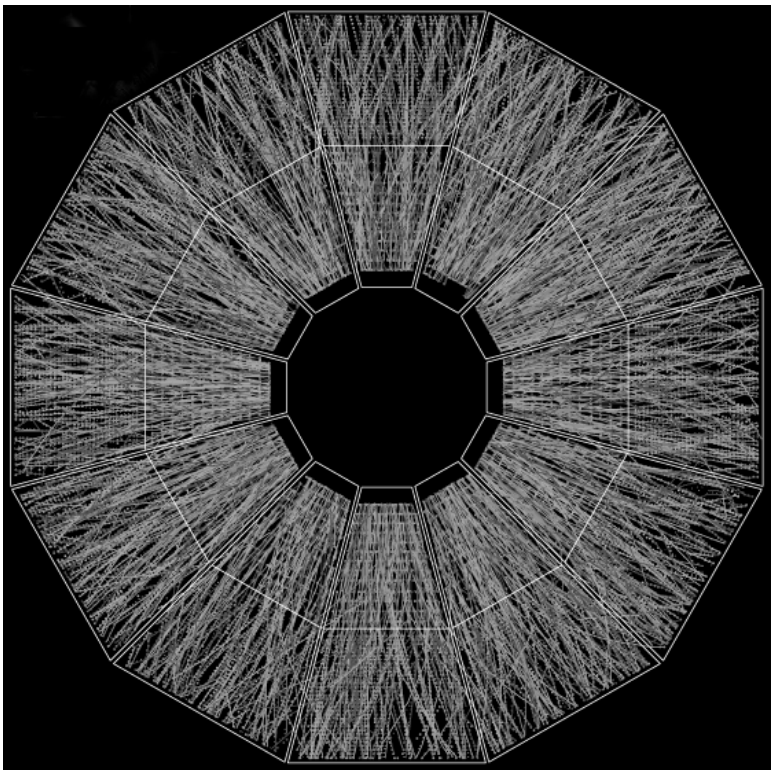


FIGURE 7 A central Au–Au event at RHIC captured in the STAR time projection chamber.

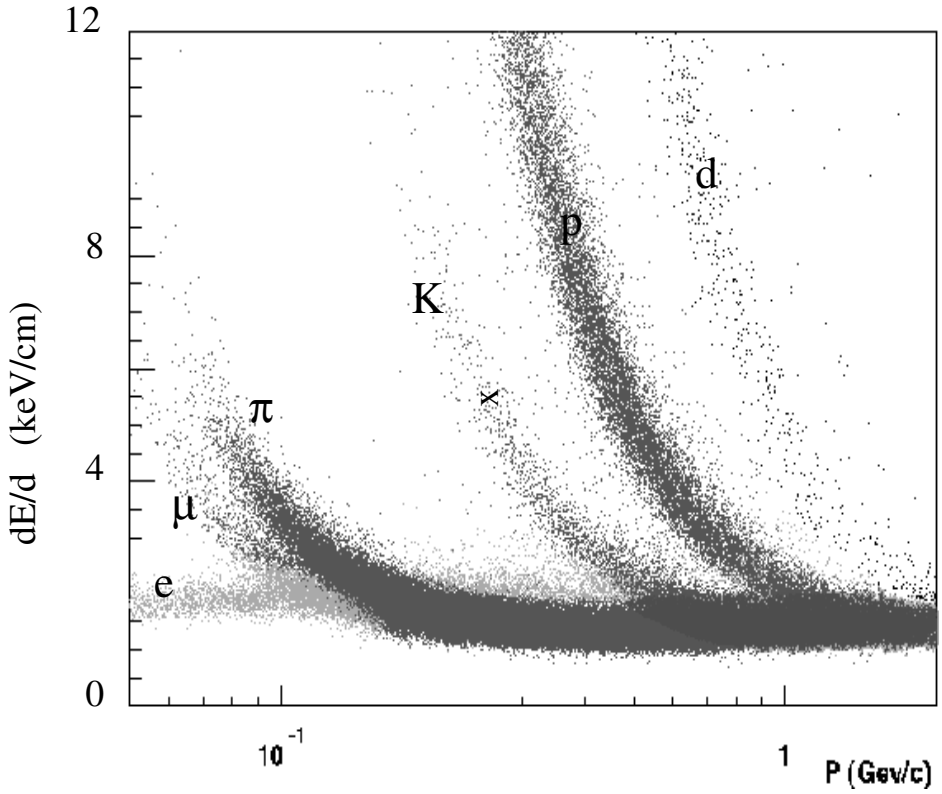


FIGURE 8 Bands of summed energy loss versus momentum for tracks measured in the STAR time projection chamber.

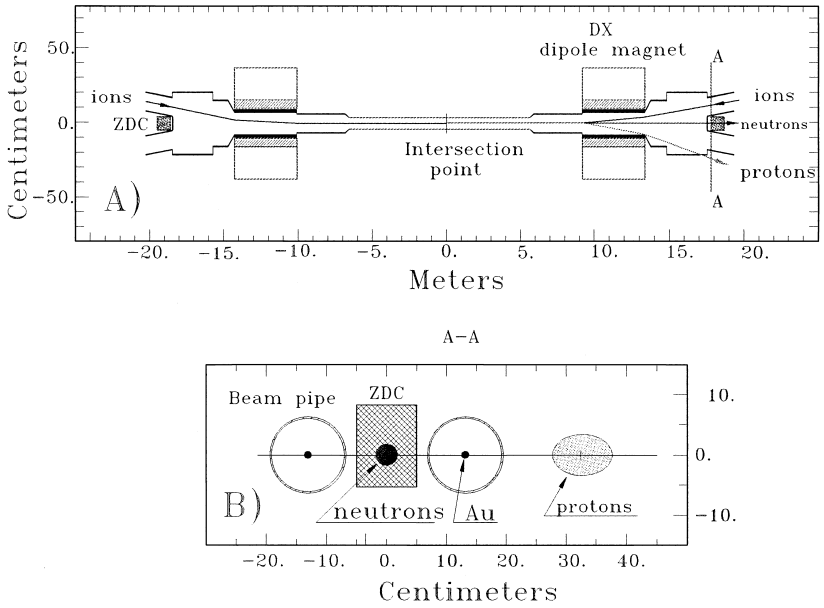


FIGURE 9 Layout of the zero-degree calorimeters (ZDCs) at a RHIC interaction region; (A) Plan view; (B) beam's-eye view showing the location of beam-rapidity neutrons, ions with the charge-to-mass ratio of the beam, and protons.

10.7 barns at $\sqrt{s_{NN}} = 130$ GeV,[9] which contains both the nuclear interaction cross section of 7.2 barns and a contribution from mutual Coulomb dissociation processes in which electromagnetic interactions between the two nuclei in a grazing collision result in the emission of at least one neutron from each nucleus. In addition, the precise pattern of energy deposition in the ZDCs can be used to infer the *impact parameter* of the collision—that is, the minimal distance between the centers of the two nuclei transverse to the beam axis. This ability is of crucial importance in separating trivial aspects of the initial nuclear collision geometry from actual dynamical effects.

IV. RHIC PHYSICS

During the initial physics run in the summer of 2000, RHIC reached approximately 10% of its design luminosity with Au ions at a per-nucleon energy of 65 GeV for each beam. Each experiment recorded several million Au–Au collisions under these conditions. The careful commissioning of the detectors and their extensive simulation prior to the running period led to a rapid analysis and publication of first results, which are discussed here.

A. Charged Multiplicity

As previously noted, the rapidity distribution of particles emitted in a heavy ion collision is a fundamental measure of the reaction process. For central collisions, where the impact parameter b is restricted to some value (typically $b < 2$ fm), the initial geometry is well controlled, as is the number of "participants," N_{part}, the total number of nucleons that participate in the collision. The maximum multiplicity observed under these conditions is then both of intrinsic interest and an important comparision to the corresponding multiplicity produced in proton–proton collisions at the same energy.

Very soon after the first collisions at RHIC, the PHOBOS collaboration provided a measurement addressing these issues.[10] Specifically, they determined the value of the charged particle pseudo-rapidity density $dn_{ch}/d\eta$ about $y = \eta = 0$ for central collisions of Au–Au nuclei at two different energies. The measured values of $dn_{ch}/d\eta_{|\eta|<1} = 408 \pm 12$ (stat) ± 30 (syst) at $\sqrt{s_{NN}} = 56$ GeV and $dn_{ch}/d\eta_{|\eta|<1} = 555 \pm 12$ (stat) ± 35 (stat) at $\sqrt{s_{NN}} = 130$ GeV were higher than any previous result from a nuclear collision, as expected for the new energy regime explored by RHIC. More importantly, the calculated value of the produced multiplicity per participating nucleon pair,

$$\frac{1}{N_{part}/2} \cdot \frac{dn_{ch}}{d\eta}$$

was significantly larger (see Fig. 10) than that observed in proton–antiproton collisions. This implies that RHIC has entered a new regime in which the previous empirical scaling derived at lower energies that $dn_{ch}/d\eta \sim N_{part}$ was no longer valid. This conclusion is supported by the increase in this quantity when studied as a function of $\sqrt{s}$, although an additional point at the RHIC top energy of $\sqrt{s_{NN}} = 200$ GeV is needed to firmly establish this result.

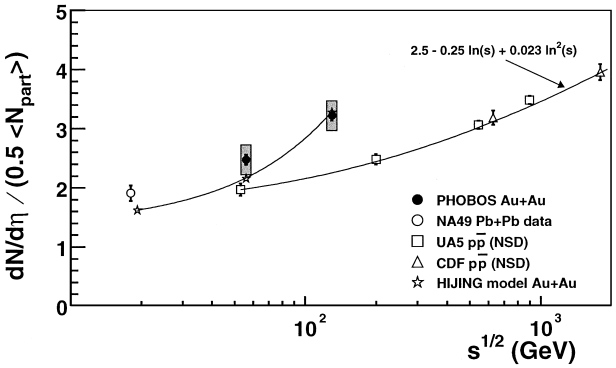


FIGURE 10 The excitation function meaured by the PHOBOS collaboration of the charged particle multiplicity per participating nucleon pair for Au–Au collisions at RHIC, in comparison to lower energy heavy ion collisions and proton–antiproton collisions.

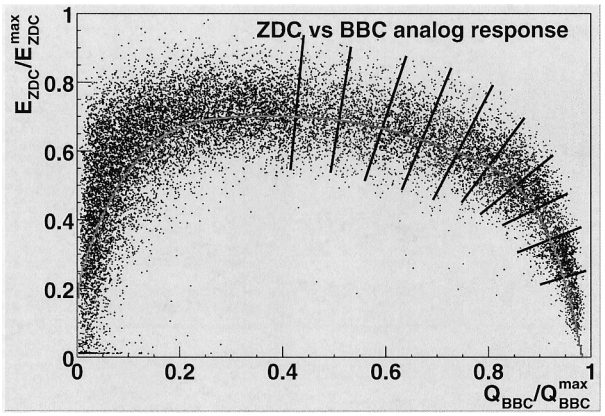


FIGURE 11 The relation between the energy deposited in the PHENIX beam–beam counters (BBCs) and the zero-degree calorimeter (ZDC) is shown together with intervals relating this information to the fraction of total cross section.

An important test of the nature of the increased yield observed by PHOBOS is provided by the study of the produced particle multiplicity versus the number of participants.[11] For instance, in models in which the increase results from the onset of hard processes (jets and/or mini-jets), the behavior is predicted to scale as:

$$\frac{dn_{ch}}{d\eta} = \frac{N_{part}}{2}\langle n\rangle_{soft} + N_{coll}\langle n\rangle_{hard}, \qquad (24)$$

where $\langle n\rangle_{soft}$ is the mean multiplicity per "soft" collision (technically, per participant pair undergoing soft interactions), and $\langle n\rangle_{hard}$ is the contribution from "hard" processes, defined as those with cross sections much smaller than the 40-mb nucleon–nucleon cross section for soft interactions. For processes governed by small cross sections, the yield is expected to scale as the number of binary collisions, N_{coll}, rather than the number of participating nucleons.

The PHENIX collaboration performed an analysis[12] aimed at determining the separate dependences on N_{part} and N_{coll}. Because neither quantity is directly measurable, extensive modeling was required to relate the observed multiplicities in the PHENIX BBCs and ZDCs to the nuclear overlap geometry. The result of that procedure is shown in Fig. 11, which demonstrates that a clear mapping is present between the fraction of the total cross section and the BBC–ZDC distribution. That relationship was then used to extract for each cross-section interval the number of participants and the number of collisions. The multiplicity was also determined bin-by-bin for the same intervals to obtain the result shown in Fig. 12. If only soft processes contributed to the multiplicity in RHIC collisions, the quantity plotted in Fig. 12 would be flat. Other models in which the gluon density saturates would also predict a horizontal line. The PHENIX data clearly rise in the region studied, with a slope consistent with models that include a contribution proportional to N_{coll}. This is a new feature of particle production in heavy ion collisions not observed at lower energies.

Closely related to these studies is the question of the energy density produced in RHIC collisions. The PHENIX collaboration has studied this,[13] using electromagnetic calorimeters located in the region $|\eta| < 0.35$. While in principle the calorimeters are optimized for response to electrons and photons, detailed test-beam measurements and simulations allow the authors to calibrate the response of the calorimters in terms of the total transverse energy E_T produced by all hadrons. This, in turn, provides an estimate of the energy density via the Bjorken formula of Eq. (19) of $\sim$5.0 GeV/fm^3. Far and away the dominant error in this estimate is the parameter τ, which is at best

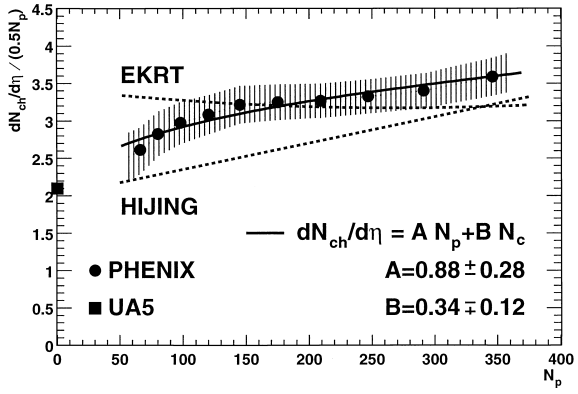


FIGURE 12 The pseudo-rapidity density per participant pair versus the number of participants as mesured by the PHENIX collaboration for $\sqrt{s_{NN}} = 130$ GeV Au–Au collisions.

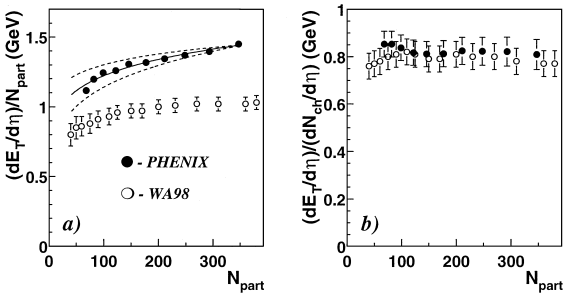


FIGURE 13 (Left) The transverse energy pseudo-rapidity density $dE_T/d\eta$ scaled by the number of participating nucleons, N_{part}, measured by the PHENIX collaboration in Au–Au collisions at $\sqrt{s_{NN}} = 130$ GeV. is compared to that measured by the WA98 collaboration in Pb–Pb collisions at $\sqrt{s_{NN}} = 17.2$ GeV. (Right) The ratio of transverse energy per charged praticle versus N_{part} measured by PHENIX and WA98.

known to a factor of two. However, the comparison with respect to previous experiments is not as sensitive to this quantity, so that it may be stated with some confidence that collisions at RHIC energies obtain energy densities at least 50% higher than those at the lower energies of the CERN SPS.[14] This ratio has a modest centrality dependence, as shown in Fig. 13a. Note the similarity of this figure to Fig. 12, which shows that the increase in transverse energy production with the number of participating nucleons closely tracks the corresponding increase in charged particle multiplicity. The constancy of the ratio, shown in Fig. 13b, and its near-equality to the value obtained at lower energies indicate that the increase in transverse energy density at RHIC energies is dominated by increased particle production, rather than an increase in the energy per particle.

B. Particle Flow

Non-central collisions of heavy ions produce an overlap region of participants that is strongly asymmetric. However, if the particles produced in each nucleon–nucleon collision do not re-interact with those from neighboring collisions, this spatial anisotropy would not be detectable in the final-state azimuthal pattern of particle emission. Conversely, the extent to which there is an emission pattern correlated with the orientation of the impact parameter is a measure of the degree of early thermalization in the collision and subsequent collectivity in the resulting expansion of the nuclear matter (see Fig. 14). The phenomenon of a momentum-space distribution correlated with the initial coordinate-space asymmetry is referred to as *elliptic flow*.

The STAR collaboration has analyzed elliptic flow in Au–Au collisions at RHIC.[15] The TPC is used to determine the azimuthal angle Ψ_R of the reaction plane, defined as the value that maximizes $\sum_i \cos 2(\phi_i - \Psi_R)$ for each

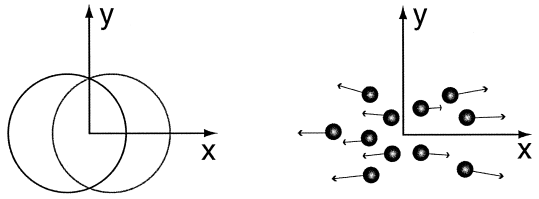


FIGURE 14 Schematic figure showing how initial-state coordinate anisotropy in the overlap of the two nuclei is translated into final-state momentum ansisotropy. The collision axis is out of the page.

event, where ϕ_i is the azimuthal angle of the ith track and the sum runs over an appropriate subset of particles in the event. The distribution of particles with respect to this reaction plane for $|\eta| < 1.3$ is then parameterized as:

$$\frac{dN}{dy\,d^2p_T d\phi} \sim 1 + 2v_2(p_T)\cos 2(\phi - \Psi_R) \qquad (25)$$

Care is taken to exclude autocorrelation effects by selecting particles from different regions, charge states, and samples from those used to determine the reaction plane.

The strength of the elliptic flow observed by STAR is parameterized by the coefficient $v_2(p_T)$. As shown in Fig. 15, the elliptic flow is strongly dependent on the centrality of the collisions (as determined by the total charged multiplicity). Moreover, the final state momentum anisotropy v_2 is roughly proportional to the inferred initial state eccentricity $\epsilon \equiv \langle x^2 - y^2\rangle/\langle x^2 + y^2\rangle$, with $v_2 \sim 0.2\epsilon$. Both the value of v_2 and its dependence on transverse momentum (not shown here) are significantly larger than at lower energies, again demonstrating that RHIC collisions develop a degree of collectivity previously unobserved.

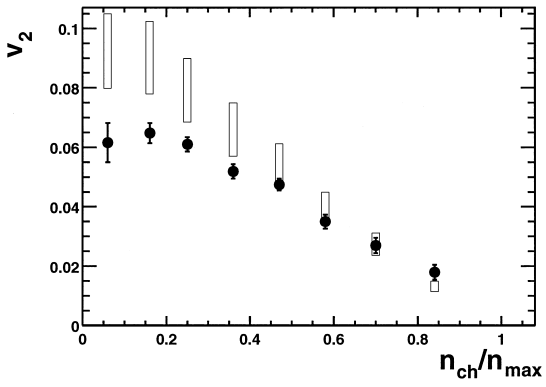


FIGURE 15 Results from the STAR collaboration on the elliptic flow parameter v_2 versus the multiplicity in the event. The multiplicity is scaled by the value observed in the most central events. The vertical boxes represent the corresponding spatial eccentricity as inferred from the centrality bin, scaled by 0.19 (bottom) to 0.25 (top) for each interval.

C. Baryons

As noted above, the central rapidity region at RHIC is expected to be substantially free of net baryon number. An observable directly related to this is the ratio of antiprotons to protons in a region near $y = 0$. While this does not directly measure the net baryon yield, it does have the advantage that many experimental systematic effects cancel in the ratio.

The STAR collaboration has reported a first measurement[16] of the ratio of antiproton ($\bar{p}$) to proton (Fig. 16). The particles are separated from pions and kaons by analyzing the measured energy loss dE/dx in the TPC gas as a function of momentum (see Fig. 8). In the interval $|y| < 0.5$ and $0.4 < p_T < 1.0$ GeV/c they find for minimum bias Au–Au collisions at $\sqrt{s_{NN}} = 130$ GeV $\bar{p}/p = 0.65 \pm 0.03$ (*stat*) $\pm$ 0.07. This value is observed to have no strong dependence on rapidity or transverse momentum and to decrease only very slightly with increasing centrality. On the other hand, comparison to data from measurements at lower energies shows that the $\bar{p}/p$ ratio is strongly dependent on the collision energy. The values found at RHIC are the highest yet seen in heavy ion collisions and indicate that these collisions are producing a central region in which the net baryon number is rather low.

This observation may be put on a somewhat more quantitative analysis via simple statistical considerations. Assume that chemical equilibrium has been established, so that particle abundances are governed by Boltzmann factors combined with the chemical potential μ. Further assume that the chemical potentials for the valence up and down quarks are equal $\mu_d = \mu_u \equiv \mu_v$, and that the density of strange quarks remains significantly lower than up and down quarks, so that $\mu_s \approx 0$. Then, noting that chemical potential for protons is given by $\mu(p) = \mu_u + \mu_u + \mu_d = 3\mu_v$ and that chemical equilibrium

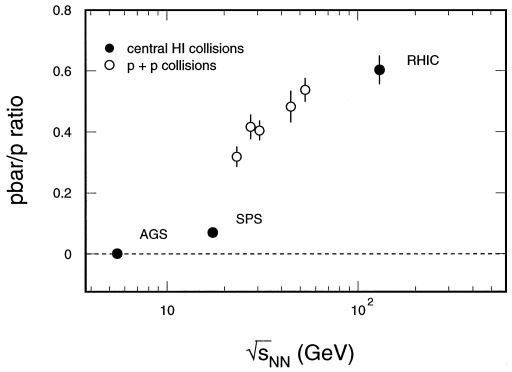


FIGURE 16 The ratio of antiprotons to protons measured by the STAR collaboration at RHIC, compared to measured values in lower energy heavy ion collisions.

implies that $\mu(\bar{p}) = -\mu(p)$, the ratio of antiprotons to protons then is given simply by:

$$\frac{\bar{p}}{p} = \frac{e^{-(E(\bar{p})-\mu(\bar{p})/T}}{e^{-(E(p)-\mu(p)/T}} = \frac{e^{-3\mu_v/T}}{e^{+3\mu_v/T}} \equiv \frac{1}{Z_v^6} \qquad (26)$$

where in the last expression we have defined the quark fugacity $Z_v \equiv e^{\mu_v/T}$. For baryons containing strange quarks such as $\Lambda = (uds)$ and $\Xi = (uss)$, this same accounting then predicts:

$$\frac{\bar{\Lambda}}{\Lambda} = \frac{1}{Z_v^4} = 0.75, \quad \frac{\bar{\Xi}}{\Xi} = \frac{1}{Z_v^2} = 0.87 = \frac{\mathrm{K}^-}{\mathrm{K}^+} \qquad (27)$$

both of which are consistent with preliminary results reported by the STAR collaboration. Such considerations argue that the final-state hadronic matter has reached the low net baryon density expected on the qualitative arguments based on the increased rapidity separation between the colliding nuclei at RHIC. The experimental confirmation of this picture has been provided by the BRAHMS collaboration,[17] which has provided essential data in the rapidity regions away from $y = 0$. They find $\bar{p}/p = 0.59 \pm 0.05$ at $y \approx 0.7$ and 0.44 ± 0.04 at $y \approx 2$, indicating that the central 2 to 3 units of rapidity at RHIC have markedly lower net baryon density than found in previous experiments at lower energies.

V. FUTURE DIRECTIONS

The first run of RHIC has resulted in a very significant advance in studies of dense nuclear matter. The previous section described only those results that have been accepted in the refereed literature at the time this article was prepared. Preliminary data from the four RHIC collaborations indicate that many more topics are under active investigation and will soon be available.[18] Forthcoming runs at RHIC will increase the existing datasets by a factor of 100 to 1000, will reach higher ($\sqrt{s_{NN}} = 200$ GeV) energies, and will explore the systematic variation of all phenomena with both beam energy and mass of the colliding species. To fully exploit this rich variety of capabilities will require a decade or more of experimental investigation.

The energy frontier in heavy ion physics does not stop at RHIC. The Large Hadron Collider (LHC),[19] scheduled for completion in 2006/7 at CERN, is planned to operate with nuclear beams for several weeks per year. The design energy of $\sqrt{s} = 14$ TeV for proton–proton collisions provides $\sqrt{s_{NN}} = \frac{Z}{A} \cdot 14$ TeV for ion–ion collisions, which is 5.5 TeV for the case of Pb–Pb collisions. This is more than 25 times the energy available at RHIC and opens a new regime of heavy ion physics in which the physics is completely dominated by hard processes There is one dedicated experiment, ALICE,[20] for the heavy ion

program that combines most of the features of the STAR and PHENIX detectors. The CMS experiment,[21] which is designed for proton–proton physics at LHC, will also operate during heavy ion running and will focus on those observables accessible through di-muon decay channels. The prospects of the focused LHC program, together with the broad RHIC physics capabilities, promise many years of studies of dense hadronic matter and of new discoveries.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC AND MOLECULAR COLLISIONS • COLLIDER DETECTORS FOR MULTI-TEV PARTICLES • HEAVY IONS (HIGH ENERGY PHYSICS) • QUANTUM CHROMODYNAMICS • X-RAY, SYNCHROTRON RADIATION, AND NEUTRON DIFFRACTION

BIBLIOGRAPHY

Groom, D. E., *et al.* (2000). "Review of particle physics," *Eur. Phys. J.* **C15,** 1.

Bjorken, J. D. (1983). *Phys. Rev. D* **27,** 140.

General information on RHIC is available from http://www.rhic.bnl.gov/.

http://www.rhic.bnl.gov/brahms/WWW/brahms.html.

http://phobos-srv.chm.bnl.gov/.

http://www.star.bnl.gov/STAR/.

http://www.phenix.bnl.gov/.

White, S. N. (1998). *Nucl. Instrum. Meth.* **A409,** 618.

Baltz, A. J., Chasman, C., and White, S. N. (1998). *Nucl. Instrum. Meth.* **A417,** 1.

Back, B. B., *et al.* (PHOBOS collaboration) (2000). *Phys. Rev. Lett.* **85,** 3100.

Wang, X., and Gyulassy, M. (2001). *Phys. Rev. Lett.* **86**, 3496.

Adcox, K., *et al.* (PHENIX collaboration) (2001). *Phys. Rev. Lett.* **86,** 3500.

Adcox, K., *et al.* (2001). *Phys. Rev. Lett.* **87**, 052301.

Aggarwal, M. M., *et al.* (2001). *Eur. Phys. J.* **C18,** 651.

Ackermann, K. H., *et al.* (2001). *Phys. Rev. Lett.* **86,** 402.

Ackermann, K. H., *et al.* (2001). *Phys. Rev. Lett.* (submitted).

Bearden, I. G., *et al.* (2001). *Phys. Rev. Lett.* '(submitted), nucl-ex/0106011.'

For example, see "Proceedings of Quark Matter 2001," to appear in *Nuclear Physics A*.

http://www.cern.ch/LHC/.

http://www.cern.ch/ALICE/.

http://cmsinfo.cern.ch/Welcome.html.



Plasma Science and Engineering

J. L. Shohet
University of Wisconsin

I. Introduction
II. Basic Plasma Properties
III. Plasma Physics
IV. Plasma Diagnostics
V. Plasma–Surface Interactions
VI. Conclusion

GLOSSARY

Attachment Phenomenon in which a neutral particle and an electron combine, producing a negatively charged ion.

De-excitation Inverse of excitation. Radiation is usually emitted.

Diffusion Phenomenon in which particles diffuse in position space, velocity space, and time.

Discharge machining The use of plasmas to provide a cutting surface between a thin wire and the work to be cut, usually by passing an arc between them through water.

Excitation Process in which neutral particles or ions (that are not fully stripped) gain energy, which is evident by orbital electrons moving to higher energy states.

Fusion plasmas Plasmas with relatively high temperatures composed of light atoms, particularly hydrogen or its isotopes.

Industrial plasmas Plasmas whose composition is generally of masses above that of hydrogen (they can have molecular weights of several thousand), and are usually of two types: thermal (equilibrium) plasmas, in which the electron and ion temperatures are approximately equal, and nonequilibrium (glow-discharge) plasmas, which tend to have relatively high electron temperatures compared to the ion temperatures. Many chemical reactions can only take place in the plasma state.

Ionization Inverse of recombination.

Ion milling The use of ions to cut or "mill" narrow regions of materials to great accuracy.

Plasma Collection of charged particles, usually of opposite sign, that tends to be electrically neutral. Often referred to as the fourth state of matter. Adding energy to a solid melts it and it becomes a liquid; adding energy to a liquid boils it and it becomes a gas; adding energy to a gas ionizes it and it becomes a plasma.

Plasma-assisted chemical vapor deposition (PACVD) The use of plasmas to deposit various chemicals on surfaces, either by treating the surface before deposition, or by providing a chemical pathway for successful deposition.

Plasma electronics Devices that exploit plasmas directly such as arc melters, microwave sources, switchgear, plasma displays, welders, analytical instrumentation, arc lamps, and laser tubes.

Plasma etching Process in which plasmas are used to etch materials: chemical reactions occuring on surfaces create a volatile compound from the surface, material, and the plasma, which can then be pumped away.

Plasma polymerization Production of polymers by ionizing a monomer gas, which can be deposited as coatings on various materials. This process often occurs during etching, since by changing the ratio of the various plasma components, etching can be turned into polymerization.

Plasma processing The use of plasmas or particle beams (charged or neutral) to alter an existing material, as in plasma etching, ion milling, ion implantation, or surface modification through plasma cleaning, hardening, or nitriding.

Plasma spray Coating process that sputters heavy particles (clumps) from an arc system and then directs the spray of these particles to a surface for coating.

Plasma synthesis The use of plasmas to drive or assist chemical reactions to synthesize compounds, alloys, polymers, or other complex species starting from simpler starting materials.

Recombination Phenomenon in which ions and electrons recombine to form neutral particles; radiation is sometimes emitted.

Sheath Generally a region near a surface in which the plasma is not electrically neutral. Some sheaths can form in the main body of the plasma and are called double layers.

Sputter deposition The use of plasmas to sputter or knock off from a target electrode particles that are then deposited on a particular material.

Surface modification The use of plasmas to modify the properties of materials by interacting on the surface of those materials in various ways. For example, tool steel can be hardened considerably by subjecting the tools to a nitrogen plasma. Turbine blades can be plasma coated for improved mechanical and thermal properties.

PLASMAS are composed of mixtures of electrons and positive and/or negatively charged ions as well as neutral particles. They are affected by electric and magnetic fields, which can be used to modify their properties. The temperature of such plasmas can be quite high. As a result, many interactions between particles are substantially different when they are in the plasma state. Thus, new materials can be manufactured that can have improved properties, new chemical compounds may be produced, and surfaces of existing materials can be altered. These aspects will have an increasingly significant role in technology.

I. INTRODUCTION

Industrial use of plasmas has applications that cover a broad range of activities and has a multibillion dollar yearly impact in the economy. We first describe what a plasma is, how it behaves under the influence of electric and magnetic fields, and how it is characterized. Normally, we specify the following quantities when we describe a plasma: composition, electron and ion temperatures, and electron and ion densities. We divide plasmas into two general types as follows.

1. *Industrial plasmas*. The ions in these plasmas are generally composed of masses above that of hydrogen (the molecular weights can reach several thousand), and are usually of two types: thermal (equilibrium) plasmas, in which the electron and ion temperatures are approximately equal, and nonequilibrium (glow-discharge) plasmas, which tend to have relatively high electron temperatures compared to the ion temperatures.
2. *Fusion plasmas*. These plasmas have much higher temperatures than industrial plasmas and are composed of light atoms, particularly hydrogen or its isotopes, and are designed to produce energy by means of a thermonuclear reaction.

Figure 1 indicates some general groupings of industrial plasmas according to their application and how they compare with fusion plasmas. One axis of the graph is proportional to temperature and the other to density.

At present, industrial plasma applications are largely empirical in nature. Further progress will require a much more thorough understanding of plasma behavior as well as of the interaction of the plasma with solid materials. Design tools, transportable diagnostics, and models are needed.

We characterize industrial applications into three broad and somewhat overlapping areas:

1. *Plasma processing*, which encompasses applications in which plasmas or particle beams (charged or neutral) are used to alter an existing material, as in plasma etching, ion milling, ion implantation, or surface modification through plasma cleaning, hardening, or nitriding.

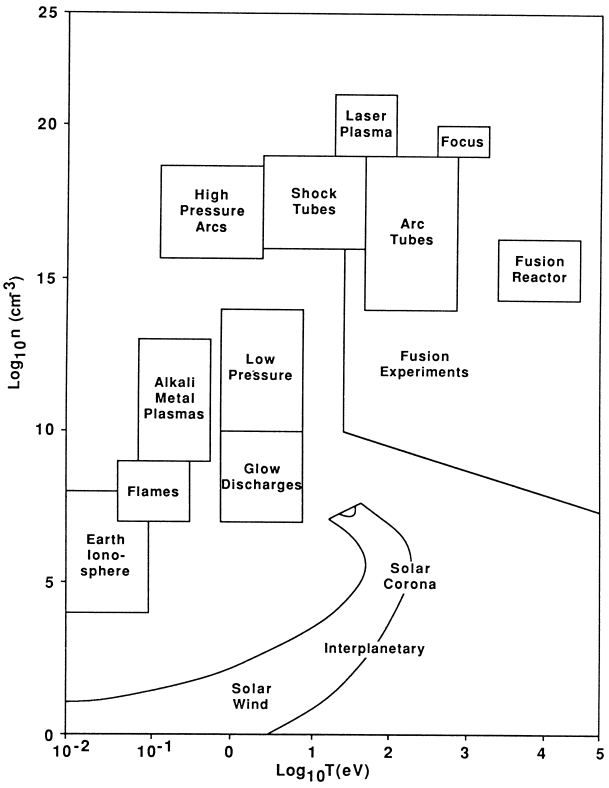


FIGURE 1 Types of plasmas.

2. *Plasma synthesis*, which refers to applications in which plasmas are used to drive or assist chemical reactions to synthesize compounds, alloys, polymers, or other complex species starting from simpler starting materials. This could also include the inverse processes of plasma decomposition. Many chemicals and/or chemical reactions can only exist or take place in the plasma state.
3. *Plasma electronics*, which includes applications in which the unique properties of plasmas are used directly in devices such as arc melters, microwave sources, switchgear, plasma displays, welders, analytical instrumentation, arc lamps, or laser tubes.

There are four common fundamental requirements needed for progress in industrial applications of plasmas:

1. Theory, modeling, and systems concepts
2. Plasma chemistry and interactions
3. Plasma diagnostics and characterization
4. Plasma generation and confinement

It is the purpose of this article to provide a general introduction to plasma science and engineering in order to show both the applications and how advances in this field can be made. The following short description of industrial applications of plasma processing and technology shows the widespread impact of plasma technology.

Plasma polymerization. By ionizing a monomer gas, certain types of polymers can be made which can be deposited as coatings on various materials. There is an important application of this work in the biotechnology field since biocompatible polymers can be used to coat various implant materials that would otherwise be rejected by the body. Various pharmaceuticals and other "exotic" chemicals can only be made with this process, which is often a result of the combination of ion and free radical generation by the plasma.

Plasma-assisted chemical vapor deposition (CVD). Here, plasmas are used to provide a mechanism to deposit various chemicals on surfaces either by treating the surface before deposition or providing a chemical pathway for successful deposition.

Sputter deposition. In this case, plasmas are used to sputter particles from a target electrode that are then deposited on a particular material.

Plasma etching. The major application of this technique is in the semiconductor industry, As the spacing between lines in integrated circuits shrinks to 0.1 μm and below, conventional "wet" etching using chemicals begins to fail because such processing acts in a spherical direction and undercuts the walls between the etch regions. Appropriately designed plasma etching (dry etching), perhaps combined with electric and magnetic fields or ion beams, offers a dramatic improvement in the etch process, and it is believed that the future of the semiconductor fabrication industry will continue to rest with plasma processing for a long time to come.

Ion milling. Beams of ions can be used to cut or "mill" narrow regions of materials to great accuracy.

Surface modification. Plasmas can be used to modify the properties of materials by interacting on the surface of those materials in several ways. For example, tool steel can be hardened considerably by subjecting the tools to a nitrogen plasma. Turbine blades can be plasma coated for improved mechanical and thermal properties.

Welding. The use of plasmas in welding, especially in arc welding, has been known for some time. However, many problems exist with welding, and largely due to the lack of understanding of the plasma composition, the plasma temperature and density, and the electric field and current distribution in the welding are.

Discharge machining. In this process, plasmas are used to provide a cutting surface between a thin wire and the work to be cut, usually by passing an arc between them through water. This process is used, for example, to cut a magnet coil that is required to be in a particular twisted shape. The tooling and resulting magnet coils are shown in Figs. 2 and 3.

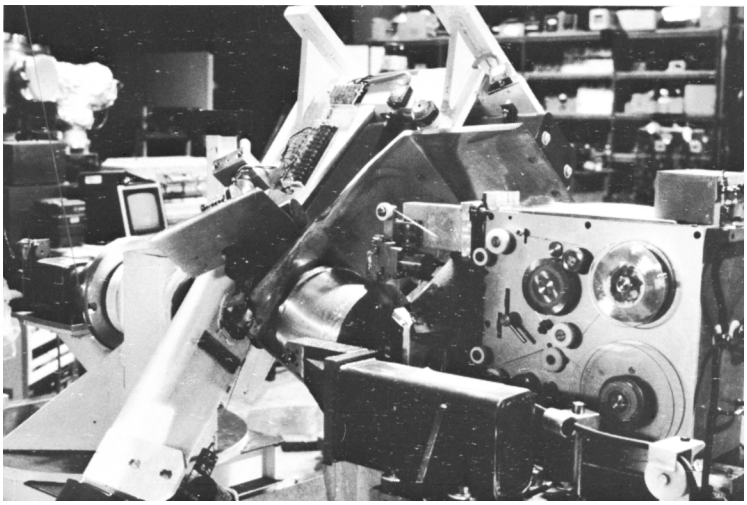


FIGURE 2 Electrical discharge machining of a magnet coil.

Plasma displays. Many new uses for such displays are being found. A portable computer whose screen is a full plasma display has recently been introduced.

Arc devices. A major component of American industry has involved the use of arc technology in the electrical power system field. Switchgear today is still being designed empirically without the understanding of plasma–surface interactions. The U.S. switchgear industry is suffering greatly from foreign competition as a result. Area illumination is a major application of this technology as well, with activity above $20 billion per year.

Arc melting. Arc furnaces have been in use for many years, with many applications to the refining and extraction of ores. Yet there can be major improvements based on the proper understanding of the interaction of plasmas with the ores and metals. For example, in the melting of iron ore, 20 lb of graphite electrode is used up per ton of ore. If just 1 lb of graphite were saved, a savings of $20 million per year would result.

Plasma spray. This is a coating process that sputters heavy particles (clumps) from the cathode of an arc system and then directs the spray of these particles to a surface for coating. It has applications where thick coatings are required.

Much needs to be done to formulate appropriate understanding of the plasma and plasma–surface interactions. Measurements often contradict intuitive understandings of plasma behavior. However, if the measurement system

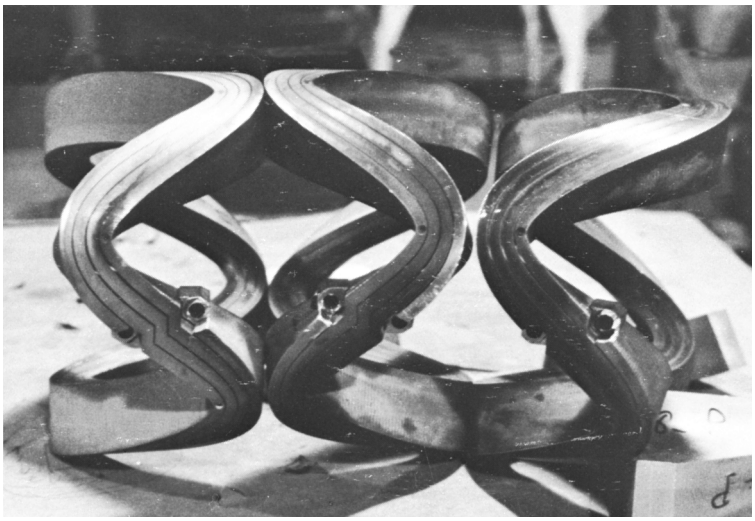


FIGURE 3 Three-dimensional twisted magnet coils made by electrical discharge machining. There is not a single bend in the coils since they are made by "slicing" a solid block of aluminium in this shape.

itself results in a perturbation of the plasma, the results may be unclear, and thus noninvasive diagnostics need to be developed.

II. BASIC PLASMA PROPERTIES

A. Density, Temperature, Composition

The mixture of ions, electrons, and neutral particles making up a plasma must be describable in a way that can provide a useful characterization of its properties. Such quantities and processes include the following:

1. *Density*, described by n_e, n_{ix}, and n_x, which refer to the electron density, the ion density of species x, and the neutral density of species x, respectively. It is important to note that most plasmas contain ions of several different species (positively and/or negatively charged), and the number density (usually expressed in units of particles per cubic centimeter) is a very important quantity. The density is usually a function of both position and time. It should be measured experimentally and is then often modeled theoretically, depending on the nature of the various processes that act to change them. In particular, we refer to the following processes:

a. *Attachment*, in which a neutral particle and an electron combine, producing a negatively charged ion.

b. *Diffusion*, in which particles diffuse in position space or velocity space. Thermal diffusion is related to particle diffusion, but refers to energy, not particle transport.

c. *Recombination*, in which ions and electrons recombine to form neutral particles; radiation is sometimes emitted.

d. *Ionization*, the inverse of recombination.

e. *Excitation*, in which neutral particles or ions (that are not fully stripped) gain energy, which is evident by orbital electrons moving to higher energy states.

f. *De-excitation*, the inverse of excitation. Often, radiation is emitted.

There are many ways in which these processes can occur, such as ionization by electron impact, chemical ionization, or radiation absorption.

It is often surmised that a plasma is electrically neutral, but such a condition usually does not occur when a plasma is in contact with a surface. Under these circumstances, a "sheath" is developed in which either electrons or ions are the dominant species. Usually, this results in a net electric field in the sheath. Sheaths have considerably different properties than do the neutral plasma, and care must be taken to understand them. Figure 4 shows the electrostatic potential between a plasma and a conducting material in the sheath region. The conductivity of a plasma may actually be quite high, often greater than that of metals.

2. Another important quantity that is needed to characterize a plasma is the *temperature* of the individual components, i.e., T_e, T_{ix} and T_x. Temperature is also a quantity that is not usually constant in space or time.

3. The *composition* of the plasma is of paramount importance. Here, one needs to know the mass numbers of all of the ions and neutral particles in the plasma. In many

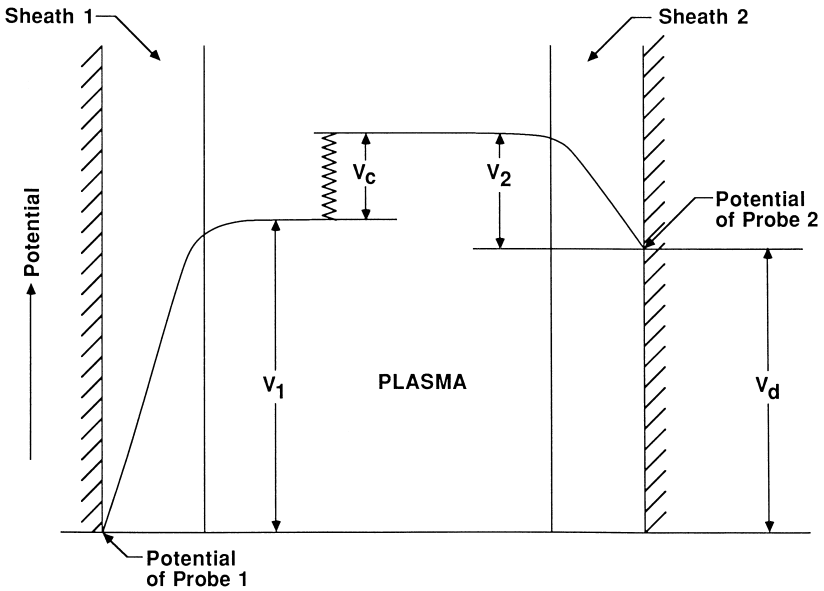


FIGURE 4 Electrostatic potential between the plasma and an electrode in the sheath region.

cases, it is desirable to know this as a function of both time and position, and the nature of the diagnostic devices needed to determine this ranges from very simple to very sophisticated.

B. Plasma Production

Plasmas may be generated by passing an electric current through a gas. Normally, gases are electrical insulators, but there are always a few charge carriers present that can be accelerated by the electric field and can then collide with neutral particles, producing an avalanche breakdown, thus making the plasma. The electric field needed for breakdown can be made with a potential set up between a pair of electrodes, with an "electrodeless" radiofrequency (rf) induction coil, with shock waves, with lasers, or with charged or neutral particle beams. The latter processes can also produce gaseous plasmas if they impinge on a solid target. In addition, heating various materials (usually alkali metals) in ovens or furnaces will cause not only evaporation of neutral particles, but also ionization, and plasmas may be made in this way. Many chemical processes can also cause ionization.

III. PLASMA PHYSICS

A. Plasma Dynamics

The dynamics of the motion of the charged particles in a plasma is governed by the fundamental equation of motion

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{1}$$

where $\mathbf{F}$ is the force on the charged particle, $\mathbf{E}$ is the electric field, $\mathbf{B}$ is the magnetic field, and $\mathbf{v}$ is the charged particle's velocity. If one knew all of the values of the electric and magnetic fields, including those produced by all of the particles (self-consistent fields) at the location of each particle, and if collisions between particles could be neglected, then the trajectories of all the particles could be obtained simultaneously with a computer. However, as can be seen from Fig. 1, such a computer cannot exist, since literally billions of equations would need to be solved simultaneously. In addition, collisions between particles, including charged-particle/neutral-particle collisions, must occur, so such a method cannot be practicable, However, much understanding can be achieved with numerical simulation of "clumps" or "clouds" of plasmas using modern supercomputers, where upward of 1 million simultaneous equations can be solved over a reasonable time period.

Thus, we need to develop a means to consider collisional interactions between particles making up the plasmas. Such interactions are classified into two types—elastic and inelastic collisions. In elastic collisions, kinetic energy, linear momentum, and angular momentum of the two colliding particles are conserved. In inelastic collisions, some of the energy and momentum is changed into or from internal vibration energy, chemical energy (such as chemical bonds), or to conduct the processes of ionization, excitation, recombination, or de-excitation and potentially generate electromagnetic radiation during the process as well.

The most common representation for such interactions is called the collision cross section. We develop this formulation as follows. Let us assume that a beam of particles of density n particles/cubic centimeter traveling with a velocity v and of cross section A passes through the plasma a distance dx. Let N be the number of plasma particles/cubic centimeter. The number of beam particles colliding with plasma particles per unit time may then be written as

$$dn/dt = -[(N\sigma A\,dx)/(A\,dx)]nv. \tag{2}$$

The term $N\sigma A\,dx$ is the probability of collision in the volume $A\,dx$, and nv is the particle current density of incoming beam particles; σ is the collision cross section for this particular process. We may also write $v\,dt = dx$ and $N\sigma = p_0 p_c$ and we can rewrite Eq. (2) as

$$n = n_0 \exp(-p_0 P_{cx}). \tag{3}$$

It is also convenient to write

$$n = n_0 \exp(p_0 P_c v t) = n_0 \exp(-\nu_c t), \tag{4}$$

introducing an average collision frequency $\nu_c = p_0 P_c v$. In the above equations, P_c is the probability of collision for a particular process and p_0 is the "reduced" pressure $= 273p/T$, which expresses a concentration of $N/V = 3.54 \times 10^{16} p_0$ molecules/cm^3. The term $p_0 P_c$ has units of 1/length, or $1/(p_0 P_c) = \lambda$, the mean free path.

Each process has its own cross section, which in many cases can only be determined experimentally.

1. Particle Diffusion

a. Free diffusion. To examine this condition, we first consider particles as free to move in a plasma of similar particles. For simplicity, we assume particle flow in one dimension along which a density gradient dn/dx has been established. By using Newton's second law, we can write, for the change in momentum due to elastic collisions between the particles, each of mass m,

$$\frac{d(nmv_x)}{dt} = -mv_x n \nu_m. \tag{5}$$

The velocity v_x is acquired by the particle due to random collisions with the other particles and is therefore

independent of the coordinates as long as the particles are all assumed to be in energy equilibrium. Thus, we may write

$$\frac{d(nv_x)}{dt} = v_x \frac{d(nv_x)}{dx} = v_x^2 \frac{dn}{dx} = -nv_x\nu_m. \tag{6}$$

When we average over v_x define the particle *current to* be $\Gamma_x = nv_x$, and make the approximation that $v_x^2 = v_T^2/3$, where v_T is the average thermal speed, we obtain

$$\Gamma_x = \frac{-v_T^2}{3\nu_m}\frac{dn}{dx} = -D\frac{dn}{dx}, \tag{7}$$

which is the free diffusion equation, with D the free diffusion coefficient,

$$D = v_T^2/3\nu_m = v\lambda/3. \tag{8}$$

The value of D is not normally the same for electrons and ions since, if they have the same temperature, their average thermal speed v_T is not the same.

Written in three dimensions, if the plasma is isotropic, we obtain $\mathbf{\Gamma} = -D\nabla n$, where ∇ is the gradient operator. In this case $\mathbf{\Gamma}$ is a vector. If there are spatial variations of the diffusion coefficient, we should write

$$\mathbf{\Gamma} = -\nabla(Dn). \tag{9}$$

The diffusion coefficient for a given species of particles could have components that are due to several processes, each of which give spatial diffusion. For example, electrons are simultaneously colliding with other electrons, ions, and neutral particles. Each process contributes a "mean free path" and a collision frequency, which must be added together, reciprocally in the case of the former and directly in the case of the latter, to obtain the combined mean free path and free diffusion coefficient for all of the combined processes.

If we consider that the diffusion has not reached a steady state, then we can calculate the time variation in concentration in any element of space, if there is no source or sink of particles, from the equation

$$\frac{\partial n}{\partial t} + \nabla \cdot \mathbf{\Gamma} = 0, \tag{10}$$

which leads directly to the time-dependent diffusion equation

$$\frac{\partial n}{\partial t} = D\nabla^2 n. \tag{11}$$

This equation is separable and can often be solved as a boundary value problem. Inclusion of other effects that generate or destroy particles in an element of space requires a modification of Eq. (11) to

$$\frac{\partial n}{\partial t} = D\nabla^2 n + \alpha n^2 - \beta n^2, \tag{12}$$

where α and β are the source and sink terms, respectively. These can be ionization (source), recombination (sink), etc.

b. Mobility and time-varying fields. In the presence of time-varying electric fields and collisions, we may approximate Eq. (1) as

$$m\left(\frac{dv}{dt}\right) + (mv\nu_m)v = qE_0\exp(j\omega t), \tag{13}$$

where q is the charge of the particle, negative for electrons and either positive or negative for ions, depending on their nature. We have

$$v = \left(\frac{q/m}{j\omega + \nu_m}\right)E. \tag{14}$$

We may take the quotient of v/E, which is defined as the mobility μ,

$$\mu = \frac{V}{E} = \left(\frac{q/m}{j\omega + \nu_m}\right). \tag{15}$$

If we are interested in the case of a dc electric field where $\omega = 0$, or a case at high pressure where the collision frequency is much larger than the ac frequency of the electric field, then the mobility reduces to

$$\mu = \frac{q}{m\nu_m}. \tag{16}$$

In many cases, the mobilities of the different species of particles in a plasma must be measured. Neutral particles, being uncharged, do not have a mobility.

The ratio of the diffusion coefficient to the mobility is an important quantity because it is a measure of the average particle energy. In a dc electric field, the ratio is

$$\frac{D}{\mu} = \frac{v_T^2 m}{3q} = \frac{2}{3q}\left(\frac{1}{2}mv_T^2\right) = \frac{2}{3q}u_\mathrm{avg}. \tag{17}$$

The quantity u_avg is the average kinetic energy of the particles, and the numerical constant depends upon how the averaging is carried out. The value 2/3 is correct for a Maxwellian distribution of velocities.

If we write the particle flow equation for the particles discussed previously and add the contribution through the mobility from the electric field, we obtain

$$\mathbf{\Gamma} = -D\nabla n - n\mu\mathbf{E}. \tag{18}$$

Assuming that the electric field is in the z direction and that a steady-state condition applies ($d/dt = 0$), we may apply Eq. (16) to the continuity equation to obtain

$$\nabla^2 n = -\frac{\mu}{D}\mathbf{E}\nabla n. \tag{19}$$

2. Ambipolar Diffusion

So far we have assumed that the ions and electrons diffuse freely with no interaction between them. When the density is high, this is not the case. That is, the free diffusion coefficients, not normally being equal, will result in the enhanced transport of one species of particle. Charge separation will occur, and, as a result, an ambipolar electric field is established that will affect the motion of the particles directly through Eq. (17). We may write, for the positively charged particles,

$$\Gamma_+ = -D_+\nabla n_+ + \mu_+ \mathbf{E}_s n_+ = n_+ \mathbf{V}_+, \tag{20}$$

where $\mathbf{E}_s$ is the space-charge field. In terms of the velocity,

$$\mathbf{V}_+ = -\frac{D_+}{n_+}\nabla n_+ + \mu_+ \mathbf{E}_s. \tag{21}$$

By writing a similar expression for the negatively charged particles, we obtain

$$\mathbf{V}_- = -\frac{D_-}{n_-}\nabla n_- + \mu_- \mathbf{E}_s. \tag{22}$$

If we eliminate $\mathbf{E}_s$ between these equations and set $n_+ = n_- = n$ as well as the gradients and velocities equal, we obtain the result

$$\mathbf{V} = -\left(\frac{D_+\mu_+ + D_-\mu_-}{\mu_+ + \mu_-}\right)\nabla n. \tag{23}$$

The quantity in the parentheses is a diffusion coefficient for the two signs of particles interacting with each other so *that they both diffuse together*. This quantity is called the *ambipolar diffusion coefficient* and is defined as

$$\mathbf{D}_{\text{amb}} = -\left(\frac{D_+\mu_+ + D_-\mu_-}{\mu_+ + \mu_-}\right). \tag{24}$$

The ambipolar time-dependent diffusion equation (for both species of particles) is thus

$$\frac{\partial n}{\partial t} = D_{\text{amb}}\nabla^2 n. \tag{25}$$

3. Heat Transport

The above discussion centered on particle transport. In addition to the transfer of particles by diffusion, *energy* can also be transported. We write the energy flux in a way similar to the particle flux, as

$$Q = n\left(\frac{1}{2}mv^2\right)v. \tag{26}$$

Let E denote the thermal energy of a particle. Then $\bar{E}$, the mean thermal energy of particles at a given point, is a function of the local temperature T at that point, and, as in the particle diffusion case, we shall assume only a one-dimensional variation for this development. The *specific heat* of the plasma is given by the relation

$$c_v = \frac{d}{dT}\left(\frac{\bar{E}}{m}\right). \tag{27}$$

The net *particle* flux passing a given point for particles passing the origin in the x direction and leaving the region between x and $x + dx$ without having made a collision is

$$d\Gamma_{0x} = +\frac{v}{6\lambda}n(x)e^{-(x/\lambda)}\,dx \qquad \text{if} \quad x > 0. \tag{28}$$

Each of the particles passing the origin carries with it a thermal energy equal, on the average, to the value of $\bar{E}$ at the region between x and $x + dx$. We may write a similar expression for the particle flux from the points where $x < 0$.

We previously expanded the density $n(x)$ in a Taylor series about the origin. We shall do a similar thing to obtain the energy flux, and expand the local energy in terms of the energy at the origin. In order to do this, we need the average velocity in a given coordinate direction, which is in, say, the positive or negative direction of that coordinate. Note that the average velocity otherwise would be zero. The number density of those particles is, on the average, half the actual density. Assuming a Maxwellian distribution of velocities, we may write the total average velocity to be

$$n\bar{V}_+/2 = \int Vf\,d^3V = (2kT/\pi m)^{1/2} = V_T/2, \tag{29}$$

where V_T is the thermal velocity $(3kT/m)^{1/2}$. Thus, the total thermal energy that the particles from x and $x + dx$ carry across the origin is

$$\frac{nV_T\bar{E}}{4} = \frac{1}{4}nV_T\left(E + \lambda\mu\frac{\partial\bar{E}}{\partial x}\right), \tag{30}$$

and for particles produced in the region to the left of the origin, the energy flux is

$$\frac{nV_T\bar{E}}{4} = \frac{1}{4}nV_T\left(E - \lambda\mu\frac{\partial\bar{E}}{\partial x}\right). \tag{31}$$

In these last two equations, the energy is now written in terms of the energy at the origin. In addition, λ is the mean free path, and μ is a numerical constant roughly of order unity. We now obtain the net rate of flow of energy from the negative to the positive side as

$$\frac{1}{4}nV_T\left(E - \lambda\mu\frac{\partial\bar{E}}{\partial x}\right) - \frac{1}{4}nV_T\left(E + \lambda\mu\frac{\partial\bar{E}}{\partial x}\right) = \frac{1}{2}nV_T\lambda\mu\frac{\partial\bar{E}}{\partial x}. \tag{32}$$

We can now obtain this result in terms of the specific heat and the temperature:

$$Q = -\tfrac{1}{2}mnV_T\lambda\mu c_v. \tag{33}$$

Thus, heat flow is proportional to the temperature gradient, rather than the density gradient. The thermal conductivity χ is then

$$\chi = \frac{1}{2} mn V_T \lambda \mu c_v, \tag{34}$$

so the heat flux is

$$\mathbf{Q} = -\chi \nabla T. \tag{35}$$

4. Effects of AC Electric and Magnetic Fields

In order to determine the response of the plasma to an electromagnetic wave, we must first examine its response to ac electromagnetic fields. We assume that the plasma is "cold" in that the motion of the individual charged particles can be considered to be entirely due to the electric fields they experience, in this case, externally imposed fields only. We do this by reexamining the equation of motion of a charged particle:

$$\mathbf{F} = q\mathbf{E} = m\frac{d\mathbf{v}}{dt}. \tag{36}$$

If we assume that the electric field is driven by an ac source, we may adopt a "phasor" notation for the ac part of the signal. That is, we assume all quantities vary as

$$\exp(-j\omega t) \tag{37}$$

with regard to their time variation, where ω is the driving frequency. Thus, the electric field, for example, is

$$\mathbf{E} = \mathbf{E}_0 \exp(-j\omega t). \tag{38}$$

We only need to solve for the spatial part of the variation. Thus, Eq. (38) becomes

$$q\mathbf{E}_0 = -j\omega \mathbf{v}_0. \tag{39}$$

We use Eq. (39) to find the ratio of velocity to electric field, that is, the mobility. It is simply

$$\mu = |\mathbf{v}_0/\mathbf{E}_0| = -q/(j\omega m). \tag{40}$$

The *conductivity* of the plasma is determined by noting that the electric current density may be written as

$$\mathbf{J} = nq\mathbf{v} = nq\mu\mathbf{E} = -\frac{nq^2\mathbf{E}}{j\omega m} = \sigma\mathbf{E}, \tag{41}$$

where σ is the conductivity of the plasma in the cold plasma approximation. If a plane monochromatic wave propagates through a plasma, we find that the wave will be cut off (no longer propagate) where the density of the plasma reaches a certain value. To determine this, we examine Maxwell's equation:

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}. \tag{42}$$

If we continue to use the phasor time notation, then we may write Eq. (42) as

$$\nabla \times \mathbf{H}_0 = \sigma\mathbf{E}_0 - j\omega\varepsilon_0\mathbf{E}_0. \tag{43}$$

or

$$\nabla \times \mathbf{H}_0 = -j\omega\varepsilon_0\left(1 - \frac{\sigma}{j\omega\varepsilon_0}\right)\mathbf{E}_0. \tag{44}$$

Using Eq. (41), we finally obtain

$$\nabla \times \mathbf{H}_0 = -j\omega\varepsilon_0\left(1 - \frac{n_e q_e^2}{m_e\varepsilon_0\omega^2}\right)\mathbf{E}_0. \tag{45}$$

The effective permittivity of the plasma is then

$$\varepsilon = \varepsilon_0\left(1 - \frac{n_e q_e^2}{m_e\varepsilon_0\omega^2}\right) = \varepsilon_0\left[1 - \left(\frac{\omega_{pe}^2}{\omega^2}\right)\right]. \tag{46}$$

Equation (46) shows that whenever the electron plasma frequency $\omega_{pe} = [n_e q_e^2/(m_e\varepsilon_0)]^{1/2}$ is greater than the driving frequency ω, the effective permittivity becomes negative and the wave no longer propagates. Note that it is always less than the permittivity of free space as well. If $\varepsilon/\varepsilon_0 < 0$, then we do not have normal propagation of waves, but evanescence.

B. Types of Plasmas

In examining Fig. 1, we can now consider the properties of those plasmas that are of direct interest to modern industrial problems. We concentrate on two types: glow (nonequilibrium) and arc (thermal) discharges. In general, if one considers a set of electrodes across which a dc potential is applied, one may classify the glow and arc discharges roughly according to Fig. 5. The vertical axis is the voltage across the discharge and the horizontal axis is the current. Note that as the current increases, the discharge goes from non-self-sustaining, to a glow discharge, to an abnormal glow, to an arc discharge. The voltage across the

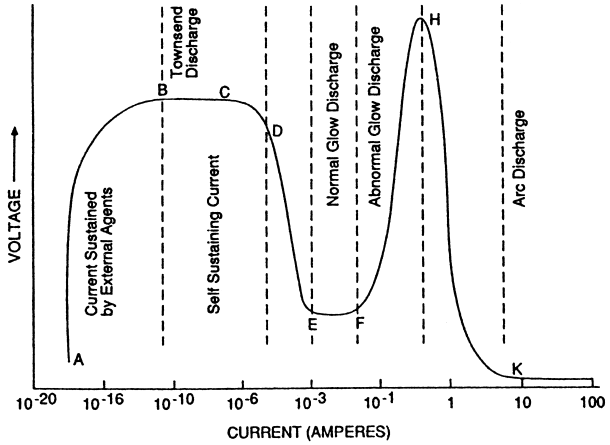


FIGURE 5 Glow and arc discharges.

electrodes drops as the abnormal glow region is entered, rises, and drops again as the arc region is entered.

1. Glow Discharges

Figure 6 displays the typical behavior of a glow-discharge plasma between planar electrodes. The appearance of this discharge is complicated. It is maintained by electrons produced at the cathode by positive-ion bombardment. In the *Aston dark space* there is an accumulation of these electrons, which gain energy through the *Crookes dark space*. The cathode glow results from the decay of excitation energy of the positive ions on neutralization. When the electrons gain sufficient energy in the Crookes dark space (also called the cathode fall, cathode dark space, or Hittdorf dark space) to produce inelastic collisions, the excitation of the gas produces the negative glow. The end of the negative glow corresponds to the range of electrons with sufficient energy to produce excitation, and in the *Faraday dark space* the electrons once more gain energy as they move to the anode.

The positive column is the ionized region that extends from the Faraday dark space almost to the anode. It is not an essential part of the discharge, and for very short discharge tubes it is absent. In long tubes it serves as a conducting path to connect the Faraday dark space with the anode. This portion of the discharge is nearly electrically neutral, and the main electron and ion loss occurs by ambipolar diffusion. In the last few mean free paths, the electrons may gain energy high enough to excite more freely as the positive ions are forced away from the anode, producing the anode glow.

It is important to note in Fig. 6 the curves for electric field strength and net space charge. In general, most plasma processing tends to be done with the object to be processed placed on the cathode, other electrodes, or a "target" electrode. As a result, knowledge of the electric field and potential distributions near these electrodes is very important for an understanding of the nature and energies of the particles as they traverse the sheath region around the cathode.

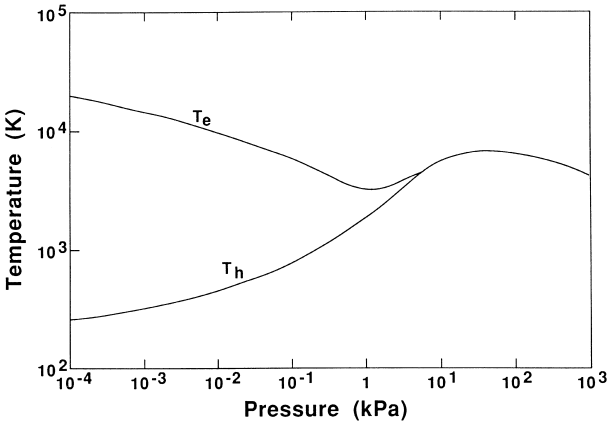


FIGURE 6 Transition between thermal and nonequilibrium plasmas.

2. Arc Discharges

A true arc discharge is characterized by a low cathode fall of the order of the ionization potential of the neutral atoms in the plasma. In glow discharges, the cathode fall tends to be much higher, perhaps of the order of 200 V. From Fig. 5, one can see that the current–voltage characteristic is falling and the current density of the arc is very high. From this information it is clear that electron emission from the cathode must be governed by some mechanism other than positive-ion bombardment of the cathode. Usually the transition from glow to arc is a rapid and discontinuous process.

Arc discharges may be classified by the emission process that occurs at the cathode. Four types of arcs may be defined.

1. The thermionic arc, in which the cathode is heated by the discharge and the arc is self-maintaining.
2. The thermionic arc, in which the cathode is heated by an external source and the arc is not self-maintaining.
3. Field-emission arcs, in which the electron current at the cathode is due to a very high electric field at the cathode surface.
4. Metal arcs.

In addition, arcs are frequently classified as high- or low-pressure arcs if the gas pressure is roughly above or below 1 atm, respectively. A high-pressure arc is characterized by a small, intensely brilliant core surrounded by a cooler region of flaming gases, sometimes called the aureole. If the arc occurs between highly refractory electrodes such as carbon or tungsten, both the anode and the cathode are incandescent.

At low pressures, the appearance of the column depends upon the shape of the discharge tube. A constricted tube gives rise to a highly luminous column even at low pressures. The most important difference between low- and high-pressure arcs is in the temperature of the positive column. The high-pressure column is at a very high temperature (5000 K or higher). The ions, electrons, and neutral atoms of the high-pressure positive column are in *local thermal equilibrium*, and therefore such plasmas are often called *thermal plasmas*.

The neutral gas temperature of the low-pressure arc is never more than a few hundred kelvins, whereas the

electron temperature may be of the order of 40,000–50,000 K (1 eV = 11,600 K). The difference between a glow and an arc plasma is shown in Fig. 6.

The current density at the cathode of an arc is very much greater than that of the glow discharge. In many cases, the cathode current density is practically *independent* of the arc current. As a result, the plasma tends to concentrate in a small area near the cathode and produces *cathode* spots.

Thus, low-pressure arc and glow discharges are sometimes called *nonequilibrium* (*nonthermal*) *discharges*.

3. RF Discharges

In most cases, rf and/or microwave radiation can be used to break down and maintain a discharge. The advantage of this process is that it is not necessary to have electrodes in contact with the plasma.

An important use of rf discharges occurs when we might wish to coat an electrode with an electrically insulating material (dielectric) by means of a plasma sputtering or chemical vapor deposition process. Normally, the cathode of the discharge is the electrode that receives the positive ions for the coating process. However, if a dc glow discharge is used to produce the plasma, the dielectric that is deposited on the cathode will charge up, and the fluxes of both ions and electrons to the surface will become equal, regardless of the potential applied to the electrode itself. The result is that electrons and ions will combine at the dielectric and no current will be drawn through the electrode to sustain the discharge and it will go out.

An ac discharge, if its frequency is high enough, can be used to maintain the discharge because as the potential on the electrode backing the dielectric is reversed, the charge on the dielectric will leak off. Then, when the cycle is reversed again, the deposition process will reoccur. Typically, rf frequencies greater than about 1 MHz are required to maintain a discharge. Below this frequency, the discharge will be extinguished before the potential is reversed.

4. Breakdown

As the electric field in a discharge tube increases from zero, a small dark current (*Townsend discharge*) is drawn, but at some point there is a sudden transition to one of the several forms of self-sustaining discharge. Figure 7 shows the critical breakdown voltage as a function of the pressure–distance product, where the distance is measured between the electrodes.

From Fig. 7, it is seen that in all cases shown a minimum value of the voltage required for breakdown appears at a critical value of the pressure–distance product. At low pressures, below the critical value of the pressure–distance product (pd), the discharge will take place in the *longer* of two possible paths. This means that bringing electrodes closer together at the lower pressures will provide better insulation.

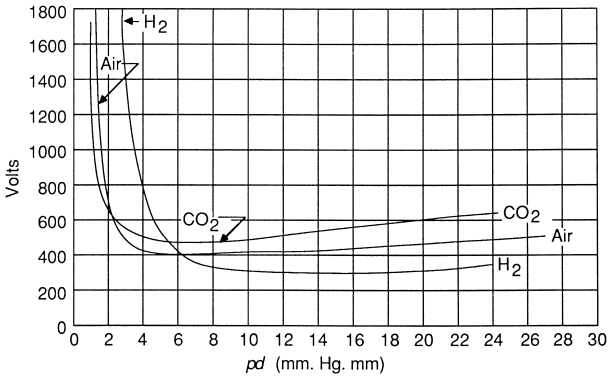


FIGURE 7 Breakdown voltage as a function of pressure–distance product.

However, in bringing metal parts closer together in order to provide a higher breakdown voltage, care must be taken in the disposition of solid insulating material because the electrostatic field on the surface of the electrode will tend to increase. Above a certain value, cold or "field" emission of electrons will take place, and a vacuum arc (metal arc) will begin.

At high pressures, the minimum breakdown distance is so short that it can be virtually impossible to make any measurements. The minimum breakdown voltage is nearly equal to the normal cathode fall potential in the glow discharge.

C. Magnetic Fields

If a dc magnetic field is superimposed on the plasma, the basic equation of motion [Eq. (1)] now includes the magnetic field vector **B**, which will greatly affect the motion of the particles. First, the equation of motion will now be different depending on whether the component equation is parallel or perpendicular to the magnetic field. In the direction parallel to the field, the motion is nearly as though there were no magnetic field. This is not totally the case, since if the energy and magnetic moment of the charged particles are conserved, the orbit in the spatially varying magnetic field will result in a coupling between the parallel and perpendicular components of motion.

The basic effect in a uniform field is that the particles tend to follow the magnetic field lines and orbit around them. The result is a collimation along the field lines, which has important consequences. The plasma conductivity becomes a tensor, and hence the plasma in a dc magnetic field is anisotropic.

D. Plasma Potential

In a nonthermal glow-discharge plasma, we can make several conclusions about the temperatures and densities of the three components, electrons, ions, and neutrals. The electron and ion densities will tend to be equal. Usually, glow-discharge plasmas are not highly ionized, so the neutral density tends to be much larger than the plasma (ion plus electron) density. The thermal speed of the electrons tends to be much greater than that of the ions or the neutrals.

Suppose we suspend a small, electrically isolated dielectric material into the plasma. Initially it will be struck by electrons and ions with respective current densities

$$J_e = q_e n_e v_e \qquad \text{and} \qquad J_i = n_i q_i v_i . \tag{47}$$

However, if $v_i \ll v_e$, the dielectric tends to build up a negative charge and its potential becomes negative with respect to the plasma. The resulting electric field surrounding this material will have a marked effect on the motions of the particles near the material. Since the potential is negative, electrons will tend to be repelled from and ions attracted to the dielectric.

The dielectric tends to keep charging up negatively until the electron flux is reduced and the ion flux increased sufficiently so that the electron and ion fluxes balance. We call the potential of the material at this point V_P or the plasma potential.

E. Debye Length

The Debye length is a measure of how far into the plasma the potential of an electrode or probe is observed. It can be expressed as

$$\lambda_{De} = \left(\varepsilon_0 k T_e / n_e q_e^2\right)^{1/2} \tag{48}$$

for the electrons, where ε_0 is the permittivity of free space, k is Boltzmann's constant, and q is the charge of the electron. A similar expression for the Debye length of the ions can be written with the appropriate ion quantities. Normally, electrons tend to congregate around a positive potential, so the electron Debye length will be important under these conditions, and vice versa for the ions. Often $\lambda_{De} = \lambda_{Di}$.

F. A Plasma Reactor

Figure 8A shows a conceptualized plasma reactor. It consists of two electrodes through which electrical energy is coupled to the gas. A substrate where various components of the plasma may strike is shown at the bottom of the figure. We first examine some of the processes that occur in the gas phase. First, we see that electrons may cause ionization of neutral particles and/or fragmentation of chemical bonds, producing free radicals that are chemically active. The free radicals may either recombine with each other or combine with free radicals produced from other species, thus making new chemical compounds, many of which cannot be made any other way.

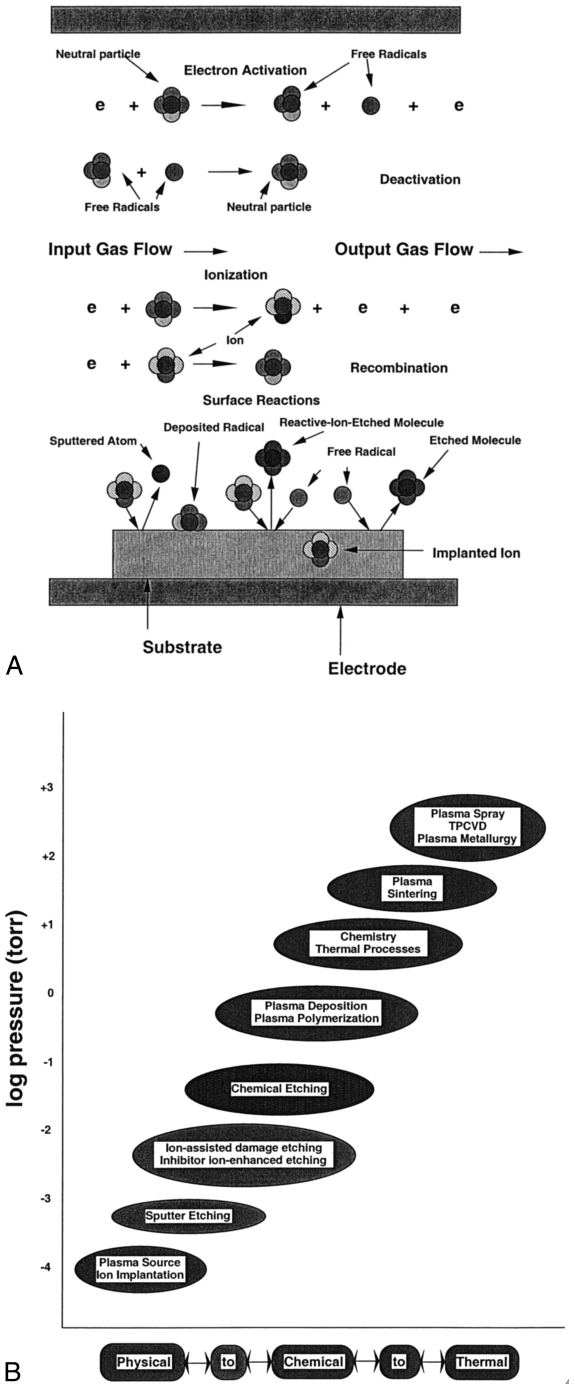


FIGURE 8 (A) Inside a plasma reactor. (B) Types of plasma processes.

The ionized particles are, of course, affected by the presence of electric and magnetic fields. At the substrate, several reactions can occur. First, positive ions can be accelerated by setting a dc bias between the plasma and the substrate so that they are implanted beneath the surface. The ions may also sputter off material from the substrate, which can then be ionized or fragmented by the charged particles in the plasma. Second, free radicals may drift to the substrate and chemically combine with free radicals on the substrate. If the resulting product is a solid, then chemical deposition has occurred. If the resulting product is a vapor, then etching has occurred.

Third, for the appropriate chemical reaction to occur, it may be necessary that both an ion and a free radical strike the substrate. The ion provides energy so that the chemical reaction between the free radical and the substrate can take place. This can result in either deposition or etching. However, the substrate may also become electrically charged during this process, and, if no mechanism is provided for the removal of the charge, arcing or similar current flows, especially through insulators, can result in undesired damage to the substrate.

Fourth, the plasma is often a copious source of photons, ranging in energy from the microwave to the X-ray portion of the spectrum. These photons can also impinge on the substrate and can create desirable or undesirable effects.

In general, all of these processes can be occurring at once. The goal of plasma processing is to maximize the desirable processes and minimize the undesirable processes so that the resulting end product meets specific requirements.

Figure 8B shows a graph of the nature of the plasma processes described above as a function of plasma pressure. As can be seen, at very low pressures, the processes tend to be primarily physical in nature and thus implantation and sputtering tend to dominate. As the pressure increases, chemical reactions in the plasma tend to take place and thus deposition and etching can occur. Finally, at very high pressures, thermal effects dominate and the processes tend to be rapid, high-temperature coating processes.

IV. PLASMA DIAGNOSTICS

In order to understand what is happening in a plasma, it is necessary to "diagnose" it. We can break down the various diagnostic measurement techniques into invasive and noninvasive techniques. Table I lists some of these.

A noninvasive technique only "listens" to or collects what comes out of the plasma. In this case, either radiation or particles are expelled, and we design instrumentation that can detect and analyze them. Invasive diagnostics require either the insertion of a probe or the injection of a particle or radiation beam into the plasma. All other things being equal, noninvasive diagnostics are the most desirable since they will perturb the plasma the least. However, some of the beam and radiation probes usually perturb the plasma so slightly that for many practical purposes they can be considered noninvasive diagnostics.

TABLE I Diagnostic Techniques

Invasive	Noninvasive
Langmuir probes	Radiation spectroscopy
Magnetic probes	Optical
Current probes	Microwave
Beam probes	X-ray
Absorption spectroscopy	Infrared
Optical	Ultraviolet
Microwave	Particle collectors
X-ray	Energy analysis
Infrared	Mass analysis
Ultraviolet	

A. Probes

Figure 9 shows a sketch of a Langmuir probe and its associated circuit. The battery supplies a potential to the probe, and the return circuit to the battery must come from the plasma. The method by which this happens is not always obvious. Such probes are usually very simple devices, consisting only of an insulated wire. The problems with such probes are that sheaths can form around the wire and the

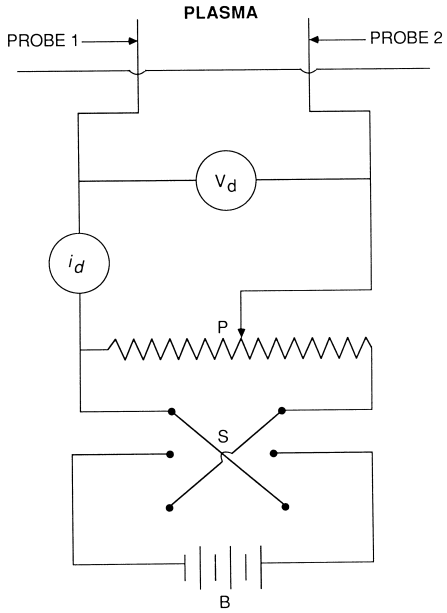


FIGURE 9 A Langmuir probe circuit.

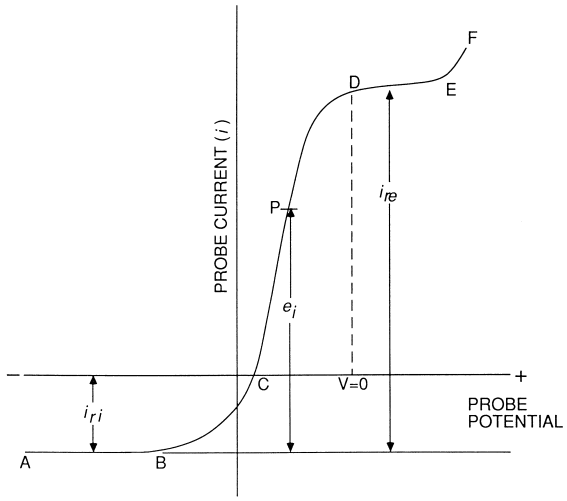


FIGURE 10 Current–voltage characteristic of a Langmuir probe.

resulting measurements will be significantly changed by their presence. A "typical" plot of the probe current versus probe voltage characteristic is shown in Fig. 10.

The characteristic is taken either continuously, by varying the battery voltage, or point by point in a pulsed discharge, if the probe bias is changed from pulse to pulse. In addition, the probe voltage can be swept rapidly to obtain a characteristic dynamically.

At the point V_s the probe is at the same potential as the plasma. The electric field across the sheath is zero at this point, and charged particles travel to the probe based entirely on their own thermal velocities. Since electrons move *much* faster than ions because of their small mass, if the temperatures of the electrons and ions are about the same, then what is collected is primarily electron current.

If the probe voltage is made positive with respect to the plasma, electrons are accelerated toward the probe and the ions are repelled. Thus, the small ion current present at potential V_s now vanishes and an electron-rich sheath builds up until the total net negative charge in the sheath is equal to the positive charge on the probe. The thickness of the sheath is of the order of the Debye length, and outside of it there is very little electric field, so that the bulk of the plasma is undisturbed. The electron current is the current that enters the sheath through random thermal motions, and since the area of the sheath is relatively constant as the probe voltage is increased, we have the fairly flat portion A of the characteristic. This is called the electron saturation current region.

If the probe is now made negative relative to V_s, we begin to repel electrons and accelerate ions. The electron current falls as V_p decreases in region B, which we call the transition region. If the electron velocity distribution were Maxwellian, the shape of the curve after the contribution from ions is subtracted would be exponential.

At large negative values of probe potential, almost all of the electrons are repelled. We then have an ion *sheath* and ion *saturation current*, as shown in region C. This is similar to region A, but there are two differences between the ion and electron saturation currents caused by the mass difference. The first is that often the ion and electron temperatures are not equal, and as a result the Debye lengths are unequal, and the sheath widths are considerably different. The second problem is that if a magnetic field is present, the motion of the electrons is affected much more than the motion of the ions.

The shape of part B of the characteristic is related to the distribution of electron energies and can be used to determine T_e if the distribution is Maxwellian. The magnitude of the electron saturation current is proportional to $(nkT_e/m_e)^{1/2}$, from which n can be obtained if T_e is previously found. The magnitude of the ion saturation current depends on n and kT_e but only slightly on kT_i if $T_i < T_e$, so ion temperature is not easily measured with probes.

The space potential is found by locating the "knee" or junction between regions A and B of the curve, or by locating the point V_p (the zero-current floating potential) and calculating V_s.

Experimental complications for the successful use of probes are many. A partial list of them is as follows:

Surface layers
Perturbation of the plasma
Change of probe area
Reflections
Macroscopic gradients
Metastable atoms
Secondary emission and arcing
Effect of the probe shield
Oscillations
Photoemission
Negative ions
Ion trapping

B. Radiation Probes

An example of a radiation probe of a plasma is a microwave interferometer. Figure 11 shows a drawing of such an interferometer. The plasma introduces a phase shift of the microwave signal that passes through it. By comparison with the reference arm of the interferometer, it yields a signal propositional to the phase difference, which can then be calibrated to yield density directly. The reason for the phase shift is easily seen if we adopt an exponential notation for the time and spatial variation of a plane monochromatic wave. That is, a wave will propagate as

$$\exp j(kz - \omega t), \tag{49}$$

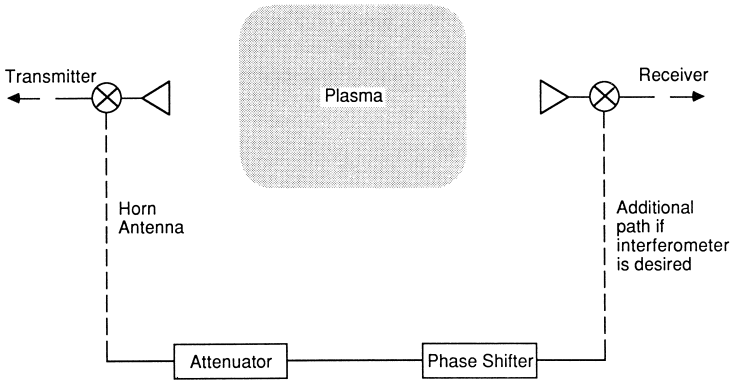


FIGURE 11 A microwave interferometer.

where k is the wave number. The wave equation

$$\frac{\partial^2 E}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} \tag{50}$$

then becomes

$$-k^2 E = -(\omega^2/c^2)E, \tag{51}$$

and thus the relationship between k and ω can be determined to be

$$k = \frac{\omega}{c} = \omega(\varepsilon\mu)^{1/2} \tag{52}$$

or

$$k = \omega(\varepsilon_0\mu_0)^{1/2}\left[1-(\omega_p/\omega)^2\right]^{1/2}. \tag{53}$$

The phase shift is the product of the wave number k in the presence of the plasma, as defined above, times a fixed path length L, minus the free-space wave number $k_0 = \omega(\varepsilon_0\mu_0)^{1/2}$ times that same path length. That is,

$$\Delta\phi = (k - k_0)L, \tag{54}$$

which is directly related to density. The interferometer will detect density as it increases until cutoff is obtained, whereupon no further information is available. The higher the density required to be measured, the higher the frequency of the interferometer.

There are other restrictions with such a system. For one thing, we are assuming plane monochromatic waves. If the plasma has a dimension that is comparable to the wavelength, the plane-monochromatic wave approximation may not be valid. The plasma is also assumed to be cold. Hot plasmas can significantly affect the results. Finally, the interferometer measures *integrated* line density along its path, so the extent of spatial resolution is severely limited. Nevertheless, a microwave interferometer is one of the more important diagnostic tools.

C. Radiation Spectroscopy

The ultimate object in using spectroscopy in a plasma is the interpretation of all spectroscopic observations in terms of a fully consistent theoretical plasma model so as to determine the composition and density and temperature variations (both temporally and spatially). Because spectra often arise from atomic processes, the model must emphasize the particulate nature of the plasma, rather than the fluid aspects that characterize many of its properties.

Development of an appropriate model must take account of numerous atomic collision processes, so that it depends on atomic physics for its basic data. The step from atomic physics to a theoretical model for the plasma consists in discovering ways of taking into account the numerous possible atomic processes to give a composite picture of the spectrum. In many ways, the problems of laboratory spectroscopy are similar to those encountered in interpreting astrophysical spectra. A major difference is that laboratory plasmas have properties that change rapidly with time, so that time-dependent, rather than steady-state, solutions are required.

It is convenient to divide the treatment into two parts: (1) consideration of the intensities of lines and continuum and (2) the shape of the lines.

1. Plasma Models

The spectroscopic radiation of interest is emitted when an electron makes a transition in the field of an atom or ion. The observed intensity of the radiation thus emitted depends on three processes:

1. The probability of there being an electron in the upper level of the transition.
2. The atomic probability of the transition in question.
3. The probability of the photons thus produced escaping from the volume of the plasma without being reabsorbed.

Considerable simplification is achieved if the effect of the interaction of radiation (process 3) with the plasma is considered separately. There are, in fact, physically realizable circumstances where this effect may be neglected (optically thin plasmas).

2. The Local Thermal Equilibrium Model

In the local thermodynamic equilibrium (LTE) model, it is assumed that the distribution of electrons is determined exclusively by particle collision processes and that the latter take place sufficiently rapidly so that the distribution responds *instantaneously* to any change in LTE plasma conditions. In such circumstances, each process is accompanied by its inverse, and these pairs of processes occur at equal rates by the principle of detailed balance. Thus, the distribution of energy levels of the electrons is the same as it would be in a system in complete thermodynamic equilibrium. The population of energy levels is therefore determined by the law of equipartition of energy and does not require knowledge of atomic cross sections for its calculation.

Thus, although the plasma temperature and density may vary in space and time, the distribution of population densities at any instant and point in space depends entirely on *local* values of temperature, density, and chemical composition of the plasma. The uncertainties in predictions of spectral line intensities from an LTE model plasma depend on the uncertainties in the values of the plasma parameters and of the atomic transition probabilities.

If the free electrons are distributed among the energy levels available to them, then, according to statistical mechanics, their velocities have a Maxwellian distribution. The number of electrons of mass m and with velocities between v and $v + dv$ is

$$dn_e = n_e 4\pi (m/2\pi kT_e)^{3/2} \exp(-mv^2/2kT_e) v^2\, dv \quad (55)$$

where n_e is the total density of free electrons and T_e is the electron temperature. For the bound levels, the distributions of electrons are given by the Boltzmann and Saha equations, respectively,

$$\frac{n(p)}{n(q)} = \frac{\omega(p)}{\omega(q)} \exp\left(\frac{\chi(p,q)}{kT_e}\right), \quad (56)$$

$$\frac{n(z+1)n_e}{n(z,g)} = \frac{\omega(z+1,g)}{\omega(z,g)} 2\left(\frac{2\pi kT_e}{h^2}\right)^{3/2} \exp\left(\frac{\chi(z,g)}{kT_e}\right), \quad (57)$$

where $n(p)$, $n(z+1, g)$, and $n(z, g)$ are the population densities of various levels designated by their quantum numbers p, q, and g (the last for the bound level) and ionic charge $z+1$ and z. The term $\omega(z, p)$ is the statistical weight of the designated level, $\chi(p, q)$ is the energy difference between levels p and q, and $\chi(z, g)$ is the ionization potential of the ion of charge z in its ground level g. Equations (55)–(57) describe the state of the electrons in an LTE model plasma.

If the plasma is optically thin, then the intensity $I(p, q)$ of the spectral line emerging from a transition between bound levels p and q is given by

$$I(p,q) = \frac{1}{4\pi} \int n(p) A(p,q) h\nu(p,q)\, ds, \quad (58)$$

where $A(p, q)$ is the atomic transition probability and $h\nu(p, q)$ is the photon energy. The integration is made over the depth of the plasma that is viewed by the detector, and the intensity of radiation $I(p, q)$ is measured in units or power per unit area per unit solid angle.

D. Particle Analysis

Determination of the composition of plasma and neutral particles is important in plasma science and engineering since the composition determines the nature of the physical and chemical processes that can occur. In addition, analysis of the products of the reaction is important to determine the effectiveness of a particular process.

Several different methods are currently in use for this purpose. We discuss one particular method, which is a procedure for analyzing the masses of different ions based on the fact that the cyclotron frequency in a dc magnetic field is proportional to the charge-to-mass ratio of that ion. In this case, the ions can be either single ionized atoms or complex molecules.

Figure 12 shows the diagram of the analysis device, called a Fourier transform mass spectrometer (FTMS). A sample of material is placed at the right-hand side of the

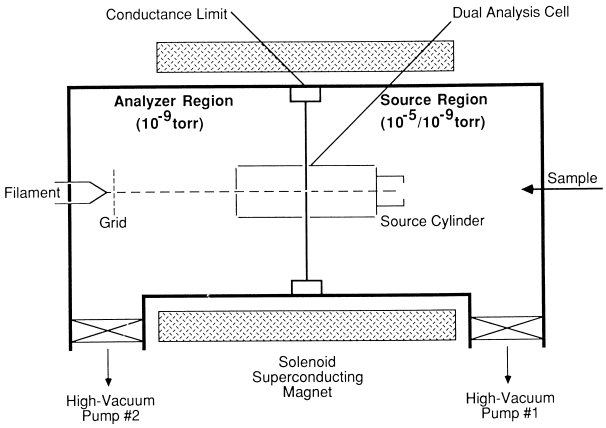


FIGURE 12 Fourier transform mass spectrometer.

cell and an electron beam or a laser is directed at the material. Ions produced by this reaction can be electrostatically confined between the "trap" plates shown in the figure if a small voltage is applied (approximately 1–2 V).

The ions orbit at their appropriate cyclotron frequencies. An rf excitation pulse that is "swept" over a large range of possible cyclotron frequencies is applied between the "excitation" plates in the figure. This excites the ions so that their orbits increase in radius and ions of *equal charge-to-mass ratio move coherently*, so that a "tube" of ions moves around the cell. This tube induces an alternating charge pulse between the "detection" plates in the figure, and the output signal from these plates is Fourier analyzed and the resulting spectrum is displayed on an oscilloscope.

These devices are capable of extremely high resolution, which means that particles of very high mass numbers that are similar (such as mass numbers 1000 and 1001) can be displayed and analyzed. Other mass spectrometers, using different principles, are used in other applications.

V. PLASMA–SURFACE INTERACTIONS

Surfaces in contact with plasmas are bombarded by electrons, ions, neutral particles, and photons. Electron and ion bombardment are particularly important, especially because the particles are so energetic compared with neutral particles.

Several effects occur at once during this process: nondissociative chemisorption, physical adsorption, surface diffusion, dissociative chemisorption, and formation of product molecules. Figure 13 shows a schematic representation of these processes.

A. Plasma-Assisted Chemical Vapor Deposition (PACVD)

In this case, we desire to deposit a thin film of some material on the surface of another material, which we call a substrate. Although with earlier methods vaporizing the material to be deposited and allowing the vapor to come in contact with the substrate under vacuum resulted in the deposition of a material film on the substrate, in order for good deposition to occur, it was often necessary to heat the substrate to elevated temperatures.

PACVD allows lower substrate temperatures. This is particularly important in electronics applications where coatings are deposited onto device structures. *In* this process, a plasma is placed in contact with the substrate and either the plasma particles themselves or neutral particles form the coating. In general, the method of deposition is far from being well understood.

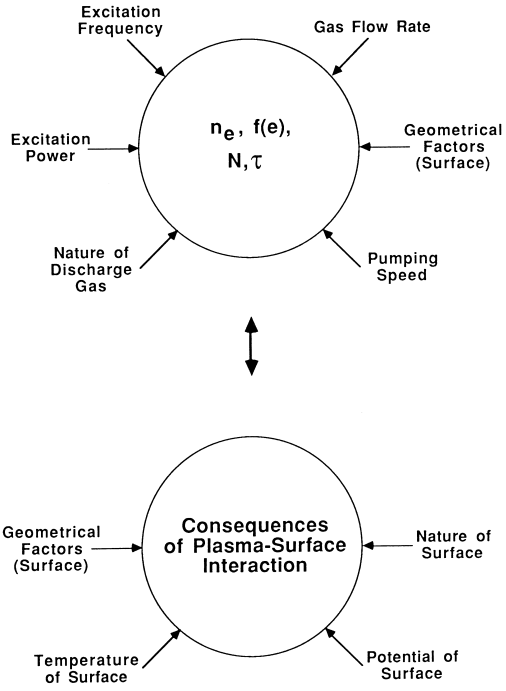


FIGURE 13 Processes involved in the plasma etch procedure.

B. Etching

1. Physical Aspects

Figure 14 shows cross sections of films etched with liquid or plasma etchants. The isotropic profile represents no *over-etch* and can be generated with either wet or dry

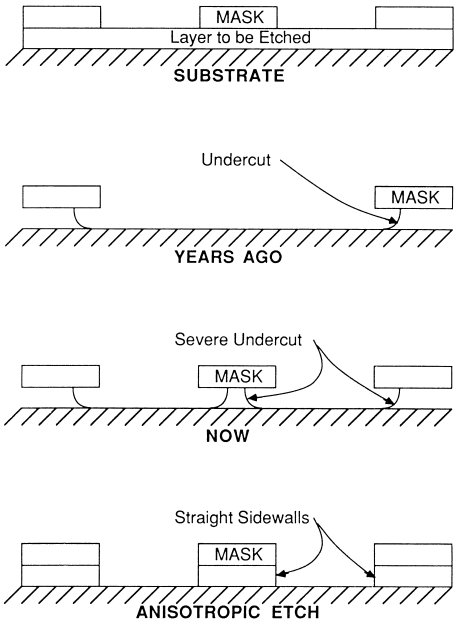


FIGURE 14 Variation of etching conditions.

(plasma) etching techniques. The anisotropic profile requires plasma etching.

Until recently, liquid (wet) etching techniques were the main method of pattern delineation. This is because of two reasons: (1) the technology involved in liquid etching is well established and (2) the selectivity (ratio of the etch rate of the film being etched to the etch rate of the underlying film or substrate) is generally infinite with typical liquid etchant systems.

Unfortunately, wet etching presents several problems for micrometer and submicrometer geometries. Because of the acid environments of most etchant solutions, photoresists can lose adhesion, thereby altering patterning dimensions and preventing linewidth control. In addition, as etching proceeds downward, it proceeds laterally at essentially an equal rate. This undercuts the mask and generates an isotropic etch profile, as shown in Fig. 14. Because the film thickness and etch rate are often nonuniform, a certain degree of *over-etching* is required. If the film thickness is small relative to the minimum pattern dimension, undercutting is inconsequential. However, if the film thickness is comparable to the lateral film dimension, as is the case for current and future devices, undercutting can be intolerable.

In addition, as device geometries decrease, spacings between stripes of resist also decrease. With micrometer and submicrometer patterns, the surface tension of etch solutions can cause the liquid to bridge the space between two resist stripes, and etching of the underlying film is eliminated.

Plasma etching has demonstrated viable solutions to essentially all the problems encountered with liquid etching. Adhesion does not seem to be critical with dry etching techniques. Undercutting appears to be controllable by varying the plasma composition, the gas pressure, and the electrode potentials.

Two additional considerations favor dry etching. First, wet etching requires the use of relatively large volumes of dangerous acids and solvents, which must be handled and ultimately recycled or disposed of. Dry etching uses relatively small amounts of chemicals, although many of the gases used in these processes are also toxic.

In making plasmas for dry etching, a fill gas is broken down by means of application of an external electric field. As the electric field increases, free electrons, whose velocities increase by the action of the field, gain energy. However, since they lose this energy by collisional processes, an increase in pressure, which decreases the mean free path, then decreases the electron energy. What is important, therefore, is a measurement of the velocity of an electron or ion as a function of the ratio E/p (electric field divided by pressure). Figure 15 is a graph of the drift velocity (in an electric field) as a function of this ratio E/p.

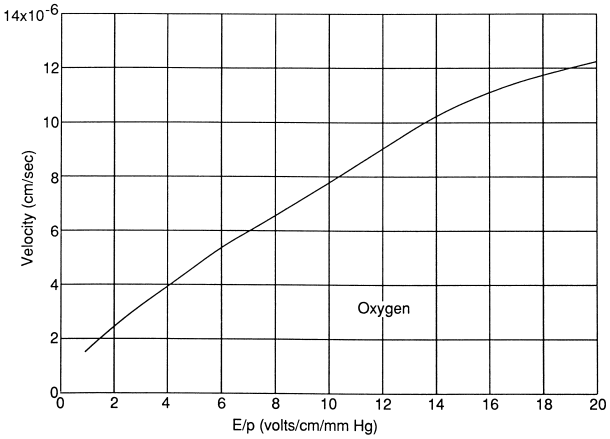


FIGURE 15 Electron drift velocity as a function of E/p.

The glow-discharge plasmas currently used for microelectronic applications can be characterized by the following parameters: pressure = 50 mtorr to 5 torr; $n_e = 109$–$1012\ \text{cm}^{-3}$; $T_e = 1$–10 eV.

Usually the electron temperature is greater than the ion temperature by a factor of about 10. The plasmas are usually very weakly ionized, well below 1%.

These characteristics give the plasma special properties. The electron temperatures are high enough so that chemical bonds can be broken by electron–neutral collisions. As a result, highly reactive chemical species can be produced for etching or deposition. In addition, the surface chemistry occurring in glow discharges is generally modified by the impingement of ions and electrons (and photons) onto the film being etched. The combination of both of these processes results in etch rates and etch profiles unattainable with either process individually.

The energy of ions and electrons striking surfaces in a glow discharge is determined by the potentials established within the reaction chamber. Etching and deposition are generally carried out in a plasma produced by rf, which is often capacitatively coupled to the plasma. The important potentials are the plasma potential, the floating potential, and the potential of the externally biased (powered) electrode.

Usually, under these circumstances, the electrode surfaces are at a negative potential with respect to the plasma. The result is that positive ions bombard the surfaces. The energy of the bombarding ions is established by the difference in potential between the plasma and the surface which the ion strikes. Because these potentials may range from a few volts to a few thousand volts, surface bonds can be broken, and, in certain instances, sputtering of film or electrode material can occur.

In addition, exposure of materials to energetic radiation can result in radiation damage. Positive ions can cause implantation or displacement damage, while electrons,

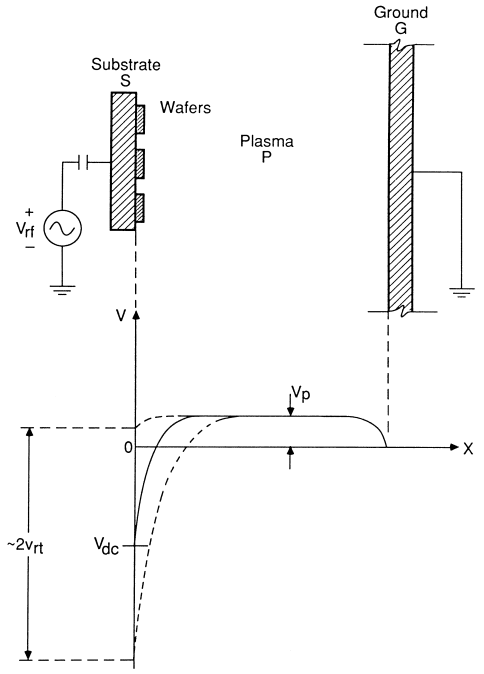


FIGURE 16 Capacitively coupled plasma etching reactor.

X-rays, or ultraviolet photons can result in ionization. Defects thus created can serve as trapping sites for electrons or holes, resulting in an alteration of the electrical properties of the materials. Such alterations may be very beneficial and can improve the surface properties of many materials considerably.

Figure 16 shows a capacitatively coupled configuration used for etching. Consider an rf field established between the two plates. On the first half cycle of the field, one electrode is negative and attracts positive ions; the other is positive and attracts electrons. At the frequencies used (50 kHz to 40 MHz), and because the mobility of electrons is greater than that of the ions, the electron current is much larger than the ion current. This causes a depletion of electrons in the plasma and thus a positive plasma potential. On the second half cycle, a large flux of electrons flows to the electrode that received the small flux of ions. Further, since plasma etching systems generally have a dielectric coating on the electrodes and/or have a series (blocking) capacitor between the power supply and the electrode, no net charge can be passed. Therefore, on each subsequent half cycle, negative charge continues to build on the electrodes (and on other surfaces in contact with the plasma), so that electrons are repelled and positive ions are attracted to the surface.

When a sufficiently large negative bias is achieved on the electrodes, such that the fluxes of electrons and positive ions striking these surfaces are equal, then the transient situation ceases. At this point, the time-averaged (positive) plasma and (negative) electrode potentials are established.

The plasma potential is essentially uniform throughout the observed glow volume in an rf discharge. Between the glow and the electrode is a narrow region wherein a change from the plasma potential to the electrode (or surface) potential occurs. This is the sheath or dark space, and ions that reach the edge of the glow region are accelerated across the potential drop and strike the electrode or substrate surface.

Because of the series capacitor and/or the dielectric coating of the electrodes, the negative potentials established on the two electrodes in the system may not be the same. For example, the ratio of the voltages on the electrodes has been shown to be inversely proportional to the fourth power of the ratio of the relative electrode areas:

$$\frac{V_1}{V_2} = \left(\frac{A_1}{A_2}\right)^4. \tag{59}$$

If V_1 is the voltage on the powered electrode and V_2 is the voltage on the grounded electrode, the voltage ratio is the inverse ratio of the electrode areas to the fourth power. However, for typical etch systems, the exponent is generally less than 4. Although the actual electrodes in a plasma reactor often have the same area, in Eq. (59) A_2 is the *grounded* electrode area, that is, the area *of all grounded* surfaces in contact with the plasma. Because of this, the average potential distribution is similar to that shown in Eq. (59).

In this case, the energy of the ions striking the powered electrode will be higher than that of ions reaching the grounded electrode.

Other plasma parameters can also affect the electrical characteristics. For example, rf power levels and frequency can radically change things. As the frequency is raised, there will be a point at which the ions can no longer follow the alternating voltage, so that an ion cannot traverse the sheath in one half-cycle. Above this frequency, ions experience an accelerating field that is in average over a number of half-cycles. Such an average motion is described with the oscillation center approximation. The drift for **such** a motion is given by

$$\ddot{\mathbf{r}}_0 = -\nabla\Phi, \tag{60}$$

where Φ is the *ponderomotive potential*,

$$\Phi = \frac{q^2}{m^2\omega^2}\left(\frac{\mathbf{E}_0^2}{2}\right). \tag{61}$$

Thus, the not drift, which is *independent* of the sign of the charge, is toward lower average electric fields. The oscillation center drifts in the high-frequency field as if subjected to this potential.

At lower frequencies, the ions are accelerated by instantaneous fields and can attain the maximum energy corresponding to the maximum instantaneous field across the sheath. As a result, for a constant sheath potential, ion bombardment energies are higher at lower frequencies.

2. Chemical Aspects

Figure 13 showed the primary processes occurring during a plasma etch. There are six required steps, and if any one of them does not occur, the entire processing stops. They are (1) generation of reactive species, (2) diffusion to surface, (3) adsorption, (4) reaction, both chemical and physical (such as sputtering), (5) desorption, and (6) diffusion into bulk gas.

The reactive species must be generated by electron–molecule collisions. This is a vital step because many of the gases used to etch thin-film materials do not react spontaneously with the film. For instance, carbon tetrafluoride, CF_4, does not etch silicon. However, when CF_4 is dissociated via electron collisions to form fluorine atoms, etching of silicon occurs rapidly.

The etchant species diffuse to the surface of the material and adsorb onto a surface site. It has been suggested that free radicals have fairly large sticking coefficients compared with relatively inert molecules such as CF_4, so adsorption occurs easily. In addition, it is generally assumed that a free radical will chemisorb and react with a solid surface. Surface diffusion of the adsorbed species or the produce molecule can also occur.

Product desorption is a crucial step in the etch process. A free radical can react rapidly with a solid surface, but unless the product species has a reasonable vapor pressure so that desorption occurs, no etching takes place. For instance, when an aluminum surface is exposed to fluorine atoms, the atoms adsorb and react to form AlF_3. However, the vapor pressure of AlF_3 is approximately 21 torr at 1240°C, and thus etching is precluded at room temperatures.

The chemical reactions taking place in glow discharges are often extraordinarily complex. However, two general types of chemical processes can be categorized: (1) homogeneous gas-phase collisions and (2) heterogeneous surface interactions. In order completely to understand and characterize plasma etching, one must understand the fundamental principles of both processes.

Figure 17 shows how two etch processes may result in a synergism in which the resulting etch rate is greater than the sum of the two. In this case a XeF_2 plasma and an argon ion beam are used together.

a. Homogeneous gas-phase collisions. These collisions generate reactive free radicals, metastable species, and/or ions. As shown in Table II, electron impact can result in a number of different reactions.

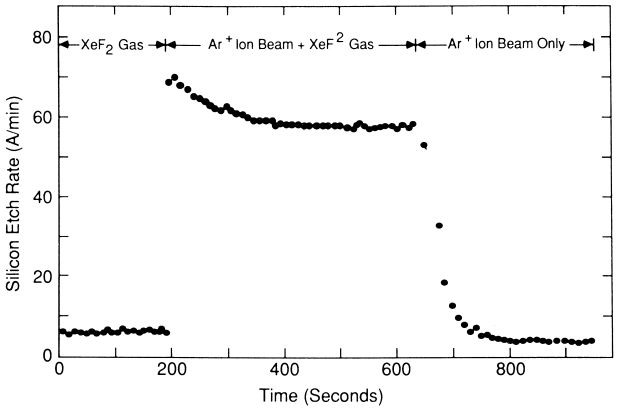


FIGURE 17 Etching synergism of two processes.

Due to the electronegative character of many of the etch gases currently used (O, CF_4, CHF_3, CCl_4, BCl_3, etc.), electron attachment often takes place, thereby generating negative as well as positive ions in the plasma. Although these negative ions affect the plasma energetics, they probably have little, if any, effect on surface reactions because they are repelled by the negative electrode potential.

C. Plasma Polymerization

In the plasma etching process, a competing process that can dominate over etching can occur which is called *polymerization*. A polymer is defined as a high-molecular-weight compound made up from a small repeating organic unit called a monomer. The magnitude of the molecular weight ranges from 1000 to several million atomic mass units (amu) and, depending on conditions, the reaction product can have a statistical distribution of molecular weights.

In order for monomers to form polymers one of two reactions must occur—condensation or addition. The condensation reaction usually results in the loss of a small portion of the original starting molecules. Addition polymers are those that result from the reaction of an unsaturated monomer with an initiator that begins a chain reaction at an activated site to start the growing polymer chain.

TABLE II Results of Electron Impact

Excitation (rotational, vibrational, electronic)	$e + A_2 \rightarrow A_2 + e$
Ionization	$e + A_2 \rightarrow A_2^+ + 2e$
Dissociative ionization	$e + A_2 \rightarrow A^+ + A + 2e$
Dissociation	$e + A_2 \rightarrow 2A + e$
Dissociative attachment	or $A^+ + A^- + e$ $e + A_2 \rightarrow A^- + A$

As the ratio of fluorine to carbon is increased, polymerization ceases and etching begins at a critical value which depends on the potential applied to the surface.

Two general schemes have been proposed for organizing the chemical and physical information on plasma etching and polymerization. Both have dealt primarily with carbon-containing gases, but with slight modifications can be easily applied to other gases. Figure 18 is a schematic depiction of the influence of the fluorine-to-carbon ratio and electrode bias on etching and polymerization.

F/C ratio determines whether etching or polymerization is favored. If the primary etchant species for silicon (F atoms) is consumed either by a loading effect or by reaction with hydrogen to form HF, the F/C ratio decreases, thereby enhancing polymerization. Such effects are caused primarily by enhanced energies of the ions striking these surfaces.

In the *etchant–unsaturate model* described by the following equations, specific chemical species derived from electron collisions with etchant gases are considered:

$$e + \text{halocarbon} \rightarrow \text{saturated radicals} + \text{unsaturated radicals} + \text{atoms} \tag{62}$$

$$\left.\begin{matrix}\text{reactive atoms}\\ \text{reactive molecules}\end{matrix}\right\} + \text{unsaturates} \rightarrow \text{saturates} \tag{63}$$

$$\text{atoms} + \text{surfaces} \rightarrow \left[\begin{matrix}\text{chemisorbed layer}\\ \text{volatile products}\end{matrix}\right. \tag{64}$$

$$\text{unsaturates} + \text{surfaces} \rightarrow \text{films}. \tag{65}$$

This model does not consider the specific chemistry occurring in a glow discharge, but rather views the plasma as a ratio of fluorine to carbon species which can react with a silicon surface. The generation or elimination of these active species by various processes or gas additions then modifies the initial fluorine-to-carbon ratio of the inlet gas.

The F/C ratio model accounts for the fact that in carbon-containing gases, etching and polymerization occur simultaneously. The process that dominates depends on etch-gas stoichiometry reactive-gas additions, amount of material to be etched, the electrode potential, and on how these factors affect the F/C ratio. For instance, as seen in Fig. 18, the

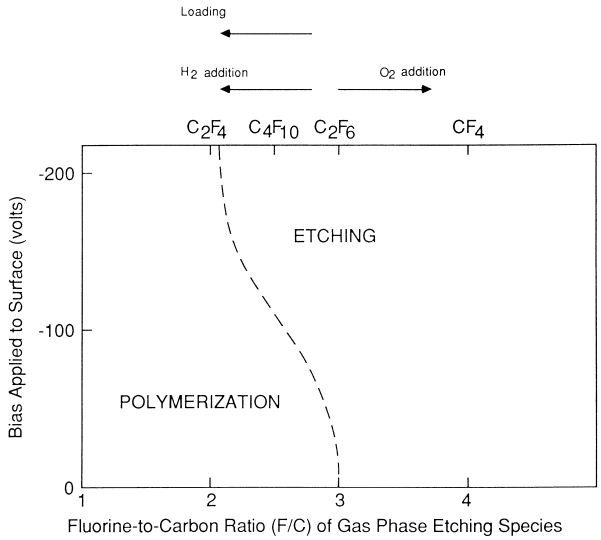


FIGURE 18 Etching rate as a function of electrode bias.

Application of this model to a CF_4 plasma results in the chemical scheme described by

$$2e + 2CF_4 \rightarrow CF_3 + CF_2 + 3F + 2e, \tag{66}$$

$$F + CF_2 \rightarrow CF_3, \tag{67}$$

$$4F + Si \rightarrow SiF_4, \tag{68}$$

$$nCF_2 + \text{surface} \rightarrow (CF_2)n. \tag{69}$$

Depending on the particular precursors generated in the gas phase, etching, recombination, or film formation (i.e., polymerization) can occur. Also, gas-phase oxidant additives (A, F_2, Cl_2, etc.) can dissociate and react with unsaturate species. As an example, O_2 can undergo the following reactions in a CF_4 plasma:

$$e + O_2 \rightarrow 2O + e, \tag{70}$$

$$O + CF_2 \rightarrow COF_{2-x} + xF. \tag{71}$$

Mass spectrometer studies of oxidant additions in fluorocarbon and chlorocarbon gases have demonstrated that the relative reactivity of atoms with unsaturate species in a glow discharge follows the sequence F = O > Cl > Br. Of course, the most reactive species present will preferentially undergo saturation reactions that reduce polymer formation and may increase halogen atom concentration. Ultimately, determination of the relative reactivity of the plasma species allows prediction of the primary atomic species in a plasma of specific composition.

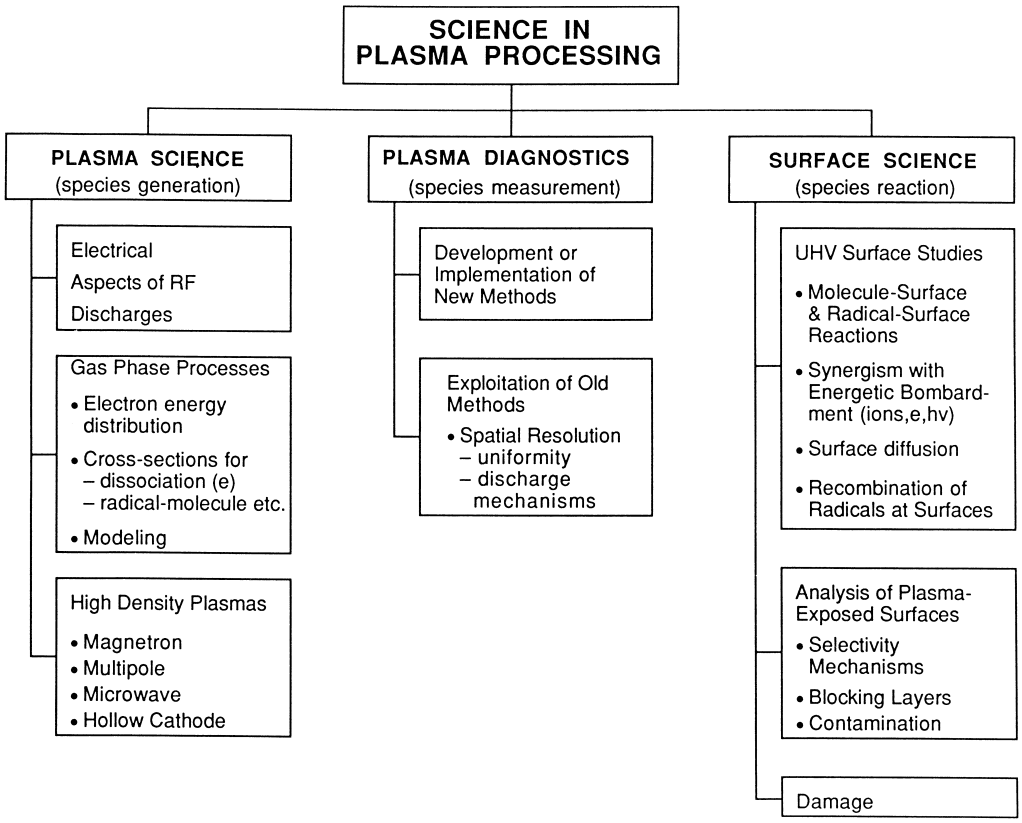


FIGURE 19 Factors that need to be elaborated for the plasma processing of semiconductors.

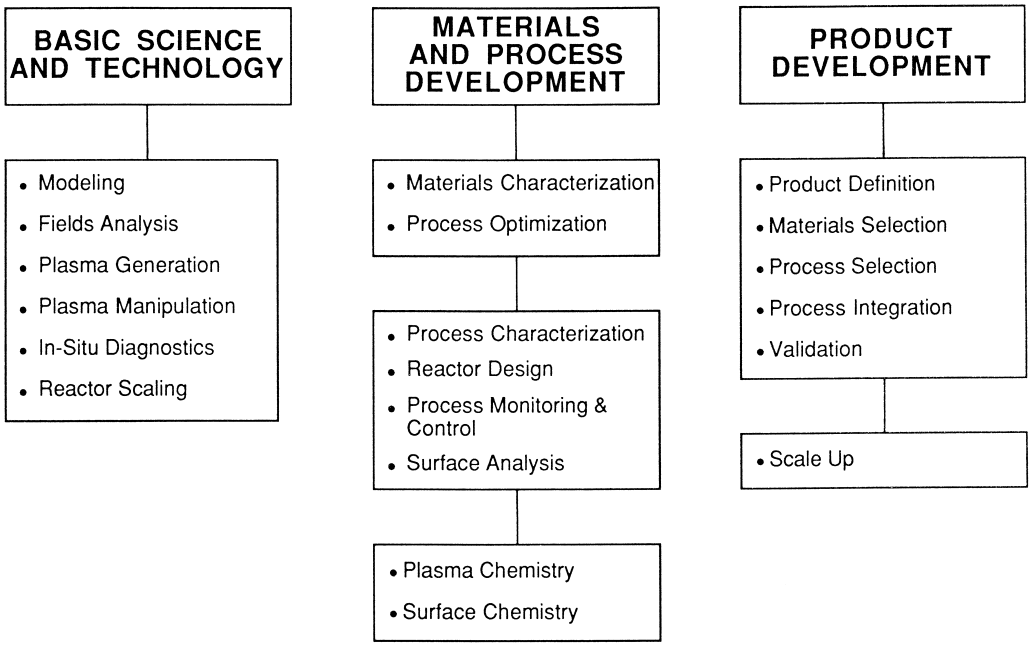


FIGURE 20 Factors that need to be elaborated for plasma processing in the pharmaceutical industry.

D. Ion Implantation

Originally, ion implantion used energetic ion beams (5 kV or higher) directed to the surface of a material to modify the surface characteristics of the substrate while still preserving its bulk characteristics. In recent years, a process of implantation has been used in which a high negative potential is placed between a plasma and the substrate. The resulting electric field accelerates ions from the plasma to the substrate. This process is particularly useful when it is desired to implant nonplanar substrates, such as tools structures, artificial hip joints, and many other metallic items.

VI. CONCLUSION

There are many areas of importance for successful progress in these fields. Figures 19 and 20 show the various problems that need to be addressed in the semiconductor and pharmaceutical industries, respectively.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC AND MOLECULAR COLLISIONS • CHEMICAL VAPOR DISTRIBUTION • DIELECTRIC GASES • ION KINETICS AND ENERGETICS • NUCLEAR FUSION POWER • PARTICLE PHYSICS, ELEMENTARY • PLASMA DIAGNOSTIC TECHNIQUES • POLYMERS, SYNTHESIS

BIBLIOGRAPHY

Blaustein, B. C. (ed.). (1969). "Chemical Reactions in Electrical Discharges," American Chemical Society, New York.

Brown. S. C. (1959). "Basic Data of Plasma Physics," Wiley, New York.

Bunshah, R. F. (ed.). (1992). "Deposition Techniques for Films and Coatings," Noyes, Park Ridge, NJ.

Chen, F. F. (1974). "Introduction to Plasma Physics," Plenum Press, New York.

Coburn, J. W. (1980). *Plasma Chem. Plasma Processes* **2,** 2.

Conrad, J. R., Dodd, R. A., Han, S., Madapura, M., Scheuer, J., Sridharan, K., and Worzala, F. (1990). *J. Vac. Sci. Tech. A* **1990,** 1820.

Holloban, J, R., and Bell, A. T. (eds.). (1974). "Techniques and Applications of Plasma Chemistry," Wiley, New York.

Knights, J. C. (ed.). (1994). "The Physics of VLSI," American Institute of Physics, New York.

Lieberman, M. A., and Lichtenberg, A. J. (1994). "Principles of Plasma Discharges and Materials Processing," Wiley, New York.

Mucha, J. A., and Hess, D. W. (1983). "History of Dry Etching," American Chemical Society, New York.

Mucha, J. A., and Hess, D. W. (1983). "Plasma Etching," American Chemical Society, New York,

National Materials Advisory Board. (1985). "Plasma Processing," National Academy Press, Washington, DC.

Nowogrozki. M. (1984). "Advanced III–V Semiconductor Materials Technology Assessment," Noyes, Park Ridge, NJ.

Shohet, J. L. (1971). "The Plasma State," Academic Press, New York.

Venugapalan, M. (ed.). (1971). "Reactions under Plasma Conditions," Wiley, New York.



Space Plasma Physics

Larry Lyons
University of California, Los Angeles

I. Introduction
II. Basic Concepts
III. Solar Wind and Interplanetary Magnetic Field
IV. Solar Wind and Interplanetary Field Interactions with the Geomagnetic Field
V. Particle Access to, and Transport within, the Magnetosphere
VI. Auroras and Auroral Currents
VII. Geomagnetic Disturbances
VIII. Conclusions

GLOSSARY

Aurora Emissions from the upper atmosphere by constituents that have been excited by the impact of energetic particles from the magnetosphere.

Auroral oval The prime region of visible auroral emissions, which consists of approximately circular zones surrounding each geomagnetic pole that are a few degrees in latitude wide and centered near 70° geomagnetic latitude.

Convection Flow of plasma throughout the magnetosphere that is driven by the solar wind.

Geomagnetic latitude Latitude based on the earth's magnetic axis.

Gyroradius Radius of the circular motion of charged particles about a magnetic field.

Interplanetary magnetic field Magnetic field from the sun that is carried throughout interplanetary space by the solar wind.

Ionosphere Region of enhanced ionization that surrounds the earth at altitudes between ~75 and 500 km altitude.

Magnetopause Current layer that to a large extent separates the interplanetary magnetic field from the geomagnetic field.

Magnetosphere Region of space within the magnetopause that is dominated by the geomagnetic field.

Plasma An ionized gas in which electric forces maintain approximate charge neutrality (the excess of negatively or positively charged particles is everywhere much smaller than the total ion density).

Plasma sheet Energetic plasma region that occupies the outer portions of the magnetosphere.

Precipitation Loss of magnetospheric particles to the

atmosphere by collisions at the low-altitude ends of magnetic field lines.

Radiation belts Region of high fluxes of very energetic electrons and ions that encircles the earth in the inner portion of the magnetosphere.

Solar wind Plasma that flows outward from the sun and fills interplanetary space.

SPACE PLASMA PHYSICS is the study of the plasmas that originate from the sun and from the planets and moons within the solar system. These plasmas occupy interplanetary space and the magnetospheres of planets. This article gives an overall description of the plasma processes which control the large-scale structure and dynamics of the near-earth space plasma environment. This includes the formation of the solar wind and interplanetary plasma disturbances. It also includes the interaction of the solar wind plasma and magnetic field with the magnetic field of the earth and how this interaction leads to the interesting and dynamic space plasma environment which exists in the vicinity of the earth. Topics include energy transfer to and within the earth's magnetosphere, formation of magnetospheric structure, and disturbances of the magnetosphere–ionosphere system which constitute what has recently been termed "space weather." Space plasma physics also includes the interaction of the solar plasma with other planets, the mixing of solar and planetary plasmas, and a wide range of wave modes associated with plasma oscillations in space.

I. INTRODUCTION

The sun continuously emits a stream of ionized particles, which is referred to as the solar wind and is the primary component of the plasma which fills interplanetary space. The average speed of this stream in the ecliptic plane is ~400 km/sec, so that it takes about 4 days for particles to reach the earth. Solar wind speeds, however, can be quite variable. They typically range from ~300 to ~800 km/sec, with speeds exceeding 1000 km/sec being occasionally observed. The earth's internal magnetic field is approximately that of a dipole. However, the interaction of the solar wind particles with the earth's magnetic field compresses the earth's field on the dayside and draws the field out into a long tail on the nightside. This interaction also confines most of the magnetic field of the earth to a region referred to as the magnetosphere (see Fig. 1, which is a sketch of the magnetosphere in the noon–midnight

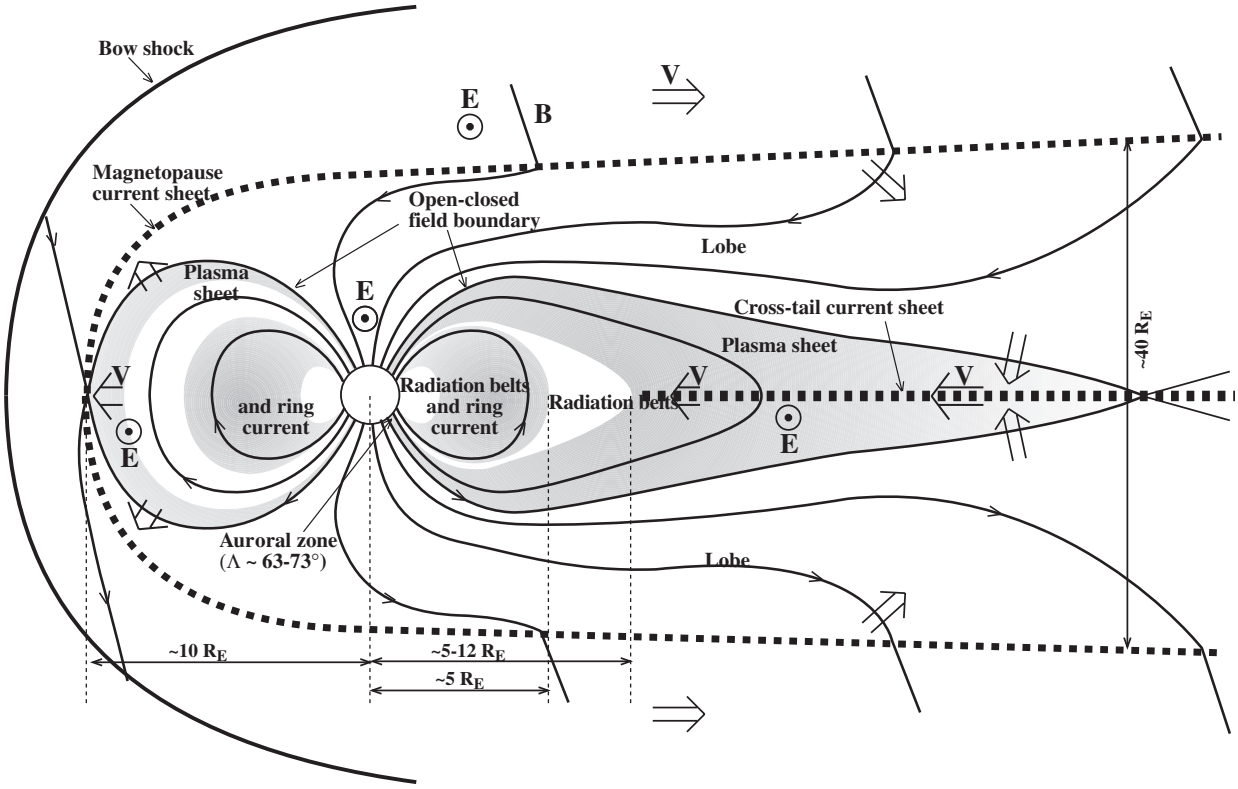


FIGURE 1 Schematic illustration of the magnetosphere in the noon–midnight meridian plane.

meridian plane). The outer boundary of the magnetosphere is called the magnetopause, which typically lies $\sim$10 R_E (earth radii) above the noon equator and $\sim$15 R_E away from the earth within the dawn–dusk meridian plane. On the nightside, the magnetosphere flares outward with increasing distance away from the earth, eventually becoming approximately cylindrical with a diameter of 40–50 R_E. Solar wind speeds are supersonic, so that a shock (called the "bow shock") lies several R_E upstream of the dayside magnetopause.

Plasma particles move in circles around magnetic field lines and thus can become trapped within the magnetosphere. Major regions of trapped energetic particles within the magnetosphere are the plasma sheet and the radiation belts. As illustrated in Fig. 1, the plasma sheet on the nightside is displaced from the high-latitude magnetopause by what are referred to as the tail lobes and extends along the entire magnetospheric tail inward to an equatorial radial distance from the center of the earth $r \approx 5$–12 R_E. The plasma sheet also extends around the earth to other local times and has an outer boundary that lies adjacent to the magnetopause on the dayside. Typical energies of plasma sheet particles are $\sim$1–50 keV for ions and $\sim$0.2–10 keV for electrons. Earthward of the plasma sheet lies a population of more energetic electrons and ions which encircle the earth and are referred to as the "radiation belts." These particles form a current encircling the earth which is referred to as the "ring current."

Particles from the magnetosphere can move along magnetic field lines and strike the upper atmosphere. Those that reach an altitude of $\sim$100–200 km undergo collisions with the neutral atmosphere resulting in a loss of their energy to the neutral atmosphere and their loss from the magnetosphere. Such loss of energetic magnetospheric particles is referred to as precipitation. The energy from precipitating particles excites constituents of the upper atmosphere, and the relaxation of the upper atmospheric constituents back to their ground state gives off emissions, which when sufficiently intense are referred to as the aurora. Such precipitation is most intense from the plasma sheet, leading to an approximately circular region of emissions surrounding each magnetic pole that is referred to as the auroral oval. The oval is typically a few to several degrees in latitude wide and is centered near 70° geomagnetic latitude.

II. BASIC CONCEPTS

Space plasma physics often requires that dynamics be analyzed in terms of both the motion of individual particle and in terms of macroscopic moments such as temperature T, density n, and pressure P. Individual particle motion

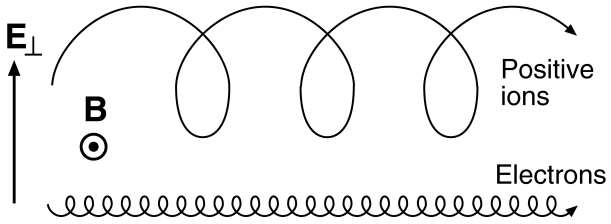


FIGURE 2 The motion of electrons and positive ions in uniform and constant electric and magnetic fields.

is based on considering the force $\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$ acting on a particle of charge q, mass m, and moving with a velocity $\mathbf{v}$ in an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$. Particle motion is generally separated into components $v_\parallel$ parallel to $\mathbf{B}$ and $v_\perp$ perpendicular to $\mathbf{B}$. With $\mathbf{E} = 0$ and a uniform, time-independent magnetic field, $v_\parallel$ is a constant and $v_\perp$ is circular motion about $\mathbf{B}$ with a frequency $|qB/m|$, which is referred to as the gyrofrequency, and radius $mv_\perp/|qB|$, which is referred to as the gyroradius. The direction of gyration is right (left)-handed with respect to the direction of $\mathbf{B}$ for electrons (positive ions) as illustrated in Fig. 2. Except very near current sheets, particle gyroradii are generally very much less than the scale length for field and plasma variations in space plasmas. Also, particle gyroperiods are generally very much less than space plasma time scales for transport and for changes in plasma and field properties.

Figure 2 shows the motion of electrons and positive ions in uniform and constant electric and magnetic fields. The acceleration by the perpendicular component of the electric field $\mathbf{E}_\perp$ alternatively increases and decreases the particle gyroradius once each gyration, so that, in addition to the gyromotion about $\mathbf{B}$, particles move with an average velocity $\mathbf{V}_E = (\mathbf{E} \times \mathbf{B})/B^2$, which is referred to as the electric field drift speed. This relation can be rewritten as

$$\mathbf{E}_\perp = -\mathbf{V}_E \times \mathbf{B}. \tag{1}$$

When $|\mathbf{V}_E| \ll |\mathbf{v}|$, as is the case throughout most of space, we can separate the particle motion into its gyration about $\mathbf{B}$ and a drift of the gyrating particle with velocity $\mathbf{V}_E$. Any electric field parallel to $\mathbf{B}$ simply gives constant particle acceleration along $\mathbf{B}$.

As shown in Fig. 2, electric fields perpendicular to $\mathbf{B}$ cause all charged particles to drift with the same velocity, which does not lead to currents. Spatial variations in magnetic field also give particle drifts. However, the drifts from magnetic field variations are oppositely directed for negatively and positively charged particles, so that a current is formed. This current is azimuthal in the region of the radiation belts, giving rise to the ring current in this region. Such a current also flows within the tail plasma

sheet and is directed across the tail from the dawn side to the dusk side.

The current formed by an individual charged particle gyrating about **B** can be represented as a magnetic dipole with magnetic moment $\mu = K_\perp/B$, where the perpendicular energy $K_\perp = (mv_\perp^2)/2$. As long as magnetic field changes experienced by a particle are small during the course of one gyration about the magnetic field, μ is conserved for particles undergoing electric and magnetic drifts perpendicular to **B** and motion parallel to **B**. This is important because it generally means that particle energies increase (decrease) when particles undergo a drift across magnetic field lines into regions of increasing (decreasing) B. On the other hand, motion along magnetic field lines results in a conversion of parallel energy to (from) perpendicular energy as B increases (decreases) with the total particle energy being conserved.

When we think of the plasma as whole, we deal with macroscopic variables that are defined per unit volume. We consider the forces acting on the plasma per unit volume, which we write as

$$\rho\, d\mathbf{V}/dt = -\nabla P + \mathbf{J}\times\mathbf{B}, \tag{2}$$

where ρ is the total plasma mass per unit volume, **V** is the mass-averaged velocity for all particles within a unit volume, P is the plasma pressure, and **J** is the current density (current per unit area normal to the current). Equation (2) assumes charge neutrality (essentially equal numbers of positive and negative charges), an assumption which is nearly always valid for space plasmas, and neglects gravity, an assumption which is valid for most space plasmas (neglect of gravity is not valid, for example, near the sun).

When plasma and field changes are small in the direction of **B**, (2) can be rewritten using the Maxwell equation

$$\mathbf{J} = \nabla\times\mathbf{B}/\mu_0 - \varepsilon_0\partial\mathbf{E}/\partial t \tag{3}$$

to obtain

$$\rho d\mathbf{V}/dt = -\nabla(P + B^2/2\mu_0). \tag{4}$$

Here the constants are $\mu_0 = 4\pi\times10^{-7}$ H/m and $\varepsilon_0 = 8.85\times10^{-12}$ F/m, and the term $\varepsilon_0\partial\mathbf{E}/\partial t$ in (3) was neglected in (4) because it is small for the large-scale phenomena discussed here. Assuming steady state and no changes in the direction of **V**, we obtain for (4)

$$P + B^2/2\mu_0 = \text{const.} \tag{5}$$

Equation (5) is referred to as pressure balance and is usually applied to regions, such as the geomagnetic tail, where changes in the direction of **B** are small. The quantity $B^2/2\mu_0$ can be thought of as magnetic pressure, so that (5) states that the total pressure (plasma + magnetic) is constant. This tells us, for example, that the magnetic pressure in the lobes, where plasma pressure is low, is greater than the magnetic pressure in the plasma sheet, where the plasma pressure is high.

Equation (3), with $\varepsilon_0\partial\mathbf{E}/\partial t$ neglected, relates currents and magnetic field structure. For example, it tells us that a change in magnetic field strength $\Delta\mathbf{B}$ across a plane perpendicular to **B** must be associated with a current within the plane across which **B** changes. Such a planar current is referred to as a current sheet and has a magnitude per unit distance normal to the current direction of $I = \Delta B/\mu_0$ [A/m]. For the magnetospheric tail, this current is directed in the dawn-to-dusk direction across the tail and is referred to as the cross-tail current sheet (see Fig. 1).

III. SOLAR WIND AND INTERPLANETARY MAGNETIC FIELD

The sun is a large ball of gas held together by its own gravity. The gases are about 90% hydrogen and 10% helium with minor amounts of other constituents. Due to the high temperature of the sun, the solar gases are mostly ionized. The sun appears to have a visible surface because the steep radial gradient in solar density gives a sharp transition between lower regions, where photons are absorbed and reemitted by solar gases without traveling very far, and higher regions, where most photons move away from the sun along straight trajectories without collisions. Solar radiation appears to arise from this narrow transition region (only a few hundred kilometers thick), which is referred to as the photosphere.

Solar gases extend outward well beyond the photosphere with very high temperatures, forming the solar corona. The solar corona is sufficiently hot that many coronal particles have outward-directed velocities with a magnitude that exceeds the speed for gravitational escape from the sun. Such particles stream outward from the sun forming the solar wind. The solar wind escapes toward the near vacuum of interstellar space at supersonic speeds, filling the entire solar system with coronal plasma.

A. General Structure

If it were not for the solar magnetic field, the escaping plasma from the solar corona would form an approximately spherically symmetric solar wind flowing radially outward from the sun at ~750 km/sec. However, the escaping plasma is strongly affected by the magnetic field of the sun because of the tendency for particles to move more easily along magnetic field lines than across them. The solar magnetic field is highly variable, which makes the solar wind and the magnetic field it carries with it into interplanetary space highly variable in space and in time.

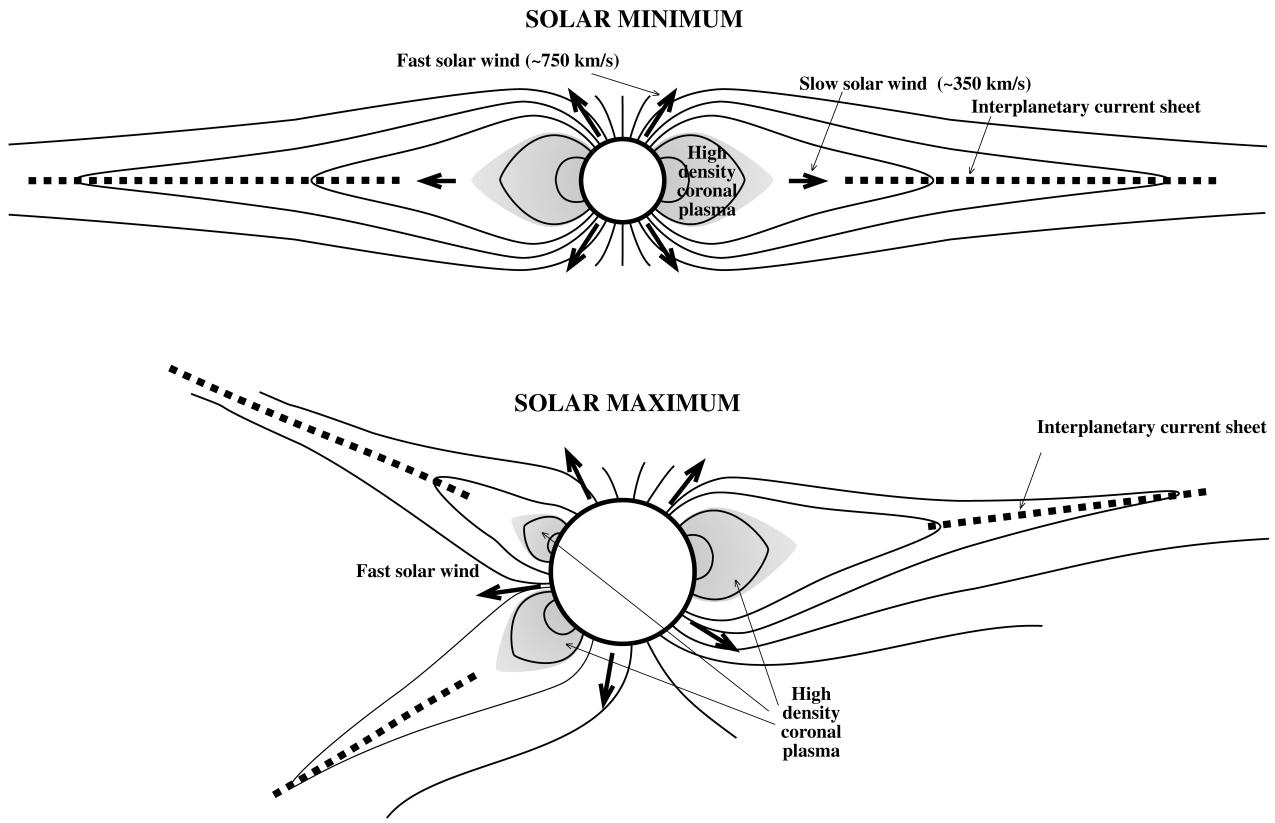


FIGURE 3 Sketch of coronal magnetic field and plasma structure during solar minimum (upper panel) and solar maximum (lower panel).

To a first approximation, the magnetic field near the visible solar surface can be regarded as a dipole like that of the earth. This field is carried outward into interplanetary space from the sun by the solar wind, giving a solar magnetic field configuration (sketched in a plane perpendicular to the ecliptic plane in the upper panel of Fig. 3) which is like a dipole near the sun, but is highly stretched away from the sun. At radial distances of more than a few solar radii, the stretched solar magnetic field reverses direction across a narrow region near the ecliptic plane forming a current sheet surrounding the sun. In the regions where the magnetic field lines are approximately dipolar and do not extend well away from the sun, plasma accumulates giving regions of high-density coronal plasma as illustrated in Fig. 3. Such regions are clearly visible in images of the solar corona. For the magnetic field configuration shown in the upper panel of Fig. 3, a region of high-density corona extends around the sun in the vicinity of the equator. Free escape of plasma only occurs well within the regions where magnetic field lines extend large distances into interplanetary space, where the solar wind speed reaches ~750 km/sec. Near the boundaries between the approximately dipolar field lines and the field lines that extent out large distances, the solar wind escapes, but with a lower average speed of ~350 km/sec. The earth, which is near the ecliptic plane, is exposed more often to this slow solar wind than to the fast solar wind that covers most of the region away from the ecliptic plane.

The magnetic field and solar wind flow configuration illustrated in the upper panel of Fig. 3 corresponds to periods when the solar magnetic field is more dipolar than at other times. During such periods, which are referred to as solar minima, magnetic and sunspot activity on the sun is low. Solar minima occur every 11 years. After solar minimum, the dipolar magnetic field structure of the sun gradually is destroyed. This process takes a few years, leaving the solar magnetic field in a much more disorganized state, as illustrated in a plane perpendicular to the ecliptic plane in the bottom panel of Fig. 3. Magnetic and sunspot activity on the sun is high during these periods, which are referred as solar maxima. At solar maximum, localized regions of magnetic field lines that return to the solar surface without extending far into space and contain high coronal densities can occur over almost any portion of the sun. After solar maximum, the dipole field of the sun returns with the direction of the dipole reversed from what is was during the previous solar minimum. This gives an 11-year solar cycle from one solar minimum to the next,

with the direction of the magnetic field reversing every cycle. While the solar cycle is 11 years, the total period for this magnetic field variation of the sun is 22 years.

B. Solar Wind Disturbances and Their Relation to Solar Magnetic Structure

Both the dynamic pressure of the solar wind and the interplanetary magnetic field (IMF) are important for the interaction of the solar wind with the earth's magnetosphere, and both of these are generally quite variable. In addition to a variety of wave phenomena, there are two types of large-scale disturbances of the solar wind plasma that are known to have large effects on the magnetosphere. The first type is related to the shape of the interplanetary current sheet. Rather than being precisely a disc within the ecliptic plane, the current sheet is generally tilted somewhat with respect to the ecliptic plane and also has a wavy structure as function of azimuthal angle around the sun. Within azimuthal regions where the tilt and/or wavy structure displaces the current sheet sufficiently far from the ecliptic plane, fast solar wind can be emitted near the ecliptic plane. Due to the rotation of the sun (with an ~27-day period) locations near the ecliptic plane can thus be exposed to periods of slow solar wind followed by fast solar wind. This is illustrated in the ecliptic plane in the upper panel of Fig. 4. In this plane, the solar rotation also imparts a spiral shape to magnetic field lines. Within azimuthal regions where fast solar wind follows slow wind, the fast solar wind will catch up with the slow solar wind. The interaction of the fast solar wind with the slow solar wind (referred to as "stream–stream interactions") causes a compression of the solar wind plasma and magnetic field within the interface region between the fast and slow streams. Such compressions give regions of greatly enhanced solar wind densities and magnetic field strengths which can significantly affect the magnetosphere. These stream–stream interactions are particularly important during the period after solar maximum, but before the next solar minimum, when the solar dipolar magnetic field is reforming.

The second large-scale disturbance of the solar wind plasma is associated with the localized high coronal density regions that form during solar maximum. These regions can become buoyant and break away from the sun, carrying the high-density coronal plasma and associated magnetic fields radially away from the sun as illustrated in the bottom panel of Fig. 4. These ejections of coronal material, referred to as coronal mass ejections, can have dramatic affects on the earth's magnetosphere.

As discussed later, the component of the IMF directed parallel to the earth's magnetic dipole (which is directed from the northern polar cap to the southern polar cap), rather than the total interplanetary magnetic field magnitude, is most important for activity within the earth's magnetosphere. This component of the magnetic field is referred to as the southward component. Thus the magnetic field enhancements associated with stream–stream interactions and coronal mass ejections most strongly affect the magnetosphere when the enhanced magnetic field happens to be directed southward.

IV. SOLAR WIND AND INTERPLANETARY FIELD INTERACTIONS WITH THE GEOMAGNETIC FIELD

A. The Magnetopause

The solar wind can be generally viewed as highly conducting, so that to a first approximation the interplanetary and geomagnetic fields do not mix. This requires that the solar wind be diverted around a cavity that has an outer boundary which approximately separates the geomagnetic and interplanetary fields. This boundary is referred as the magnetopause, which is a current sheet of appropriate intensity to separate the interplanetary and geomagnetic fields.

The location of the magnetopause can be estimated by balancing the dynamic pressure of the incoming solar wind ($n_{sw}m_pV_{sw}^2$, where n_{sw} is the solar wind density, m_p is the proton mass, and V_{sw} is the solar wind speed) with the pressure of the geomagnetic field, giving

$$n_{sw}m_pV_{sw}^2\cos^2\psi = B_{in}^2/2\mu_0. \tag{6}$$

This neglects the interplanetary magnetic pressure, which is generally a reasonable assumption. In (6), B_{in} is the magnetic field just inside the magnetopause, ψ is the angle between the solar wind velocity vector and the normal to the magnetopause, and the pressure of the IMF and of the magnetospheric plasma are neglected. For typical solar wind parameters ($n_{sw} = 5 \times 10^6 \mathrm{m}^{-3}$; $V_{sw} = 400$ km/sec), and a dipole geomagnetic field, (6) places the noon, equatorial magnetopause (a location referred to as the "nose" of the magnetosphere) at a distance $r = 10R_E$ from the center of the earth, which agrees very well with its average observed position. [The earth's dipole field strength is $3.1 \times 10^{-5}/r^3$ T. However, twice this value is typically used for B_{in} in Eq. (6) in order to include contributions from the magnetopause current.] Equation (6) also shows that the location of the magnetopause moves further from the earth with increasing distance from the nose. This is because $\cos^2\psi$ decreases away from the nose so that B_{in} must also decrease.

If only the earth's dipole field and the magnetic fields of the magnetopause current sheet were included, the magnetopause would not extend significantly tailward of the earth, and there would not be a magnetospheric tail.

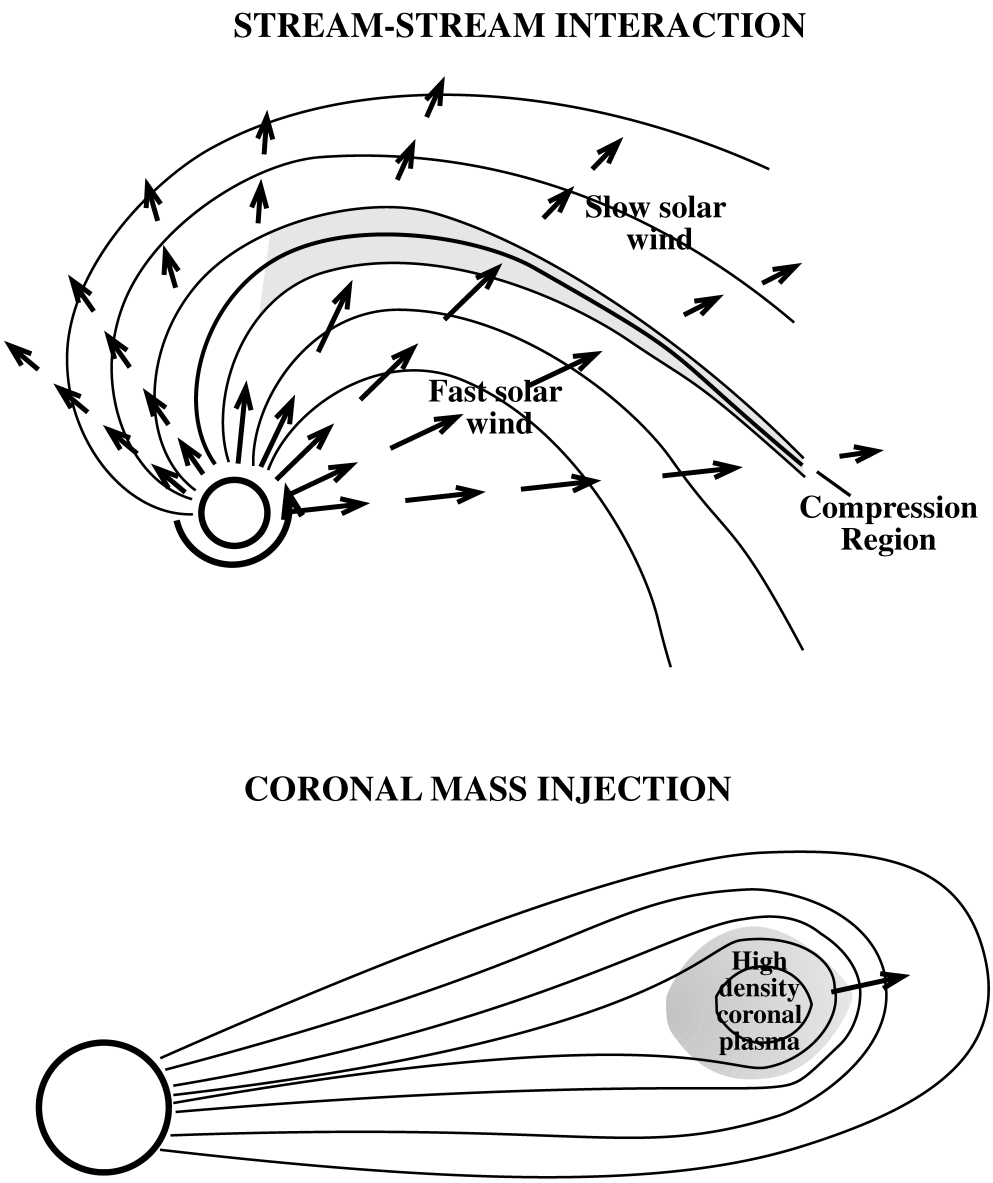


FIGURE 4 Sketch of two important large-scale disturbances of the solar wind plasma, stream–stream interactions (upper panel) and coronal mass ejections (lower panel).

However, the large plasma pressures of the tail plasma sheet form the cross-tail current sheet which is identified in Fig. 1, and the magnetic field of this current allows the tail magnetopause to extend up to several hundred R_E away from the earth in the anti-sunward direction.

On the dayside, B_{in} varies as r^{-3}. With this variation, Eq. (6) shows that the distance to the dayside magnetopause is proportional to $(n_{sw})^{1/6}$. Solar wind densities can be quite variable, and a large solar wind disturbance can increase n_{sw} to $\sim 50 \times 10^6\ \mathrm{m}^{-3}$. A disturbance of this magnitude compresses the magnetosphere significantly, and brings the nose of the magnetosphere to $r < 7R_E$, which is well earthward of its average position. Such an earthward displacement of the magnetopause corresponds to more than a factor of three increase in the magnitude of the magnetopause current. This illustrates one important way in which the solar wind disturbances mentioned in the previous section can cause large dynamic changes to the earth's magnetosphere.

B. Closed and Open Field Lines

The above pressure calculation, with the inclusion of the earth's dipole magnetic field and of the magnetic fields

from the magnetopause and cross-tail current sheets, gives an accurate description of the shape of the magnetosphere, which is illustrated in Fig. 1. Because the calculation assumes that the interplanetary and geomagnetic fields do not mix, the calculation gives geomagnetic and interplanetary magnetic fields that are parallel to the magnetopause at all locations directly adjacent to the magnetopause. However, this is not strictly valid because the magnetopause current sheet has large, but finite, conductivity, which allows a small portion (∼10–20%) of the IMF to cross the magnetopause and connect with the geomagnetic field. Such penetration of the IMF into the magnetosphere connects the interplanetary and geomagnetic fields and is critical to magnetospheric dynamics, though it does not significantly affect the shape of the magnetosphere.

The connection of the interplanetary and geomagnetic fields is illustrated in Fig. 1 for an IMF that is directed primarily southward (i.e., nearly parallel to the earth's magnetic dipole). The figure shows how the penetration of a small portion of the interplanetary field into the magnetosphere modifies the geometry of magnetic field lines emanating from the polar regions of the earth. Without a penetrating field, all magnetic field lines would leave the earth and return to the earth after crossing the equatorial plane. Such field lines are referred to as "closed." With a penetrating field, closed magnetic field lines do not extent all the way to the magnetic pole. Instead there is an approximately circular region centered near each magnetic pole where field lines cross the magnetopause and enter interplanetary space. Such polar-cap field lines are referred to as "open." For the earth, the boundary between open and closed field lines is at ∼73° magnetic latitude. Open field lines allow for a tapping of energy directly from the flowing solar wind plasma, and such energy drives a wide variety of phenomena within the magnetosphere.

C. Mapping of Interplanetary Electric Field into the Magnetosphere

The solar wind flows radially outward from the sun carrying the IMF with it. In general the magnetic field is not parallel to the solar wind, so that there is an electric field in interplanetary space that is related to the solar wind and IMF by

$$\mathbf{E} = -\mathbf{V}_{sw} \times \mathbf{B}. \tag{7}$$

Since particles can move easily along magnetic field lines, it is generally appropriate to assume that magnetic field lines are so highly conducting that there can be no electric fields parallel to the magnetic field lines. This assumption implies that magnetic field lines are equipotentials so that the interplanetary electric field given by (7) maps along open polar-cap field lines through the magnetosphere down to the ionosphere. (The ionosphere is a region of ionized upper-atmospheric constituents that surrounds the earth at altitudes between ∼75 and ∼500 km. For the purposes here, the ionosphere is at the low-altitude ends of the magnetic field lines shown in Fig. 1 and marks the lowest altitude to which the interplanetary electric field has significant effects.) For the orientation of the IMF shown in Fig. 1, the mapping of the interplanetary electric field into the magnetosphere gives an electric field throughout the open field line region of the magnetosphere that points from the dawn side of the magnetosphere toward the dusk. For other orientations of the IMF, the mapping into the magnetosphere is similar, but the orientation of the electric field can have some differences from that shown in the figure.

Figure 5 shows the mapping of the interplanetary electric field into the magnetosphere along open field lines in the dawn–dusk meridian plane, where the coordinate system used has x directed from the earth to the sun, y directed from the dawn to the dusk side of the earth, and z directed from the south to the north magnetic pole. This mapping gives an anti-sunward flow of plasma all along the open field region of the polar caps. At the boundary between open and close magnetic field lines, the mapped interplanetary electric field and the anti-sunward flow terminate. This boundary thus becomes charged as indicated in Fig. 5, giving an electric field that extends into the closed field line region of the magnetosphere. This electric field is oriented so as to give sunward flow on closed field lines.

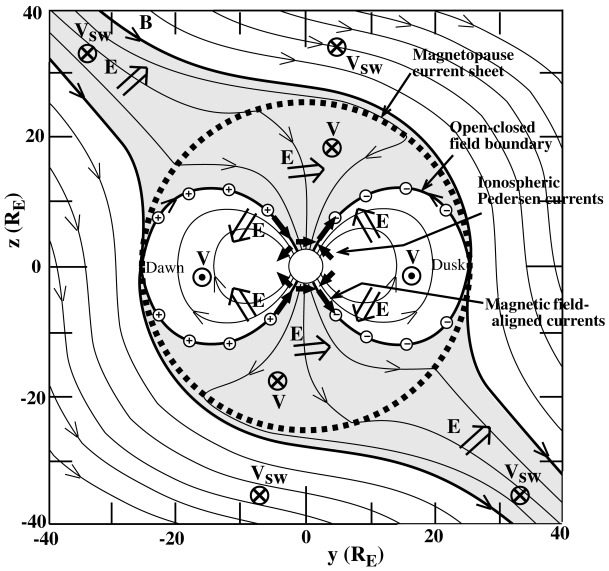


FIGURE 5 Schematic illustration of the magnetosphere in the dawn–dusk meridian plane. Plus and minus signs indicate charges along the boundary between open and closed magnetic field lines, and shading identifies the region of open, polar-cap field lines.

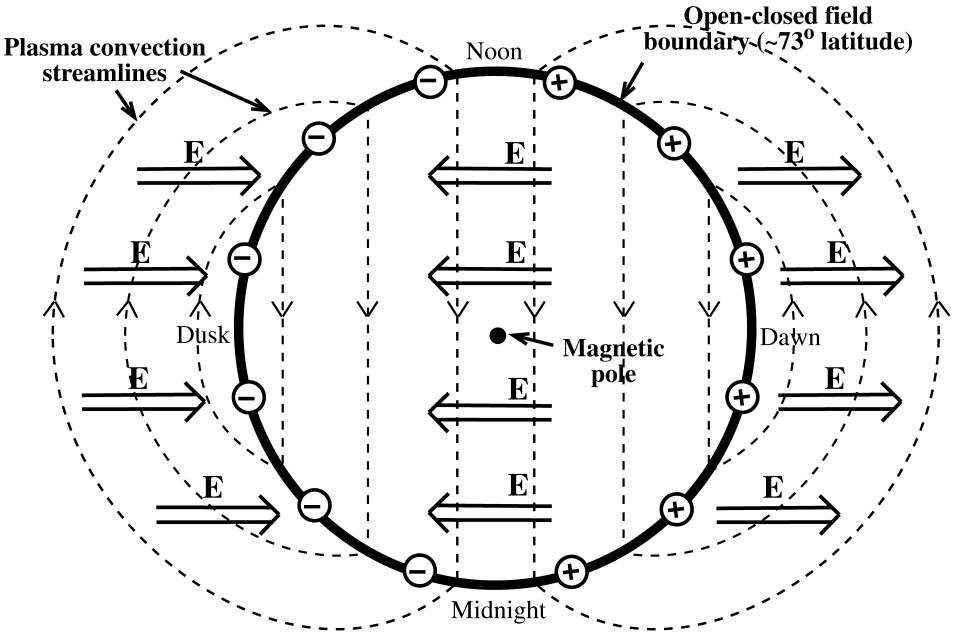


FIGURE 6 Electric fields and plasma flows driven by the mapping of the interplanetary electric field into the magnetosphere as seen looking down onto the ionosphere from above one of the polar caps.

The total electric field pattern drives a three-dimensional circulation of plasma, which is illustrated in the noon–midnight meridian plane in Fig. 1. This flow is referred to as magnetospheric convection. The x component of the flow is anti-sunward across open, polar-cap field lines and the flow returns in the sunward direction within the closed field line region. The flow also moves poleward on the dayside and equatorward within the tail, completing the convective circulation. This magnetospheric convection pattern is often viewed by looking at the electric field, or equivalently the plasma flow, since the two are related by Eq. (1), as mapped to the ionosphere. Such a mapping, illustrated in Fig. 6, gives a complete picture of magnetospheric convection because magnetic field lines are approximately equipotentials. Figure 6 shows electric fields and plasma flow streamlines as seen looking down onto the ionosphere from above one of the polar caps. The figure shows how the flow moves in the anti-sunward direction over the polar caps, crosses the boundary between open and closed field lines, and returns to the dayside within the closed field line region.

The strength of magnetospheric electric fields and the resulting convection depends upon the magnitude of the interplanetary electric field, which varies with the magnitude of the y and z, components of the IMF and with the solar wind speed. Generally, variations in the IMF are much greater than are variations of solar wind speed, so that variations in the IMF are generally more important in modifying the strength of convection. The efficiency of the mapping of the interplanetary electric field into the magnetosphere also depends significantly on the orientation of the IMF, the efficiency increasing as the orientation becomes increasingly southward (i.e., increasing toward the negative-z direction). Thus variations in the z component of the IMF when this component is negative (and thus antiparallel to the equatorial magnetospheric field) have the largest effects on the strength of convection, though variations in the magnitude of the y component are also important.

V. PARTICLE ACCESS TO, AND TRANSPORT WITHIN, THE MAGNETOSPHERE

A. Access

Particles within the magnetosphere come from both the solar wind and the ionosphere. While ionospheric particles make important contributions, the solar wind source generally dominates throughout most regions of the magnetosphere.

The entry of solar wind particles and their transport to and within the tail plasma sheet is illustrated in Fig. 7. Because the geomagnetic and interplanetary magnetic fields are connected, particles with a finite $v_{\parallel}$ are able to flow across the magnetopause. This is most effective across the dayside magnetopause. There some of the solar wind particles, which have been heated after crossing the bow

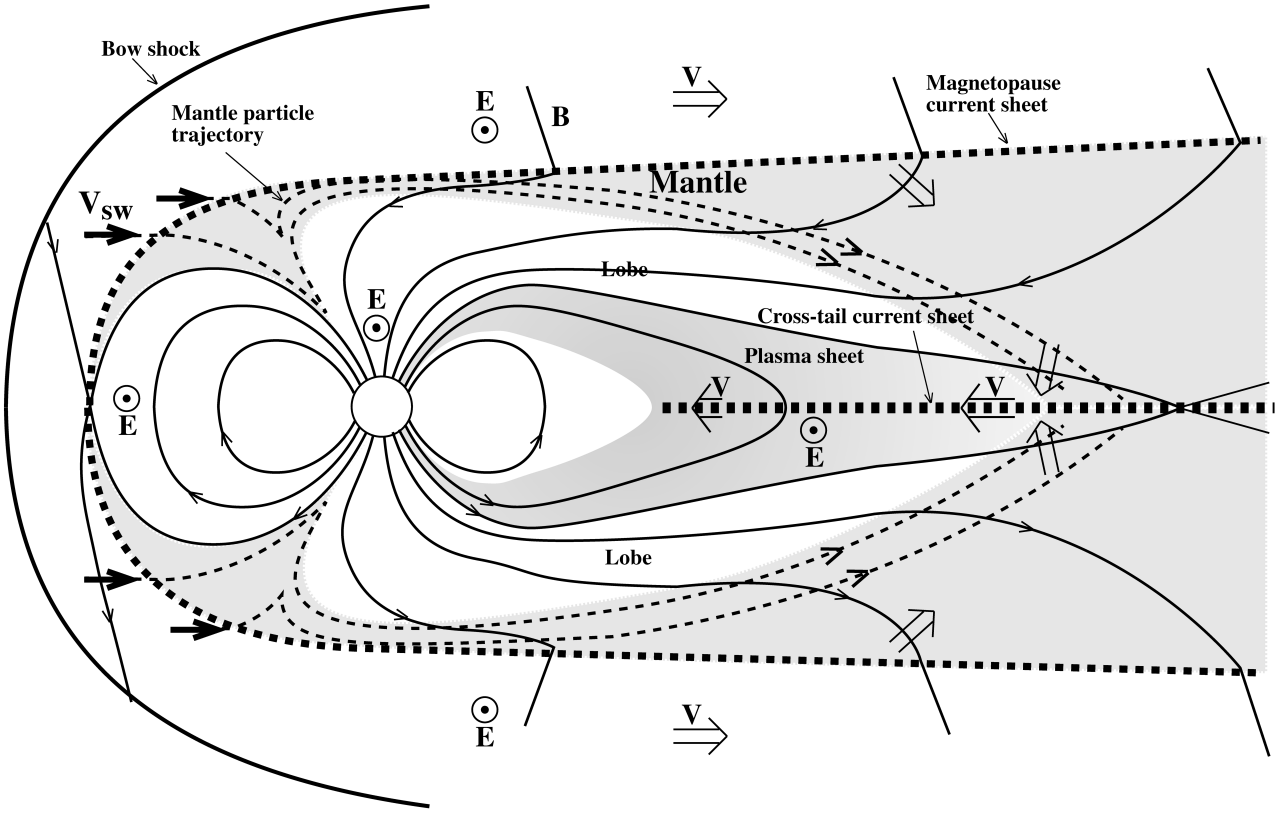


FIGURE 7 Illustration of the entry of solar wind particles into the magnetosphere and their transport to and within the tail plasma sheet.

shock, cross the magnetopause and enter the portion of the open field line region that is near noon. These particles initially flow primarily along magnetic field lines toward the ionosphere, but they also drift slightly poleward across magnetic field lines due to the magnetospheric electric field. As the particles flow toward the ionosphere, the magnetic field strength strongly increases, so that conservation of μ requires that parallel energy is converted into perpendicular energy. For the vast majority of particles, all parallel energy is lost before the particles reach the ionosphere. The parallel component of velocity for these particles then reverses, and the particles move along field lines in the direction away from the earth. At the same time' the particles also continue to move poleward due to the electric field drift across field lines. The reversal of the particles' parallel velocity as they flow toward increasing B is referred to as "magnetic mirroring."

B. Transport to and within the Plasma Sheet

After mirroring, these solar wind particles form a magnetospheric layer that lies adjacent to the magnetopause that is referred to as the "mantle." Mantle particles flow away from the earth along open field lines of the magnetospheric tail. As they flow down the tail, the electric field across the tail causes them to also drift toward the plasma sheet. This causes a significant fraction of the particles to cross the boundary between open and closed magnetic field lines and enter the region of the distant plasma sheet. When they reach the cross-tail current sheet, they are significantly energized by the cross-tail electric field and become an important contributor to the energetic particle population of the plasma sheet. The particles are then carried earthward by electric field drift, gaining energy as they move earthward into regions of increasing magnetic field strength. This inward motion continues until the spatial variation in magnetic field becomes sufficiently strong to deflect the particles around the earth toward the dayside portion of the plasma sheet, which is identified in Fig. 1. This deflection forms the inner edge of the nightside plasma sheet, which typically lies at $r \approx 5$–$12 R_E$.

C. Formation of the Radiation Belts and Stormtime Ring Current

The location of the inner edge of the plasma sheet depends on the strength of convection as well as on the strength of magnetic drift. As the strength of convection increases,

particles are able to move closer to the earth before being deflected around the earth by magnetic drift. This leads to significant temporal variations in the location of the inner edge of the plasma sheet, variations which are very important components of geomagnetic activity. Occasionally, when there is a very large negative-z component of the IMF, convection becomes so strong that particles convect into the $r \approx 2$–$5 R_E$ region of the magnetosphere before being deflected around the earth. Particles that reach this region of high magnetic fields gain significantly more energy than normally occurs and cause significant increases in particle intensities in the region labeled "radiation belts and ring current" in Fig. 1. When the strength of convection reduces back to normal, particles left behind in the $r \approx 2$–$5 R_E$ region begin to move in complete circles around the earth and become part of the radiation belts. The current carried by these particles as they circle the earth, ions in one direction and electrons in the other, increases with the number of energetic particles within this region. Periods when this ring current is sufficiently strong are referred to as magnetic "storms." The ring current during storms causes significant magnetic field changes on the surface of the earth at low and mid-latitudes, and these changes are the primary means by which storms are identified and monitored.

Particles also have access to the region of the ring current during periods of weaker convection as a result of fluctuations of the convection electric field. Resonant interactions between these fluctuations and the azimuthal drift of particles around the earth give small perturbations in the radial position of individual particles. The sum of many of these perturbations can viewed as diffusion in the radial position, and the balance between this "radial diffusion" and particle losses (precipitation to the upper atmosphere and, for positive ions, charge exchange with neutral hydrogen which extends from the upper atmosphere into the region of the radiation belts) forms a permanent distribution of energetic particles within the radiation belts. The discovery of the radiation belts by James Van Allen and colleagues in 1958 using instrumentation on-board the first two successfully launched U.S. satellites received widespread national and international attention, and the radiation belts became popularly known as the Van Allen radiation belts.

VI. AURORAS AND AURORAL CURRENTS

Most visible auroras are formed by the precipitation of magnetospheric electrons into the atmosphere. Such auroras can be divided into two general classes. The first is diffuse auroras, which are formed primarily by the direct loss of electrons by precipitation into the atmosphere. Diffuse auroras tend to be broad in latitudinal extend and to not have strong spatial structure. Such auroras are thus generally visually unimpressive. Discrete aurora, on the other hand, result from the precipitation into the atmosphere of electrons which have been energized as they moved toward the ionosphere by electric fields aligned parallel to the magnetic field. The "field-aligned" electric fields responsible for this energization are associated with currents flowing upward from the ionosphere to the magnetosphere. Discrete auroral displays can be intense and dynamic and are generally the most dramatic type of aurora.

Discrete auroras occur because the magnetospheric electric field driven by the solar wind maps to the ionosphere, and the ionosphere is a good conductor of current. The relation between $\mathbf{E}$ and $\mathbf{V}$ given by (1), which would not allow for differential motion between electrons and ions within the horizontal plane of the ionosphere, only holds in the absence of collisions. In the lower regions of the ionosphere, between about 100 and 150 km in altitude, collisions with atmospheric neutral constituents disrupt the electric field drift of ions but do not significantly disrupt the electric field drift of electrons. As a result, the mapping of magnetospheric electric fields to the ionosphere gives rise to currents in the horizontal plane of the ionosphere. These horizontal currents have components parallel and perpendicular to the applied electric field, which are referred to a Pedersen and Hall currents, respectively. Hall currents generally flow along closed paths within the ionosphere and are responsible for significant magnetic perturbations on the ground that are observable from within the auroral oval and the polar caps. Pedersen currents, on the other hand, generally have regions of strong convergence and divergence. Because of the requirement for current continuity, regions of convergence (divergence) of horizontal ionospheric currents are connected to currents that flow along magnetic field lines to (from) the magnetosphere from (to) the ionosphere. These field-aligned currents are an important aspect of coupling that occurs between the magnetosphere and the ionosphere, and they are important for the formation of the auroral arcs.

Large-scale coupling between the magnetosphere and ionosphere that leads to significant field-aligned currents is illustrated by the heavy filled arrows in Fig. 5. The arrows parallel to the earth's surface indicate ionospheric Pedersen currents flowing parallel to the direction of the ionospheric mapping of the magnetospheric electric field. These currents converge on the dusk side of the polar cap regions of open field lines and diverge on the dawn side of these regions. This gives a large-scale field-aligned current system which is upward on the dusk side and downward on the dawn side. This current system extends

along the boundary between open and closed magnetic field lines to all local times, as indicated by the converging and diverging electric fields along the open–closed field line boundary in Fig. 6. The field-aligned currents are upward where the ionospheric electric fields converge and the open-closed field line boundary is negatively charged and downward where the ionospheric electric fields diverge and the boundary is positively charged. Another field-aligned current system lies near the inner edge of the plasma sheet. This current system is oppositely directed from the one near the open–closed field line boundary. There are also a variety of smaller scale field-aligned currents, some of which occur within the nightside plasma sheet and are an important component of geomagnetic activity.

The field-aligned electric fields which are responsible for discrete auroras occur where the convergence of horizontal ionospheric currents gives rise to field-aligned currents flowing upward out of the ionosphere that are too large to be carried by the precipitation of electrons that cause the diffuse aurora. The field-aligned electric fields enhance the upward field-aligned current by increasing the number of downgoing electrons which reach the upper atmosphere before mirroring. Field-aligned electric fields are generally less important in regions of downward field-aligned currents than in regions of upward field-aligned currents because downward currents can readily be carried by ionospheric electrons moving from the ionosphere to the magnetosphere. (The ionospheric ion contribution to upward field-aligned currents is generally not as large because the heavy mass of ions limits the rate at which ions can be extracted from the ionosphere.)

Most of the aurora within the auroral oval is diffuse aurora. Regions where upward field-aligned currents within the aurora become large enough for the development of discrete aurora include the large-scale upward-aligned currents that lie on the dusk side of the polar caps near the boundary between open and closed field lines. There is nearly always at least some discrete auroral activity along this boundary. Upward currents also become large enough for the formation of the discrete aurora within smaller scale regions of the nightside plasma sheet in association with geomagnetic activity. The large-scale field-aligned current system which lies near the inner edge of the plasma sheet is generally not sufficiently intense for the formation of very much discrete aurora.

VII. GEOMAGNETIC DISTURBANCES

Transient enhancements of auroral emissions and ionospheric currents often occur within the auroral ovals and are good indicators of geomagnetic activity. These disturbances occur along magnetic field lines that extend to the plasma sheet, and they are observable at high latitudes via intense auroral activity and significant ground magnetic field perturbations associated with enhanced ionospheric currents. There are different types of such disturbances, having time scales ranging from a few minutes to 1 hr or so, and they occur within the ionospheric extension of the plasma sheet. Magnetic storms are a fundamentally different phenomenon from these disturbances within the auroral oval. They occur when fluxes of energetic particles within the region of the radiation belts cause a significantly enhanced ring current, leading to magnetic field depressions at the earth's surface at latitudes equatorward of the auroral oval. Auroral oval disturbances often occur during magnetic storms, but they are not directly related to the injection of particles into the radiation belts that leads to the formation of the stormtime ring current.

Auroral oval disturbances are related to the energy and dynamics of the plasma sheet, which are highly variable and depend strongly on the solar wind dynamic pressure and the IMF. Solar wind dynamic pressure exerts control by compressing the entire magnetosphere, including the tail plasma sheet. The IMF affects the plasma sheet by controlling the strength of convection. The strength of convection strongly affects both the heating of particles within the cross-tail current sheet and the earthward penetration of the plasma sheet. When convection is enhanced, the inner edge of the plasma sheet moves earthward. This corresponds to an equatorward motion of the equatorward boundary of the plasma sheet as mapped to the ionosphere. This leads to an increase in the latitudinal width of the ionospheric mapping of then plasma sheet, and thus to an increase in the latitudinal width of the auroral oval. In addition, significant enhancement and earthward penetration of the cross-tail current occurs when convection is enhanced.

Three different types of auroral oval disturbances have been identified that are related to large-scale disturbances of the magnetosphere–ionosphere system: poleward boundary intensifications (PBIs), substorms, and effects of solar wind dynamic pressure enhancements referred to here as "dynamic pressure disturbances." Each of these types of disturbance has unique characteristics and reflects distinctly different physical processes occurring within the magnetosphere. The signatures of each within the auroral oval are illustrated in Fig. 8. In that figure, lightly shaded regions indicate regions of undisturbed auroral emissions, darkly shaded regions indicate regions with strong discrete aurora, and regions with medium shading indicate regions of enhanced diffuse aurora which may contain some discrete auroral features.

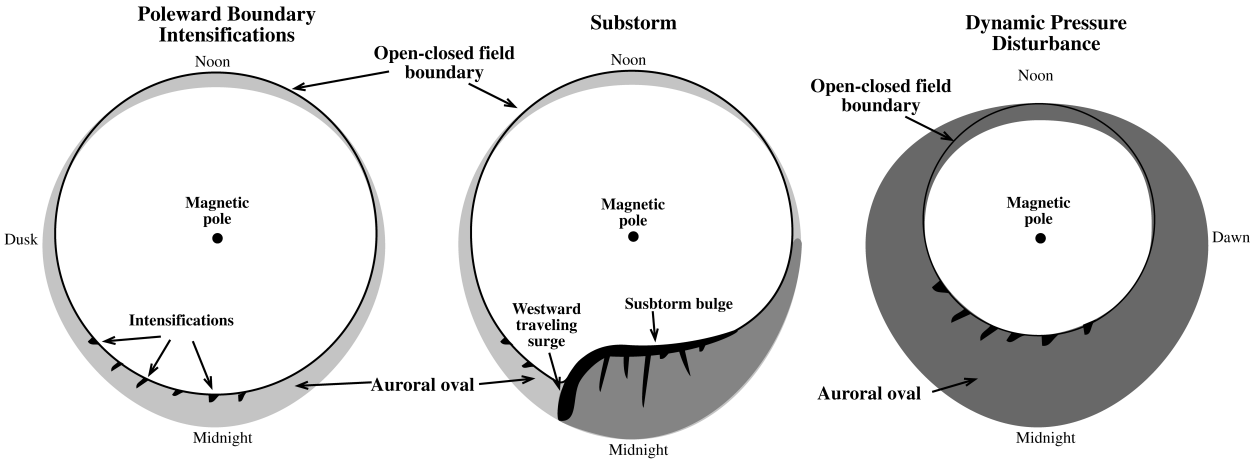


FIGURE 8 Illustration of the auroral oval signatures of the three different types of auroral oval disturbance discussed in the text.

A. Poleward Boundary Intensifications

The most common type of auroral-zone disturbance is the PBI. PBIs occur repetitively with a period on the order of 10 min. They can occur independently from other types of disturbances, though their intensity and frequency of occurrence tend to increase with the strength of convection. They have an auroral signature that often can be seen to move equatorward from the poleward boundary of the auroral oval, which, on the nightside, lies very near the boundary between open and closed fields. There can be several such disturbances within the nightside auroral oval at one time, disturbances typically occurring from near dusk to 1–2 hr past midnight. They also extend varying distances through the auroral oval, some staying confined to very near the polar-cap boundary and others extending through a large portion of the auroral oval and becoming elongated in the north–south direction. PBIs are typically associated with ground magnetic perturbations of a few tens of nanoteslas, but perturbations can be as high as ∼500 nT during periods of strongly enhanced convection. The time scale for individual intensifications is typically a few minutes.

Individual PBIs are longitudinally localized and are associated with longitudinally localized bursts (about a few minutes in duration) of enhanced plasma flow that are often observed within the tail plasma sheet. Such flow bursts transport significant mass and energy within the plasma sheet and are thus an important component of the dynamics of the plasma sheet. The flow bursts extend only a small distance across the tail. They lead to enhance auroral emissions because they are associated via Eq. (1) with localized enhancements in electric fields across the tail. The mapping of these electric fields to the ionosphere results in longitudinally localized regions of enhanced dawn-to-dusk-directed electric fields, which on their western edge give rise to converging Pedersen currents. These converging ionospheric currents are connected to upward field-aligned currents which are often sufficiently strong to lead to the formation of the discrete auroral forms which are observed as PBIs. Sometimes individual PBI structures observed at low altitudes traverse essentially the entire latitudinal extent of the plasma sheet, which would correspond to flow bursts that extend from the distant tail plasma sheet (∼50–100R_E) all the way to the vicinity of synchronous orbit.

B. Substorms

Substorms are a far more dramatic and large-scale, but far less common, disturbance than PBIs. The substorm occurrence rate is highly variable, but there are typically several per day. Auroral activity during substorms typically initiates within a local time sector of ∼1–2 hr near the equatorward boundary of the nightside auroral oval and then expands both poleward and azimuthally. Very intense discrete aurora lies along the poleward and westward boundaries of this expanding region. The poleward expansion can bring strong aurora well into the region which is normally occupied by the polar cap, forming what is known as the auroral "bulge." The westward-expanding region of strong discrete aurora is referred as the "westward traveling surge" and can continue for up to ∼30 min after the substorm onset. Ground magnetic disturbances associated with substorms are typically a few hundred nanoteslas, but can range from ∼50 to ∼2000 nT.

Substorm onsets are preceded by a $\gtrsim$30-min growth-phase period of enhanced convection that is typically associated with a moderate to large southward-directed IMF. The onsets are often associated with the impact on the

magnetosphere of an IMF change, such as a reduction in the southward component of the IMF, that cause a reduction in the strength of convection. During the growth phase, the plasma sheet moves earthward and particle energization increases, leading to an increase in the cross-tail current. This increase is particularly strong within the earthward portion of the plasma sheet. In addition to being associated with auroral activity, the expansion phase is associated with a reduction in strength of the cross-tail current and a displacement of the inner edge of the plasma sheet and the cross-tail current away from the earth, and thus with a large release of energy from the inner portion edge of the plasma sheet.

The reduction of cross-tail current during a substorm is initially localized and does not extend across a large distance of the tail. As with auroras at low altitudes, the current reduction region expands azimuthally within the plasma sheet with time after substorm onset. The edges of the current reduction region are connected to field-aligned currents extending to the ionosphere. These currents are upward and large on the dusk side of the current-reduction region. The large upward currents connect along magnetic field lines to the strong discrete aurora that forms the westward traveling surge, and their azimuthal motion within the plasma sheet corresponds to the westward motion of the surge.

While many substorms occur in response to IMF-driven reductions in the strength of convection, and thus can be viewed as being triggered by appropriate IMF changes, the extent to which substorms result from such convection reductions has not yet been determined. There are various ideas of how substorms might result from a large-scale, internal instability of the plasma sheet during periods of steady, enhanced convection. It is known that substorms are infrequent during periods of steady enhanced convection. However, it has not yet been determined whether or not substorms are absent during such periods. If they are absent, then the idea of substorm onset by internal instability will have to be discarded. If not, then it will be necessary to determine why some substorms require appropriate IMF changes and some do not.

C. Dynamic Pressure Disturbances

It has recently been found that enhancements in solar wind dynamic pressure increases can cause large auroral zone disturbances during periods of strong magnetospheric convection. Dynamic pressure enhancements affect the entire auroral zone simultaneously (unlike substorms and PBIs). Within a few minutes of the time an enhancement in dynamic pressure hits the magnetopause, the poleward boundary of the auroral oval moves poleward and the intensity of auroral emissions increases throughout essentially the entire auroral oval. The equatorward boundary of the oval is not significantly affected, so that the latitudinal width of the auroral oval increases. Most of the increase in auroral intensity is in the diffuse aurora, but increases in discrete aurora most likely also occur. The enhancement in diffuse auroral emissions results from the heating of the plasma sheet plasma as the entire magnetosphere is compressed by the increase in solar wind dynamic pressure. The poleward motion of the poleward boundary of oval can be as much as 10° in latitude, which corresponds to a large broadening of the auroral oval. It also corresponds to a large reduction in the area of open, polar-cap magnetic field lines, since, except near local noon, the poleward boundary of the oval lies at the boundary between open and closed field lines.

Dynamic pressure enhancements compress the entire magnetosphere and enhance the entire magnetospheric current system. As discussed in Section IV.A, increases in the current within the dayside magnetopause can be over a factor of three. It is also known that the global field-aligned current system that lies along the open–closed field-line boundary, the ionospheric current system driven by magnetospheric convection, and the cross-tail current significantly increase in response to enhancements in solar wind dynamic pressure. However, the relationship between the full magnetospheric current system and the solar wind dynamic pressure is not well understood. The cause of the large reduction in the area of the polar cap driven by increases in solar wind dynamic pressure is also not well understood, but it is likely related to the large enhancements in magnetospheric currents.

VIII. CONCLUSIONS

The interaction of the interplanetary plasma with the magnetic field of the earth has only been studied in detail since the beginning of the space age in the late 1950s. With the use of ground measurements of ionospheric phenomena and limited point measurements from spacecraft within the solar wind and the magnetosphere, much is now understood about how this interaction leads to the interesting features and dynamics of the magnetosphere–ionosphere system. This is a remarkable accomplishment. However, much remains to be learned. We do not yet have a full observational description of the magnetosphere in the absence of disturbances or of the various disturbances which occur within the magnetosphere-ionosphere system. We are also far from being able to make quantitatively accurate predictions of how the magnetosphere responds to the highly variable solar wind and IMF which impacts the magnetosphere. Only with continuing observational programs, extensive analysis of existing and future

datasets, and innovative theory and modeling studies will a far more quantitative understanding of the space plasma physics of the earth's environment become possible.

SEE ALSO THE FOLLOWING ARTICLES

AURORA • GEOMAGNETISM • IONOSPHERE • PARTICLE PHYSICS, ELEMENTARY • PLASMA SCIENCE AND ENGINEERING • RADIATION, ATMOSPHERIC • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS

BIBLIOGRAPHY

Crooker, N. U. (2000). "Solar and heliospheric geoeffective disturbances," *J. Atmos. Solar-Terr. Phys.* **62,** 1071.

Hultqvist, B., Oieroset, M., Paschmann, G., and Treumann, R. (eds.). (1999). "Magnetospheric Plasmas Sources and Losses," Kluwer Academic Publishers, Boston.

Kivelson, M. G., and Russell, C. T (eds.). (1995). "Introduction to Space Physics," Cambridge University Press, New York.

Lyons, L. R. (2000). "Geomagnetic disturbances: Characteristics of, distinction between types, and relations to interplanetary conditions," *J. Atmos. Solar-Terr. Phys.* **62,** 1087.

Lyons, L. R., and Williams, D. J. (1984). "Quantitative Aspects of Magnetospheric Physics," Reidel, Dordrecht, Holland.

Richmond, A. D., and Lu, G. (2000). "Upper-atmosphere effects of magnetic storms: A brief tutorial," *J. Atmos. Solar-Terr. Phys.* **62,** 1115.

Schulz, M., and Lanzerotti, L. J. (1974). "Particle Diffusion in the Radiation Belts," Springer-Verlag, New York.

Wang, Y.-M., Sheely, Jr., N. R., Socker, D. G., Howard, R. A., and Rich, N. B. (2000). "The dynamical nature of coronal streamers," *J. Geophys. Res.* **105,** 25,133.