ENCYCLOPEDIA OF

Physical Science and Technology

THIRD EDITION

Astronomy

Astrochemistry

Steven N. Shore
Indiana University South Bend

GLOSSARY

Astrochemistry The theoretical study of chemical processes in cosmic environments and the observational determination of physical parameters through the study of abundances of molecular species. This review concentrates on the recent results concerning circumstellar envelopes and the interstellar medium. The field deals, however, with synthesis of molecules in cometary nuclei and planetary atmospheres, as well as stellar photospheres.

Fractionation Process by which isotopes are included in molecules, either by the process of direct transfer or by charge exchange.

Large-velocity gradient approximation (also *called Sobolev approximation*) the assumption that the line width due to thermal broadening is small compared with the large-scale velocity field in a medium. It is basic to the assumption that the optical depth of a line depends only on the velocity gradient.

Molecular cloud The densest phase of the interstellar medium. Clouds have large size (of order 1–10 pc) and high densities ($\geq 10^3$ cm^{-3}).

Polycyclic aromatic hydrocarbons (PAHs) Hydrocarbons in complex chains and agglomerated rings thought to be responsible for the diffuse emission lines observed in dust nebulae in the near-infrared. These molecules form the lowest-mass end of the dust distribution and are responsible for ubiquitous diffuse emission in the 1- to 25-μm galactic background radiation.

Units Parsec (pc), 3.1×10^{18} cm; solar mass ($M_\odot$), 2×10^{33} g; Jansky (Jy), 10^{-23} erg^{-1} cm^{-2} Hz^{-1}.

Vibronic Transition Molecular transition involving states which are split by rotation that is induced through rotation–vibrational coupling; the fine-structure states of electronic transitions.

I. INTRODUCTION

Astrochemistry is a field that spans virtually all cosmic environments, from comets and planetary atmospheres to the interstellar medium. As such, it is more concerned with processes and what they reveal about the physical nature of the medium than with the specific arena in which these processes occur. It is one of the few astrophysical fields in which laboratory work is possible, and in which conditions similar to those studied astronomically can be simulated. In this review, we shall concentrate on the most recent work, concerned mainly with stellar mass outflows and the interstellar medium. For want of space, planets and stellar photospheres have been excluded.

Observational astrochemistry is accomplished primarily with infrared (IR) and millimeter techniques, areas of technology which have been seen explosive expansion in the past decade. With the improvement of superconducting detectors and the development of interferometric arrays, this field is one which will surely change significantly in the next decade.

We will begin by examining the physical processes needed to diagnose the conditions in astrochemical environments and then examine some of the chemical products thus detected. It is important to keep in mind that, with the exception of solar system objects, astrochemical analyses are quintessentially remote sensing, studied by observations of spectral lines emanating from distant sources through the applications of radiative transfer theory and molecular line formation.

II. PHYSICAL PROCESSES

A. Basic Molecular Spectroscopy

1. Electronic Transitions

The electronic states in a molecule, analogs of the atomic states and characterized by total electronic angular momentum and spin, are generally separated in energy by about the same order of magnitude as for isolated atoms, usually several electron volts (eV). Thus the transitions from molecular electronic states, which also correspond to different potentials, are best observed in the ultraviolet (UV) region. Excitation depends on the presence of UV radiation, and electronic transitions are usually seen in absorption, as in the ${}^1\sum_g^+$ state of H_2. The strength of the transition depends on the dipole (heteronuclear molecules and ions) or the quadrupole (homonuclear) moments. Vibrational states dominate the optical and infrared, and rotational states are best observed in the millimeter and centimeter portions of the spectrum.

For diatomic molecules, these states are classified by the projection of the electronic angular momentum along the internuclear axis, Λ, and the projected combined spin $\sum$ and are grouped into multiplets according to the coupling between these states and the rotational angular momentum J. For $\Lambda \neq 0$, the states are split by $\Omega = \Lambda + \sum$ into fine-structure levels, called Λ-doubling. Interaction with nuclear spins produces hyperfine splitting of the rotational levels, with quantum number F. These hyperfine transitions in OH are responsible for the observed maser emission.

2. Rotational Transitions

The nuclei, which are the massive components of molecules, are free to rotate and precess about the center of mass. Thus, a molecule with a moment of inertia, I, has a rotational angular momentum J which is quantized and thus takes on only discrete values. The energies of these states can be shown to be

$$E_J = \frac{h}{4\pi I} J(J+1) \equiv B_v J(J+1), \tag{1}$$

so that transitions between states of unequal J show a stepladder pattern in the separation of lines of the same series. Because of the large value of the moments of inertia, due to the mass of the nucleus, the separation of these states is small, of order 0.01 eV, increasing with increasing J. The simple representation that has just been used, however, is only appropriate for diatomic species, which have only one rotational axis that is degenerate in the two axes orthogonal to the internuclear axis about the center of mass.

If the molecule is more complex, for example, H_2O, then two or three axes are needed to fully describe the rotation. Each of these has an associated moment of inertia, depending on the details of the electronic states and the internuclear distances and masses. The projection of the rotation along the body axis, K, now appears in the terms for the rotational splitting. For instance, for a symmetric top molecule line CH_3,

$$F_{|v|} = B_{|v|} J(J+1) + \left(B_{|v|} - A_{|v|}\right) K^2 \pm 2A_{|v|}\zeta K, \tag{2}$$

where ζ describes the coupling of the vibrations and the rotations (vibronic states) and splits the otherwise degenerate K levels. The constants A and B are the moments of inertia about the parallel and perpendicular axes of rotation, relative to the body axis. The rotational lines will be distributed in a way that depends on the ratio of the moments of inertia of the principal axes of the molecule. Thus there will be multiplets for lines which are closely spaced in energy and which can be strongly radiatively coupled (see discussion of masers).

3. Vibrational Transitions

Vibrational states distribute as those of a harmonic oscillator, with

$$E_v = h\nu_0\left(v + \tfrac{1}{2}\right), \tag{3}$$

where ν_0 is the vibrational frequency of the ground state. Polyatomic molecules have additional vibrational states due to the multiple modes presented by different configurations. For instance, bending modes in water are states for which the O moves and the H remains fixed, while others have both the O and H moving oppositely (the so-called ν_2 and ν_3 modes), in addition to the fundamental ν_1 mode which involves the O–H bond stretch. Vibrational coupling produces an angular momentum which splits rotational states, as mentioned above under vibronic transitions.

Coupling of vibrational and electronic states—that is, between Λ and v—produces an angular momentum K, which for polyatomic molecules depends on the axis about which the rotation is executed. For instance, in H_2O there is a prolate and oblate rotational axis for the molecule, so that a state J is split by two values of K and labeled by $J_{K_+K_-}$. More complex states are possible, depending on the complexity of the molecule. For example, inversion transitions of molecules such as NH_3, which occur at centimeter wavelengths, result from small perturbations of rotational states by vibrational transition between mirror molecular conformations.

B. Radiative Transfer: Observational Astrochemistry

1. Line Radiative Transfer

Molecular observations are almost always concerned with specific discrete transitions. These are generally observed at millimeter or centimeter wavelengths. The intensity of a source is determined by the rate of collisional versus radiative transitions between levels. Because of the extremely low densities usually associated with molecular environments, whether in a circumstellar envelope or a molecular cloud, pressure broadening is unimportant. Instead, the molecule radiates at its local velocity into the line of sight. This dispersion of velocity may be due strictly to the thermal motions of the particles, or it may be due to the presence of turbulence or large-scale chaotic motions within the medium. Either way, the local profile, $\phi(\nu)$ is a Gaussian with a finite width in frequency.

The absorption coefficient for a state can be written as

$$\kappa_\nu = (n_1 B_{12} - n_2 B_{21})\phi(\nu), \tag{4}$$

where n_j is the population of the upper (n_2) or lower (n_1) state, and B_{12} and B_{21} are the Einstein transition probabilities for stimulated transitions. The emission coefficient, due to spontaneous transitions, is given by

$$j_\nu = n_2 A_{21}\phi(\nu), \tag{5}$$

where A_{21} is the Einstein spontaneous transition probability and ϕ_ν is the line profile function that describes the frequency dependence of the line. In its most general form, ϕ_ν is the convolution of the intrinsic line profile due to radiative and collisional broadening of the upper and lower states (usually a Lorentzian profile) and the extrinsic broadening due to the random motions of the molecules (a Gaussian profile whose width depends on the thermal and turbulent velocities added quadratically). The equation of radiative transfer is

$$\frac{dI_\nu}{dl} = -\kappa_\nu I_\nu + j_\nu, \tag{6}$$

where I is the intensity and l is the path length through the medium. Collisions dominate most molecular excitation, and if the emission rate is low, as usually occurs in molecular clouds, then the populations can be assumed to be in local thermal equilibrium, hence given by the Boltzmann distribution:

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} e^{-E_{12}/kT_{\text{ex}}}, \tag{7}$$

where T_{ex} is the excitation temperature, which is assumed to be of the order of the kinetic temperature of the exciting particles, and g is the statistical weight of the states, which are separated by E_{12}. The optical depth for a line is proportional to the path length through the medium, so that

$$\kappa n dl = \kappa n dv \bigg/ \left(\frac{dv}{dl}\right) = \frac{c}{\nu}\kappa_0 n \, \frac{\Delta\nu}{dv/dl}, \tag{8}$$

where $\Delta\nu$ is the line width in frequency, κ_0 is the opacity at line center, n is the number density, and dv/dl is the velocity gradient along the line of sight. The so-called *large velocity gradient* or *Sobolev* approximation (also called the "on the spot" approximation because the emission and/or absorption is assumed to depend on only the local conditions) assumes that this gradient is larger than the thermal speeds so that the optical depth of the medium is small. For molecular clouds, observed line widths are usually a few kilometers per second, while the thermal speed is about 0.1 km s^{-1}, so this approximation seems to be valid for all but the most abundant species.

Like their atomic counterparts, molecular lines saturate when the populations have reached the values associated with strict equilibrium with the incoming radiation. This occurs first at the line center. Any motion in the medium, ordered or random, will broaden the line and thus the molecules will "see" radiation at other wavelengths against which they can absorb, or into which they can emit. If the medium is optically thick at line center but the velocity dispersion is large, the overall optical depth can be considerably reduced by spreading out the line in frequency.

The most abundant molecules, because of the low velocities observed in the clouds and high column densities, cannot be interpreted by simple optically thin models. For CO (any isotope), the ratio of the $(2 \rightarrow 1)$ to $(1 \rightarrow 0)$ transitions should be 3:1 in strength, if completely optically thin, because of the ratio of the statistical weights and transition probabilities and the temperature known to exist in the clouds. However, $^{12}C^{16}O$ $(1 \rightarrow 0)$ is often observed to show flat-topped profiles, not the Gaussian form which would be typical of a randomly moving optically thin molecular gas. Also, the intensity ratio of the transitions is often seen to depart from that expected for such a medium. The implication is that ^{12}CO is optically thick, and that the densest parts of the cloud may not be observable in the ground-state transition. Lower-abundance species (for instance, the isotopes ^{13}CO and $^{12}C^{18}O$, or more highly excited states of $^{12}C^{16}O$ (such as $3 \rightarrow 2$) may, however, probe denser parts of the cloud. Further, the higher transitions require higher densities for excitation, so there is a delicate interplay between chemistry and radiative transfer which enters into the interpretation of abundances. This is crucial to the understanding of the formation of the molecules.

The abundance of a species is related to the observed line intensity by the *antenna temperature*, the temperature which an equivalent blackbody radiator would have to have at the line frequency to equal the observed line intensity:

$$I_{\text{line}} = \int_0^\infty \frac{T_A}{\nu^2}\, d\nu, \tag{9}$$

which is integrated over the velocity width of the line. The column density, by number, in the lower level is defined as

$$N_l = \frac{4\pi^{3/2}}{(\ln 2)^{1/2}} \frac{k\nu}{hc^2} \frac{(2J_l + 1)}{(2J_u + 1)} \frac{1}{A_{ul}} \int_{\text{line}} T_B\, dv, \tag{10}$$

where k is the Boltzmann constant and v is the velocity width of the line. Then the total column density of the species is found using the Boltzmann distribution for the levels:

$$N_{\text{tot}} = N_l \frac{Q(T)}{(2J_l + 1)} \exp\left(\frac{E_l}{kT}\right), \tag{11}$$

where $Q(T)$ is the molecular partition function, the sum over the population probabilities for all of the rotational levels,

$$Q(T) = \sum_{J,K} g_{J,K} e^{-E_{JK}/kT}, \tag{12}$$

which can often be found in closed form. For instance, for a symmetric top molecule,

$$Q(T) = \left(\frac{\pi}{A_v B_v^2}\right)^{1/2} \left(\frac{kT}{hc}\right)^{3/2} \exp\left(\frac{B_v hc}{4kT}\right), \tag{13}$$

where $B_v = C_v$ and A_v are the moments of inertia for the rotational states and $E_l = hc[B_v J(J+1) + (A_v - B_v)K^2]$. In the optically thin, LVG approximation, this suffices to determine the abundance of the species of interest. It assumes that all emission is due to thermal equilibrium prevailing due to collisions among the levels.

2. Masers

Because of the low densities, molecules in cosmic environments can show populations of many levels which are inverted. That is, the higher levels sometimes have higher populations than the lower ones. The primary reason for this is that transitions take place between high states which are only weakly collisionally coupled to lower levels, and for which radiative transitions are long. If there is a strong background radiation field at shorter wavelength than the transition of interest, upper states of the molecule may be radiatively excited with subsequent overpopulation of some of the lower states by radiative and collisional deexcitation. Thus, for masers to occur, more than two states must be involved in strongly coupled transitions.

Masers also serve as a warning that the intensity of a spectral line is not necessarily a direct measure of its abundance. Population inversions enhance the brightness temperatures, leading to overestimates of excitation and abundance in those species in which masing occurs. Because not all molecules undergo maser amplification, the assumption of thermal equilibrium is usually not bad, but should be employed with caution.

The emission and absorption coefficients for the system can be defined as before. Now assume that there are a total of n levels, and that they are coupled via collisions and radiation to the levels 1 and 2. Then the time-dependent populations of 1 and 2 are given by

$$\begin{aligned}\frac{dn_1}{dt} &= \mathcal{P}_1(n - n_1 - n_2) - (n_1 B_{12} - n_2 B_{21})\frac{\Omega}{4\pi} I \\ &\quad - n_1 C_{12} - n_1 \Gamma; \end{aligned} \tag{14}$$

$$\begin{aligned}\frac{dn_2}{dt} &= \mathcal{P}_2(n - n_1 - n_2) - (n_2 B_{21} - n_1 B_{12})\frac{\Omega}{4\pi} I \\ &\quad - n_2(A_{21} - C_{21}) - n_2 \Gamma. \end{aligned} \tag{15}$$

Here $\mathcal{P}_1$ and $\mathcal{P}_2$ are the pump rates from the higher-lying levels through radiation and collisions, Ω is the solid angle, and Γ is the rate at which the masing levels are depopulated.

For molecular masers, it can be assumed that the two masing levels have the same statistical weight. Thus, $B_{12} = B_{21} = B$. The number of levels involved in the particular population inversion is small. This implies strong radiative coupling between states which, for some reason, are selectively pumped by the external sources of radiation. The OH molecule is an excellent example of this behavior.

If the absorption coefficient is *negative*, in other words, if the populations are *sufficiently inverted*, the radiation in the $2 \rightarrow 1$ transition will amplify along its path until the maser saturates, that is, until the populations do not change along the path length. The amplification selects out the line center, and the line gradually narrows as a result of increasing path length. This behavior is of great importance, because the brightness temperature increases as the line gets narrower. Further, the radiation has a finite amplification length; the maser can saturate. In the absence of collisions and in steady state, Γ can be replaced by $\Gamma = 2BI(\Omega/4\pi)$, so that the brightness temperature of a saturated maser is given by

$$T_{B,s} = \frac{h\nu}{2k}\frac{\Gamma}{A}\frac{\Omega}{4\pi} \tag{16}$$

This makes masers very intense radiation sources, since the pump radiation at higher frequency has been converted both to lower frequency and narrower bandwidth by the amplification process. The emission is also highly polarized, since it is coherent.

Masing depends on the presence of a strong radiation field for excitation and maintenance. Such radiation, usually infrared, is significant in several environments, notably in circumstellar envelopes (CSEs) and in molecular clouds. In CSEs, far-infrared radiation is converted to centimeter radiation by OH, which has transitions centered around 1665 MHz. Ammonia, water, HCN, and SiO, are also important stellar maser sources. Water masers are also associated with regions of active star formation, where IR from the protostellar cores can excite the millimeter radiation in the densest parts of the cloud. Because they are strongly amplifying, the masing sites are easily distinguished from the background and their proper motions can be directly measured using VLBI techniques. Their time variability is also well observed, although it is still not fully understood theoretically.

III. DUST

A fundamental constituent of the atmospheres of the coolest stars (whether red giants or brown dwarfs) and of the interstellar medium is the solid material that has become known as *dust*. Although dust was recognized and characterized more than 50 years ago, many of its basic properties are still debated. In large measure, this is due to the very indirect way in which information about the composition and structure of the dust is obtained.

The spectral signature of dust is the presence of several very broad features in the infrared and the ultraviolet, whose strength correlates well with the extinction of visible starlight. The UV feature, near 2175 Å, is likely due to some form of solid carbon, something like graphite or an amorphous state of carbon. Its strength and shape are variable throughout the galactic plane, although not entirely absent along most lines of sight through the plane, and these also are variable from one galaxy to another. The identification of the infrared band at 10 μm is more secure, being due to silicates and at 11 μm due to SiC. Dust is normally virtually transparent at this wavelength because of the size of the grains, and in the diffuse interstellar medium the column densities are insufficient to produce appreciable absorption in the IR band; it is seen in the atmospheres of highly evolved red supergiants. In these stars, because of the low emission from the envelope as a whole and the large spatial extent of the outer stellar layers, the feature is often seen in emission. Unless the star happens to be sufficiently cold that the outer atmosphere is emitting significantly at these wavelengths, the feature will always be seen in emission; a few very dense shells show absorption at the same wavelength.

The presence of silicates in both the interstellar medium (ISM) and stellar envelopes is certain, but the precise physical state of the silicate is not well known. As is typical of solids, most of the detailed information about the internal structure of the radiating species is lost due to the complexity of the lattice structure and the effects of nearest-neighbor perturbations to the energy states. These result in broad diffuse absorption or emission bands.

Other IR features in the 3- to 10-μm region have been identified with both water ice (near 3.2 μm) and with polycyclic aromatic hydrocarbons (PAHs) (several bands, especially near 3.3, 6.2, 7.7, 8.6, and 11.3 μm). They are identified with C–H and C–C bending and stretching modes of complex organic molecules, although specific identifications are insecure. The water is presumed to condense onto the grains in dense environments, like molecular clouds. The PAHs are more like small grains than molecules, but are likely associated with the formation and destruction of dust, forming the small particle end of the size spectrum. Because they are nearly molecular, they are not in equilibrium with the radiation field—that is, they do not radiate like blackbodies. Instead, they deexcite from UV radiative absorption from the diffuse interstellar radiation field (DIRF) via vibronic transitions in the near-IR. This means that their emission requires some UV excitation, which is supported by their presence in photodissociation regions at the boundaries of molecular clouds and planetary nebulae. The observed IR diffuse bands, which have optical analogs seen in absorption against background sources, are likely due to the C–H bond stretch, the analog of the Si–O vibration responsible for the 10-μm feature.

Dust radiates in the infrared. By Kirchhoff's law, solid material in thermal equilibrium radiates like a blackbody. Most of the incident energy falling on the grain is scattered, hence the blueness of reflection nebulae and the reddening of starlight. The absorbed photons are reradiated at a rate approximated by $j_\nu = \kappa_\nu B_\nu(T)$, where κ_ν is the monochromatic absorption coefficient and $B_\nu(T)$ is the Planck function at temperature T. The equilibrium temperature of the grain depends, therefore, on its size and composition. Carbon grains absorb effectively in the UV due to the diffuse 2175-Å band, but radiate inefficiently in the IR, so they are hotter than silicates, for which the reverse holds. The PAHs are distinguished by two effects. They show a much higher color temperature than the larger grains and, in addition to their lines, they cannot be in thermal equilibrium since their specific heats are temperature dependent (see below).

Grains provide a solid surface on which chemical reactions take place. In fact, they are the primary site for H_2 synthesis. In addition, metallic ions deplete onto the grains. This is evident from the lower-than-stellar abundances observed for the heavy metals, such as iron and calcium, in the diffuse interstellar medium. It appears that most of the heavy metals in the diffuse and molecular cloud phases of the ISM may be tied up in the grains, which nonetheless constitute only about 10^{-6}, by number, of the ISM. CO and H_2O may also stick to the grain surfaces, and models and laboratory simulations show a host of complex organic molecules can be synthesized in the resultant mantle. It remains to be determined whether these simulations are relevant for interstellar conditions; they do seem to mimic many of the reaction products observed *in situ* in comet Halley.

An important problem in dust chemistry is the precise determination of both the formation mechanism and size spectrum of the grains. At the smallest-particle end, the grains behave like large molecules. The PAHs are stable against UV radiation and also can be cleaved from the larger graphite grains. The signature of small grains is that they cannot come into equilibrium with the UV radiation field, and do not radiate like blackbodies. Instead, their specific heats depend explicitly on the number of available modes, N, proportional to the number of constituent molecules. They radiate with an excitation temperature depending on the incident photon energy as $T = h\nu/NC$. Hence, they have color temperatures which are of order 1000 K in the presence of the DIRF, in spite of the fact that they are never in equilibrium and so have no true kinetic temperature. The resulting emission bands are vibrational transitions which redistribute the incident UV radiation on short time scales.

IV. MOLECULAR ENVIRONMENTS

A. Circumstellar Envelopes

Molecules are frequently observed in the outer envelopes of red supergiants with surface temperatures less than about 5000 K. These stars have strong stellar winds, of order 10^{-7} to 10^{-6} $M_\odot$ yr^{-1}, with velocities typically less than 50 km s^{-1}. A few hotter stars, such as 89 Herculis and HD 161796, have also been found to show CO emission; these and related stars are proto-planetary nebula objects in the process of becoming white dwarf stars. For some very highly evolved dusty stars, strong far-IR emission, indicative of dust, is accompanied by maser emission. These are the so-called OH/IR stars, which are most frequently Mira variables. Some evolved stars also show SiO masers. A few stars, such as the extreme supergiant IRC + 10216, are veritable chemical factories, displaying almost all of the molecular species observed in comets and in dense interstellar molecular clouds.

B. The Environment of the Interstellar Medium

The interstellar medium is a very inhomogeneous, disequilibrated place. The diffuse medium has densities of 0.01 to 1 cm^{-3} in the ionized phase and about 1 to 10^3 cm^{-3} for the cooler neutral phase. The medium is heated by supernova stellar wind shocks, and has a sufficiently long cooling time that it never becomes any colder than about 10^6 K. This is because the atoms are very inefficient coolants. In denser regions, the temperature is reduced to about 10^6 K, and cooling due to neutral hydrogen recombination becomes efficient. For lower temperature, the cooling increases dramatically, first due to hydrogen line emission, which reduces the temperature to 10^4 K, and then from atomic fine-structure transitions, which reduce the temperature to several hundreds of degrees. To lower this temperature further takes two additions to the medium—dust grains and molecules.

In diffuse regions, the dust will always be colder than the gas, primarily because of the larger number of modes available for the redistribution of the energy. In fact, the temperature, or equivalently the excitation, of the dust is sensitive more to the spectrum of the radiation incident on it than to its dilution. The cooling of the dust is strictly radiative, efficiently absorbing in the UV and radiating in the IR. The harder the UV radiation which is incident on the grain, the warmer the grain will be, regardless of the intensity of the radiation. The grains are the critical shield for the cloud material from background radiative heating. The presence of dust also effectively cools the gas because of surface atom interactions which promote the formation of molecules, especially H_2. Molecular cooling is due to collisional excitation of abundant species, which then reradiate their energy in the far-IR and millimeter wavelengths, at which the grains are optically thin. Deep in the cores of molecular clouds, the situation is reversed. Here the grains are warmer than the gas, and actually heat the gas through collisions of the particles with the grain surface. As such, they serve as the fuel for the chemistry. Collisions of molecular hydrogen with, for instance, CO excite the latter, which radiates its energy from the cloud at the expense of the gas kinetic energy. Thus, the IR which penetrates the cloud and heats the grains can be transferred to the excitation of the various chemical constituents of the medium effectively, serving to power the reactions which build complex molecular species.

V. CHEMICAL PROCESSES

A. Surface Chemistry: Formation of Molecular Hydrogen

The basic problem in the study of the interstellar medium is the formation of molecular hydrogen, H_2. This is of primary importance because, at the low temperatures characteristic of the cloud environments, this molecule is responsible for the excitation of CO. One might initially expect reactions of the form

$$2\mathrm{H} \rightarrow \mathrm{H}_2 + \gamma, \tag{17}$$

where γ is an emitted photon. This process, the so-called *radiative association* mechanism, is important for the formation process. The rate is, alas, many orders of magnitude short of that required to produce the molecule. In fact, it appears that the formation of H_2 can only proceed via one of two possible avenues: (1) if there is a sufficient abundance of free electrons,

$$\mathrm{H} + e \rightarrow \mathrm{H}^- \tag{18}$$

$$\mathrm{H}^- + 2\mathrm{H} \rightarrow \mathrm{H}_3^- \tag{19}$$

$$\rightarrow \mathrm{H}_2^* + \mathrm{H}^-; \tag{20}$$

or (2) via some form of surface interaction where radiative association is replaced by reaction on a solid surface of two neutral atoms in the presence of a UV radiation field which is capable of exceeding the appropriate binding energy of the molecule to the grain surface. The first of these is independent of the abundance of metals, while the latter is critically dependent on the existence of interstellar grains on which the reactions can take place. Because it appears that the solid lattice is far more efficient, and because it exists in the interstellar environment that now is observed in molecular clouds, we shall concentrate on the second mechanism. The first type of mechanism has been implicated in star formation processes in the early universe.

Assume that the grain consists of a simple lattice, like graphite. Should an atom of neutral hydrogen strike the surface, there is a *sticking probability*, S, such that the atom will become bound to the surface rather than be reflected back into the diffuse medium. The rate of impact on the surface of hydrogen atoms is given by the mean collision time of a H atom with a grain having a geometric cross section $\sum$. The velocity dispersion in the gas is $v_{\mathrm{th}} \sim T^{1/2}$, so that

$$\frac{dn}{dt} = n_H n_g S \pi a_g^2 v_{\mathrm{th}} \tag{21}$$

and the rate of hopping, or migration, among lattice sites is t_{mig}. Then, if K_{HH} is the reaction rate, the rate of formation of H_2 is approximately given by

$$n_{\mathrm{H}_2} \approx K_{\mathrm{HH}} n_{\mathrm{H_g}} S \sum \sigma t_{\mathrm{mig}} \beta^{-1}, \tag{22}$$

where β is the rate of release of H_2 from trapping sites back into the medium. An empirical rate for H_2 formation is

$$\frac{dn(\mathrm{H}_2)}{dt} \approx 3 \times 10^{-17} \mathrm{cm}^3\,\mathrm{s}^{-1}\, n_{\mathrm{H}}^2. \tag{23}$$

Here n_{H} is the ambient gas density, with the grains scaling as a fixed fraction of n_{H}. There is, however, reason to believe that at the lowest neutral hydrogen densities, the formation process depends on the random rate of arrival of molecules on the surface and there is an exponential threshold for the molecular formation (see Caselli *et al.*, 1998). Grains are a prerequisite for the formation of molecular hydrogen in the present galaxy, but gas-phase reactions in the dense regions that are typical of molecular clouds will otherwise produce all of the species which are observed. Therefore, in what follows, we shall concentrate on the work which done in the 1990s on the problem of gas-phase chemistry.

B. Gas-Phase Chemistry

The first observations of diatomic species in the diffuse interstellar medium several decades ago posed serious challenges for theorists because of the extremely low densities which are found there. Radiative association seemed unable to produce any of the observed species, most importantly CH, and this meant that exotic mechanisms were initially held responsible for the presence of such molecules. Work on the abundance of H_2, following the observation of the molecule in the diffuse medium in the ultraviolet by the *Copernicus* satellite in the mid-1970s, and the discovery of elemental depletion along many lines of sight in the interstellar medium, led to the suggestion that grains were also the fundamental site for the chemistry required to produce even the simplest diatoms. The *Fuse* mission, launched in 1999, covers the same spectral range (900–1200 Å) as did *Copernicus*, but with significantly higher sensitivity and resolution, and is now being used to study more thoroughly the molecular hydrogen component of the galaxy. In addition, the *ISO* mission detected large abundances of H_2 in the far-IR even from regions that have very low CO abundances.

The discovery of large, cold molecular cloud complexes dramatically altered this view, providing the necessary conditions for low-temperature, high-density gas-phase reactions to occur. The development of many computational schema for handling enormous reaction networks, often involving thousands of reactions and hundreds of reacting species, also spurred theoretical work on this subject. This section is meant only to serve as a guide to these calculations. The basic physical input is really quite simple; it is the computational complexity that makes for the differences found among various workers in the field.

The chemistry of gas-phase reactions, either in the interstellar medium or in stellar atmospheres, is mediated by the abundance of ions. These can be formed in several ways: by cosmic-ray ionization, or by the direct photoionization of the atoms involved in the reactions with subsequent charge transfer to the molecules. Ionic reactions are generally exothermic and so occur efficiently at low temperature. In the presence of an ion, a neutral molecule or atom develops an induced dipole which increases its capture cross section. Thus the reactions occur quickly and lead to stable states, in addition to allowing the molecule to form in a radiatively unstable excited state which, upon decaying, radiates the energy of formation away from the site of the reaction.

1. Reaction Rates

In general, reactions depend on two factors, the activation energy Δ and the temperature. Cross sections can be observed directly in the laboratory at some controlled temperature, usually near room temperature (about 300 K), and then scaled to the temperatures found in clouds. They can also be deduced from first principles. Generally, these reactions have the form $K = \langle \sigma v \rangle$ where the average of the cross section, σ, is taken over the velocity distribution of the interacting particles where the relative velocity is v. For this reason, assuming the reacting particles are thermalized, the rates depend on temperature. For ionic reactions, where the potential is the coulomb interaction, the so-called *Langevin approximation* applies, and it can be shown that the rates are approximately constant. Thus measurements at room temperature suffice for the determination of the rate coefficients, $K = k$, where usually k, the reaction constant, is of order 10^{-8} to 10^{-13} cm^3 s^{-1}. Most ionic reactions have little or no potential barriers, being generally exothermic. Hence there is usually a simple constant volumetric rate which is assumed to be a constant. Neutral reactions are most likely to involve substantial activation energies which greatly inhibit their rates of formation. If a neutral channel is important in a network of reactions, it will likely be a bottleneck for the formation of the product species. These have strong temperature dependences because of their activation energies and usually are several orders of magnitude slower than the ion-neutral channels. Electronic recombination reactions typically scale as $T^{-1/2}$.

Before proceeding with a discussion of specific results, one point should be emphasized. In many of the reaction networks, among all of the rates which must be tabulated and all of the reactions which must be tracked, many of the rates have to be approximated by guesses or simple fits to laboratory data. Few measurements (except on the *Space Shuttle*) at interstellar conditions are available, and this is a very significant challenge for future laboratory astrophysics. In many of the networks, perhaps as few as 10% of the rates are known to within a factor of 50%, and perhaps as many as half are bald guesses and may be uncertain to a factor of 10. This is a field still in its infancy, where only the dominant channels are well understood, but many of the details are still extremely important because of the physical conditions that can be probed by trace species.

2. Ionization

Ionization in the densest parts of molecular clouds depends on the penetration of cosmic rays and UV radiation, as well as the presence of shocks generated by such processes as cloud–cloud collisions and internal star formation. In order to probe the electron density in the clouds, it is important to be able to account for the presence of complex polyatomic molecules, whose formation requires ion gas-phase reactions.

The rate for cosmic-ray (CR) ionization, ζ_{CR}, is about $(3 \pm 1) \times 10^{-17}$ s^{-1}. This is an integral over the collisional ionization cross section for low (MeV)-energy CR protons, but it is approximately a constant for most of the species of interest. An obstacle in our understanding of the detailed structure of molecular clouds is our ignorance of the precise specification of this rate. The low-energy end of the cosmic-ray spectrum is difficult to determine empirically from terrestrial observation, because these particles propagate diffusively through the interplanetary medium, scattering off of turbulence in the solar wind; their spectrum cannot be observed directly , even with *in situ* measurements from the *Voyager* and *Ulysses* spacecraft, and must be inferred from models for their motion through the heliosphere. The more easily observed cosmic ray protons and electrons, in the GeV and higher range, have little or no effect on the ionization of the interstellar medium because of the small interaction cross sections for atoms at such high energies.

In molecular clouds, atomic species with ionization energies greater than 13.6 eV must be predominantly neutral because of the shielding effects of neutral hydrogen. It is mainly the heavier elements, such as C, N, and O, which are observed in the peripheral portions of the clouds to be in the partially ionized state. For circumstellar envelopes, cosmic rays lose out to photo processes and the chemistry is mediated by the input of stellar photospheric radiation (in the hotter stars and in novae and supernovae) and from the diffuse interstellar radiation field.

The basic equations for two body interactions can be written in the form

$$\frac{dN_i}{dt} = \sum_{j,k\neq i} K_{ijk} N_j N_k - \sum_j K'_{ij} N_i N_j, \qquad (24)$$

where K_{ijk} is the formation rate for the ith molecular species, while K'_{ij} is the destruction rate for the molecule. The inclusion of UV photo processes is accomplished by the photodissociation rate:

$$R_{pd} = \int_{\nu_0}^{\infty} \kappa_\nu F_\nu e^{-\tau_\nu} \frac{d\nu}{h\nu}, \qquad (25)$$

where F_ν is the incident photon flux, τ_ν is the opacity of the ambient medium (presumed to be from dust), κ_ν is the continuous absorption coefficient for the dissociative continuum, and the dissociation energy is $h\nu_0$.

An aspect in which circumstellar environments differ from interstellar is the net mass advection through the medium. Abundances become time dependent—and hence space dependent—in the envelope, due both to the implicit time dependence of the reactions and to the transport of matter through different radii via stellar wind flow. The atomic abundances are fixed at stellar photosphere, rather than having to be assumed for some mixture of physical parameters of temperature and pressure as they must for molecular clouds. It is then essentially an initial-value problem to compute the abundances which will be a function of radius in the envelope. For a steady-state wind, the abundances become strictly a function of radius. Also, unlike a molecular cloud, the density profile of the envelope is specified from the assumption of steady mass loss at the terminal velocity for the wind, so that $\rho(r) = \dot{M}/(4\pi r^2 v_\infty)$, where $\dot{M}$ is the mass loss rate and v_∞ is the terminal velocity of the wind.

An interesting aspect of stellar envelopes is that they may have two different sources of UV radiation, internal and external. Work on the envelope of two extreme, low-temperature, evolved supergiants, IRC + 10216 and α Ori, showed that the outer limit of the molecular envelope is determined by the DIRF, which destroys the outermost molecular species by photodissociation, while the inner boundary is set by both the temperature and UV emission from the stellar chromospheres. In this respect, since the dynamics can be probed in exquisite detail for several of the nearer supergiants through molecular observations, and since the input abundances are known and atomic in nature, it is possible to use these stars as very well-conditioned laboratories for the study of the same processes which must be involved in at least some aspects of molecular cloud chemistry. For the densest envelopes, which are completely optically thick and hence very similar to molecular clouds, cosmic rays are significant in governing the ion fractions but can be neglected in thin envelopes (low mass loss rates).

3. Cooling Processes

Chemistry also feeds back into the thermal balance of the clouds. Molecules radiate in portions of the spectrum where the medium is usually optically thin. Since this radiation can escape from the cloud, it is the primary means whereby the clouds cool. Star formation requires that otherwise hydrostatic clouds become gravitationally unstable, a process which can be affected by the rate of energy loss as well as by external perturbations. Thus time-dependent processes, those which cause the stability of the clouds to alter with time, are extremely important, since the time scale for molecular formation is not too short (of order 10^6 years) compared with the estimated lifetimes of the clouds ($\leq 10^8$ years). For example, the cooling rate for CO depends on the abundance of both dust and of H_2 and CO by

$$\Lambda_{\mathrm{CO}} = \frac{1.1 \times 10^{-30} n(\Delta v/v_{\mathrm{th}}) T^{1/2}}{1 + 1.4 \times 10^{-4} n T^{1/2}(1 + N/N_c)} \mathrm{erg\ cm^{-3} s^{-1}}, \qquad (26)$$

where $N_c = 2 \times 10^{18} T$ cm^{-2} and N is the column density, related to the extinction. Molecular species are therefore quite efficient in radiatively removing energy from the clouds and, literally, refrigerating the medium.

4. Shock Chemistry

Hydrodynamic and magnetohydrodynamic (MHD) shocks are important in the time-dependent chemistry of the diffuse interstellar medium. The time scales are very short for shock passage through a region, typically less than 10^5 years per parsec, but they can cause considerable abundance variations and produce long-lived products. CH^+ is primarily produced this way, and CH and CN also seem to have input from shocks.

The role of shocks in the chemistry of clouds is best seen in the effect it can have on molecular hydrogen. While many reaction products remain unchanged in abundance, CH_4, H_2O, and HCO can be greatly enhanced due to the increased production of these molecules in the hotter, and denser, shock environments. The chemistry is also dependent on the role of magnetic fields. Magnetic shocks can have sizable compression without significant increases in the temperature, due to the pressure provided by the magnetic field. As a result, the relative abundance of shock-produced molecules serves as a probe of the nature of the shock producing the enhancement in the reaction rates.

Chemistry in the post-MHD shock environment is dominated by the separation between neutral and ionic species, the former being less affected than the latter by the magnetic field. In consequence, the reaction sites are calculated to show abundance stratification depending on the reaction channels. Since the fronts may be broad enough to be spatially resolved, it is possible to study the diffuse-phase ISM shock chemistry observationally in some detail.

5. Specific Molecular Reactions

a. Hydrogen. The most important ion for all molecular reactions is H_3^+. It is formed by several channels, the primary one being capture of low-energy cosmic-ray protons by molecular hydrogen, which is itself formed on grains. It is stable at the densities and temperatures which are typical of molecular clouds. The capture of a carbon atom to form CH_3^+ is a critical step in the chemistry of the interstellar medium, especially in the generation of the ions of the cyanopolyyne series, such as $HC_{11}N$. H_3^+ has been observed directly in several dense clouds along with its deuterated phase, H_2D^+. One can therefore assert with confidence the role of this ion in molecular chemistry. Subsequent to carbon capture, interactions with H_2 can form all of the hydrocarbons observed in molecular clouds. An example of this chemistry is given by the network

$$H_3^+ + C \rightarrow CH^+ + H_2, \tag{27}$$

$$CH^+ + H_2 \rightarrow CH_2^+ + H, \tag{28}$$

$$CH_2^+ + H_2 \rightarrow CH_3^+ + H, \tag{29}$$

$$CH_3^+ + H_2 \rightarrow CH_5^+ + \gamma, \tag{30}$$

$$CH_5^+ + e \rightarrow CH_4 + H, \tag{31}$$

$$\rightarrow CH_3 + H_2, \tag{32}$$

$$\rightarrow CH_2 + H_2 + H, \tag{33}$$

$$\rightarrow CH + 2H_2, \tag{34}$$

which also illustrates the reason for the complexity of many of the reaction calculations: there are many product states for electron-capture reactions, due to the role of dissociative recombination.

b. Carbon. Carbon chemistry is both interesting and important for the determination of the detailed structure of molecular clouds. Early observations of formaldehyde were an indication that some form of gas-phase chemistry must occur in clouds, and the reaction mechanism for the production of H_2CO is

$$CH_3 + O \rightarrow H_2CO + H, \tag{35}$$

$$H_3CO^+ + e \rightarrow H_2CO + H. \tag{36}$$

For CO, there are several reaction channels. All are initiated by the formation of the HCO^+ ion:

$$H_3^+ + CO \rightarrow HCO^+ + H_2, \tag{37}$$

$$C^+ + H_2O \rightarrow HCO^+ + H, \tag{38}$$

$$CH_3^+ + O \rightarrow HCO^+ + H_2, \tag{39}$$

$$C^+ + H_2CO \rightarrow HCO^+ + CH, \tag{40}$$

$$HCO^+ + e \rightarrow CO + H. \tag{41}$$

The CO is subsequently excited by nonreacting collisions with H_2 which produces the observed line emission. Another possible pathway involves

$$H_3^+ + O \rightarrow OH^+ + H_2, \tag{42}$$

$$OH^+ + H_2 \rightarrow H_2O^+ + H, \tag{43}$$

$$H_2O^+ + H_2 \rightarrow H_3O^+ + H, \tag{44}$$

$$H_3O^+ + e \rightarrow OH + H_2, \tag{45}$$

$$\rightarrow H_2O, \tag{46}$$

$$OH + C^+ \rightarrow CO^+ + OH, \tag{47}$$

$$H_2O + C^+ \rightarrow HCO^+ + H, \tag{48}$$

$$HCO^+ + e \rightarrow CO + H, \tag{49}$$

$$CO + H_3^+ \rightarrow HCO^+ + H_2. \tag{50}$$

This illustrates the mediating role of oxygen in the formation of CO, and also the fact that the pathway can be blocked by photodissociation of H_2O to form OH in all but the densest regions of the clouds.

A most important aspect of CO is that it is self-shielding. Should it be possible to build up a significant column density of the molecule, it will form a photodissociative block to incoming UV. In the interstellar medium, this means that the formation of C^+ is important in the outer layers of the clouds, which are irradiated by the diffuse interstellar UV photons, but that within a short distance (a layer that is perhaps no more than 1 pc in thickness), all of the photons have been absorbed. The presence of ions is therefore a good indication that low-energy cosmic rays can penetrate the deepest regions of the cloud without being lost due to either grain interactions or energy loss by scattering off of internal MHD turbulent eddies in the cloud cores. Further input of electrons from grain ionization may contribute in the intermediate layers of the cloud as well, but these are unlikely to be important in the innermost regions.

Temperature and density profiles are derived from CO observations. This is true for both circumstellar envelopes and molecular clouds. The optical depth is measured using the $^{13}CO/^{12}CO$ $(1 \rightarrow 0)$ transition, while the excitation temperature is given by the ratio of the ^{12}CO $(2 \rightarrow 1)$ to $(1 \rightarrow 0)$ intensities. The excitation is presumed to be due to H_2 collisions, so the populations should reflect the local thermal properties.

c. Nitrogen. Nitrogen is another important species for reaction kinetics. Here the primary initiating reaction is

$$H_3^+ + N_2 \rightarrow N_2H^+ + H_2. \qquad (51)$$

The most important reaction network is initiated by the ionization of nitrogen by charge exchange with He^+ or by cosmic-ray ionization of N:

$$N^+ + H_2 \rightarrow NH^+ + H, \qquad (52)$$

$$NH^+ + H_2 \rightarrow NH_2^+ + H, \qquad (53)$$

$$NH_2^+ + H_2 \rightarrow NH_3^+ + H, \qquad (54)$$

$$NH_3^+ + H_2 \rightarrow NH_4^+ + H, \qquad (55)$$

$$NH_3^+ + e \rightarrow NH + H_2, \qquad (56)$$

$$\rightarrow NH_2 + H, \qquad (57)$$

$$\rightarrow NH_3, \qquad (58)$$

$$NH_4^+ + e \rightarrow NH_3 + H, \qquad (59)$$

$$\rightarrow NH_2 + H_2. \qquad (60)$$

The formation of NH^+ by H_2 capture is only probable at temperatures exceeding about 20 K; its slight endoergic nature at lower temperature inhibits this reaction channel. The presence of metals can also be of importance for the production of ammonia, because charge-exchange reactions can occur which will neutralize the ammonium ion.

Many complex polycyanoacetylenes are observed in both circumstellar envelopes and dense molecular clouds, the heaviest being $HC_{11}N$, one of the cyanopolyynes. As mentioned in the section on carbon chemistry, these molecules are likely built through the formation of HCN, HNC, and CN, all though interactions initiated by the formation of CH_3^+. Heavier molecules are likely formed through the incorporation of cyanogen and HCN into the molecule, but because of the large number of species involved in these calculations, the mechanisms are still not well understood. Many of the rates have yet to be calculated in detail.

The discovery of molecules containing heavier elements such as PN and iron compounds opened the field of heavy-ion chemistry. It is well known that phosphorus is important in the evolution of biota on the earth, and its discovery in the interstellar medium is a major clue to the development of chemical processes during the early stages of life's development on planets.

d. Sulfur. Sulfur chemistry is also of great interest because of the abundance of CS. An important initiating step is

$$SH^+ + H_2 \rightarrow SH_3^+ + \gamma, \qquad (61)$$

$$SH_3^+ + e \rightarrow H_2S + H, \qquad (62)$$

$$\rightarrow SH + H_2. \qquad (63)$$

For carbon compounds, CS for example, there are several very important reactions:

$$CH + S \rightarrow CS + H, \qquad (64)$$

$$CS + O \rightarrow CO + S, \qquad (65)$$

$$H_3^+ + CS \rightarrow HCS^+ + H_2. \qquad (66)$$

The helium ion is also important because charge transfer can lead to disintegration of the CS molecule:

$$He^+ + CS \rightarrow C^+ + S + He. \qquad (67)$$

Observations of CS are important because they probes dense portions of the molecular clouds, where collisional excitation produces strong emission lines. However, since these dense interior parts of the cloud cores are also sites of very complex chemistry, the detection of CS and of related molecular species may be very dependent, in ways still not well understood, on the chemistry of these sites. SO_2, HS, and SO have also been detected in molecular clouds, and it is also possible to study isotopic fractionation among very heavy molecules such as ^{13}CS and $C^{33}S$. As with most isotopic species, these are optically thin and permit study of the ionization structure of the cloud.

e. Water as a special case. The H_2O molecule is central to much organic chemistry, even in the interstellar medium. It has been observed in maser sources, in deuterated form in molecular clouds, and in its ion, but the neutral molecule has not been observed in the cores of molecular complexes. This aspect of the current observational picture is puzzling, because according to models the abundance should be of order $H_2O/H_2 \geq 10^{-6}$. This is the about 10% of the CO fraction, indicating that much of the oxygen in the clouds may be tied up in water. The difficulty is that so far, this species has not been observed directly. On the basis of the chemical fractionation (see next section) and the abundance of the deuterated form of water, this abundance may be a lower limit.

The importance of H_2O to cloud chemistry is seen from the reaction sequence:

$$CH_3^+ + H_2O \rightarrow CH_3OH_2^+, \qquad (68)$$

$$CH_3OH_2^+ + e \rightarrow CH_3OH + H, \qquad (69)$$

which competes with the reaction $CH_3^+ + HCN \rightarrow CH_3CNH^+$ in the destruction of the methyl ion. Water is also important in the formation of another species of interest,

$$C^+ + H_2O \rightarrow HCO^+ + H, \qquad (70)$$

$$\rightarrow HOC^+ + H, \qquad (71)$$

$$HOC^+ + CO \rightarrow HCO^+ + CO, \qquad (72)$$

the last being a rearrangement collision due to the environmental CO. In both cases, these rates compete with the formation due to H_3^+ and will enhance the formation of HCO^+, an easily observed species. The reason for thinking this important is that the diagnosis of physical conditions in the densest parts of molecular clouds is affected through the use of complex chemical species. We do not have direct access to the cores via CO because of the high optical depths and self-absorption by other abundant molecules. However, we know from maser observations that the water is being produced in at least some environments.

What this means for the structures of clouds is not presently clear. It is possible that the clouds are fluffy—that UV radiation penetrates deeper into the clouds than previously thought. There is some indication that the ionization fractions for the clouds are higher than we would have expected.

f. Isotopic fractionation. The fingerprint of stellar nucleosynthesis, and of the chemical history of the galaxy, is most clearly seen in the abundances of the isotopes. For example, deuterium, D, is easily destroyed in stellar interiors via low temperature ($\approx 10^6$ K) reactions, but was synthesized in the Big Bang via nonequilibrium nucleosynthesis in the expanding universe during the first few minutes of its existence. Thus its abundance was fixed primordially. The rates of change of temperature and density during this initial epoch were fixed by the rate of expansion, which depends on the amount of mass in the universe; the slower the expansion rate in this early period, the closer to equilibrium will be the proton process products and the lower the D abundance. Because of the overwhelming abundance of H, it is very difficult to observe the D abundance directly, especially in stellar atmospheres but also in the wings of the Lyα line from the interstellar medium. With an anticipated abundance ratio $D/H \approx 10^{-5}$, current observations of interstellar absorption lines cannot place good limits on this cosmologically important number.

It is possible, however, to determine the D/H ratio through a different avenue, the isotopic shift caused by the mass ratio of D/H in the rotational and vibrational spectrum of the H_2 molecule compared with HD. This method is not free from difficulties, though, because of the many possible routes available for depletion of HD through molecular formation. The reaction energies for the isotopic species are slightly different, by several tens of degrees (hundredths of an electron volt), which at interstellar temperatures produces substantial differences in reaction rates. A detailed calculation of the isotope's mobility and reaction among the different species is required.

Both H_2 and HD are formed on grain surfaces. The subsequent reactions that involve these molecules can enhance the D abundance in the more complex species which are formed, and also open up new channels. For example,

$$H_3^+ + HD \rightarrow H_2D^+ + H_2, \qquad (73)$$

which then competes with H_3^+ in the formation of hydrocarbons. This reaction is especially important because of the influence of cosmic-ray ionization and dissociative recombination on the fractionation process. The ratio of H_3^+ to H_2D^+ is given by

$$H_2 + H(CR) \rightarrow H_3^+, \qquad (74)$$

$$H_3^+ + e \rightarrow H_2 + H, \qquad (75)$$

$$H_3^+ + HD \rightarrow H_2D^+, \qquad (76)$$

$$H_2D^+ + H_2 \rightarrow H_3^+ + H, \qquad (77)$$

$$H_2D^+ + e \rightarrow HD + H, \qquad (78)$$

$$\rightarrow H_2 + D, \qquad (79)$$

where H(CR) represents a cosmic-ray proton.

One well observed species, HCO^+, is observed to yield a higher than expected D/H ratio. This appears to be due to the protonation reaction chain

$$H_3^+ + CO \rightarrow HCO^+ + H, \qquad (80)$$

$$H_2D^+ + CO \rightarrow HCO^+ + D, \qquad (81)$$

$$\rightarrow DCO^+ + H, \qquad (82)$$

where both DCO^+ and HCO^+ are destroyed via dissociative recombination to yield CO.

Other means have been determined for solving for the electron density, for instance, the ratio of H_2O to HDO and of H_2D^+ to H_3^+. These are all uncertain because of the incompleteness of the abundance catalogs for various clouds; that is, many of the intermediate reactants are not well determined, and this hampers understanding of the conditions in the clouds.

Similar behavior is observed for the CNO isotopes. That is, their relative abundances in molecular combination depart from that expected on the basis of either the terrestrial abundance ratios or from nucleosynthesis. The effects are caused by fractionation reactions, usually through ion-transfer reactions with CO. In general, the enhancement of isotopic abundances of the CNO group seems to be due to charge-transfer reactions with isotopic ions. For example,

$$^{13}C^+ + {}^{12}CO \rightarrow {}^{13}CO + {}^{12}C^+ \tag{83}$$

will produce a net enhancement of ^{13}CO, since it will then free the carbon atom for other channels. At low density, where the molecules are exposed to UV as well as cosmic-ray ionization, this is an important mechanism. The abundances of the CNO isotopes will therefore not reflect the initial conditions in the cloud, but will rather reflect the processing which goes on during the time it takes for the molecules to come into equilibrium, about 10^6 years.

Because of the importance of the CNO isotopes for the study of stellar evolution and of the chemical history of the galaxy, this is one of the most important areas for astrochemistry. In addition, because of the dependence of these reactions on the ion fraction, which in turn depends on the electron fraction in the dense parts of the clouds, the isotopes become useful tools for studying the ionization of the environment, probing the electron density in portions of the cloud otherwise hidden from view.

g. Polycyclic aromatic hydrocarbons. We have repeatedly stressed that the smallest grains behave like very large molecules, never coming into strict equilibrium with the radiation field and radiating in diffuse bands in the near-IR. There are several likely candidates as identifications, all so-called PAHs. These are either linear or ring molecules, which are very stable in the presence of UV radiation. For instance, coronene ($C_{20}H_{20}$) is especially well studied in the laboratory, although it is not specifically implicated in any of the emission lines. These molecules produce broad optical absorption bands, like the diffuse interstellar features which have been known since Merrill's discovery of the λ4430 Å band in the 1930s, but are important mainly for their vibrational transitions in the 3- to 20-μm region, which in aggregate match virtually every unidentified feature. They can be built in the process of forming dust grains, in the atmospheres of carbon-rich giants and supergiants, and also from the process of grain photodestruction. At interstellar conditions, their partial pressures allow them to remain stable against evaporation. They should, however, also be strong absorbers in the UV; this contradicts many available observations, which place strong limits (less than a few percent) on the absorption-band strengths for these features. On the other hand, the IR *cirrus*, the ubiquitous diffuse emission detected by the *IRAS* and *ISO* missions throughout the galactic plane, cannot be explained any other way. The PAHs could incorporate as much as 10% of the available carbon in the ISM, making them a major component of the dust.

Recent infrared observations with the *ISO* satellite have detected fine structure in several of the bands, which indicates variable PAH composition throughout the medium. However, it is too early to begin identifying any of the particular species. These species have been detected in the IR emission from the comma of comet Halley as well, and it appears that in at least some environments, the chemistry required to produce the PAHs is sufficiently efficient that large abundances can be achieved. It has been suggested that fullerenes (such as C_{60}) are responsible for some of the small grain population, but the spectral signatures—particularly in the visible and ultraviolet—have yet to be seen.

A remaining problem with the PAH explanation for the diffuse bands is the lack of an UV signature of these species. Complex organic molecules produce deep absorption bands in the vacuum UV, between 2000 and 3000 Å. To date, these have not been observed. An important result is the discovery of deep λ4430 Å absorption against SN 1987a in the Large Magellanic Cloud. This argues that at least this galaxy may have the same small grain component as our galaxy, despite its lower than solar metal abundances.

VI. COSMOLOGICAL CHEMISTRY

On area that will certainly develop in the early years of the twenty-first century is the study of chemistry in the early universe. The need for this is obvious. Galaxies formed at a very early epoch within the much larger cluster and supercluster scale masses that grew from primordial seed density fluctuations. This requires energy dissipation to promote the collapse toward progressively lower masses, precisely as we expect in star formation. However, since stars did not form before this time, there were no heavy elements (in particular CNO) to form the major coolant molecules such as CO. It is possible, however, to form H_2 and H_3^+ in the expanding gas, provided sufficient ionization remains after the recombination that forms the cosmic background radiation. The efficiency of formation is very low and so is the abundance of the molecular species, but

TABLE I Some Molecules Detected in Cosmic Environments[a]

Molecule	Wavelengths	Transitions	Environment
H_2	UV, IR	E, VR	CSE, PS, ISM, MC
CO	UV, IR, mm	R, VR, R	CSE, PS, ISM, MC
CH	Opt, IR, cm	E, R, Λ-doubling	ISM, MC
OH	UV, IR, cm	E, R, Λ-doubling	CSE, PS, MC; masers
SiO	IR, mm	VR, R; maser	CSE, PS, MC
CS	mm	R	CSE, MC
HCN	IR, mm	VR, R	CSE, PS, MC
HNC	mm	R	CSE, MC
H_2O	mm, cm	R	CSE, MC
HCO^+	mm	R	CSE, MC
H_2D^+	sub-mm	R	MC
NH_3	IR, mm, cm	R, maser	CSE, MC, PS
H_2O	IR, mm	R, maser, ice	CSE, PS, MC
H_2CO	mm, cm	R	MC
$HC_{11}N$	cm	R	CSE, MC

[a] Key: Transitions: E, electronic; R, rotational; VR, vibrotational. Environments: CSE, circumstellar envelopes; PS, protostars; ISM interstellar medium; MC, molecular clouds.

any cooling will lead to some structure formation. Other molecules, particularly LiH, are expected to form from the chemical mix emerging from pregalactic nucleosynthesis.

VII. CONCLUSION

The field of observational astrochemistry is at an important stage in its development. Several large-millimeter telescopes are currently operating. With increased spatial resolution, the sites of complex molecular processing can be isolated from the overall structure of molecular clouds. More important is the fact that several interferometers are either operating (such as BIMA in California and IRAM near Grenoble), or under construction (ALMA in Chile). Many large-aperture millimeter and submillimeter telescopes are now available (such as the JCMT on Mauna Kea). These provide detailed comparative maps of clouds in the lines of specific molecules. Extragalactic astrochemistry is coming of age, and many galaxies have now been studied in diatomic and even more complex species. The crucial step in determining abundances, and whether these reflect local chemistry or excitation conditions, can only be accomplished through the comparison between species which trace each. The coming years promise to realize this goal.

SEE ALSO THE FOLLOWING ARTICLES

Cosmic Radiation • Galactic Structure and Evolution • Infrared Astronomy • Interstellar Matter • Planetary Atmospheres • Quantum Chemistry • Galactic Structure and Evolution • Surface Chemistry • Ultraviolet Space Astronomy

BIBLIOGRAPHY

Balin, R., Encrenaz, P., and Lequeux, J., eds. (1974). “Atomic and Molecular Physics and the Interstellar Medium: Les Houches Summer School XXVI,” North-Holland, Amsterdam.

Burton, W. B., Elmegreen, B. G., and Genzel, R. (1992). “The Galactic Interstellar Medium,” Springer-Verlag, Berlin.

Caselli, P., Walmsley, C. M., Terzieva, R., and Herbst, E. (1998). “The ionization fraction in dense cloud cores,” *Astrophys. J.* **499,** 234.

Draine, B. T., and Katz, N. (1986). “Magnetohydrodynamic shocks in diffuse clouds: I. Chemical processes,” *Astrophys. J.* **306,** 655.

Duley, W. W., and Williams, D. A. (1984). “Interstellar Chemistry,” Academic Press, New York.

Dyson, J. E., and Williams, D. A. (1997). “Physics of the Interstellar Medium,” 2nd ed., IOP Press, Bristol, U.K.

Elitzur, M. (1992). “Astrophysical Masers,” Kluwer, Dordrecht, The Netherlands.

Flower, D. R., and Pineau des Forêts, G. (1990). “Thermal and chemical evolution in interstellar clouds,” *MNRAS* **247,** 500.

Graedel, T. E., Langer, W. D., and Frerking, M. A. (1982). “The kinetic chemistry of dense interstellar clouds,” *Astrophys. J. Suppl.* **48,** 321.

Herzberg, G. (1971). “The Spectra and Structures of Simple Free Radicals,” Cornell University Press, Ithaca, NY; reprinted by Dover Books, New York.

Hollenbach, D. J., and Thronson, H. A., Jr., eds. (1987). “Interstellar Processes,” D. Reidel, Dordrecht, The Netherlands.

Millar, T. J., Farquhar, P. R. A., and Willacy, K. (1996). “The UMIST database for astrochemistry 1995” (and subsequent updates). *Astron. Astrophys. Suppl.* **121,** 139.

Millar, T. J., and Raga, A., eds. (1995). “Shocks in Astrophysics,” Kluwer, Dordrecht, The Netherlands.

Pauzat, F., Talbi, D., and Ellinger, Y. (1997). “The PAH hypothesis: A computational experiment in the combined effects of ionization and dehyrogenation on the IR signatures,” *Astron. Astrophys.* **319,** 318.

Rohlfs, K., Wilson, T. L., and Huettmeister, S. (1998). “Tools of Radio Astronomy,” Springer-Verlag, Berlin.

Spitzer, L., Jr. (1978). “Physical Processes in the Interstellar Medium,” Wiley, New York.

Spitzer, L., Jr. (1983). “Searching between the Stars,” Yale University Press, New Haven, CT.

Turner, B. E., and Ziurys, L. M. (1988). *In* “Galactic and Extragalactic Radio Astronomy” (G. Verschuur and K. Kellermann, eds.), p. 200, Springer-Verlag, New York.

van Dishoek, E. F., and Black, J. H. (1986). “Comprehensive models of diffuse interstellar clouds: physical conditions and molecular abundances,” *Astrophys. J. Suppl.* **62,** 109.

Vardya, M. S., and Tarafdar, S. P., eds. (1987). “Astrochemistry: IAU Symposium Nr. 120,” D. Reidel, Dordrecht, The Netherlands.

Celestial Mechanics

Steven N. Shore
Indiana University, South Bend

GLOSSARY

Astronomical unit (AU) Distance of the earth from the sun.

Ecliptic Apparent orbit of the sun, the plane of the earth's orbit.

Epicyclic frequency Rate at which an orbiting body sees radial and angular oscillations in nearly comoving orbits due to variations in eccentricities of the orbits.

Mean anomaly Mean rate of motion of a body in an ellipse, relative to the center of the osculating circle.

Osculating circular orbit Literally, the "kissing" orbit; the circular reference orbit that precisely matches an inscribed ellipse along the major axis

Perigee, perihelion Distance of closest approach to the central body, the earth or sun in this instance, respectively, in a conic section. The opposite of apogee or aphelion.

CELESTIAL MECHANICS is the study of dynamics in gravitational fields of cosmic bodies. In recent years, this has come to include gravitational statistical mechanics, galactic dynamics, nonlinear stability theory and chaos, and the practical field of satellite dynamics and attitude control.

I. HISTORICAL INTRODUCTION

The history of celestial mechanics is essentially the history of classical physics. The problem of predicting the motion of the planet is the central problem of most of the past two thousand years of physical investigation.

The first attempts to construct mathematical or physical cosmologies were those of the Platonists during the fourth century B.C. These included the model of Eudoxus, who introduced the homocentric spheres, and Aristotle, who included physical arguments in the Eudoxian system. The primary assumption of the immobility of the earth leads to very complex motions, which must be reproduced using compound circular motions in a series of nested inclined spheres. In effect, this early approach is like a Fourier analysis of the planetary motions into a set of periods and inclinations of the spheres.

The first alteration in the basic schema was introduced in Apollonius, who added the eccentricity, thereby allowing the motion to be regular about a point displaced from the center of the earth. In addition, he introduced the epicycle into the system. This is a small secondary path on which a planet is transported and which has a period that may not be the same as the period on the deferent circle. The deferent is the path along which the planet moves with the mean period; the epicycle causes periodic accelerations and decelerations relative to this mean value. Apollonius also proved a theorem for the determination of the relative radius of the epicycle compared with the deferent using the stationary points in the orbit, those points at which the motion of the planet appears to halt before it reverses its projected direction of motion.

The equant was added by Ptolemy (second century A.D.) as a modification of the eccentric, a point on the opposite side of the center of the deferent circle that carries the planet. This added point was also permitted to move, as required for the motion of the moon and Mercury. All of these constructions were justified under the rubric of "preserving the appearances," a dictum attributed to Plato. In addition, the physical basis for the model derived from the Aristotelian doctrine of geocentricity and contact forces. These Hellenistic astronomers were doing celestial mechanics, according to their lights.

The discovery of precession of the equinoxes by Hipparchos (second century B.C.) provoked little theoretical activity until the eleventh century. Al Bitruji introduced the *trepidation*, a mechanism that permitted the multiple periodicity that appeared to be required to explain the variation in the rate of precession. The mechanism, repeated in Copernicus, derives from the erroneous determination of the period of precession in which the rate of precession of the poles varied with time from about 1° per century to 0.75° per century. The mechanism demanded something like an equant in the polar motion. As retained by Copernicus, this introduced a substantial complication into the theory of rotation of the earth and also the calculation of celestial motions and the correction of star catalogs. It was not until the seventeenth century that the error was recognized and quietly suppressed, but the trepidation represents one of the few innovations in the basic dynamical theory of the heavens in the period between the Alexandrian school of astronomy and the early Renaissance.

To Copernicus (1472–1543) is ascribed the first concerted effort to break with the geocentricity of the Greek constructions. While Aristarchus had proposed a heliocentric system in the second century B.C., the system was never carried through to include the computation of planetary orbits. Copernicus introduced the machinery for the determination of planetary phenomena but also argued physically for the added effects attendant on the overthrow of the geocentric picture. This especially included the physical explanation for the precession and the alteration of the ascending node of the lunar orbit.

Celestial mechanics as we now think of it really started with Kepler (1571–1630), although in a rather oblique form. It was Kepler who first pointed to a physical driving influence of the sun and argued that its position at the center of the system was more than coincidence and of more than kinematical significance. He attributed the planetary orbits to magnetic influence by the sun. Kepler's basic point, however, that some kind of action at a distance is necessary for planetary motion, and not contiguous geocentric spheres, provided the critical insight for the later development of celestial mechanics.

Kepler's principal contribution is summarized in his laws of planetary motion. Originally derived semiempirically, by solving for the detailed motion of the planets (especially Mars) from Tycho's observations, these laws embody the basic properties of two-body orbits. The *first law* is that the planetary orbits describe conic sections of various eccentricities and semimajor axes. Closed, that is to say periodic, orbits are circles or ellipses. Aperiodic orbits are parabolas or hyperbolas. The *second law* states that a planet will sweep out equal areas of arc in equal times. This is also a statement, as was later demonstrated by Newton and his successors, of the conservation of angular momentum. The *third law*, which is the main dynamical result, is also called the "Harmonic Law." It states that the orbital period of a planet, P, is related to its distance from the central body (in the specific case of the solar system as a whole, the sun), a, by $P^2 \sim a^3$. In more general form, speaking ahistorically, this can be stated as $G(M_1 + M_2) = a^3\Omega^2$, where G is the gravitational constant, $\Omega = 2\pi/P$ is the orbital frequency, and M_1 and M_2 are the masses of the two bodies. Kepler's specific form of the law holds when the period is measured in years and the distance is scaled to the semimajor axis of the earth's orbit, the *astronomical unit* (AU).

The dynamical foundations of celestial mechanics derive from Newton's (1642–1716) discovery of the Universal Gravitational Law. This law states that, by action-at-a-distance, every mass, M, exerts a force proportional to the inverse square of its distance, r, from every other mass, m, in the proportion $F_{\text{grav}} = -GMm/r^2$. For a distended mass, this amounts to the summation over the individual mass elements to determine the gravitational acceleration, $g_{\text{grav}} = -G\int[dm(r)/r^2]$, and Newton produced a proof that only the interior mass attracts a body in a homogeneous spheroid. He also added the formalism that the gravitational force derives from a potential, Φ (a term introduced nearly 150 years later by George Green, although employed in all earlier extensions of Newtonian formalism) $\mathbf{F} = -\nabla\Phi$. It was through the application of this general principle, and the assertion that G is a universal constant independent of the composition of the body, that permitted the generalization of dynamics and enabled the computation of planetary orbits. Newtonian methodology was quickly extended even to cosmological problems of large-scale distribution of mass in space. The principle of inertia, originally discussed by Galileo and Descartes, was elevated to a basic axiom in the laws of motion and, combined with the gravitational law, served as the basis of modern dynamical theory. Newtonian gravitation thus provided a basis for celestial mechanics by yielding the means for deriving Kepler's third law from first principles and for identifying the proportionality constant with the mass of the bodies.

The two centuries following Newton's statement of his dynamical laws in the *Principia* saw an explosive development of the mathematical formalism required to extend the theory to many celestial mechanics problems. Laplace (1749–1827), in the *Mechanique Celeste*, the first comprehensive treatment of gravitational mechanics, succeeded in developing the concept of "field" into a complex structure capable of calculating such diverse phenomena as orbital resonances and tidal acceleration. He also introduced the general equation for the gravitational field external to a massive body, which solves the equation $\nabla^2\Phi = 0$, the *Laplace equation*. Lagrange (1746–1813) stated the general method for incorporating energy (*vis viva*) into the system of dynamical equations, was the first to find a solution to the restricted three-body problem, and developed a method for treatment of generalized coordinate systems by variational principles. Clairaut (1713–1765) calculated the first example of the inverse problem, the figure and density distribution of the interior of the earth. Bessel (1784–1846) extended the calculation of planetary motions. Gauss (1777–1855), among his numerous contributions to mathematical physics, developed the method of least squares for the calculation of orbits in the presence of observational uncertainty and introduced spherical harmonics (also explored by Legendre) into the description of the gravitational field of finite bodies. Poisson (1781–1840) and Green (1793–1841) created general methods for the calculation of gravitational potentials. Maclauren, Riemann (1826–1866), and Jacobi (1804–1851) succeeded in finding general solutions for the gravitational equilibrium of rotating incompressible fluid masses, work that was extended by George Darwin (1845–1912) and James Jeans (1877–1946) to the problem of the formation of the moon and the description of the tides.

But perhaps the pivotal event of nineteenth-century celestial mechanics was the successful prediction of the orbit of Neptune through the analysis of orbital perturbations on Uranus. Accomplished by Adams and Le Verrier, the discovery in 1846 of this hitherto unknown planet served to inspire much of the revived interest in orbital dynamics and the detailed exploration of perturbation theory. Significant contributions were made by Delaunay, Hill, Hansen, Brown, and Airy in this regard, especially concerning the motion of the moon (a problem that had even perplexed Ptolemy and still one of the most challenging dynamical calculations). Almost immediately on their invention, Hamilton applied quaternions and Gibbs applied vectors to orbital computations with considerable success. The energy principle became more prominent after the development of potential theory, spurred by a renewed interest in tides and the configurations of strongly perturbed rotating bodies. In fact, virtually every physicist of the past century was involved in the exploration of the properties of gravitational fields for bodies of different shape and mass.

Modern theoretical celestial mechanics was founded in the last quarter of the nineteenth century by Liapounov and Poincaré, who were the first to detail the conditions required for general orbital stability. Newcomb, Plummer, Moulton, and Brouwer were among those who made substantial contributions to this problem. Much of the theoretical development was driven by thoroughly applied results—the need to predict planetary positions for time keeping and navigation, the need for a consistent geodetic reference frame, calculation of tides—although occasionally it came from the simple desire to push the calculation to yet another decimal point (a spirit best described by the American poet Walt Whitman in *The Learned Astronomer*).

A major advance in the twentieth century has been the generalization of celestial mechanics to the problem of orbits of stars in galaxies. Statistical mechanics was first applied by Schwarzschild and Kapteyn to the velocity distribution of stars in the vicinity of the sun. The differential rotation of the galaxy was demonstrated by Oort. The discovery of internal orbital motion in spiral nebulae by Slipher and Hubble, generalized by many later investigations, permitted the determination of the masses of these enormous stellar aggregates. Tidal interaction between galaxies was discovered observationally by Vorontsov-Veliaminov and Arp, inspiring considerable theoretical effort in the past two decades. Chandrasekhar and von Neumann in the 1940s began the study of dynamical interaction between stars and their neighbors, a field that has blossomed in recent years. Jeans and Contopoulos, among others, have explored the role of resonances and integrals of motion in the dynamics of stars in galaxies. Density wave theory, originally developed to explain spiral structure of galaxies by Lin and Shu, has been generalized to the study of wave phenomena in planetary rings. And the list of applications of gravitational dynamics to astronomical problems expands each year.

The spur in this century to developing methods for the computation of orbits has been the advent of spaceflight. For the first time, gravitational mechanics has played a central role in engineering, and many of the theoretical developments concerning control theory, nonlinear mechanics, and orbit prediction have resulted from this need.

We here conclude this broadbrush overview of the history of developmental work in celestial mechanics. Keep in mind that we have merely cataloged a few of the enormous advances achieved in this field. More detailed histories can offer the full extent of this spectacular, and ongoing, intellectual effort. More details will be found in the subsequent sections of this article.

II. CONIC SECTIONS AND ORBITAL ELEMENTS

In the classical two-body problem, the orbit is described by a finite number of *elements*. The eccentricity, the measure of the flatness of the conic, is given by:

$$e = \left(1 - \frac{b^2}{a^2}\right)^{1/2} \tag{1}$$

where a is the semimajor axis, or half the length of the major axis, and b is the half-width of the minor axis. For a circle, $e = 0$, for a parabola $e = 1$, and for a hyperbola $e > 1$. The quantity $a(1 - e^2)$ occurs frequently and is called the *semilatus rectum*. The inclination, i, is defined relative to some reference plane, usually taken to be the solar orbit for solar system celestial mechanical calculations, which is also called the *ecliptic*. The mean anomaly, $M = n\Delta t$ depends on the mean angular rate $n = 2\pi/P$, where P is the period, of the orbit about the center of the conic. The osculating circle is the one that just touches the major axis at its extrema and represents the circle that can be transformed into the ellipse of the observed eccentricity by inclining the plane of reference. The *eccentric anomaly*, E, measures the difference between the mean motion, referred to the osculating circle, and the motion about the center of the conic. The mean motion is related to the eccentric anomaly by the *Kepler equation*, one of the earliest transcendental equations of mathematical astronomy:

$$M = n\Delta t = E - e \sin E \tag{2}$$

where Δt is the elapsed time from some initial epoch t_0. The *true anomaly, f*, is the angular motion relative to the focus. We will return to this momentarily. The *longitude of perihelion, ω*, is measured in the plane of the orbit relative to the ascending node. For the solar system, this point is the vernal equinox, the point where the sun's orbit crosses the earth's equator. This point is, however, an arbitrary element that depends on the definition of the reference plane. The same is true of the *argument of perihelion*, Ω, which is measured in the reference plane (in the solar system, along with ecliptic).

For a conic section, the radial position is given by:

$$r = a(1 - e \cos E) \tag{3}$$

and therefore the *true anomaly, f*, is given by:

$$\cos f = \frac{\cos E - e}{1 - e \cos E} \tag{4}$$

This angular motion is referred to as the longitude of closest approach and is given by $\phi - \omega$ where ϕ is the angular motion about the focus and the coordinate appropriate to the description of the angular momentum of the orbiting body. Another form for this equation is:

$$\tan\frac{f}{2} = \left(\frac{1+e}{1-e}\right)^{1/2} \tan\frac{E}{2} \tag{5}$$

These are purely geometric relations, having only a kinematic basis. It is one of the results of classical celestial mechanics to provide a firm dynamical foundation for these relations, especially the Keplerian equal areas law and the harmonic law. For more complex orbits, those often found in multiple systems, it is usually best to calculate the motion of the body exactly. However, the relative ephemeris of an object can be expressed in terms of these orbital elements, which may change in time depending on the orbital stability. For instance, earth satellite orbits (e.g., the space shuttle or the Hubble Space Telescope) are referred to the earth's equatorial plane, while deep space satellites (like Voyager) are referred back to the ecliptic.

III. THE DYNAMICAL PROBLEM IN GRAVITATIONAL PHYSICS

A. The Two-Body or Central Field Problem

The central field problem distinguishes celestial mechanics from other areas of classical dynamics. This deals with the motion of a test particle, whose mass is negligible with respect to the central body, in the gravitational field of a point mass. The extended version of this problem is to allow the central mass to have a finite spatial extent, to depart from spherical symmetry, and perhaps to rotate. The basic Newtonian problem is the following.

The equation of motion for a single particle under the action of an arbitrary force, $\mathbf{F}$, is $m\ddot{r} = \mathbf{F}$, where m is the mass of the particle and $\mathbf{r}$ is its position vector. In most of what follows, we will assume a cylindrical or spherical coordinate system (which will be specified), the most appropriate one for use with planar motion. The radial component of the momentum equation is

$$\frac{d^2r}{dt^2} - \frac{v_\phi^2}{r} = a_r \tag{6}$$

and for the angular motion

$$\frac{d}{dt}\left(r^2\frac{d\phi}{dt}\right) = a_\phi \tag{7}$$

and we will ignore the z component for the time being. In the absence of any angular torques, that is, for centrally symmetric problems, the angular motion can be collapsed into a conserved quality, the specific angular momentum, which is

$$j = rv_\phi = r^2\dot{\phi}, \tag{8}$$

and therefore the equation of motion is strictly radial for the central force problem

$$\ddot{r} = -\frac{\partial\Phi}{\partial r} + \frac{j^2}{r^3} \tag{9}$$

Here the solution to the central field problem makes specific use of the gravitational force. The substitution of

$$a_r = -\frac{GM}{r^2} \tag{10}$$

completes the specification of the central field problem. Since there is a conserved quantity for the angular coordinate, we can substitute $d/dt = (j/(d/r^2)/d\phi)$, so that defining $u = 1/r$, the equation for radial motion reduces to

$$\frac{d^2u}{d\phi^2} + u = -\frac{GM}{j^2}. \tag{11}$$

This is the equation for a *harmonic oscillator* in the radial direction. Physically, this is an important result, indicating that the motion remains bounded within a finite range of radius only for a limited range of j and binding energy, and repeats periodically for the cyclically varying angle ϕ. In other words, for small enough values of j, the orbit closes on itself precisely; if j is large enough, the orbit is unbounded. The radial distance from a point mass is therefore

$$r = a(1 - e^2)/[1 + e\cos(\phi - \phi_0)] \tag{12}$$

where we can now identify

$$e = \left(1 - \frac{j^2}{GMa}\right)^{1/2} \tag{13}$$

as the relation between the angular momentum and the eccentricity (note that for $a = \infty$, as in a parabola, $e = 1$). The term $\phi - \phi_0$ is the true anomaly, f as earlier. The two retrograde for a period until the viewpoint from the earth reaches a tangent point, at which time the planet's motion reverses. Dynamically, a moving reference point in a differentially revolving system of masses, like the solar system viewed from the earth, knows of a natural frequency at which bodies commit epicyclic or retrograde motion. The only force that "drives" this motion is the Coriolis effect. The frequency, called the epicyclic frequency, is given by $k = 2\Omega$. The importance for understanding the kinematics within the solar system, or in any rotating system, is that any moving observer will see epicyclic motion as long as the frame of observation is assumed to be stationary. We shall generalize this result below.

B. Gravitational Potential Theory

The primary problem in central field dynamics is the calculation of the gravitational potential $\Phi(\mathbf{r})$. It is defined by Poisson's equation

$$\nabla^2\Phi = -4\pi G\rho \tag{14}$$

where ρ is the density. For a vacuum, this reduces to the Laplace equation. The solution for this equation is

$$\Phi(\mathbf{r}) = -G\int \frac{\rho(\mathbf{r}')\,d\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|} \tag{15}$$

where the kernel of this equation is called the Green function. For nearly spherically symmetric bodies, this reduces to a series in Legendre polynomials, P_{nm}, whose coefficients depend on the density (mass) distribution.

When the central gravitating body is not spherical, additional terms are required for this power series, and the equations of motion reflect the presence of a torque. For the external gravitational field, this potential is generally expressed by the series in the radius, R, latitude, θ, and longitude ϕ:

$$\begin{aligned}\Phi(r,\theta,\phi) = -\frac{GM}{r} + \frac{GM}{r}\Bigg\{\sum_{n=2}^{\infty}\Bigg[\left(\frac{R}{r}\right)^n J_n P_{n0}(\cos\theta)\\ + \sum_{m=1}^{n}\left(\frac{R}{r}\right)^n (C_{nm}\cos m\phi + S_{nm}\sin m\phi)\\ \times P_{nm}(\cos\theta)\Bigg]\Bigg]\Bigg\}\end{aligned} \tag{16}$$

Here R and M are the radius and mass of the planet, respectively. The coefficients C_{nm} and S_{nm} are called *tesseral* ($m \neq n$) or *sectoral* harmonic coefficients ($m = n$), and the J_n are the *zonal* harmonic coefficients. These coefficients are determined through fits to satellite orbits, and the same description can be applied to any planetary gravitational field for which *in situ* measurements have been accomplished. For instance, the coefficient J_2 measures the gravitational quadrupole moment of the body and is responsible for the largest contribution to the precession of the orbit of a planet about a nonspherical central body. The application of this formalism to planetary gravitational fields can also be accomplished by the solution of satellite orbits, but this is likely only to provide the lowest order terms. The specification of the potential for a planet provides important information about the internal mass distribution, and in this sense the study of the interior structure of a planet through its external gravitational potential is a classic *inverse problem*, attempting to solve uniquely for the distribution in the density that gives rise to the observed gravitational field.

The definition of the potential for various classes of distended bodies began in the eighteenth century with the work of Clairault, Maclauren, Laplace, and Roche. Classic work on potentials involved some of the most illustrious mathematicians of the past century, including Gauss, Riemann, Jacobi, and Poincaré. These were mainly applications of gravitation theory to the figure of the earth. Recent discussions of the problem of calculating these potentials are due to Kellogg and Chandrasekhar. For astrophysical systems the problem is harder because the potentials are often composite, consisting of spheroids and disks in combination and often requiring bars or triaxial ellipsoids.

One extreme example of a nonspherical gravitational field is that of the *three-body problem* (see Section III.F). Here one assumes that a point mass orbits in the corotating frame of two (possibly unequal) masses. The motion of the large masses has a period P and is assumed to be circular about the center of mass. The third body, whose mass is negligible, orbits with any arbitrary inclination i to the (m_1, m_2) plane and with some initial energy, also called the Jacobi constant. The particle is constrained by minimum energy to move along zero velocity or equipotential surfaces if the body is initially at rest in the corotating frame. The precise form of the potential was first discussed by Roche, who also derived the limiting radius of a self-gravitating body being acted upon tidally by a perturber that now bears his name. The centrifugal term is at the rate of rotation of the massive components and is taken relative to the center of mass. The individual gravitational terms are referred to the locations of the massive bodies:

$$\begin{aligned}\Phi_{\text{Roche}}(x, y, z) = &-\frac{1}{2}(x^2 + y^2)\Omega^2 \\ &- \frac{GM_1}{\left[(x - x_1)^2 + y^2 + z^2\right]^{1/2}} \\ &- \frac{GM_2}{\left[(x - x_2)^2 + y^2 + z^2\right]^{1/2}} \qquad (17)\end{aligned}$$

where x_j is the position of the jth body relative to the center of mass. Although this potential involves point masses, it can be generalized to finite bodies.

C. Perturbation Theory

Gravity is a force of infinite range, and it is impossible for any pair of objects to be truly isolated and subject to a point mass central field. The closed form solution of the two-body problem thus represents an idealized orbit. The departures from this trajectory are treated by perturbation theory. The action of any additional mass in a system can be thought of as a perturbation on the central field problem. The basic assumption of perturbation theory is that the magnitude of the disturbance is small, so that the dynamical equations remain linear. In the presence of massive, nearby objects, or in the vicinity of resonances, nonlinear techniques must be applied.

Perturbations introduced by the action of external bodies fall into several categories. For examle, the effect of finite size of the central object in a two-body problem introduces precession in an orbit that can be treated as a perturbation above the point central field. These orbital perturbations represent simple time-independent and periodic departures from the closed ellipse. They will cause the orbiting body to evolve toward a stable trajectory if the central body is not rotating. Rotation of the central body introduces an additional time scale into the problem and can produce secular instabilities in the orbit. The basic starting point of a perturbation calculation is that one already knows what the orbit is for a particle. One is interested in finding out whether it is stable against small perturbations, due to other bodies, and what the evolution will be for the orbit.

One of the best examples of the effects of perturbations in the solar system is provided by the gravitational interaction between comets and the Jovian planets, especially Jupiter itself. New comets, that is, those coming into the inner solar system for the first time, begin their decent toward the sun in nearly parabolic orbits. As they come close to Jupiter, the acceleration provided by the planet changes the orbital angular momentum through a torque that, depending on the phase of the kick from the interaction, can either increase or decrease the eccentricity of the orbit. For capture, the eccentricity is decreased below unity. Comet Encke is one of the best examples of this, being trapped in an orbit that is nearly resonant with the Jovian period. Comet Halley is in a near resonance with Neptune. On the other hand, the eccentricity can be increased and the comet sent out of the solar system with an increased total energy and angular momentum in any hyperbolic orbit. The key factor is whether the perturbation is leading or trailing in the orbit. Several asteroid families are trapped in resonant orbits with Jupiter, notably the Apollo asteroidal group. The Trojans are trapped in orbits near the triangular Lagrangian points (L_4 and L_5) of Jupiter. These changes in the orbital properties occur in real time, that is to say in the course of a single orbit.

Tidal perturbations are important for the orbits of many satellites. In the course of time, a satellite gains angular momentum through interaction with the sun and moon, as well as because of the nonspherical gravitational field of the earth. Orbital ephemerides must be frequently updated to take these changes into account, especially for geosynchronous satellites.

The interaction of the sun and moon with the earth is responsible for several important physical effects, notably the precession of the rotation axis with time (the phenomenon first described by Hipparchos) and for the change in the length of the day due to tidal friction and dissipation of rotational energy. The tidal term results from the finite size of the earth relative to its orbital radius and to that of the moon. The differential gravitational acceleration across the body produces a torque that accelerates the moon outward and slows the earth's rotation.

As we have discussed previously, an orbit is characterized by a finite number of orbital elements. Under the influence of an external force $\Re$, called the *disturbing function*, all of these may vary. In particular, torques change a and e but also cause the orbit to precess so that Ω and ω vary in time. The presence of additional mass in the system changes the orbital frequency, n, through changes in M. If the angular momentum changes, it is possible for the total energy of the particle to change as well. The disturbing function has components in the cylindrically symmetric coordinate system we have been using: $R = \partial\Re/\partial r$ for the radial component, $S = \partial\Re/r\,\partial\phi$ for the azimuthal torque, and $W = \partial\Re/\partial z$ for the force perpendicular to the orbital plane. The full system of evolution equations for the orbital constants under the action of these forces is given by

$$\frac{da}{dt} = \frac{2}{a(1-e^2)^{1/2}}\left(Re\sin f + S\frac{a(1-e^2)}{r}\right) \tag{18}$$

$$\frac{de}{dt} = \frac{(1-e^2)^{1/2}}{na}\left\{eR\sin f + \left[\frac{a(1-e^2)}{r} - \frac{r}{a}\right]S\right\} \tag{19}$$

$$\frac{di}{dt} = [na(1-e^2)^{1/2}]^{-1}\frac{r}{a}\cos(\omega + f)W \tag{20}$$

$$\sin i\,\frac{d\Omega}{dt} = i\tan(\omega + f) \tag{21}$$

$$\begin{aligned}\frac{d\omega}{dt} = \frac{(1-e^2)^{1/2}}{nae}\Bigg[&-R\cos f \\ &+ S\left[\frac{r}{a(1-e^2)} + 1\right]\sin f\Bigg] \\ &- \frac{er\sin(f+\omega)\cos i}{a(1-e^2)}W\Bigg]\end{aligned} \tag{22}$$

$$\begin{aligned}\frac{dM}{dt} = n + \frac{1}{na}\Bigg[&\left(\frac{(1-e^2)\cos f}{e} - \frac{2r}{a}\right)R \\ &- \frac{(1-e^2)\sin f}{nae}\left(1 + \frac{r}{a(1-e^2)}\right)S\Bigg]\end{aligned} \tag{23}$$

Notice that the variation of the semimajor axis depends on both the radial and the torque, but because the orbit can be taken in the two-body problem as planar, there is no dependence on W. Changes in ω and Ω are equivalent to orbital precession. All of these may be periodic or secular, depending on the details of $\Re$. For a given disturbing function, this system of equations can be well explored using numerical methods.

D. Lunar Theory

Because of its centrality in the development of perturbation theory, the motion of the moon has its own terminology and also a special set of treatments. More is known about the motion of this body than of any other object in the solar system. After the placement of laser retroreflectors on the surface during the Apollo mission, regular ranging from the earth has determined the lunar distance as a function of time to an accuracy of tens of meters.

Historically, the departures of the moon's motion from circularity have been among the driving anomalies that celestial mechanics was designed to address and explain. The inequalities were expressed as variations in the longitude of the moon relative to that expected in the simplest theory in which its rate about the center of the deferent circle, or later around the eccentric. The development of theory by Newton, and workers in the past two centuries, has derived the precise forms of all of the classical inequalities. All motion is referred to both the position of the moon along the ecliptic. λ and that of the sun λ'. The most important terms are

1. the *evection*, discovered by Ptolemy, due to the motion of the longitude of perigee $n'^2a^2e^2\cos(2\omega - 2\lambda')$ as a departure from the eccentric deferent;
2. the *variation*, which depends on $n'^2a^2\cos(2\lambda - 2\lambda')$, where n' is the mean solar motion;
3. the *annual equation*, which depends only on the longitude of the perigee: $n'^2a^2e'\cos(\lambda' - \omega')$, where e' is the eccentricity of the solar orbit;
4. the *parallactic inequality*, $n^2a^3e\cos(\lambda' - \omega)/a'$, where a' is the semimajor axis of the solar orbit;
5. the *principal perturbation in longitude*, which depends on the inclination of the lunar orbit relative to the ecliptic $n^2\cos^2 i\,a^2\cos(2\lambda' - 2\Omega)$.

The chief problem is that the rotation of the earth and the disturbing tidal accelerations from the sun change the lunar orbit with time. Each of these motions must be treated separately in the system of equations that describe the lunar orbit.

The detailed study of the lunar motion has implications for geophysics as well. The length of the day is affected by the tidal interactions between the moon, sun, and earth. In the course of the millennia, the day has steadily lengthened; the change in the lunar orbital parameters reflect this because of the variation in the tidal torque on the moon. These variations yield important information about the tidal coupling mechanism and the rate of dissipation of the angular momentum of the earth–moon system.

E. General Relativity and Celestial Mechanics

One of the first tests of the General Theory of Relativity (GRT) was the calculation of the precession of the orbit of Mercury. A longstanding problem at the end of the nineteenth century was that it observed orbital precession rate exceeded the value of 5557 arcsec per century, produced by the combined effect of all of the planets, by about 43.11 ± 0.45 arcsec per century. The planetary precession value was based on the known masses of the planets and on the solar potential and assumed a spherical shape for the sun. It is a tribute to the efforts of theorists in the past century that a discrepancy of this tiny amount, less than 0.1%, was considered not only significant but compelling. Several solutions were suggested, including as light modification to the gravitational force law by making the law very slightly weaker at large distance giving the sun a quadrupole moment due to rotational distortion of about 40 milliarcsec, or the presence of a planet within the orbit of Mercury. The last of these suggestions, perhaps the most widely accepted solution at the close of the century, was made by Leverrier, whose successes in the prediction of the place of Neptune on the basis of perturbation theory led to the eventual universal acceptance of Newtonian methodology and represented the triumph of the past century in classical dynamics. This hypothetical planet. Vulcan, was actually reported during several eclipses toward the end of the century, but subsequent investigations have since eliminated it as a possible member of the solar system.

Einstein discovered the celestial mechanical consequences of GRT in 1914, just before his completion of GRT. General relativity produces a change in the gravitational field in the vicinity of a massive body. The distortion is equivalent to introducing an additional term in the harmonic oscillator equation

$$\frac{d^2u}{d\phi^2} + u = \frac{GM}{j^2} + 3GMu^2 \tag{24}$$

so that the magnitude of the advance of the perihelion is predicted to be $3GM/(a(1-e^2)) = 12\pi^2 a^2/(c^2 P^2(1-e^2))$, where c is the speed of light. Although Mercury is about 0.4 AU from the sun, this is only about 100 solar radii, and the precession predicted by relativity is 43 arcsec per century. For the earth, in contrast, the rate is predicted to be only about 4 arcsec per century. Along with the deflection of starlight observed during the 1919 eclipse expedition, this prediction stands as one of the great triumphs of the GRT.

Two types of GRT corrections are required for orbits. One is due to the precession of the orbit, the other is the radiation of gravitational waves. While not important in the vast majority of dynamical systems, it plays a role in binary star systems in which the components are massive, compact, and revolving with very short periods. The rate of gravitational wave radiation varies as Ω^6, where Ω is the orbital frequency, so that for binaries like PSR $1913 + 21$ (the best studied of the binary pulsars) it produces a secular change in the semimajor axis that can be measured from several years of observation. These effects can be incorporated into the standard perturbation evolution equations through modifications to the distubing function.

F. The Three-Body Problem

The most celebrated problem in celestial mechanics is the so-called *three-body problem*. First elucidated by Lagrange, this problem focuses on the determination of the allowed class of periodic motions for a massless particle orbiting a binary system. In this case, the motion is determined by the gravitational and centrifugal accelerations and also the Coriolis force. A closed form analytic solution is possible in only one case, that of equal masses in a circular orbit. This so-called *restricted three-body problem* can be specified by the curves of constant potential, also called the zero velocity surfaces. Consider a binary with a coplanar orbit for the third mass. In this case, a local coordinate system (ζ, η) is defined as centered at $(a, 1-a)$ so that the equations of motion are

$$\ddot{\zeta} + 2\Omega\dot{\eta} = -\left.\frac{\partial\Phi}{\partial\zeta}\right|_{\eta} \tag{25}$$

$$\ddot{\eta} - 2\Omega\dot{\zeta} = -\left.\frac{\partial\Phi}{\partial\eta}\right|_{\zeta} \tag{26}$$

Here the gravitational potential, Φ, is the Roche potential already discussed. The assumption required for this potential is that the two massive bodies are in a circular orbit about the center of mass. In the absence of eccentricity, stable orbits are possible in several regions of the orbital plane. These are defined by the condition that $\nabla\Phi = 0$ and are critical points in the solution of the equations of motion. These are stationary in the rotating frame. In the presence of eccentricity, they oscillate and produce a loss of stability, as we shall explain in Section IV.C.

Several critical points in the three-body potential dominate the motion of particles. They are specified as the points at which the gravitational acceleration vanishes.

Called the *Lagrangian points*, two lie perpendicular to the line of centers (the L_4 and L_5 points) and are due to the balance between centrifugal and gravitational forces modified by the Coriolis acceleration that governs the angular momentum of the particles. These points are actually potential maxima, and therefore the particle orbits are only quasi-stable. These points lose their stability for large mass ratio or in the presence of an eccentricity in the massive binary. Two Lagrangian points, the L_2 and L_3 points, lie along the line of centers but beyond the two massive bodies. These are minima in Φ and form a barrier to mass loss. Finally, the most important is L_1, the point of balance between gravitational accelerations of the two bodies, which lies along the line of centers between the massive members of the system. This point corresponds to the Roche radius, the separation at which the gravitational perturbation of the companion dominates over the self-gravitation of a deformable, compressible, self-gravitating body. Orbits are unconditionally unstable at L_1, L_2, and L_3. They are asymptotic at L_4 and L_5. If the central orbit is eccentric, the tidal component of the gravitational field at the Lagrangian points oscillates in the course of a single period. Depending on the local orbital frequency of a body about the L_4 or L_5 point, this oscillation may render the orbit unstable on a short time scale, transferring angular momentum to a trapped particle and sending it out of the system.

G. The Few-Body Problem

Few-body problems can be handled by conventional integrators, such as Runge–Kutta or Adams–Moulton methods. Here one calculates the position and velocity for each particle and then the precise two-body interaction for that body with every other particle in the system. Both methods are predictor-corrector procedures in which the next step is computed and corrected iteratively. Leapfrog methods, which use the velocity from one step and the positions from the previous step to compute the new positions, are also computationally efficient and stable. The basic problem is to solve the equations of motion for a particle at position $\mathbf{r}_i$,

$$m_i\ddot{\mathbf{r}}_i = -\sum_{j\neq i}\frac{Gm_im_j}{|\mathbf{r}_i-\mathbf{r}_j|^3}(\mathbf{r}_i-\mathbf{r}_j), \qquad (27)$$

supplemented by initial positions and velocities of the particles. Thus, the angular momentum is initially specified by the set of velocities and coordinates $(\mathbf{r}_i(0), \mathbf{r}_i(0))$. Efficient algorithms for the solution of this problem have been presented by Aarseth and have become standard in astrophysical calculations. For only a few bodies, Runge–Kutta procedure can be used. For more than about a dozen particles, this becomes computationally expensive, and leapfrog and predictor-corrector integrators are employed.

H. *N*-Body Problems

When the number of objects becomes large, more than a few dozen, then conventional integration techniques for orbit calculation become inadequate, and new procedures have to be introduced. The primary reason is not any change in the physics. Rather, it is the enormous number of individual quantities that must be tracked for the constituent particles (three coordinates, three velocities). Conventional few-body integrators require $\frac{1}{2}N(N-1)$ calculations per step to determine the motion of the particles. So the rate of calculation scales like N^2. For complex systems, like galaxies, this is prohibitively expensive and slow.

Instead, assume that the field is given by a statistical gravitational field, Φ_0, derived from the instantaneous density distribution. This is given, symbolically, by a convolution of the density distribution with the Green function $\Phi_0 = \rho\star\mathcal{G}$. Therefore, the Fourier transform of the potential is the product of the transforms of the density and the Green function, both of which are known at each step. The Green function depends on the coordinate system, but once computed can be tabulated to be reused without modification at every step. The potential calculation thus requires only a set of fast Fourier transforms (FFTs) for each step. The density is determined by some initial guess, the potential is computed, and the particles are then pushed within this potential. In the next step, the density is recalculated for the particle distribution, and the process is repeated. Because the FFT is used for the potential, the complexity of the computation grows slowly with particle number, N In N, and therefore handily wins over the more conventional N-body methods for large systems. While there are some problems with this method due to the aliasing and gridding effect of the particles (for instance, the particles must be resampled at each step to a uniform grid for the FFTs), experiments with systems up to 10^6 bodies show that the method is stable and reproducible.

A technique, called smoothed particle hydrodynamics (SPH) has been introduced to the N-body problem. This involves the combination of the two methods of conventional few-body integrators for each body with its immediate nearest neighbors and the overall computation of the distant gravitational potential via FFTs. SPH and FFT methods have also been merged with hierarchical tree searching techniques to improve their speed. Computation of many-body effects, while not important for satellite dynamics, which depend primarily on the central field approximation, may be important in the study of asteroid and comet orbital evolution and also applies to the computation of evolution of self-gravitating planetary ring systems. Refinements of these methods have been rapidly making inroads into astrophysical calculations and also plasma and solid state physics.

IV. APPLICATIONS: A FEW INTERESTING EXAMPLES

A. Satellite Dynamics: Transfer Orbits

The simplest practical application of celestial mechanics is in the computation of satellite dynamics, in particular, the transfer orbit between two planets. This orbit is the one that has the minimum energy and therefore has an aphelion at the inner planet at r_i and perihelion at the outer planet, r_{o}, so $a = (r_i + r_{\mathrm{o}})/2$. The eccentricity is $e = r_i/a$, and thus the period of the orbit is given by $(a^3/4\pi^2 GM)^{-1/2}$ and the binding energy can be calculated using $E = -GM/2a$ as before. The total change in the energy is $(r_{\mathrm{o}} - r_i)(v_{\mathrm{o}}^2 + v_i^2)/(r_{\mathrm{o}} + r_i)$, where v is the orbital velocity at each of the radii. The increase in the angular momentum required for a parking orbit is smaller than that required from the surface of the body. As a side piece of trivia, since 1925 this transfer orbit has been known as a *Hohmann ellipse*.

B. Solar System

Long before spacecraft encounters, celestial mechanics had been employed to determine the masses of those planets that possess moons. With the exceptions of Mercury and Venus, for which the arguments were more indirect, the masses of all the planets are now known from satellite observations. Detailed examination of the periodicities of their moons also reveals that they interact through resonant orbits, which causes the structuring of the radial distribution of the planetary satellite systems. Detailed observations of satellite motion also permit the determination of internal mass distribution and oblateness for most of the planets. These determinations have been augmented for the outer planets by direct flybys with the Voyager 1 and 2 spacecraft. Finally, mutual phenomena of the moons of several of the major planets provide the determination of satellite masses through the solution of the motion under mutual perturbations for the satellite systems.

The evolution of cometary orbits—especially Comet Halley for which there is a significant historical record—shows evidence for nongravitational forces. These are presumed to arise from the mass loss from the comet produced by interaction with the solar wind and radiation-induced outgassing. The change in the mass of the comet carries angular momentum because of the finite escape velocity for the lost gas. The effects of these “rocketlike” forces complicate the interpretation of cometary orbits. This problem is not merely of theoretical interest—accurate orbits are essential for predicting trajectories for artificial satellite encounters with comets like Halley.

C. Orbital Resonances

One of the oldest problems in celestial mechanics is that of resonances, the near coincidence of two frequencies or their rational multiples. Also called the problem of small divisor, it first appeared in the theory of the solar system. Two bodies are said to have commensurate periods when the ratio of their periods is the ratio of integers, $m : n$. The problem produced by such orbits is that their mutual interaction will enhance phase-realted torques and lock bodies into specific periods. The best example of this resonance phenomenon is a child in a swing. The period of the system is given by the length of the supports for the swing. But the child controls the amplitude of the swing by kicking in phase, or in antiphase, with respect to the maximum velocity along the arc. Thus, with the proper phasing, a periodic kick delivered at the minimum in the potential energy (maximum velocity) will produce an increasing amplitude. The same is true for bodies in gravitational orbits. These resonances grow without limit in the linear theory because the forces are always in phase, thereby producing a secular acceleration. In the case of two bodies in near resonance, the tendency will be for the bodies to attract to the resonance and lock. To see this analytically, consider a single particle one-dimensional harmonic oscillator subject to a periodic force: $m\ddot{x} + m\omega_0^2 x = F(t) = a \sin \omega t$ where ω is the frequency. The amplitude as a function of time is the Fourier transform of this frequency-dependent amplitude, so that if there is a near commensurability, there will be amplitudes which will grow linearly with time. The natural frequency of the oscillator is ω_0 so that the amplitude is given by

$$x(\omega) \sim \frac{a}{\omega_0^2 - \omega^2} \tag{28}$$

Notice that as $\omega \to \omega_0$ the energy if the mode $|x_\omega|^2$ grows without limit. In celestial mechanics as elsewhere, the periods are generally in integer ratios if they resonate.

In the case of two orbiting bodies, the largest torques are delivered at the point of closest approach. If the two bodies are not in circular orbits, then the phasing of the kick is important, and should the body be going at its maximum speed at the time the acceleration is delivered, its angular momentum will be increased and the eccentricity of the orbit should increase secularly.

Such resonances were first realized in the problem of the so-called “Great Inequality” of the orbits of the outer planets. Jupiter and Saturn have a period ratio of 5:2 for the ratio of their frequencies. Other resonances were noted in the gaps in the Saturn ring system (the Cassini and Encke divisions being the most famous) and in the ratio of rotation and revolution periods for Mercury (2:3). But perhaps the most spectacular examples are noted in the asteroid belt, manifested as the so-called *Kirkwood Gaps*. These are low population zones in the asteroidal distribution at points in resonance with the outer planets, especially Jupiter. The distribution of these gaps is complex because all possible commensurabilities should be present, and in fact there is denumerable infinity of them. The asteroid distribution, like the ring system of Saturn, represents a good example of a chaotic system because of the resonance and near resonance conditions.

The primary cause of resonant trapping is the beating between the orbital frequency of the body being acted upon and that of the perturber. The phasing of the perturbation produces an acceleration for a portion of the orbit and a compensating deceleration elsewhere in the orbit.

For a rotating coordinate system, we can introduce a new frequency, the *epicyclic frequency*, which is the rate at which bodies appear to revolve around the point of observation. This frequency results from the gradient in the central gravitational field and is important for resonant orbits. To see where it comes from, consider a moving observer in a coordinate system (r, ϕ) where ϕ is the azimuth. Assume that this observer revolves about a central body in a gravitational field $\Phi(r)$ at a distance r_0 and with an angular frequency Ω. The radial velocity is therefore assumed to vanish for this observer, and the angular speed of the frame is also constant. Now assume that we can create a comoving locally Cartesian coordinate system (ζ, η) around this orbit. A body with higher angular momentum than required for the circular orbit will therefore have a radial speed $\dot{\zeta}$ and a variable angular speed $r_0\dot{\phi} = \dot{\eta}$. For a corotating frame of reference, the equations of motion are

$$\ddot{\zeta} = 2\Omega\dot{\eta} - \frac{\partial^2\Phi}{\partial r^2}\zeta \tag{29}$$

for the radial equation and

$$\ddot{\eta} + 2\Omega\dot{\zeta} = 0 \tag{30}$$

for the angular equation. Therefore, there is a characteristic frequency

$$\kappa^2 = 3\Omega^2 + \frac{\partial^2\Phi}{\partial r^2} \tag{31}$$

which is the *epicyclic frequency* we were seeking. This frequency depends on the steepness of the gravitational field gradient. For instance, for a Keplerian orbit, $\Phi \sim r^{-1}$ and $\Phi'' > 0$, so that the epicyclic frequency is greater than the orbital frequency. It should be remarked that $\kappa = 0$ is also allowed for any central body comoving with the frame. In terms of the orbital frequency of the reference circular orbit:

$$\kappa^2 = 4\Omega^2\left(1 + \frac{r}{2\Omega}\frac{d\Omega}{dr}\right) \tag{32}$$

The importance of this frequency for the resonance problem is that any force that is periodic on this time scale will produce a resonant interaction. For the more general problem of galactic structure, the points at which the perturbation frequency is the same as the epicyclic frequency are called the Lindblad resonances. These play a central role in *density wave theory*, which has been applied to a wide variety of astrophysical problems, including planetary ring systems, accretion disks, and the structure and evolution of spiral disk galaxies.

Resonances of special interest in the solar system include the lock between Neptune and Pluto ($\lambda_\mathrm{P} - 2\lambda_\mathrm{N} - \omega_\mathrm{P} = 180^\circ$), the Trojan asteroids, which are trapped at the triangular Lagrangian point of the Jovian orbit, and the moons of Jupiter. These last are perhaps the most accessible examples, because they can be readily observed by anyone with the patience to watch the moons for a few days with binoculars. The best is the coupling of Europa and Ganymede, $\lambda_\mathrm{E} - 2\lambda_\mathrm{G} + \omega_\mathrm{E} = 0$, but the three moons, including Io, display $\lambda_\mathrm{I} - 3\lambda_\mathrm{E} + 2\lambda_\mathrm{G} = 180^\circ$.

D. Planetary Ring Dynamics

Planetary ring systems represent the best available celestial mechanics laboratories. Here we can observe all of the phenomena of resonance, trapping, collisions, and both periodic and nearly periodic motion. The number of particles in a typical planetary ring system like Saturn or Jupiter is great enough, and their masses low enough, that they serve as tracers of the gravitational dynamics. Thus, they constitute the best observable examples of the conditions required for the solution of many mechanical problems.

The first detailed examination of ring dynamics was accomplished by James Clerk Maxwell in his 1857 Smith Prize essay, a study that still repays reading. He dealt with the question of the composition and stability of Saturn's rings, demonstrating that they must consist of a swarm of small particles trapped in planar orbits. The demonstration of differential rotation in the Saturn rings by the observation of Doppler shifts in a reflected solar spectrum was first performed by Keeler about 30 years later.

In the best studied case, Saturn and Uranus, predictions of the placement and evolution of rings and gaps due to resonances predicted the great number of separate rings observed by satellite flybys. However, observations of Uranus during occultations of stars by the planet produced a remarkable result. The thickness of the ring system was such that several separated rings could be observed from the immersion and emersion of stars behind the planet. Five distinct rings were discerned from ground-based observations, each with remarkable optical depth but extremely geometrically thin. The theoretical development of the “shepherding moon” model has successfully accounted for nearly all of the properties of these systems.

The basic idea is that moons with near commensurabilities radially flank the orbit of a particle ring. As the particle encounters the inner one, it is sped up and its orbit starts slightly to increase. It then encounters the inner one, and it is torqued down back into a stable orbit, around which it executes a radial oscillation. In other words, the fact that there are several larger bodies which are nearly co-orbiting with the test particle serve to keep it confined within a narrow radial distance. The basic theory, first developed by P. Goldreich and S. Tremaine to explain the ε-ring of Uranus, works especially well for a number of ring systems. A notable and recently observed exception is the Neptune ring system. This planet displays what can best be described as ring-arc structures, since the particle concentrations display resonances that enhance the density of particles in confined banana-shaped regions and also form continuous rings around the planet. There is some evidence that these are also resonant, but the models have not been fully developed for this system.

Another possibility is that waves may be produced through resonant excitation of self-gravitating modes in the rings. The Saturn rings show complex, wavelike patterns that move with a fixed pattern speed around the planet. These may be density waves, collisionless modes that produce a periodic variation in the gravitational potential and thereby produce self-consistent density variations that support the waves. While there are serious problems remaining in the theory of spiral galaxies, and no unambiguous evidence has been represented for the existence of such waves in galaxies, the possible application of density wave theory to some planetary ring systems has been an encouraging indication that the basic physical mechanism is possibly realized in nature.

E. Binary Stars

The first test of Newtonian mechanics outside the solar system was the discovery of dynamically bound multiple star systems, first by Herschel and later by Bessel and his school in the nineteenth century. The observation and determination of orbits for visual binaries have been especially important for understanding the masses of many star systems. In addition, after the discovery of spectroscopic binaries, the measurement of stellar masses became routine through the observation of eclipsing binary stars. However, binaries provide many challenging problems for dynamical theory.

Close binaries interact via tides, and the stars are deformable. Therefore, they are able to show changes on relatively short time scales in the structure of the stellar envelope and to provide important clues to the origin of tidal coupling.

Most stellar masses have been calibrated through the use of double line eclipsing binary stars. Knowing the orbital parameters, such as the inclination, eccentricity, and period of an orbit, one can determine the masses of the two stars from their velocity amplitudes, the ratio of which is the inverse ratio of the stellar masses, and the period through

$$\begin{aligned} f(m) &= \frac{m_1^3 \sin^3 i}{(m_1+m_2)^2} \\ &= 1.04 \times 10^{-7}(1-e^2)^{3/2} K_1^3 P \quad M_\odot \end{aligned} \tag{33}$$

where K is the velocity amplitude of star m_1, and $M_\odot$ is the solar mass. Therefore, since $m_1/m_2 = K_2/K_1$, for eclipsing systems for which $i = 90^\circ$, the individual masses can be determined.

These measurements constitute the corner stone of the edifice of stellar evolution. The simple applicability of Kepler’s third law to the motion of bodies highlights the importance and astonishing success of classical mechanics in yielding modern and diverse results.

F. Stellar and Galactic Dynamics

Large stellar systems behave much like a collisionless gas in which the interactions are all at large distances between the particles but the forces are always attractive. Work during the past half century has made considerable strides in the elucidation of the detailed structure of stellar clusters and galaxies. The basic premise is that, much like a gas, the particles prossess a velocity distribution function $f(\mathbf{v})$ that depends only on a stable distribution in velocity. Unlike a gas, however, there is a characteristic maximum velocity for the system, the escape velocity v_∞. In the case of a cluster, two conditions govern the evolution. One is that the individual stars do not actually collide, but that they can exchange momentum at large distance via gravitational interactions. The other is that the cluster obeys the virial theorem, which states that $2T + \Phi = 0$ for a stable cluster of stars. Here T is the kinetic energy of the cluster as a whole, and Φ is the potential energy. Since $E = T + \Phi$ is the total energy, $E = \frac{1}{2}\Phi$ and $\Phi < 0$ implies that $E < 0$.

The consequences of the virial theorem are important and differentiate the behavior of a gravitating cluster from a normal gas. Because the total energy is negative, a loss of mass causes the cluster to contract. This is familiar already from the argument earlier that a loss of energy causes the orbit of a particle in a central field to contract. Thus, the orbital frequency is increased, which is equivalent to the statement that the velocity dispersion in a cluster increases. This phenomenon, known as the "negative specific heat problem" or the *gravitothermal catasrophe*, is essential for understanding the evolution of gravitationally dominated systems. If mass or energy is lost, the cluster must contract, leading to an increase in the collision frequency and enhanced loss of bodies. Eventually, the cluster becomes so low mass that the process stops, but one could think of the body as cooling so rapidly that it heats catastrophically (a seemingly paradoxical situation encountered as a result of the binding energy).

In celestial mechanics, or in this case the newer field of "gravitational statistical mechanics." the lessons learned from more traditional areas of celestial mechanics are applicable. For instance, in the case of the collision between two bodies, there is a gravitational deflection, and momentum is exchanged between the bodies. If a single star moves with respect to a background of more distant and perhaps more slowly moving stars, momentum is transferred from the moving star to the background. This results in slowing the star down, much like a viscosity. Hence, the process is called *dynamical friction*. The basic theory was developed by S. Chandrasekhar. More recent work has included the incorporation of low velocity interactions, tides, and more realistic specifications of the orbital dynamics during the encounter. Put differently, there will be a finite time in any stellar system over which the bodies making up the system will collisionally relax through distant encounters with the background stars of the system.

Single particle stellar orbital mechanics is not unlike planetary ring theory. Bodies move under the influence of a gravitational field created by the mutual interactions of all of the stellar masses in the system, a spatially extended and complicated potential.

Galactic differential rotation was discovered by J. Oort in 1927, and the introduction of rotation into the interpretation of stellar kinematics and dynamically fundamentally changed discussions of galactic structure. What Oort noted was the epicyclic frequency in the radial and transverse velocities for stars in the solar neighborhood (that is, stars within about 100 parsecs of the sun). For nearby stars, the transverse velocities can be determined through proper motion and parallax measurements. The components of the velocity parallel to the galactic plane, and normal to the direction toward the galactic center, and the radial velocities alone. Since stars have different angular momentums, their orbits will have slightly different eccentricities, even for the same total energy. These deviate from precise corotation with the sun and show a periodic dependence on twice the galactic longitude. From the amplitude of this effect, one can deduce the mass of the galaxy. The *Oort constants* are determined from the transverse and radial velocities of stars in the solar neighborhood. They are defined from the relation: $v_{\mathrm{T}} = d(A\cos 2l_{II} + B)$ and $v_{\mathrm{R}} = dA\sin 2l_{II}$, where l_{II} is the galactic longitude and the constants A and B are given by

$$A = \frac{1}{2}\left(\frac{\Theta_0}{r_0} - \frac{d\Theta}{dr}\right)_{r_0} \tag{34}$$

and

$$B = -\frac{1}{2}\left(\frac{\Theta_0}{r_0} - \frac{d\Theta}{dr}\right)_{r_0} \tag{35}$$

where Θ_0 is the orbital velocity of the sun about the galactic center. It should be noted that these coefficients are the same as those found for epicyclic motion. The gradient in the galactic gravitational potential is represented by the change in Θ with radius.

G. Chaos Theory and Celestial Mechanics

The first manifestation of chaos theory was in celestial mechanics. In his fourteenth-century arguments against astrology, Oresme pointed out that the planetary periods were incommensurate, and therefore every planetary configuration throughout all time was unique. The idea that a particle could be trapped within a broad region of phase space—where in the course of time a particle could have virtually any momentum at a given position within a bound energy—was first demonstrated by Poncaré. He showed that single particle orbits could have properties much like the ensemble of particles in a gas at very high entropy. This state is called *ergodic*. This result was extended by the ergodic theorem of Birkhoff, von Neumann, and Wiener and by the recurrence theorem of Poincaré. Chaos is the stochastic behavior of deterministic physical systems, resulting from extreme sensitivity of the dynamical evolution of such a system to small changes in the initial conditions. Strongly perturbed orbits, like those of planetary ring systems or stellar orbits in a galaxy, will display chaotic behavior. Two different orbits in a Roche potential, for instance, started out near the L_4 or L_5 point, will evolve along totally different trajectories for infinitesimal changes in the initial conditions. Perhaps stellar orbits in a strongly barred potential, such as those observed in the centers of some spiral galaxies, or the orbits of comets in the outer solar system, constitute among the best examples of such behavior. During galaxy collapse, after star formation has occurred, the rapidly changing collective gravitational field causes strong perturbations in each of the stellar orbits and thoroughly mixes them in velocity and position, or phase space, and produces a random initial velocity distribution. This process, called *violent relaxation*, is one of the most important areas of study in stellar dynamics and may have analogs in the formation of orbits in planetary ring systems and the asteroid belt.

V. CLOSING REMARKS

Celestial mechanics, begun as an applied area of physics, has broadened into one of the most fruitful and exciting fields of theoretical mathematics and physics. The introduction of new computing techniques has made it practical to calculate the dynamical evolution of quite complex systems. For the first time it is possible to study questions related to the long-term stability of the solar system, the structure of planetary rings, and the evolution of spiral and elliptical galaxies. Problems related to the stability of orbits of stars in complicated galactic potentials have revived entire areas of classical theory with dramatic results. Renewed interest in triaxial bodies has been spurred by the observations of tidal interactions between galaxies. Even the large-scale structure of the universe, a huge N-body problem involving the expansion of the universe as a whole in addition to the gravitational interactions between individual galaxies, has almost become a routine mechanical calculation. Now with the discovery of extrasolar planetary systems we can learn from new examples about the complex interactions of small bodies by natural experiments. The riches of celestial mechanics, the most "classical" area of all physical theory, are far from being mined out.

SEE ALSO THE FOLLOWING ARTICLES

Binary Stars • Chaos • Cosmology • Gravitational Wave Astronomy • Mechanics, Classical • Moon (Astronomy) • Perturbation Theory • Planetary Satellites (Natural) • Relativity, General • Solar Physics • Solar System, General • Stellar Structure and Evolution

BIBLIOGRAPHY

Arnold, V. (1983). "Mathematical Methods of Classical Mechanics, 2nd ed.," Springer-Verlag, Berlin.

The Astronomical Ephemeris (formerly *The American Ephemeris and Nautical Almanac*), published annually by the U.S. Naval Observatory and the U.K. Alamanac Office contains the most current available data on orbital dynamics of solar system bodies.

Barow-Green, J. (1996). "Poincaré and the Three Body Problem," (A. Math Soc., Providence.)

Binney, J., and Tremaine, S. (1987). "Galactie Dynamics," Princeton University Press, Princeton.

Brouwer, D., and Clemence, G. M. (1961). "Methods of Celestial Mechanics," Academic Press, New York.

Brown, E. W. (1896). "An Introductory Treatise on the Lunar Theory," Dover, New York.

Chandrasekhar, S. (1969). "Ellipsoidal Figures in Equilibrium," Dover, New York.

Danby, J. (1988). "Fundamentals of Celestial Mechanics," William-Bell, Baltimore.

Diacu, F., and Holmes, P. (1996). "Celestial Encounters: The origins of Chaos and Stability," Princeton University Press, Princeton.

Greenberg, R., and Brahic, A., eds. (1984). "Planetary Rings," University Arizona Press, Tucson.

Hagihara, Y. (1970). "Celestial Mechanics" (5 vols.), MIT Press, Cambridge, Massachusetts.

Mouton, F. (1914). "Celestial Mechanics," Dover, New York.

Peale, S. J. (1976). "Orbital resonances in the solar system," *Ann. Rev. Astron. Astrophys.* **14,** 215.

Plummer, H. C. (1918). "An Introductory Treatise on Dynamical Astronomy," Dover, New York.

Roy, A. E. (1982). "Orbital Motion," Adams Hilger, Bristol.

Saslaw, W. C. (1985). "Gravitational Physics of Stellar and Galactic Systems," Cambridge University Press, Cambridge.

Spitzer, L., Jr. (1987). "Dynamical Evolution of Globular Clusters," Princeton University Press, Princeton.

Sternberg, S. (1969). "Celestial Mechanics" (2 vols.), W. A. Benjamin, Reading, Massachusetts.

Szebehely, V. G., and Mark, H. (1998). "Adventures in Celestial Mechanics," J. Wiley, New York.

Wertz, J. R. (ed.) (1980). "Spacecraft Attitude Determination and Control," D. Reidel, Dordrecht.

Wilson, C. (1985). "The great inequality of Jupiter and Saturn: from Kepler to Laplace," *Arch. Hist. Exact Sci.* **33,** 15.

Winter, A. (1949). "Analytic Foundations of Celestial Mechanics," Princeton University Press, Princeton.

Cosmic Radiation

Peter L. Biermann
Max-Planck Institute for Radioastronomy
and University of Bonn

Eun-Suk Seo
University of Maryland

GLOSSARY

Active galactic nuclei When massive black holes accrete, their immediate environment, usually thought to consist of an accretion disk and a relativistic jet, emits a luminosity often far in excess of the emission of all stars in the host galaxy put together; this phenomenon is called an active galactic nucleus.

Antimatter All particles known to us have antiparticles, with opposite properties in all measures, such as charge.

Big Bang Our universe is continuously expanding, and its earliest stage reachable by our current physical understanding is referred to as the Big Bang.

Black holes Compressing a star to a miniscule size, in the case of our sun to a radius of 3×10^5 cm, makes it impossible for any radiation to come out; all particles and radiation hitting such an object disappear from this world. This is called a black hole.

Chemical elements In atoms the number of protons Z in the nucleus, equal to the number of electrons in the surrounding shell, determines the chemical element.

Cosmic ray airshower When a primary particle at high energy, either a photon or a nucleus, comes in to the upper atmosphere, the sequence of interactions and cascades forms an air shower.

Cosmic ray ankle At an energy of 3×10^{18} eV, or 3 EeV, there is an upturn in the spectrum, to an approximate spectral index of 2.7 again.

Cosmic ray GZK cutoff The interaction with the cosmic microwave background is predicted to produce a strong cutoff in the observed spectrum at 5×10^{19} eV called the GZK cutoff. This cutoff is not seen.

Cosmic ray knee At about 5×10^{15} eV, or 5 PeV, there is a small bend downward in the cosmic ray spectrum by about 0.4 in spectral index, from 2.7 to 3.1.

Cosmic ray spectrum The number of particles at a certain energy E within a certain small energy interval dE is called the spectrum. Flux is usually expresssed as the number of particles coming in per area, per second, per steradian in solid angle (all-sky is 4π), and per energy interval.

Elementary particles The natural constituents of normal matter are the proton, neutron, and electron.

Gamma ray bursts Bursts of gamma ray emission coming from the far reaches of the universe, and almost certainly the result of the creation of a stellar mass black hole.

Interstellar matter The medium between the stars in our Galaxy, which is composed of very hot gas (order 4×10^6 K), various stages of cooler gas, down to about 20 K, dust, cosmic rays, and magnetic fields.

Magnetic monopoles The physics of electric and magnetic fields contains electric charges but no magnetic charges. In the context of particle physics it is likely that monopoles, basic magnetically charged particles, also exist.

Microwave background The very high temperature of the Big Bang is still visible in the microwave background, a universal radiation field of 2.73 K temperature.

Our Galaxy Our Galaxy is a flat, circular, disklike distribution of stars and gas enmeshed with interstellar dust and embedded in a spheroidal distribution of old stars.

Nuclear collisions When elementary particles collide with each other or with a photon, other particles can be created that are often unstable.

Radio galaxies In the case that an active galactic nucleus produces a very powerful and visible jet, often with lobes and hot spots, most readily observable at radio wavelengths, it is called a radio galaxy.

Spallation The destruction of atomic nuclei in a collision with another energetic particle such as another nucleus or, commonly, a proton.

Supernovae All stars above an original mass of more than 8 solar masses are expected to explode at the end of their lifetime after they have exhausted nuclear burning; the observable effect of such an explosion is called a supernova.

Topological defects It is generally believed that in the very early times of the universe, when the typical energies were far in excess of what we can produce in accelerators, there was a phase when the typical energies of what might be called particles was around 10^{24} eV, usually referred to as topological defects, or relics.

Units: Energy Electron volts (eV); 1 eV is of the order of what is found in chemical reactions; $1 \text{ eV} = 1.6 \times 10^{-12}$ erg.

Units: length Centimeters; the radius of the earth is 6.4×10^8 cm and the radius of the Sun is 7×10^{10} cm.

Units: mass Grams; the sun has a mass of 2×10^{33} g; the earth has a mass of 6×10^{27} g.

Units: time Seconds; the travel time of light from the sun to the earth is about 8 min = 480 sec; the number of seconds in a year is 3.15×10^7.

ENERGETIC PARTICLES, traditionally called *cosmic rays*, were discovered nearly a 100 years ago and their origin is still uncertain. Their main constituents are the normal nuclei as in the standard cosmic abundances of matter, with some enhancements for the heavier elements; there are also electrons. Information on isotopic abundances shows some anomalies as compared with the interstellar medium. There is also antimatter, such as positrons and antiprotons. The known spectrum extends over energies from a few hundred MeV to 300 EeV ($=3 \times 10^{20}$ eV), and shows a few clear spectral signatures: There is a small spectral break near 5 PeV ($=5 \times 10^{15}$ eV), commonly referred to as the *knee*, where the spectrum turns down; and there is another spectral break near 3 EeV ($=3 \times 10^{18}$eV), usually called the *ankle*, where the spectrum turns up again. Due to interaction with the microwave background arising from the Big Bang, there is a strong cutoff expected near 50 EeV ($=5 \times 10^{19}$ eV), which is, however, not seen; this expected cutoff is called the GZK cutoff after its discoverers, Greisen, Zatsepin, and Kuzmin. The spectral index α is near 2.7 below the knee, near 3.1 above the knee, and again near 2.7 above the ankle, where this refers to a differential spectrum of the form $E^{-\alpha}$. We will describe the various approaches to understanding the origin and physics of cosmic rays.

I. INTRODUCTION AND HISTORY

Cosmic rays were discovered by Hess and Kohlhörster in the beginning of the 20th century through their ionizing effect on airtight vessels of glass enclosing two electrodes with a high voltage between them. This ionizing effect increased with altitude during balloon flights, and therefore the effect must come from outside the earth, so the term *cosmic rays* was coined. The earth's magnetic field acts on energetic particles according to their charge, and hence they are differently affected coming from east and west, and so their charge was detected, proving that they are charged particles; at high energies near 10^{18} eV or 1 EeV, there is observational evidence that a small fraction of the particles are neutral and in fact neutrons. From around 1960 onward there has been evidence of particles at or above 10^{20} eV, with today about two dozen such events known. After the cosmic microwave background was discovered in the early 1960s, it was noted only a little later by Greisen, Zatsepin, and Kuzmin that near and above an energy of 5×10^{19} eV (called the GZK cutoff) the interaction with the microwave background would lead to strong losses if these particles were protons, as is now believed on the basis of detailed air shower data. In such an interaction, protons see the photon as having an energy of above the pion mass, and so pions can be produced in the reference frame of the collision, leading to about a 20% energy loss of the proton in the observer frame. Therefore for an assumed cosmologically homogeneous distribution of sources for protons at extreme energies, a spectrum at earth is predicted which shows a strong cutoff at 5×10^{19} eV, the GZK cutoff. This cutoff is not seen, leading to many speculations as to the nature of these particles and their origin.

Cosmic rays are measured with instruments on balloon flights, satellites, the Space Shuttle, the International Space Station, and with ground arrays. The instrument chosen depends strongly on what is being looked for and the energy of the primary particle. One of the most successful campaigns has been with balloon flights in Antarctica, where a balloon can float at about 40 km altitude and circumnavigate the South Pole once and possibly even several times during one Antarctic summer. For very high precision measurements very large instruments on the Space Shuttle or the International Space Station are used, such as for the search for antimatter.

Critical measurements are the exact spectra of the most common elements, hydrogen and helium, the fraction of antiparticles (antiprotons and positrons), isotopic ratios of elements such as neon and iron, the ratio of spallation products such as boron to primary nuclei such as carbon as a function of energy, the chemical composition near the knee, at about 5×10^{15} eV, and beyond, and the spectrum and nature of the particles beyond the ankle, at 3×10^{18} eV, with special emphasis on the particles beyond the GZK cutoff, at 5×10^{19} eV.

II. PHYSICAL CONCEPTS

Here we expand upon the terms explained briefly in the Glossary.

- *Big Bang*. Our universe is continuously expanding, and its earliest stage reachable by our current physical understanding is referred to as the Big Bang, when energy densities were extremely high. Within the first 3 min the chemical elements such as hydrogen and helium were produced, and minute amounts of deuterium (an isotope of hydrogen), ^{3}He, an isotope of helium, and ^{7}Li, an isotope of the third element, lithium.
- *Microwave background*. The very high temperature of the Big Bang is still visible in the microwave background, a universal radiation field of 2.73 K. There is a corresponding cosmic bath of low-energy neutrinos. Both photons and neutrinos have a density of a few hundred per cubic centimeter.
- *Units*: *length* Centimeters; the radius of the earth is 6.4×10^8 cm and the radius of the sun is 7×10^{10} cm. The distance from earth to sun is 1.5×10^{13} cm. One pc = parsec = 3.086×10^{18} cm, 1 kpc = 10^3 pc, 1 Mpc = 10^6 pc; 1 pc is about 3 light-years, the distance traveled by light in 3 years; the speed of light c is 3×10^{10} cm/sec, and is the same in any inertial reference frame. The basic length scale of the universe is about 4000 Mpc.
- *Units*: *mass*. Grams; the sun has a mass of 2×10^{33} g. The earth has a mass of 6×10^{27} g. A typical galaxy like our own has a mass of order 10^{11} solar masses. A proton has a mass of 1.67×10^{27} g.
- *Units*: *time* Seconds; the travel time of light from the sun to the earth is about 8 min = 480 sec; the number of seconds in a year is 3.15×10^7. The age of the solar system is about 4.5×10^9 years, and the age of our Galaxy and also of our universe is about 1.5×10^{10} years; our galaxy is younger than the universe, but we do not know the two ages well enough to determine the difference with any reliability.
- *Units*: *energy*. Electron volts (eV); 1 eV is of the order of what is found in chemical reactions, 1 eV = 1.6×10^{-12} erg; 1 MeV = 10^6 eV, 1 GeV = 10^9 eV, 1 TeV = 10^{12} eV, 1 PeV = 10^{15} eV, 1 EeV = 10^{15} eV.
- *Elementary particles*. The natural constituents of matter are the proton, neutron, and electron. Protons have a mass of about 938 MeV, neutrons of about 940 MeV, and electrons of about 0.511 MeV. This is in energy units using Einstein's equivalence $E = mc^2$, where E is the energy, m the rest mass, and c the speed of light. The proton is positively charged, the electron negatively charged, with the same charge as the proton, and the neutron is neutral. All atomic nuclei are built from protons and neutrons, where the number of protons determines the chemical element, and the number of neutrons determines the various isotopes of each chemical element. The surrounding shell of electrons has for the neutral atom exactly the same number of electrons as the nucleus has protons. Photons are another primary stable constituent, have no rest mass, no charge, and always travel at the speed of light, in any frame of reference. Neutrinos come in three varieties, and appear to continuously change among themselves; they have a very low mass.
- *Antimatter*. All known particles have antiparticles, with opposite properties in all measures, such as charge. The collision of a particle and its antiparticle always leads to a burst of radiation, when both particles are annihilated. In cosmic rays we observe antiprotons, and positrons, the antiparticles to electrons. The search for antinuclei has not been successful; any detection of even a single antinucleus, such as antihelium, would provide extremely strong constraints on the physics of matter in the universe.

The instrument AMS, first on the Space Shuttle, and then the International Space Station, will search for antimatter particles.

- *Nuclear collisions*. When elementary particles collide with each other or with a photon, other particles can be created that are often unstable, i.e., they decay into other particles and continue to do so until they reach a state where only stable particles result. Such a process is called a cascade. Common intermediate and final particles are the pion, the muon, the photon, and the neutrino. The pion comes in various forms, charged and uncharged, the muon is always charged, and the neutrino is always neutral. The neutrino is characterized by a very small interaction cross section with matter. The pions have a mass, again in energy equivalents, of about 140 MeV, the muons of about 106 MeV, and the neutrinos of about 0.03 eV, a still rather uncertain number.
- *Chemical elements*. In atoms the number of protons Z in the nucleus, equal to the number of electrons in the surrounding shell, determines the chemical element. The number of neutrons in the nucleus is approximately equal to the number of protons. Chemical elements are now known to beyond Z of 110. The first and most common elements are hydrogen ($Z = 1$), helium (2), carbon (6), oxygen (8), neon (10), magnesium (12), silicon (14), sulfur (16), calcium (20), and iron (26). The intermediate elements lithium ($Z = 3$), beryllium (4), and boron (5) are very rare. The odd-Z elements are commonly rarer than the even-Z elements. The overall abundance by mass is about 73% for hydrogen, 25% for helium, and the rest all other elements combined, with the most abundant among these being carbon and oxygen.
- *Cosmic ray airshower*. When a primary particle at high energy, either a photon or a nucleus, comes into the upper atmosphere, the sequence of interactions and cascades forms an airshower. These airshowers are dominated by electrons and photons generated in electromagnetic subshowers. Cerenkov light, a bluish light, is emitted in a narrow cone around the shower direction, produced when particles travel at a speed (always less than or equal to c; particles have exactly the speed of light at zero rest mass, such as photons) higher than the speed of light c divided by the local index of refraction (which is $4/3$ in water, for instance, and about 1.0003 in air). Observing this bluish light allows observations of high-GeV-to-TeV photon sources in the sky because Cerenkov light allows good pointing. For particles such as protons or atomic nuclei at high energy, air fluorescence can be used as a means of observation, advantageous because it is omnidirectional in its emission and gives enough light at high energy: Such fluorescence occurs when normal emission lines of air molecules are excited. The airshower includes a pancake of secondary electrons and positrons as well as muons. The ratio between Cerenkov light and fluorescence light is almost constant and independent of particle energy. Most modern observations of very high energy cosmic rays are done either by observing the air fluorescence (arrays such as Fly's Eye, HIRES, or AUGER) or by observing the secondary electrons and positrons (in arrays such as Haverah Park, AGASA, Yakutsk, or also AUGER). In the future such observations may be possible from space by observing the air fluorescence or the reflected Cerenkov light from either the International Space Station or from dedicated satellites. Fly's Eye was and HIRES is in Utah, AUGER is in Argentina, AGASA is in Japan, Yakutsk is in Russia, and Haverah Park was in the United Kingdom.
- *Spallation*. Spallation is the destruction of atomic nuclei in a collision with another energetic particle, such as another nucleus, or commonly a proton. In this destruction many pieces of debris can be formed, with one common result being the stripping of just one proton or neutron, and another common result a distribution of smaller unit nuclei. Since the proton number determines the chemical element, such debris is usually other nuclei, such as boron.
- *Cosmic ray spectrum*. The number of particles at a certain energy E within a certain small energy interval dE is called the spectrum. As a function of energy E this is usually described by power laws, such as $E^{-2.7}\,dE$; the exponent is called the spectral index, here 2.7. Heat radiation from a normal object has a very curved spectrum in photons. Cosmic rays usually have a power-law spectrum, which is called a nonthermal behavior. Flux is usually expresssed as the number of particles coming in per area, per second, per solid angle in steradians (all-sky is 4π), and per energy interval.
- *Cosmic ray knee*. At about 5×10^{15} eV, or 5 PeV, there is a small bend downward in the cosmic ray spectrum, by about 0.4 in spectral index, from 2.7 to 3.1. This feature is called the *knee*. There is some evidence that it occurs at a constant energy-to-charge ratio for different nuclei. There is also considerable evidence that toward and at the knee the chemical composition slowly increases in favor of heavier nuclei such as iron. A similar somewhat weaker feature is suggested by new AGASA and HIRES data near 3×10^{17} eV, where again the spectrum turns down a bit more.
- *Cosmic ray ankle*. At an energy of 3×10^{18} eV, or 3 EeV, there is an upturn in the spectrum, to an approximate spectral index of 2.7 again. At the same energy there is evidence that the chemical composition changes from moderately heavy to light, i.e., back to mostly hydrogen and helium.
- *Cosmic ray GZK cutoff*. The interaction with the cosmic microwave background should produce a strong cutoff

in the observed spectrum at 5×10^{19} eV called the GZK cutoff; this is provided that (a) these particles are protons (or neutrons) and (b) the source distribution is homogeneous in the universe. This cutoff is not seen; in fact, no cutoff is seen at any energy, up to the limit of data, at 3×10^{20} eV, or 300 EeV. This is one of the most serious problems facing cosmic ray physics today.

- *Black holes.* Compressing a star to a miniscule size, in the case of our sun, to a radius of 3×10^5 cm, makes it impossible for any radiation to come out; all particles and radiation hitting such an object disappear from this world. This is called a black hole. It is now believed that almost all galaxies have a massive black hole at their center, with masses sometimes ranging up 10^{10} solar masses, but usually much less. There are also stellar mass black holes, but their number is not well known, probably many thousands in each galaxy.

- *Our Galaxy*. Our Galaxy is a flat, circular, disklike distribution of stars and gas enmeshed with interstellar dust and embedded in a spheroidal distribution of old stars. The age of this system is about 15 billion ($=15 \times 10^9$) years; its size is about 30 kpc across, and its inner region is about 6 kpc across. At its very center there is a black hole with 2.6×10^6 solar masses. The gravitational field is dominated in the outer parts of the Galaxy by an unknown component, called dark matter, which we deduce only through its gravitational force. In the innermost part of the Galaxy normal matter dominates. The mass ratio of dark matter to stars to interstellar matter in our Galaxy is about 100:10:1. Averaged over the nearby universe these ratios are shifted in favor of gas, with gas dominating over stars probably, but with dark matter still dominating over stars and gas by a large factor.

- *Interstellar matter.* The medium between the stars in our Galaxy is composed of very hot gas (order 4×10^6 K), various stages of cooler gas, down to about 20 K, dust, cosmic rays, and magnetic fields. All three components, gas, cosmic rays, and magnetic fields, have approximately the same energy density, which happens to be also close to the energy density of the microwave background, about 1 eV/cm^3. The average density of the neutral hydrogen gas, of temperature a few 10^{10} K to a few 10^3 K, is about 1 partcile/cm^3, but highly clumped, in a disk of thickness about 100 pc ($=3 \times 10^{20}$ cm). The very hot gas extends much farther from the symmetry plane, about 2 kpc on either side.

- *Supernovae*. All stars above an original mass of more than 8 solar masses are expected to explode at the end of their lifetime after they have exhausted nuclear burning; the observable effect of such an explosion is called a supernova. When they explode, they emit about 3×10^{53} ergs in neutrinos and also about 10^{51} erg in visible energy, such as in shock waves in ordinary matter, the former stellar envelope, and interstellar gas. These neutrinos have an energy in the range of a few MeV to about 20 MeV. When stars are in stellar binary systems, they can also explode at low mass, but this process is believed to give only 10% or less of all stellar explosions. The connection to gamma ray bursts is not clear. It is noteworthy that above an original stellar mass of about 15 solar masses, stars also have a strong stellar wind, which for original masses above 25 solar masses becomes so strong that it can blow out a large fraction of the original stellar mass even before the star explodes as a supernova. The energy in this wind, integrated over the lifetime of the star, can attain the visible energy of the subsequent supernova, as seen in the shock wave of the explosion.

- *Gamma ray bursts*. Bursts of gamma ray emission come from the far reaches of the universe, almost certainly the result of the creation of a stellar-mass black hole. The duration of these bursts ranges from a fraction of a second to usually a few seconds, and sometimes hundreds of seconds. Some gamma ray bursts have afterglows in other wavelengths like radio, optical, and X rays, with an optical brightness which very rarely comes close to being detectable with standard binoculars. The emission peaks near 100 keV in observable photon energy, and appears to have an underlying power-law character, suggesting nonthermal emission processes.

- *Active galactic nuclei*. When massive black holes accrete, their immediate environment, usually thought to consist of an accretion disk and a relativistic jet (i.e., where the material flies with a speed very close to the speed of light), emits a luminosity often far in excess of the emission of all stars in the host galaxy put together. There are such black holes of a mass near 10^8 solar masses, with a size of order the diameter of the earth orbit around the sun and a total emission of 1000 times that of all stars in their host galaxy.

- *Radio galaxies*. In the case that an active galactic nucleus produces a very powerful and visible jet, often together with lobes and hot spots, most readily observable at radio wavelengths, it is called a radio galaxy. The radio image of such a galaxy can extend to 300 kpc or more, dissipating the jet in radio hot spots embedded in giant radio lobes. The frequency of such radio galaxies with powerful jets, hot spots, and lobes is rare, less than $1/1000$ of all galaxies, but in the radio sky they dominate due to their extreme emission.

- *Topological defects*. It is generally believed that in the very early times of the universe, when the typical energies were far in excess of what we can produce in accelerators, there was a phase when the typical energies of what might be called particles was around 10^{24} eV, usually referred to as topological defects, or relics. It is conveivable that such particles have survived to today, and some of them decay, emitting a copious number of neutrinos, photons, and also protons. These particles then themselves would have near 10^{24} eV initially, but interact strongly with the cosmic radiation field.

- *Magnetic monopoles.* The physics of electric and magnetic fields contains electric charges but no magnetic charges. In the context of particle physics it is likely that monopoles, basic magnetically charged particles, also exist. Such monopoles are a special kind of topological defect. The basic property of magnetic monopoles can be described as follows: (a) Just as electrically charged particles short-circuit electric fields, monopoles short-circuit magnetic fields. The observation of very large scale and permeating magnetic fields in the cosmos shows that the universal flux of monopoles is very low. (b) Monopoles are accelerated in magnetic fields, just as electrically charged particles are accelerated in electric fields. In cosmic magnetic fields, the energies which can be attained are of order 10^{21} eV or more. Any relation to the observed high-energy cosmic rays is uncertain at present.

III. ENERGIES, SPECTRA, AND COMPOSITION

The solar wind prevents low-energy charged articles from entering the inner solar system due to interaction with the magnetic field in the solar wind, a steady stream of gas going out from the sun into all directions, originally discovered in 1950 from the effect on cometary tails: they all point outward, at all latitudes of the sun, and independent of whether the comet actually comes into the inner solar system or goes outward, in which case the tail actually precedes the head of the comet. This prevents us from knowing anything about interstellar energetic particles with energies lower than about 300 MeV. Above about 10 GeV per charge unit Z of the particle, the effect of the solar wind becomes negligible. Since cosmic ray particles are mostly fully ionized nuclei (with the exception of electrons and positrons), this is a strong effect.

Our Galaxy has a magnetic field of about 6×10^{-6} G in the solar neighborhood; the energy density of such a field corresponds approximately to 1 eV/cm^3, just like the other components of the interstellar medium. In such a magnetic field charged energetic particles gyrate with a radius of gyration called the Larmor radius, which is proportional to the momentum of the particle perpendicular to the magnetic field direction. For highly relativistic particles this entails that around 3×10^{18}-eV protons, or other nuclei of the same energy-to-charge ratio, no longer gyrate in the disk of the Galaxy, i.e., their radius of gyration is larger than the thickness of the disk. Thus, they cannot possibly originate in the Galaxy, and must come from outside; indeed, at that energy there is evidence for a change both in chemical composition and in the slope of the spectrum.

The energies of these cosmic ray particles that we observe range from a few hundred MeV to 300 EeV. The integral flux ranges from about 10^{-5} per cm^2, per sec, per sterad at 1 TeV per nucleus for hydrogen or protons, to 1 particle per sterad per km^2 and per century around 10^{20} eV, a decrease by a factor of 3×10^{15} in integral flux, and a corresponding decrease by a factor of 3×10^{23} in the spectrum, i.e., per energy interval, which means in differential flux. Electrons have only been measured to a few TeV.

The total particle spectrum spectrum is about $E^{-2.7}$ below the knee and about $E^{-3.1}$ above the knee, at 5 PeV, and flattens again to about $E^{-2.7}$ beyond the ankle, at about 3 EeV. Electrons have a spectrum which is similar to that of protons below about 10 GeV, and steeper near $E^{-3.3}$ above this energy. The lower energy spectrum of electrons is inferred from radio emission, while the steeper spectrum at the higher energies is measured directly (Fig. 1).

The chemical composition is rather close to that of the interstellar medium, with a few strong peculiarities relative to that of the interstellar medium: (a) hydrogen and helium are less common relative to silicon. Also, the ratio of hydrogen to helium is smaller. (b) lithium, beryllium, and boron, the odd-Z elements, as well as the sub-iron elements (i.e., those with Z somewhat less than iron) are all enhanced relative to the interstellar medium (Fig. 3). (c) Many isotope ratios are quite different, while some are identical. (d) Among the cosmic ray particles there are radioactive isotopes, which give an age of the particles since acceleration and injection of about 3×10^7 years. (e) Toward the knee and beyond the fraction of heavy elements appears to continuously increase, with moderately heavy to heavy elements almost certainly dominating beyond the knee, all the way to the ankle, where the composition becomes light again. This means that at that energy we observe a transition to what appears to be mostly hydrogen and helium nuclei. At much higher energies we can only show consistency with a continuation of these properties; we cannot prove unambiguously what the nature of these particles is.

The fraction of antiparticles, ie., positrons and antiprotons, is a few percent for the positron fraction and a few 10^4 for antiprotons. No other antinuclei have been found (Figs. 3–5).

There is no significant anisotropy of cosmic rays at any energy, not even at the highest energies, beyond the GZK cutoff. Only at those highest energies is there a persistent hint that events at quite different energies occasionally cluster into pairs and triplets in the sky in the arrival direction, and this is hard to understand in almost any model.

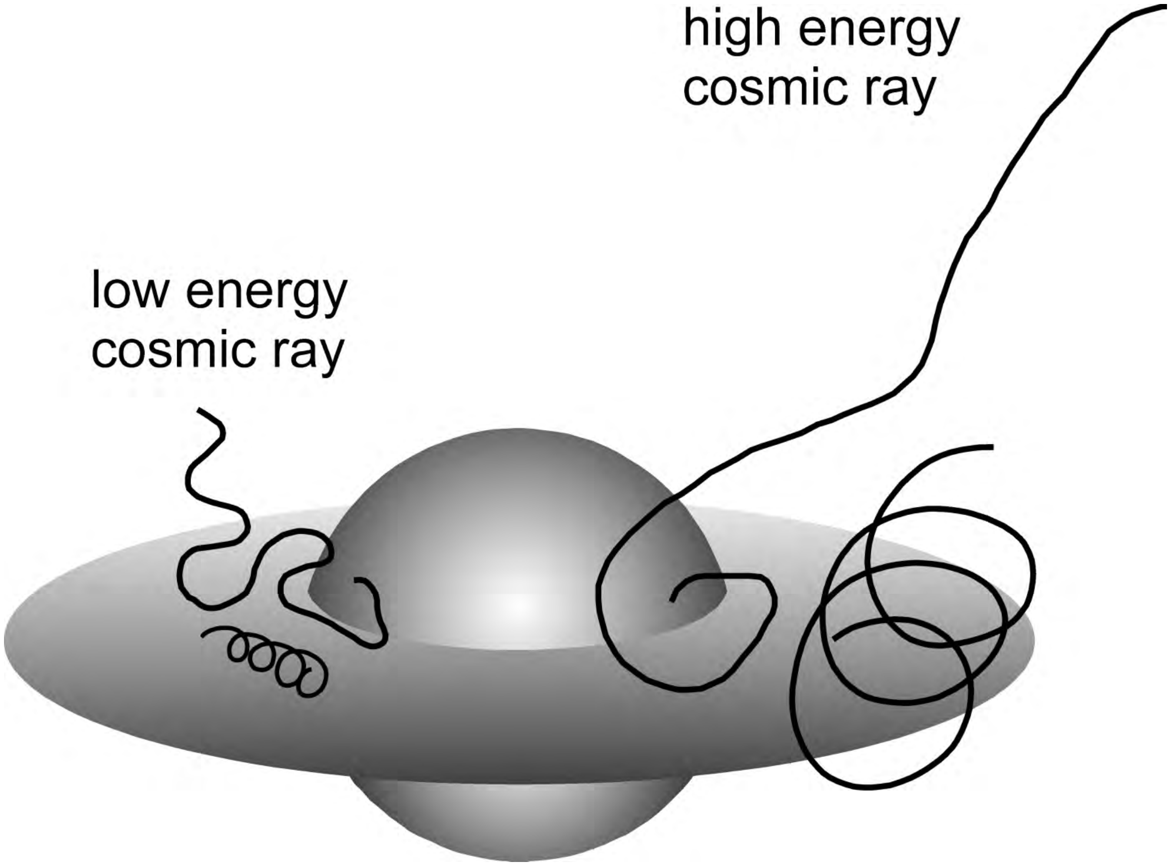


FIGURE 1 The Galaxy with some sample orbits of low energy and high energy cosmic ray particles. The Galaxy is shown with the disk and the spherodial component of older stars.

IV. ORIGIN OF GALACTIC COSMIC RAYS

It has been long surmised that supernova explosions provide the bulk of the acceleration of cosmic rays in the Galaxy. The acceleration is thought to be a kind of ping-pong effect between the two sides of the strong shock wave sent out by the explosion of a star. This ping-pong effect is a repeated reflection via the magnetic resonant interaction between the gyromotion of the energetic charged particles and waves of the same wavelength as the Larmor motion in the magnetic thermal gas. Since the reflection is usually thought to be a gradual diffusion in direction, the process is called diffusive shock acceleration, or, after its discoverer, Fermi acceleration.

For a shock wave sent out directly into the interstellar gas this kind of acceleration easily provides particle energies up to about 100 TeV. While the detailed injection mechanism is not quite clear, the very fact that we observe the emission of particles at these energies in X rays provides a good case and a rather direct argument for highly energetic electrons. Even though protons are by a factor of about 100 more abundant than electrons at energies near 1 GeV, we cannot yet prove directly that supernova shocks provide the acceleration; only the analogy with electrons can be demonstrated.

However, we observe what ought to be galactic cosmic rays up to energies near the knee, and beyond to the ankle, i.e., 3 EeV.

The energies, especially for particles beyond 100 TeV, can be provided by several possibilities, with the only theory worked out to a quantitative level suggesting that those particles also get accelerated in supernova shock waves, in those which run through the powerful stellar wind of the predecessor star. Then it can be shown that energies up to 3 EeV per particle are possible (mostly iron). An alternate possibility is that a ping-pong effect between various supernova shock waves occurs, but in this case seen from outside. In either (or any other) such theory it is a problem that we observe a knee, i.e., a downward bend of the spectrum at an energy-to-charge ratio which appears to be fairly sharply defined. The concept that stellar explosions are at the origin entails that all such stars are closely similar in their properties, including their magnetic field, at the time of explosion; this implies a specific

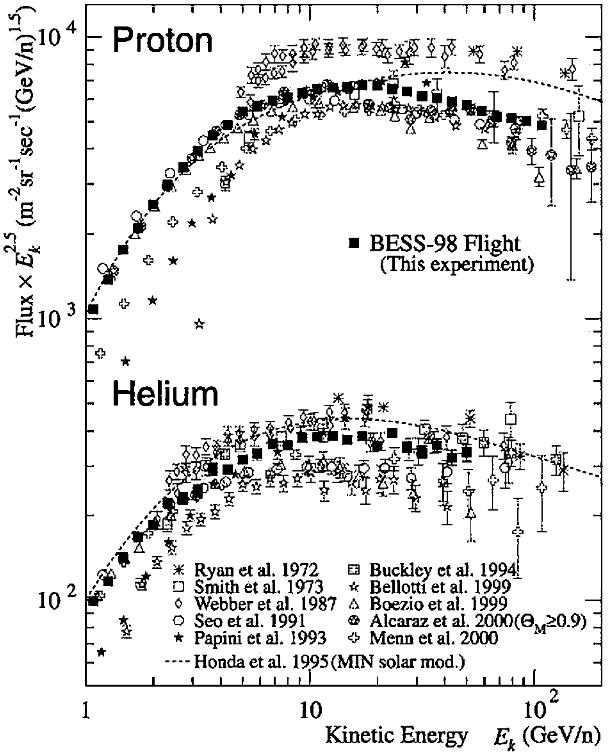


FIGURE 2 The proton and helium spectra at low energy. [From Suzuki, T., *et al.* (2000). "Precise measurements of cosmic-ray proton and helium spectra with BESS spectrometer," *Astrophys. J.* **545,** 1135–1142; The AMS Collaboration. (2000). "Protons in near earth orbit," *Phys. Lett. B* **472,** 215–226; The AMS Collaboration. (2000). "Helium in near earth orbit," *Phys. Lett. B* **494,** 193–202.]

length scale in the explosion, connected to the thickness of the matter of the wind snowplowed together by the supernova shock wave. While this is certainly possible, we have too little information on the magnetic field of pre-supernova stars to verify or falsify this. In the case of the other concept it means that the transport through the interstellar gas has change in properties also at a fairly sharply defined energy-to-charge ratio, indicating a special scale in the interstellar gas, for which there is no other evidence.

Galactic cosmic rays get injected from their sources with a certain spectrum. While they travel through the Galaxy from the site of injection to escape or to the observer, they have a certain chance to leak out from the hot magnetic disk of several kpc full-width thickness of the Galaxy; occasionally this thick disk is referred to as the halo. The escape of cosmic rays becomes easier with higher energy. As a consequence their spectrum steepens, as shown by comparing source and observed spectrum. The radio observations of other galaxies show consistency with the understanding that the average spectrum of cosmic rays at least in the GeV to many GeV energy range is always the same in various locations in a galaxy and also in different galaxies. During this travel inside a galaxy the cosmic rays interact with the interstellar gas, and in this interaction produce gamma ray emission from pion decay, positrons, and also neutrons, antiprotons, and neutrinos. The future gamma ray emission observations of the galactic disk will certainly provide very strong constraints on this aspect of cosmic rays. In these interactions secondary nuclei and isotopes are also produced by spallation. This spallation also gives secondary isotopes such as ^{10}Be, which is radioactive and decays with a half-life of 1.6×10^6 years, allowing models of cosmic ray propagation to be tested.

One kind of evidence about where cosmic rays come from and what kind of stars and stellar explosions really dominate among their sources is the isotopic ratios of various isotopes of neon, iron, and other heavy elements; these isotope ratios suggest that at least one population is indeed the very massive stars with strong stellar winds; however, whether these stars provide most of the heavier elements, as one theory proposes, is still quite an open question.

Antimatter as observed today can all be produced in normal cosmic ray interactions. However, even the detection of a single antinucleus of an element such as helium would constitute proof that the universe contains antimatter regions and would radically change our perception of the matter–antimatter symmetries in our world.

There is some evidence that just near EeV energies there is one component of galactic cosmic rays which is spatially associated in arrival direction with the two regions of highest activity in our Galaxy, at least as seen from earth: the Galactic Center region as well as the Cygnus region show some weak enhancement. Such a directional association is only possible for neutral particles, and since neutrons at that energy can just about travel from those regions to here before they decay (only free neutrons decay, neutrons bound into a nucleus do not decay), a production of neutrons in cosmic ray interactions is conceivable as one explanation of these data.

V. THE COSMIC RAYS BETWEEN 3 AND 50 EeV, THE EXPECTED GZK CUTOFF

The cosmic rays between the ankle and the expected GZK cutoff are readily explained by many possible sources, almost all outside our Galaxy.

Some, but not all of these proposals can also explain particles beyond the GZK cutoff, discussed in a separate section.

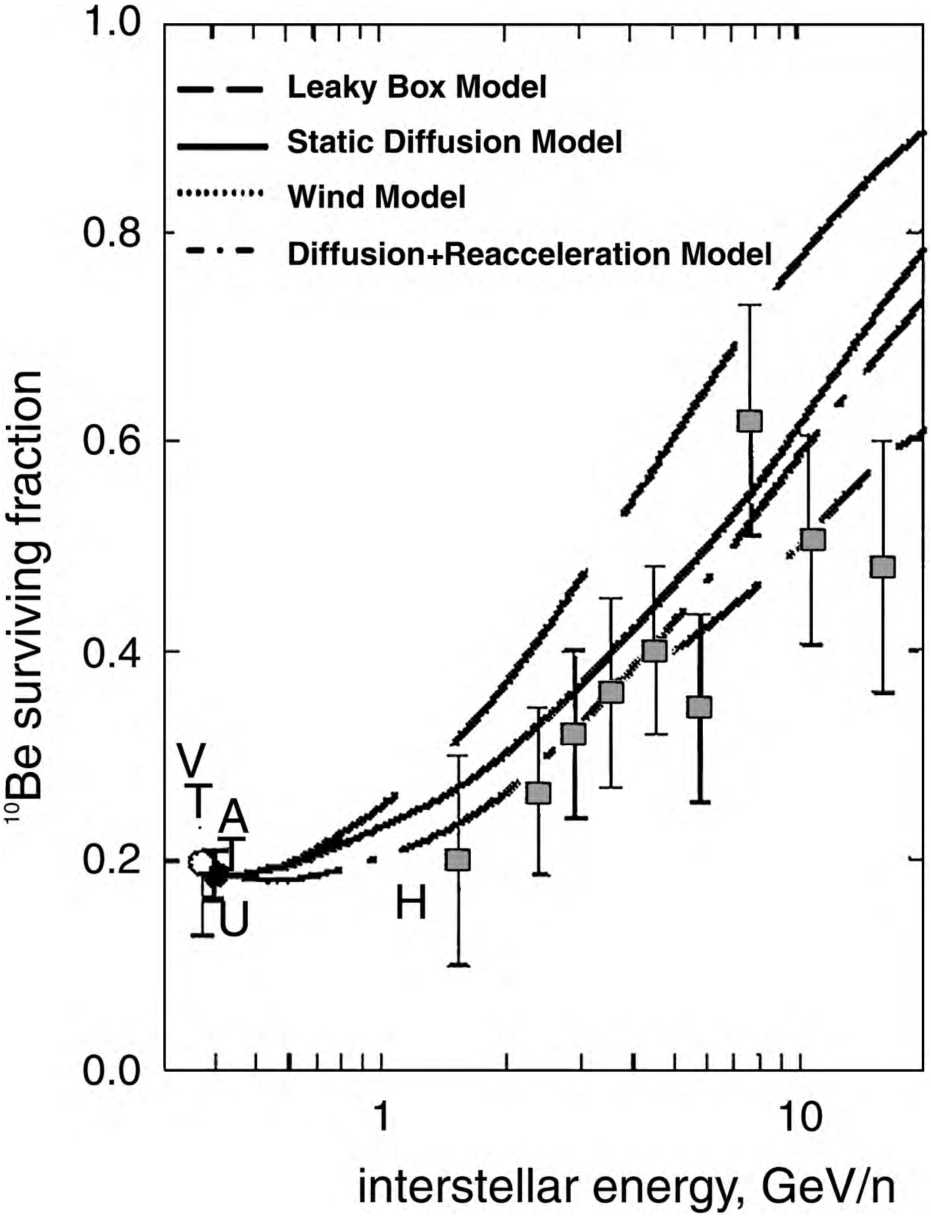


FIGURE 3 The surviving fraction of ^{10}Be in data and a comparison with models of cosmic ray propagation. [From Ptuskin, V. S. (2000). "Cosmic ray transport in the galaxy," In "ACE-2000 Symposium" (R. A. Mewaldt *et al*., eds.), pp. 390–395, AIP, Melville, NY.]

Pulsars, especially those with very high magnetic fields, called magnetars, can almost certainly accelerate charged particles to energies of 10^{21} eV. There are several problems with such a notion, one being the adiabatic losses on the way from close to the pulsar out to the interstellar gas, and another one the sky distribution, which should be very anisotropic given the distribution and strength of galactic magnetic fields. On the other hand, if this concept could be proven, it would certainly provide a very easy explanation for why there are particles beyond the GZK cutoff: for galactic particles the interaction with the microwave background is totally irrelevant, and so no GZK cutoff is expected.

Another proposal is gamma ray bursts. However, since we know that the bulk of gamma ray bursts arise from stellar explosions at cosmological distances this notion is hard to maintain. The frequency of nearby gamma ray bursts is too small to provide any appreciable flux. However, since

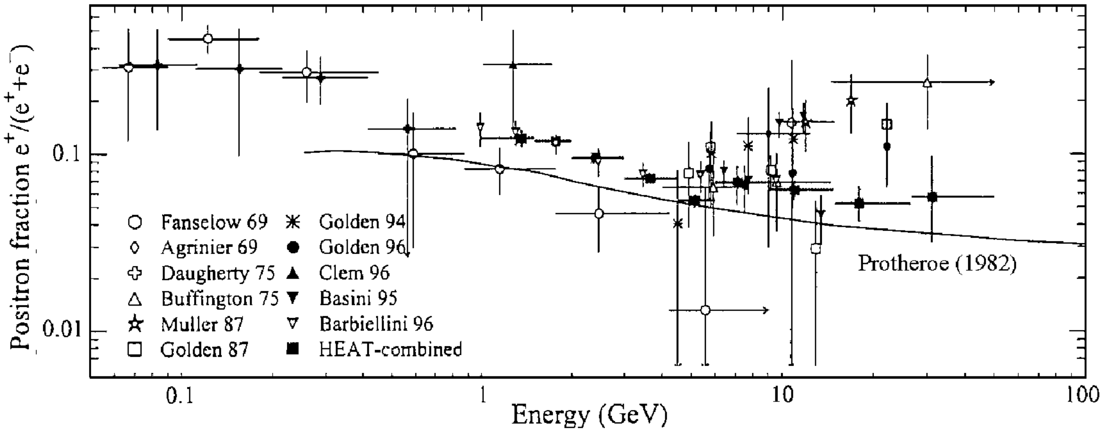


FIGURE 4 The positron fraction compared with various models of cosmic ray interaction. [From Coutu, St., *et al.* (1999). "Cosmic-ray positrons: Are there primary sources?" *Astrophys. J.* **11,** 429–435; The AMS Collaboration. (2000). "Leptons in near earth orbit," *Phys. Lett. B* **484,** 1022; Wiebel-Sooth, B., and Biermann, P. L. (1999). "Cosmic rays," In "Landolt-Börnstein," Vol. VI/3c, pp. 37–90, Springer-Verlag, Berlin.

ultimately we do not yet know what constitutes a gamma ray burst, their contribution cannot be settled with full certainty.

Shock waves running through a magnetic and ionized gas accelerate charged particles, as we know from *in situ* observations in the solar wind, and this forms the basis of almost all theories to account for galactic cosmic rays. The largest shock waves in the universe have scales of many tens of Mpc, and have shock velocities of around 1000 km/sec. These shock waves arise in cosmological large-scale structure formation, seen as a soap-bubble-like distribution of galaxies in the universe. The accretion flow to enhance the matter density in the resulting sheets and filaments is continuing, and causes shock waves to exist all around us. In such shock waves, which also have been shown to form around growing clusters of galaxies, particles can be accelerated and can attain fairly high energies. However, the maximum energies can barely reach the energy of the GZK cutoff.

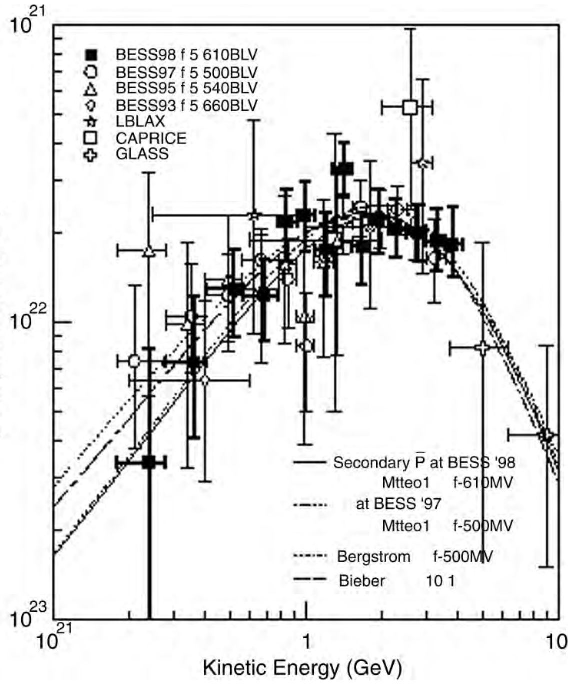


FIGURE 5 The antiproton spectrum. [From Yoshimura, K., *et al.* (2000). "Cosmic-ray antiproton and antinuclei," *Advan. Space Res.*, in press.]

The most conventional and easiest explanation is radio galaxies, of which provide the hot spots an obvious acceleration site: These hot spots are giant shock waves, often of a size exceeding that of our entire Galaxy. The shock speeds may approach several percent, maybe even several tens of percent, of the speed of light. Therefore, in this interpretation, these radio galaxy hot spots provide a very straightforward acceleration site. Integrating over all known radio galaxies readily explains the flux and spectrum as well as chemical composition of the cosmic rays in this energy range, and quite possibly also beyond.

VI. PARTICLES BEYOND THE GZK CUTOFF

Since we do not observe the expected GZK cutoff, we need to look for particles which defy the interaction with the microwave background or for a source distribution which reduces the time for interaction with the microwave background substantially.

Here we emphasize those concepts which can explain the events beyond the GZK cutoff; these proposals do not necessarily also explain those particles below the expected cutoff.

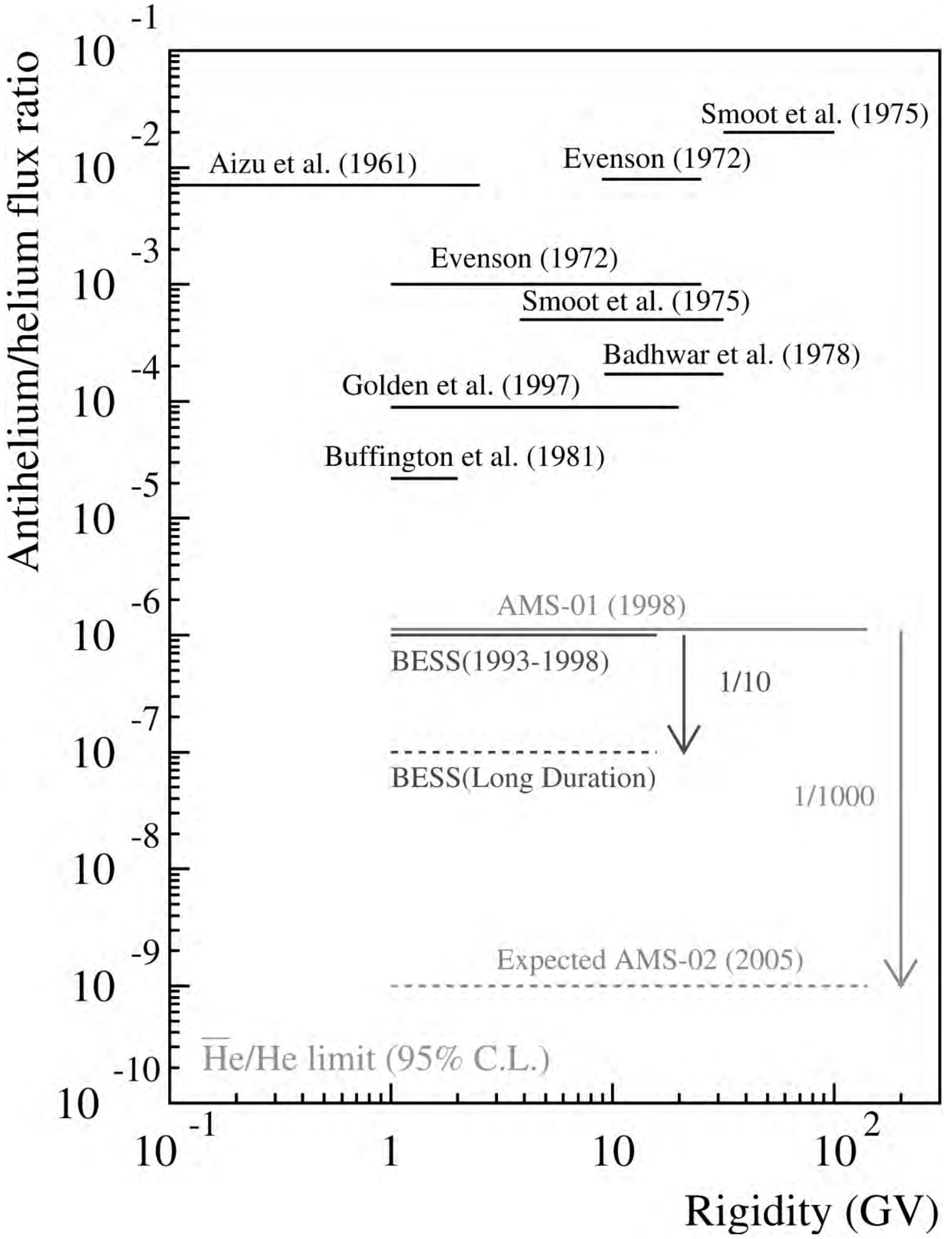


FIGURE 6 The antihelium limits at present. [From Yoshimura, K., *et al*. (2000). "Cosmic-ray antiproton and antinuclei," *Advan. Space Res*., in press.]

A source distribution which is closely patterned after the actual galaxy distribution in the nearby universe greatly enhances the expected flux of events near the GZK cutoff and may allow an interpretation of the observed events given a properly biased version of the galaxy distribution and a source population like a special class of galaxies.

However, the data are also compatible with a spectrum which suggests a rise beyond the expected GZK cutoff, and this strongly hints at a totally different and new event class. This supports an interpretation as the result of the decay of primordial relics, with an energy of 10^{24} eV. Such a decay produces a large number of highly energetic neutrinos, photons, and only 3% of ultimately protons. Since these protons start with a much flatter spectrum than normal cosmic rays, and also have such extreme initial energies, the lack of a GZK cuotff can be understood. There is, however, a strong prediction: The large number of photons produced yields a cascade that gives rise to a strong contribution to the gamma ray background. This gamma ray background adds to all the gamma rays produced by active galactic nuclei. It is not finally resolved whether observations exclude this option because of an excessive gamma ray background or whether the actual spectrum of the gamma ray background in fact supports it.

There is an interesting variant of this idea, and that is to consider neutrinos which come from large distances in the universe and interact with the local relic neutrinos (the population of Big Bang neutrinos predicted by standard Big Bang theory, akin to the microwave background). In such a theory, the low density of relic neutrinos may be a serious problem.

Obviously, if we could understand the sky distribution of events at these extreme energies, then the option of using pulsars would become very attractive. For particles with a single charge as protons, the anisotropies in arrival directions would be severe, but for iron nuclei this constraint would be much alleviated. Therefore, the option of considering pulsars entails an interpretation as iron particles, inconsistent at the transition from galactic to extragalactic cosmic rays at about 3 EeV, but conceivable at a possible transition from one source population to another at the expected GZK cutoff.

On the other hand, these particles could be something very different, as occasionally speculated, particles suggested by some versions of an extension of particle physics that may not interact with the microwave background and yet interact in the atmosphere very similarly to protons. In this case, we may ask what sources produce so extreme particles. The class of compact radio galaxies has a powerful jet, and hot spots that live currently inside a very large amount of local interstellar gas, and so provide both accelerator and beam dump, i.e., make a particle physics experiment in the sky. However, whether anything detectable comes down to us is open to question.

Finally, it has been shown that radio galaxy hot spots can in fact accelerate protons to the required energies. Here the difficulty is to identify the single most important source for the events at the highest energies. One idea, originally suggested around 1960, has been to consider the nearby radio galaxy M87. It has several advantages, and also disadvantages: First, the main problem with this notion is that the events do not show any clustering in direction with the direction to M87, which is in the nearby Virgo cluster, and so magnetic scattering or bending is required. On the other hand, M87 clearly can accelerate protons to the required energies, and in fact one needs such protons to explain its nonthermal spectrum. Intergalactic magnetic fields and also the fields in our galactic halo can bend and maybe even scatter the orbits of very energetic charged particles. These intergalactic magnetic fields clearly play the key role here, and have not been fully understood.

In conclusion, the origin of the events beyond the expected GZK cutoff remains a unsolved problem in modern high-energy physics.

VII. OUTLOOK

The origin of cosmic rays with observed particle energies to 300 EeV remains an unsolved problem.

A number of efforts related to cosmic rays will surely help our understanding:

- The determination of gamma ray spectra and maps of the Galaxy at high energy.
- More information on stellar evolution including rotation and magnetic fields. Are massive stars all converging to a small set of final states, which includes the magnetic field?
- The search to determine the strength and structure of cosmic magnetic fields, both in the halo of galaxies and in the web of the cosmological galaxy distribution. Clearly this is the key to understand the propagation of high-energy cosmic rays.
- The search for possible correlations in arrival directions of high-energy cosmic rays with astronomical sources. If there were such a subset of events, it would provide very strong constraints on the nature of the particles, as already is the case with some events near EeV energies. At higher energies it might provide new constraints on models for the fundamental properties of particles.

Progress in our understanding of cosmic rays will be mainly determined by much better data:

- Very accurate spectra, at low energies, such as now available from the first AMS data.
- Very accurate data on antimatter, positrons, antiprotons, and antinuclei, also from AMS and extended campaigns of balloon flights.
- Very accurate data on isotopic ratios, from a new generation of balloon flights.
- Accurate chemical composition near the knee and beyond.
- Accurate spectra, sky distribution, and information on the nature of particles in the three energy ranges from the knee to the ankle, through the energy range from 3 EeV to 50 EeV, the expected GZK cutoff, and beyond.

The search for the origin of cosmic rays promises to remain at the focus of research in physics in the 21st century.

ACKNOWLEDGMENTS

The authors would like to thank Ray Protheroe and Todor Stanev for very helpful comments. We thank Sera Markoff and Ramin Sina for help with the graphs.

SEE ALSO THE FOLLOWING ARTICLES

COSMIC INFLATION • COSMOLOGY • GAMMA-RAY ASTRONOMY • INFRARED ASTRONOMY • NEUTRINO ASTRONOMY • PARTICLE PHYSICS, ELEMENTARY • RADIATION PHYSICS • SOLAR PHYSICS • SUPERNOVAE

BIBLIOGRAPHY

Bhattacharjee, P., and Sigl, G. (2000). "Origin and propagation of extremely high-energy cosmic rays," *Phys. Rep.* **327,** 109–247.

Biermann, P. L. (1997a). "The origin of the highest energy cosmic rays," *J. Phys. G* **23,** 1–27.

Biermann, P. L. (1997b). "Supernova blast waves and pre-supernova winds: Their cosmic ray contribution," *In* "Cosmic Winds and the Heliosphere," (J. R. Jokipii *et al*., eds.), pp. 887–957, University of Arizona Press, Tucson, AZ.

Clay, R., and Dawson, B. (1998). "Cosmic Bullets," Addison-Wesley, Reading, MA.

Diehl, R., Kallenbach, R., Parizot, E., and von Steiger, R. (eds.) (2001). "The Astrophysics of Galactic Cosmic Rays," Kluwer, Dordrecht.

Kronberg, P. P. (2000). "Magnetic fields in the extragalactic universe, and scenarios since recombination for their origin—A review," *In* "Proceedings the Origins of Galactic Magnetic Fields," Manchester, England.

Longair, M. S. (1992/1994). "High Energy Astrophysics," Vols. 1 and 2; 2nd ed., Cambridge Univesity Press, Cambridge.

Nagano, M., and Watson, A. A. (2000). "Observations and implications of the ultra-high energy cosmic rays," *Rev. Mod. Phys*. **72,** 689–732.

Piran, T. (1999). "Gamma-ray bursts and the fireball model," *Phys. Rep.* **314,** 575.

Wiebel-Sooth, B., and Biermann, P. L. (1999). "Cosmic rays," *In* "Landolt-Börnstein," Vol. VI/3c, pp. 37–90, Springer-Verlag, Berlin.

Cosmology

John J. Dykla
Loyola University Chicago and Theoretical Astrophysics Group, Fermilab

GLOSSARY

Background radiation Field quanta, either photons or neutrinos, presumably created in an early high-temperature phase of the universe. The photons last interacted strongly with matter before the temperature fell below 3000 K and electrons combined with nuclei to form neutral, transparent hydrogen and helium atoms. Although shifted to the radio spectrum by the expansion of the universe, this "light" has been moving independently since this decoupling (in the standard model, ~700,000 years after the "big bang"), providing direct evidence of how different the universe was at this early time from its appearance now. The measured homogeneity and isotropy of the photon background are very difficult to account for in models, such as the "steady-state" theories, which avoid a "big bang."

Black hole Region of space-time with so large a curvature (gravitational field) that, classically, it prevents any matter or radiation from leaving. The "one-way" surface, which allows passage into the black hole but not out, is called an *event horizon*. Black holes are completely characterized by their mass, charge, and angular momentum.

Cosmological principle The assumption that at each moment of proper time an observer at any place in the universe would find the large-scale structure of surrounding matter and space the same as any other observer (homogeneity) and appearing the same in all directions (isotropy). Combined with the general theory of relativity, this forms the basis of the standard model in cosmology. Observational evidence for the expansion of the universe, combined with the principle of conservation of mass–energy, indicates that the large-scale average density, though everywhere the same at a given time, is decreasing as the universe evolves. The perfect cosmological principle seeks to avoid a "big bang" by asserting that the homogeneous and isotropic density of matter at all points in space is also unchanging in time. Such an assumption could be reconciled with an expanding universe only by postulating the continuous creation of matter and appears untenable in light of the microwave background radiation.

Dark matter problem The discrepancy by approximately an order of magnitude between the ordinary matter that is directly visible through astronomical observations and the mass inferred from dynamic theory applied to the observed motions of galaxies and clusters of galaxies. The calculated average density is roughly the critical value for a flat universe, which is also strongly favored by "inflationary" particle physics scenarios. Thus, various exotic forms of astronomical objects and elementary particles have been proposed to account for the dark matter assumed present.

Doppler shift Change in frequency (or, inversely because of invariant speed of propagation, wavelength) of radiation due to the relative motion of source and observer. For motion purely along the line of sight, recession implies increased wavelength (redshift) and approach implies decreased wavelength (blueshift), complicated by nonlinear relativistic effects when the relative velocity of source and observer is an appreciable fraction of the velocity of light. Purely transverse motion is associated with a relativistic and intrinsically nonlinear redshift.

Event Point in space at a moment of time. The generalization, in the four-dimensional space–time continuum of relativity theory, of the concept of a point in a space as a geometric object of zero "size."

General theory of relativity Understanding of gravitation invented by Albert Einstein as the curvature of space-time, described mathematically as a four-dimensional geometric manifold. His field equations postulate the way in which the distribution of mass–energy tells space–time how to curve and imply the equations by which curved space–time tells mass–energy how to move. With the cosmological principle, it forms the basis of the standard model in cosmology.

Nebula Latin for "cloud," this old term in observational astronomy refers to a confusing variety of extended objects easily distinguished from the pointlike stars. Some are now known to be dust or gas clouds associated with the births or deaths of stars in the Milky Way, but others are galaxies far outside the Milky Way.

Perfect fluid Idealization of matter as continuous and completely characterized by three functions of space–time: mass–energy density, pressure, and temperature. At each event, these functions are related through a thermodynamic equation of state.

Proper time Measure of time by a standard clock in the reference frame of a hypothetical observer. In describing periods in the evolution of the entire universe, the reference frame is chosen as one that at each event is comoving with the local expansion. The proper time since the "big bang" singularity is called the *age of the universe*.

Singularity Event or set of events at which some physical quantities that are generally measurable, such as density and curvature (gravitational field), are calculated to have infinite values. It represents a limit to the validity of a field theory.

COSMOLOGY as a scientific endeavor is the attempt to construct a comprehensive model of the principal features of material composition, geometric structure, and temporal evolution of the entire physically observable universe. The primary tools of the modern cosmologist are those of observational astronomy and theoretical physics. Signals are gathered over a very broad range of wavelengths of the electromagnetic spectrum, from radio waves collected by the 300-m-diameter dish at Arecibo, Puerto Rico, through visible quanta recorded by charge-coupled devices attached to the twin 10-m Keck telescopes atop Mauna Kea, Hawaii, to γ rays detected by instruments orbiting above the earth's atmosphere. In recent years, experiments in particle physics have become increasingly relevant to cosmological questions. Examples include measurements of the solar neutrino flux which challenge our understanding of energy generation in the stars and attempts to detect the superpartners, axions, strings, and other exotic objects predicted by various models in high-energy physics and possible solutions to the missing-mass problem.

The standard theoretical model in cosmology for the past several decades has been based on Einstein's general theory of relativity, supplemented by assumptions about the homogeneity and isotropy of space–time, and data from spectroscopy, consistent with understanding gained from nuclear physics, about the distribution of ordinary matter among various species. Recently, progress in understanding the early universe has been made by fusing this theory with the standard $SU(3)_C \times SU(2)_L \times U(1)_Y$ model in particle physics, which seems to describe very accurately the interactions of quarks and leptons at energies up to $\sim 2 \times 10^3$ GeV, the current limit of terrestrial accelerators. Traditional analytical techniques of the theorist are now often supplemented by the raw-number-crunching power of increasingly rapid and sophisticated computers, especially in applications to otherwise intractable calculations involving real or speculative space–times or quantum fields. Further synthesis is clearly dependent on continued advances in observational astronomy, high-energy experimentation, and theoretical physics.

Research in cosmology is subject to one obvious but fundamental limitation that no other field of scientific inquiry shares. By definition, there is only one entire physical reality that we can observe and attempt to understand. Since we cannot acquire data from outside all this, it is impossible in principle to perform a controlled experiment to test a truly comprehensive theory of the cosmos. The best we can hope for are ever more refined models, constrained by data from increasingly diverse observational sources, which become more broadly successful as we continue to measure and to think.

I. ISSUES FROM PRESCIENTIFIC INQUIRIES

A. Philosophical Speculations and Myths

Cosmological thought recorded in diverse cultures during past millennia displays a continued fascination with a relatively small number of fundamental questions about the physical universe and our place within it. Many attempts to provide satisfying answers have been made, It is not appropriate to present here a comparative survey of the myriad world views that have been proposed, but there is value in understanding those issues raised by early thought that are still at the core of modern scientific cosmology.

Logically, the first issue is that of order versus chaos. Indeed, the word *cosmology* implies an affirmative answer to the question of whether the totality of physical experience can be comprehended in a meaningful pattern. We should be aware, however, that adherents of the view that nature is fundamentally inscrutable or capricious have presented their arguments in many cultures throughout recorded history.

Is matter everywhere in the universe of the same type as that which is familiar to us on earth, or are there some distinctively "celestial" substances? Speculations favoring one view or the other have been stated with assurance as long as humans have engaged in intellectual debate.

Is the space of the universe finite or infinite, bounded or unbounded? To this, also, pure thought unfettered by empirical evidence has, in many places and times, offered dogmatic pronouncements on either side. The logical independence between the issues of finiteness and boundedness was not fully clarified until the invention of non-Euclidean geometrics in the nineteenth century. Thus, we now see that the opening question of this paragraph actually comprises two distinct but related questions. Of the issues discussed in this section, this is perhaps the only one in which progress in pure thought has raised questions not addressed by ancient philosophers and myth makers.

Is the universe immutable in its total structure, or does it evolve with the passage of cosmic time? If the latter, did it have a creation in time, or has it always existed? Similarly, will a changing cosmos continue forever, or will there be an end to time that we can experience? Such questions have sometimes been distinguished from the structural questions of cosmology and said to be of a different order called *cosmogony*. We shall see that in scientific cosmology such a separation is artificial and unproductive. Indeed, all the questions posed in this section are, in all current scientific models, inextricably interrelated.

B. Value of Some Early Questions

All modern science is predicated on the philosophical assumption that its subject is comprehensible. Although even Albert Einstein remarked that this is the most incomprehensible thing about the universe, it is difficult to see how this can be a question for scientific debate. The presumption of discoverable regularities, not meaningless chaos, is a necessary underpinning of any scientific endeavor. Although twentieth-century quantum mechanics has compelled reappraisal of earlier determinism, the work of physicists is founded upon belief in objective laws correlating observations of natural phenomena and confirmed by successful predictions.

Philosophical speculations in earlier ages about the substance of the cosmos often assumed an insurmountable qualitative distinction between the "stuff" of terrestrial experience and the matter of the celestial spheres. The heginnings of modern science included an explicit rejection of this view in favor of one in which all the matter of the universe is of fundamentally the same type, subject to laws that hold in all places and at all times. It is interesting that some recently proposed solutions to the dark matter problem in astrophysics suggest that the vast majority of the matter of the universe may be in one or more exotic forms that are predicted by elementary particle theory but for which we do not yet have experimental evidence. This is not a return to prescientific notions. It is still assumed that, no matter what the substance of the universe, all physical reality is subject to one set of laws. Most important, it is assumed that such laws can be discovered with the aid of data obtained from terrestrial experiments as well as astronomical observations.

Although most early cosmologists imagined a finite and bounded universe, some thinkers (e.g., the Greek philosopher Democritus in the fifth century B.C.) argued eloquently for an infinite and unbounded expanse of space. Until the latter part of the nineteenth century, however, a finite space was presumed synonymous with a bounded one. The non-Euclidean geometries of Riemann, Gauss, and Lobachevsky first presented the separation of the questions of extension and boundedness in the form of logically possible examples amenable to mathematical study. Originally, generalizations of the theorem of Pythagoras to curved spaces assumed that the square of the distance between two distinct points is greater than zero, described by a positive definite metric tensor. In relativity theory, time is treated as an additional "geometric" dimension, distinguished by the fact that the square of the interval between two events may be a quantity of either sign (depending on whether there is a frame of reference in which they are at the same place at different times or a frame of reference in which they are at different places at the same time) or zero (if they can be related only by the passage of a signal at the speed of light). Gravitation is understood to be the relation between the curvature of space–time and the mass–energy contained and moving within it. Modern research on the asymptotic structure of space–times is based on the analysis of four-dimensional geometries of indefinite metric within the framework of relativistic gravitational theory. All viable models consistent with minimal assumptions about our lack of privileged position describe spaces that are without a boundary surface at each moment of time, although some are finite in volume.

Myths of creation, whether based on purely philosophical speculations or carrying the authority or religious conviction, are commonplace in almost all cultures of which we have any knowledge. Some suppose the birth of a static universe in a past instant, but most describe a process of formation or evolution of the cosmos into its present appearance. In many eras, some individuals or groups have championed an eternal universe. Most of these models of the universe are static or cyclic, but occasionally they represent nature with infinite change, never repeating but with no beginning. In earlier times, a beginning in time was almost always taken to imply finiteness in a bounded space. The new geometries mentioned earlier have made possible serious consideration of a universe that began at a definite moment in the past, is not bounded in space, and may be finite or infinite in extent. The question of whether there will be an end to time or whether there will be eternal evolution is intimately related to the question of finite or infinite spatial extent in these models. At present, the empirical evidence about whether time will end is inconclusive. The "inflationary scenarios" that are currently in vogue favor a universe remarkably close to the marginally infinite and thus eternally unwinding "asymptotically flat" model, but allow the possibility of a miniscule difference from flatness of either sign, resulting in ultimate closure or eternal expansion. Although the observed density of visible matter appears much too small to halt the present expansion, there are numerous hypothetical candidates for the dark matter.

II. ORIGINS OF MODERN COSMOLOGY IN SCIENCE

A. Copernican Solar System

Centuries before the Christian era, several Greek philosophers had proposed heliocentric models to coordinate the observed motions of the sun, moon, and planets in relation to the "fixed" stars. However, the computational method of Alexandrian astronomer Claudius Ptolemy which he developed during the second century A.D. using a structure of equants and epicycles based on the earth as the only immovable point also made successful predictions of apparent motions. Although additional epicycles had to be added from time to time to accommodate the increasingly precise observations of later generations, this system became dominant in the Western world. Its assumption of a dichotomy between terrestrial experience and the laws of the "celestial spheres" stifled fundamental progress until the middle of the sixteenth century. In 1543, the Polish cleric Mikolaj Kopernik, better known by the Latin form of his name, Nicholas Copernicus, cleared the way for modern astronomy with the publication of *De Revolutionibus Orbium Coelestium* (*Concerning the Revolutions of the Celestial Worlds*). Appearing in his last year, this work summarized the observations of a lifetime and presented logical arguments for the simplicity to be gained by an analysis based on the sun as a fixed point.

More important, this successful return to the heliocentric view encouraged the attitude that a universal set of laws governs the earth and the sky. In the generations following Copernicus, the formulation of physical laws and their empirical testing were undertaken in a manner unprecedented in history: Modern science had begun. The Dane Tycho Brahe observed planetary orbits to an angular precision of 1 arc min. The German Johannes Kepler used Brahe's voluminous and accurate data to formulate three simple but fundamental laws of planetary motion. The Italian Galileo Galilei performed pioneering experiments in mechanics and introduced the use of optical telescopes in astronomy. From these beginnings, a scientific approach to the cosmological questions that humanity had been asking throughout its history developed.

B. Newtonian Dynamics and Gravitation

The first great conceptual synthesis in modern science was the creation of a system of mechanics and a law of gravitation by the English physicist Isaac Newton, published in his *Principia* (*Mathematical Principles of Natural Philosophy*) of 1687. His "system of the world" was based on a universal attraction between any two point objects described by a force on each, along the line joining them, directly proportional to the product of their masses and inversely proportional to the square of the distance between them. He also presented the differential and integral calculus, mathematical tools that became indispensable to theoretical science.

On this foundation, Newton derived and generalized the laws of Kepler, showing that orbits could be conic sections other than ellipses. Nonperiodic comets are well-known examples of objects with parabolic or hyperbolic orbits. The constant in Kepler's third law, relating the squares of orbital periods to the cubes of semimajor axes, was found to depend on the sum of the masses of the bodies attracting one another. This was basic to determining the masses of numerous stars, in units of the mass of the sun, through observations of binaries. As Cavendish balance experiments provided an independent numerical value for the constant in the law of gravitation, stellar masses in kilograms could be calculated. Later, analysis of observed perturbations in planetary motions led to the prediction of previously unseen planets in our solar system. Studies of the stability of three bodies moving under their mutual gravitational influence led to the discovery of two clusters of asteroids sharing the orbit of Jupiter around the sun. These successes fostered confidence in the view of one boundless Euclidean space, the preferred inertial frame of reference, as the best arena for the description of all physical activity. Lacking direct evidence to the contrary, theoretical cosmologists at first assumed that the space of the universe was filled, on the largest scales, with matter distributed uniformly and of unchanging density. Analysis soon disclosed that Newtonian mechanics implied instability to gravitational collapse into clumps for an initially homogeneous and static universe. This stimulated the observational quest for knowledge of the present structure of the cosmos outside the solar system.

C. Technology of Observational Astronomy

Galileo's first telescopes had revealed many previously unseen stars, particularly in the Milky Way, where they could resolve numerous individual images. Based on elementary observations and careful reasoning, the ancient concept of a sphere of fixed stars centered on the earth had been supplanted by a potentially infinite universe with the earth at a position of no particular distinction. Further progress in understanding the distribution of matter in space was dependent on actual measurements of the distances to the stars. The concept of stellar parallax, variation in the apparent angular positions of nearby stars in relation to more distant stars as seen from earth at various points in its orbit around the sun, became useful when the observational precision of stellar location surpassed the level of 1 arc sec. An important contribution was the development of the micrometer, first used by William Gascoigne, later independently by Christiaan Huygens, and systematically appearing in the telescope sights of Jean Picard and others after the latter part of the seventeenth century. The first parallax measurements were reported by Bessel in Germany, Henderson in South Africa, and Struve in Russia, in rapid succession in 1838 and 1839. By 1890, distances were known for nearly 100 stars in the immediate neighborhood of the sun, and the limits of the trigonometric parallax approach (∼100 parsec, based on limits to the useful size of an optical telescope looking through the earth's atmosphere) were being approached.

As long as sufficient light can be gathered from a star to be dispersed for the study of its spectrum, much can be learned about it even if one is not certain how far away it is. Spectroscopic studies of stars other than the sun were first undertaken by William Huggins and Angelo Secchi in the early 1860s. In 1868, Huggins announced the measurement of a Doppler redshift in the lines of Sirius indicating a recession velocity of 47 km/sec, and Secchi published a catalog of 4000 stars divided into four classes according to the appearance of their spectra. The introduction of the objective prism by E. C. Pickering in 1885 tremendously increased the rate at which spectra could be obtained. The publication of the Draper catalogs in the 1890s, based on the spectral classification developed by Pickering and Antonia Maury, provided the data that led to the rise of stellar astrophysics in the twentieth century.

D. Viewing the Island Universe

Some hazy patches among the fixed stars, such as the Andromeda nebula or the Magellanic Clouds, are clearly visible to the unaided human eye. When Galileo observed the band called the Milky Way through his telescope and was surprised to find it resolved into a huge number of faint stars, he concluded that most nebulosities were composed of stars. Early in the eighteenth century, the Swedish philosopher Emanuel Swedenborg described the Milky Way as a rotating spherical assembly of stars and suggested that the universe was filled with such spheres. The English mathematician and instrument maker Thomas Wright also considered the Milky Way to be one among many but supposed its shape to be a vast disk containing concentric rings of stars. By 1855, Immanuel Kant had further developed the disk model of the Milky Way by applying Newtonian mechanics, explaining its shape through rotation. He assumed that the nebulae are similar "island universes."

Early in the nineteenth century, however, William Herschel discovered the planetary nebulae, stars in association with true nebulosities. This reinforced a possible interpretation of nebulae as planetary systems in the making, in agreement with the theory of the origin of the solar system developed by Pierre Simon de Laplace. John Herschel's observations of 2500 additional nebulae, published in 1847, emphasized that they were mostly distributed away from the galactic plane. The "zone of avoidance" was used by some to argue for the physical association of the nebulae with the Milky Way. (It is now known that dust and gas in the plane of the galaxy diminish the intensity of light passing through it, whether from sources inside it or beyond.) In 1864, Huggins studied the Orion nebula and found it to display a bright line spectrum similar to that of a hot gas. Also, photographs of Orion and the Crab nebula did not show resolution into individual stars. As the twentieth century began, the question of the nebulae was intricately linked with that of the structure of the Milky Way. The size of nebulae was still quite uncertain in the absence of any reliable distance indicators, but most astronomers believed the evidence favored considering the nebulae part of the Milky Way, thus reducing the universe to this one island of stars.

III. REVOLUTIONS OF THE TWENTIETH CENTURY

A. New Theoretical Models

The two great conceptual advances of theoretical physics in the twentieth century, relativity and quantum mechanics, have had profound implications for cosmology. The general theory of relativity provides the basis for the evolving space–time of the now standard model, and fundamental particle identities and interactions are keys to understanding the composition of matter in the universe.

In 1905, Albert Einstein established the foundations of the special theory of relativity, which connects measurements of space and time for observers in all possible inertial frames of reference. He devoted the next decade to developing a natural way of including observers in accelerated frames of reference, using as a fundamental principle the fact that all test masses undergo the same acceleration in a given gravitational field. The result was Einstein's theory of space–time and gravitation, the general theory of relativity (GTR), completed in 1916. It was immediately apparent that the GTR would have a substantial impact on cosmological questions. Although many model space–times have been studied in the GTR, in 1922 Alexander Friedmann showed that the assumptions of spatial homogeneity and isotropy can be embodied in only three. They are presented here in the modern notation of the Robertson–Walker metric with the choice of units commonly used in theoretical research. The intimate relationship of space and time in relativistic theories is recognized through units of measure in which the speed of light is numerically 1, eliminating the letter c representing its value in units such as those of SI. We may think of the units as geometrized (e.g., time measured in meters) or chronometrized (e.g., length measured in seconds), where $1 \text{ sec} = 299,792,458$ m (exactly, by definition).

The invariant element of separation ds between two neighboring events in a homogeneous and isotropic universe can be expressed by:

$$ds^2 = -dt^2 + R^2[dr^2/(1-kr^2) + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)] \tag{1}$$

where $R(t)$ is the scale factor that evolves with cosmic time and k is the curvature constant, which may be positive, negative, or zero. The invariant separation depends on k only through the pure number kr^2, where r is a comoving space coordinate, for which matter is locally at rest. Without loss in generality, we may choose k (and hence also r) dimensionless and set the scale of r so that k is $+1$, -1, or 0. Although R carries the units of spatial length, it is not possible to interpret it as a "radius of the universe" if k is not positive, since the volume of a model with $k = 0$ or $k = -1$ is infinite.

To ensure that the physics is described in a manner independent of arbitrary choices, it is useful to introduce the scale change rate, also known as the Hubble parameter, defined as $H = (dR/dt)/R$. Since H has dimensions of velocity divided by length, $1/H$ is a characteristic time for evolution of the model. For an expanding empty universe, $1/H$ would be the time since the beginning of the expansion. Assuming that the distribution of matter on a large scale can be described in terms of a perfect fluid of total density ρ and pressure p, the coupled evolution of space-time and matter in the GTR are determined by the Friedmann equations,

$$H^2 = 8\pi G\rho/3 - k/R^2 \qquad \text{and} \qquad d(\rho R^3) = -p\, d(R^3) \tag{2}$$

where G is the Newtonian gravitational constant. (Note that the term *density* can be taken as either mass per volume or energy per volume in units where $c = 1$.)

Although the GTR implies that a homogeneous and isotropic space generally expands or contracts, at the time the theory was formulated the common presumption held that the real universe was static. This led Einstein to modify his original simple theory by introducing a "cosmological term" into the field equations. The "cosmological constant" in this term was chosen to ensure a stable static solution. When the evidence for an expanding universe became apparent in the next decade, Einstein dropped consideration of this additional term, calling its introduction the biggest blunder of his career. We shall leave it out of the presentation in this section but later remark on recent motivations for a possible reinstatement of the cosmological constant with a value other than that which makes the universe static. Thus, there are only three possible homogeneous and isotropic universes in the GTR, the Friedmann models. In view of Eq. (2), which determines the evolution of the Hubble parameter, the curvature constant can be positive only if the density exceeds the value $\rho_{\text{crit}} = 3H^2/8\pi G$. Thus, an expanding universe in which $k = +1$ must halt its growth and begin to collapse before its mean density drops below this critical value. Such a space-time is finite in volume at each moment but has no boundary. While this model is expanding, the Robertson–Walker coordinate time since the expansion began is less than $2/3H$, depending on how much the density exceeds the critical value. The classical GTR implies that a finite proper time passes between its beginning in a singularity of zero volume and its return to such a singularity. (However, Charles Misner has argued that a logarithmic time scale based on the volume of the universe is more appropriate as a measure of possible change that may occur, giving even this closed cosmology a potentially "infinite" future and past.) If the density is less than the critical value, space–time has negative curvature and will have a positive velocity of expansion into the infinite future of proper time, when H approaches zero. The time since expansion began is greater than $2/3H$, depending on how much the density is less than the critical value, but never exceeds $1/H$. Although such an open universe still had a beginning in proper time at a singularity, its volume is always infinite and it contains an infinite mass–energy. If the density precisely equals the critical value, then the curvature vanishes. Such a flat universe has been expanding for a time $2/3H$, it has infinite mass–energy in an infinite volume described by Euclidean geometry at each moment, and its expansion velocity as well as Hubble parameter will approach zero asymptotically as proper time in comoving coordinates continues toward the infinite future. Which of these three models best describes the universe we inhabit is an empirical as well as theoretical question that remains central to current research in cosmology.

B. New Observational Evidence

Beginning in the late 1920s, Vesto Slipher, Edwin Hubble, and Milton Humason used the Hooker telescope at Mt. Wilson, California, then the largest optical instrument in the world, to measure Doppler shifts in the spectra of nebulae that they realized were outside the Milky Way. Convincing reports of a relationship between recession velocities v deduced from Doppler redshifts and estimates of the distances d to these galaxies were published in 1929 and 1931. Hubble pointed out that the velocities, in kilometers per second, were directly proportional to the distances, in megaparsecs ($1 \text{ Mpc} = 3.08 \times 10^{19}$ km). The observational relation is "smooth" only when the distances considered are at least of the order of a few megaparsecs, allowing the averaging of the distribution of galaxies to produce an approximately homogeneous and isotropic density. Since the ratio v/d is clearly the current value of the scale change rate H in a Friedmann universe, this was the first observational evidence in support of GTR cosmology.

The parameter H was at first called the Hubble constant but is now recognized as a misnomer, because the early data did not extend far enough into space to correspond to looking back over a sustantial fraction of the time since the expansion began. Hubble's initial value of H was so large that the time scale for expansion derived from it was substantially less than geological estimates of the age of the earth, $\sim 4.7 \times 10^9$ years. Subsequently, the numerical value of the Hubble parameter has undergone several revisions due to reevaluations of the cosmic distance scale, yielding a universe older than originally supposed and thus a much larger volume from which signals at the speed of light can be observed at our position. Allowing for present uncertainties, H is now believed to be between 50 and 80 (km/sec)/Mpc. Thus, the time scale of the cosmic expansion, $1/H$, is between approximately 1.2 and 2×10^{10} years.

When early large values of the Hubble parameter were in conflict with geological and astrophysical estimates of the age of the solar system, it was suggested that the simple GTR cosmologies might not be viable. One solution was to reinstate the cosmological constant, not to halt the universal expansion but merely to slow it down in a "coasting" phase. Recently, some observations of supernovae suggest an extragalactic distance scale with H greater than 70 (km/sec)/Mpc. If we accept this and also believe that stellar evolutionary theory is sufficiently established to yield precise ages of globular cluster stars in excess of 1.3×10^{10} years, then invoking a nonzero cosmological constant would be a way of avoiding the unacceptable conclusion that the universe contains stars older than its expansion time. Models studied by George Lemaitre and others after 1927 often had unusual features, such as a closed space of positive curvature that could continue expanding for an infinite proper time. A more radical idea, which had other motivations as well, was to abandon conservation of energy in favor of a "steady-state" cosmology based on the continuous creation of matter throughout an infinite past. In such a theory, introduced by Hermann Bondi, Thomas Gold, and Fred Hoyle in 1948, there is no initial singularity, or "big bang," to mark the beginning of the expansion. In 1956, George Gamow predicted that a residual electromagnetic radiation at a temperature of only a few kelvins should fill "empty" space as a relic of the high temperatures at which primordial nucleosynthesis occurred soon after the initial singularity in a GTR cosmology. In 1965, engineers Arno Penzias and Robert Wilson detected a microwave background coming from all directions in space into a communications antenna they had designed and built for Bell Laboratories in Holmdel, New Jersey. Cosmologists Robert Dicke and P. J. E. Peebles at nearby Princeton quickly explained the significance of this 2.7 K blackbody spectrum to them and thus dealt a severe blow to the viability of any cosmological model that avoids a "hot big bang."

A direct determination of whether the universe is spatially closed or open would necessitate precisely measuring density over volumes of many cubic megaparsecs, detecting the sign of departures from Euclidean geometry in an accurate galactic census over distances exceeding several tens of megaparsecs, or finding the change in the Hubble parameter associated with looking back toward the beginning over distances exceeding several hundred megaparsecs. Unfortunately, these conceptually simple observations appear to be somewhat beyond the scope of our present technology. However, there are various indirect ways to estimate, or at least place bounds on, the present mean density of the cosmos. To eliminate the influence of uncertainties in the Hubble parameter on the precise value of the critical density, it is now common to describe the resulting estimates or bounds in terms of the dimensionless quantity $\Omega = \rho/\rho_{\text{crit}}$. Clearly, $\Omega > 1$ corresponds to a closed, finite universe and $\Omega \leq 1$ corresponds to an infinite universe. On the basis of the amount of ordinary visible matter observed as stars and clouds in galaxies, we conclude that $\Omega > 0.01$. Requiring the present density to be low enough so that the age of the universe (which depends inversely on H and, through a monotonically decreasing function, on Ω) is at least as great as that of the oldest stars observed, estimated to be 10^{10} years, yields an upper bound $\Omega < 3.2$. Of the infinite range of conceivable values, it is quite remarkable that the universe can so easily be shown to be nearly "flat." Further evidence and theoretical insight suggest that near coincidence (within one or two powers of 10) of the density and the critical density is not an accident.

The most severe constraints on the contribution of ordinary matter to Ω presently come from demanding that primordial synthesis of nuclei during the "hot big bang" of a standard Friedmann model produces abundance ratios in agreement with those deduced from observation. In nuclear astrophysics, it is customary to specify temperatures in energy units which corresponds to setting the Boltzmann constant equal to 1. Thus, the relation between the megaelectronvolt of energy and the kelvin of absolute temperature is

$$1 \text{ MeV} = 1.1605 \times 10^{10} \text{ K}$$

Primordial nucleosynthesis models are based on the assumption that the temperature of the universe was once higher than 10 MeV, so that complex nuclei initially could not exist as stable structures but were formed in, and survived from, a brief interval as the universe expanded and cooled to temperatures of less than 0.1 MeV, below which nucleosynthesis does not occur. Computing all relevant nuclear reactions throughout this temperature range to determine the final products is a formidable undertaking. Of the various programs written since Gamow suggested the idea of primordial nucleosynthesis in 1946, the one published by Robert Wagoner in 1973 has become the accepted standard. With updates of reaction rates by several groups since then, the numerical accuracy of the predicted abundances is now believed to be ~1%. Since the weakly bound deuteron is difficult to produce in stars and easily destroyed there, its abundance, 1×10^{-5} relative to protons as determined in solar system studies and from ultraviolet absorption measurements in the local interstellar medium, is generally accepted as providing a lower bound to its primordial abundance and hence an upper bound of 0.19 to the contribution of baryons to Ω. Analogous arguments may be applied to establishing a concordance between predicted and observed abundances of ^{3}He, ^{4}He, and ^{7}Li, the only other isotopes calculated to be produced in significant amounts during this primordial epoch. The resulting constraint on the contribution of baryons to the critical density, $0.010 \leq \Omega_{\text{B}} \leq 0.080$, shows clearly that baryons alone cannot close the universe. Of course, this does not eliminate the possibility that the universe may have positive curvature due to the presence of less conventional forms of as yet unseen matter.

IV. UNSOLVED CLASSICAL PROBLEMS

A. Finite or Infinite Space

Although concordance between the standard model and observed nuclear abundances limits baryon density to well below the critical value needed for closure of a Friedmann universe, the question of finite or infinite space remains observationally undecided owing to other complications. In principle, it should be possible to determine the curvature constant k by direct measurement of the deviation from Hubble's simple linear relationship between velocity and distance. Unfortunately, substantial deviation is not expected, until sources at distances of the order of $1/H$ are studied, and galaxies are too faint to have their spectra measured adequately at such distances by present technology. Furthermore, observing galaxies at such distances implies seeing them at earlier times, and estimates of their distance could be subject to systematic errors due to unknown evolution of galactic luminosity.

Since the discovery of quasars by Maarten Schmidt at the Palomar Observatory in 1963, cosmologists have hoped that these most distant observed objects could be used to extend the Hubble relationship to the nonlinear regime and decide the sign of the curvature. Quasars are now known with spectral features up to 6.0 times their terrestrial wavelengths, corresponding to Doppler recession velocities up to 96% of the speed of light, placing them at substantial fractions of the distance from us to the horizon of the observable universe. However, uncertainties in estimating very large distances in the universe due to insufficient understanding of the evolution of galaxies, not to mention the structure of quasars, have prevented unambiguous determination of the sign of the curvature. In fact, Hubble's law is still used to estimate distances to the quasars, rather than they being used to determine both distance and redshift and thus test their relationship. The question of whether space is finite or infinite remains unresolved by observation at this time.

B. Eternal Expansion or an End to Time

In the Friedmann cosmologies of the GTR, a finite space implies an end to proper time in the future, but this is not required in some nonstandard models. Assuming the Robertson–Walker form of the metric for a homogeneous and isotropic space–time, it is convenient to discuss the future evolution of any such expanding universe in terms of a dimensionless deceleration parameter, defined as:

$$q = -R(d^2R/dt^2)/(dR/dt)^2$$

Throughout most of its history, the dynamics of the universe have been dominated by matter in which the average energy density is very much greater than the pressure. Neglecting the pressure of nongravitational fields, space will reverse its expansion and collapse in finite proper time if and only if $q > \frac{1}{2}$. From the Friedmann equation for the Hubble parameter, it is easy to show that $q = \Omega/2$ in a space–time described by the GTR with zero cosmological constant. In nonstandard models, the deceleration parameter depends on the density, the cosmological constant, and the Hubble parameter in more complicated ways. For some choices of cosmological constant, it is even possible to have an accelerating universe ($q < 0$) with a positive density. However, a cosmological constant whose magnitude substantially exceeded the critical energy density would produce detectable local effects that are not observed. Thus, we can conclude that a sufficiently large density must imply an end to time.

Applications of the virial theorem of Newtonian mechanics to galactic rotation and the dynamics of clusters indicate that masses often exceed those inferred from visible light by about an order of magnitude. Such results push at the upper bound of the baryon density inferred from nucleosynthesis but are still far short of supporting $q > \frac{1}{2}$. However, if only one-tenth of the ordinary matter in galaxies and clusters may be visible to us, is it not possible that there exists mass–energy, in as yet undetected forms or places, of sufficient quantity to produce deceleration exceeding the critical value? Primordial black holes, formed before the temperature of the universe had dropped to 10 MeV, would not interfere with nucleosynthesis in the standard model but could substantially increase the value of Ω. (Notice that black holes due to the collapse of stars are made of matter that contributed to Ω_B during the time of primordial nucleosynthesis and thus are limited by the nuclear abundance data.) Calculations of the primordial black hole mass spectrum, such as those by Bernard Carr, have demonstrated that present observational data are insufficient to decide whether the contribution of black holes to the density will reverse expansion. More conventional astronomical candidates for the dark matter including brown dwarfs (examples have been detected in the halo of the Milky Way through gravitational lensing) and dead (radio-quiet) pulsars (almost certain to exist but not as yet detected) are unfortunately constrained by the nucleosynthesis limits on baryon density. Some speculative ideas in high-energy particle physics present other exotic candidates for dark matter that may dominate the gravitational dynamics of the universe as a whole. These include massive neutrinos (although observations of electron neutrinos from SN1987A [Supernova Shelton] constrain the electron neutrino mass to probably less than 10 eV, tau and mu neutrinos could be heavier), the axions required to banish divergences in grand unified theories (GUTs) (as yet undetected), and the supersymmetric partners of all known fermions and bosons (none yet found experimentally). Empirically, whether there will be an end to time remains an unresolved question.

C. Observations and Significance of Large-Scale Structure

The measured homogeneity and isotropy of the cosmic microwave background radiation temperature ($\Delta T/T \sim 10^{-5}$) is strong evidence that the observable universe is rather precisely homogeneous and isotropic on the largest scale ($1/H$ is ~3000 Mpc). However, it is well known that on only slightly smaller scales, up to 120 Mpc, the universe today is very inhomogeneous, consisting of stars, galaxies, and clusters of galaxies. For example, the variation in density divided by the average density of the universe, $\delta\rho/\rho$, is of the order of 10^5 for galaxies. The density of visible matter in large "voids" recently discovered is typically less than average by about one order of magnitude. Since gravitational instability tends to enhance any inhomogeneity as time goes on, the difficulty is not in creating inhomogeneity but rather in deviating from perfect homogeneity at early times in just the way that can account for the structural length scales, mass spectra, and inferred presence of dark matter that are so obvious in the universe today. Opinions about the structure of the universe at early times have run the gamut from Misner's chaotic “mixmaster” to Peebles’s quite precisely homogeneous and isotropic space-times. The issue of the origin of structure in the universe on the largest scales remains unresolved, because conflicting scenarios that adequately account for galaxies and clusters can involve so many tunable parameters that it is difficult to distinguish among competing models observationally.

The three-dimensional map of more than 1.1×10^4 galaxies within a sphere of diameter 400 Mpc centered on the Milky Way, created by Huchra and Geller on the basis of more than 5 years’ data gathered with an earthbound telescope comparable in aperture to the Hubble Space Telescope (HST returned its first images May 1990), shows voids and filamentary structures on scales up to about 120 Mpc. They strain “top down” scenarios of the evolution of cosmic structure, since it is difficult to have gravitationally bound structures so large which “later” fragmented to form quasars when the universe was less than 7% of its present age. Advocates of such pictures have questioned the statistical significance of a sample only on the order of 10^{-7} of the galaxies within our horizon. The HST, despite initial difficulties in forming optimally sharp images (corrected in 1993), has gathered spectra of a sample from roughly an order of magnitude deeper into space (implying roughly 10^3 times as many galaxies). Combined with evidence of galaxy formation at early times visible in the Hubble deep fields, this enlarged sample supports a two-component model of structure formation. Hot dark matter provides a gravitational field which fragments on the large scale of galaxy clusters, while cold dark matter pockets act as seeds for the concurrent formation of subunits of galaxies. HST will spend part of its remaining life in orbit further improving statistics to refine the interpretation of an enlarged sample. It might finally gather data that might unravel the contributions of galactic evolution from that of the cosmological deceleration to the redshift-distance relation of the most distant active galaxies and quasars.

D. Viability of Nonstandard Models

The assumption that the universe is homogeneous and isotropic on a sufficiently large scale has been called the cosmological principle. This principle, applied to the Riemannian geometry of space-time using the methods of group theory, leads to the Robertson–Walker form of the invariant separation between events. This mathematically elegant foundation for theoretical cosmology is independent of a particular choice of gravitational field equations. However, the evolution of the scale factor and the relation of the curvature constant to the matter distribution are, of course, intimately related to the structure of the gravitational theory that is assumed.

Previous mention was made of the logical possibility of complicating the equations of the GTR by introducing a cosmological constant. Though hints of a possible conflict between ages of galactic halo stars and the Hubble time might be resolved by a nonzero cosmological constant, there is abundant evidence that it cannot be significantly larger than the critical density. Attempts to derive a value from quantum theories of fundamental interactions yield estimates too large by many orders of magnitude. This embarrassing failure leaves advocates of a cosmological constant only the unpleasant option of arbitrarily adjusting it to a suitable small value. Critics then ask why it should not be chosen to vanish exactly.

Numerous alternatives to the GTR consistent with the special theory of relativity have been proposed. Some may be eliminated from further consideration by noncosmological tests of gravitational theory. For example, theories based on a space–time metric that is conformally flat, such as that published by Gunnar Nordstrom in 1912, are untenable because they fail to predict the deflection of light rays in the gravitational field of the sun, first measured by Dyson, Eddington, and Davidson during a solar eclipse in 1919. Others, such as the scalar–tensor theory published by Carl Brans and Robert Dicke in 1961, may be made consistent with current observational data by suitable adjustment of a parameter. Their cosmological consequences present distinct challenges to the standard model for some times during the evolution of the universe. However, sufficiently close to singularities, the predictions of most such theories become indistinguishable from those of the GTR. For example, in 1971 Dykla, Thorne, and Hawking showed that gravitational collapse in the Brans–Dicke theory inevitably leads to the “black hole” solutions of the GTR. Hence, cosmological models based on these theories make predictions very much like those of the GTR for the strong fields near the “big bang” and the nearly flat space-time of today. If they are very different in some intermediate regime, perhaps one important to the formation of structures such as galaxies, there currently appear to be no crucial tests that could cleanly decide in their favor. Thus, by an application of Occam’s razor, most contemporary models are constructed assuming the validity of the GTR.

An influential exception to the dominance of GTR models was the “steady-state” cosmology of Bondi, Gold, and Hoyle. The philosophical foundation of this work was the extension of the cosmological principle to the “perfect cosmological principle,” which asserted that there should be uniformity in time as well as spatial homogeneity and isotropy. As remarked earlier, observations of the microwave background presently render this model untenable. It is also unable to account for the abundances of various light nuclei synthesized when the universe was much hotter and denser than it is now. The apparent overthrow of the perfect cosmological principle encouraged questioning of the assumption of spatial homogeneity and isotropy. Since the empirical evidence in favor of spatial uniformity on very large scales is quite strong, most cosmologists today would like to deduce the cosmological principle rather than assume it. That is, we would like to demonstrate that, starting with arbitrary initial conditions, inhomogeneities and anisotropies are smoothed out by physical processes in a small time compared with the age of the universe. A serious difficulty in attempts to derive the cosmological principle was first emphasized by Misner in 1969. Relativistic space–times of finite age have particle horizons, so that at any moment signals can reach a given point only from limited regions, and parts of the universe beyond a certain distance from one another have not yet had any possibility of communicating. The correlation in conditions at distant regions that is asserted by the cosmological principle can be derived only if chaotic initial conditions smooth themselves out through an infinite number of processes, such as the expansions and contractions along different axes in "mixmaster" universes. In the standard model, the cosmological principle is regarded as an unexplained initial condition.

V. INTERACTION OF QUANTUM PHYSICS AND COSMOLOGY

A. Answers from Particle Physics

The interaction between elementary particle physics and cosmology has increased greatly since 1980, to the benefit of both disciplines. Several initial conditions of classical cosmology are given a tentative explanation in "inflationary universe" scenarios, proposed by Alan Guth in 1981 and modified by Linde, Albrecht, and Steinhardt in 1982. These models assume a time when the energy density of a "false vacuum" in a grand unified theory (GUT) dominated the dynamics of the universe. Since the density was essentially constant throughout this period, the Robertson–Walker scale factor grew exponentially in time, allowing an initially tiny causally connected region (even smaller than the small value of $1/H$ at the start of inflation) to grow until it included all of the space that was to become the currently observable universe. The original version (1981) assumed that this occurred while the universe remained trapped in the false vacuum.

Unfortunately, such a universe that inflated sufficiently never made a smooth transition to a radiation-dominated, early Friedmann cosmology. In the "new inflationary" models (1982), the vacuum energy density dominates while the relevant region of the universe inflates and evolves toward the true vacuum through the spontaneous breaking of the GUT symmetry by nonzero vacuum expectation values of the Higgs scalar. The true vacuum is reached in a rapid and chaotic "phase transition" when the universe is of the order of 10^{-35} sec old, resulting in the production of a large number and variety of particles (and antiparticles) at a temperature of the order of 10^{14} GeV. It is supposed that the universe evolves according to the standard model after this early time.

Since the entire observable universe evolves from a single causally connected region of the quantum vacuum, inflationary models obviously avoid horizon problems. The homogeneity and isotropy of the present observable universe are a consequence of the dynamic equilibrium in the tiny region. Since the density term in the Friedmann equation for the evolution of the Hubble parameter remains essentially constant throughout the inflationary era, while the curvature term is exponentially suppressed, these scenarios also offer a natural explanation for the present approximate flatness of the universe. In fact, plausible suppression of the curvature greatly exceeds that required by any astrophysical observations. Only by artificially contrived choices of parameters in an inflationary universe could we avoid the conclusion that any difference between the current value of Ω and unity is many orders of magnitude less than 1. If this result and the bounds on Ω_B from primordial nucleosynthesis are both true, we must conclude that dark matter of as yet undetermined form dominates the dynamics of the universe. Since GUTs predict the nonconservation of B (baryon number), C (charge conjugation), and CP (product of charge conjugation and parity), the decay of very heavy bosons far from thermodynamic equilibrium offers a way of dynamically generating the predominance of matter over antimatter rather than merely asserting it as an initial condition. In the absence of observation of the decay of the proton or accelerators capable of attaining energies at which GUTs predict convergence of coupling "constants," the apparent baryon asymmetry of the universe is perhaps the best empirical support for some sort of unified theory of quarks and leptons.

B. Constraints from Cosmology

Even as elementary particle theory solves some problems of cosmology, it is subject to limitations derived from cosmological data involving energies far beyond the 2×10^3 GeV limit of existing terrestrial accelerators. An important example involves the production of magnetic monopoles in the early universe. In 1931, P. A. M. Dirac showed that assuming the existence of magnetic monopoles led to a derivation of the quantization of magnetic and electric charge and a relation between them implying that magnetic charges would have to be very large. However, other properties of the hypothetical monopoles, such as mass and spin, were undetermined in his theory. In 1974, Gerhard t'Hooft and Alexander Polyakov showed that monopoles must be produced in gauge theory as topological defects whenever a semisimple group breaks down to a product that contains a $U(1)$ factor, for example,

$$SU(5) \to SU(3) \times SU(2) \times U(1)$$

All proposed GUTs, which attempt to unify the strong and the electroweak interactions, are examples of gauge theories in which monopoles are required and have masses of the order of the vacuum expectation value of the Higgs field responsible for the spontaneous symmetry breaking. The present experimental lower limit, of the order of 10^{33} years, for the mean life of the proton implies a lower limit for this mass of the order of 10^{18} GeV. Only within a time of at most 10^{-35} sec after the "big bang" was any place in the universe hot enough to produce particles of so great a mass, either as topological "knots" or as pairs of monopoles and antimonopoles formed in the energetic collisions of ordinary particles.

In addition to their enormous masses and relatively large magnetic charges, monopoles are predicted by GUTs to serve as effective catalysts for nucleon decay. Thus, if present in any appreciable abundance in the universe today, monopoles should make their presence obvious by doing some conspicuously interesting things. If monopoles were made in about the same abundance as baryons, their density alone would exceed the critical value for a closed universe by a factor of $\sim 10^{11}$. Monopoles would use up the potential energy of stationary magnetic fields, such as that of the Milky Way, by converting it to increases in their own kinetic energy. Collecting in a star throughout its history, they would render its collapsed "final state" short-lived by catalyzing nucleon decay. The observational upper bound on Ω, lower limits on the galactic magnetic field, and the life spans of neutron stars place severe constraints on the flux of monopoles at present and hence on their rate of production during the spontaneous symmetry-breaking era. In fact, unacceptably large magnetic monopole production in the simplest GUTs was one of the primary motivations for the development of the new inflationary universe models, which solve the problem through an exponential dilution of monopole density that leaves very roughly one monopole in the entire observable universe at present. While such a scenario appears to make the experimental search for monopoles essentially futile, it is important not to ascribe too much quantitative significance to this result, because the predicted number is exponentially sensitive to the ratio of the monopole mass to the highest temperature reached in the phase transition at the end of spontaneous symmetry breaking. Thus, an uncertainty of a mere factor of 10 (theoretical uncertainties are at least this large) in this ratio changes the predicted number of monopoles by a factor of the order of 10^8.

On the experimental front, Blas Cabrera claimed the detection of a magnetic monopole on February 14, 1982, after 150 days of searching with a superconducting quantum interferometer device (SQUID). If this observation were correct and even approximately corresponded to the typical distribution of monopoles in space, then neither the excessive production of a naive GUT model nor the extreme scarcity of a new inflationary model could be credible. Confirmation of this monopole detection would leave current theory totally at a loss to explain the monopole abundance, but neither Cabrera nor other observers have yet claimed another detection. After more than 3000 days had passed, most workers were of the opinion that the single "event" was due to something less exotic than a magnetic monopole.

The apparent smallness of the cosmological constant is a fact that has not yet been explained in any viable theory of particle physics or gravitation. Below some critical temperature in electroweak theory or GUTs, the effective potential function of the Higgs fields behaves like a cosmological constant in contributing a term equal to this potential function times the space–time metric to the stress–energy–momentum tensor of the universe. Empirical bounds on the vacuum energy density today imply that this potential at the spontaneous symmetry-breaking minimum was already less than 10^{-102} times the effective potential of the false vacuum. There is no derivation of this extremely small dimensionless number within the framework of GUTs. In fact, the assumption that the cosmological constant is negligible today is an unexplained empirical constraint on the otherwise undetermined scale of the effective potential in a gauge theory.

Another possible success of inflationary models is the natural development of nearly scale-independent density inhomogeneities from the quantum fluctuations in the Higgs field of GUTs during inflation. Inhomogeneities should later evolve by gravitational clumping into galaxies and clusters of galaxies. This opens the possibility of calculation from first principles of the spectrum of later structural hierarchies. Comparison of the results of such calculations with the observed large-scale structure of the universe may provide the most stringent constraints on new inflationary models.

C. Singularities and Quantum Gravity

This survey has looked back nearly 10^{18} sec from the present to the "big bang" with which the standard cosmological model claims the universe began. The attempt to understand its evolution reveals a number of significant eras. Let us review them from the present to the initial singularity in reverse chronological order, which is generally the order of decreasing direct experimental evidence and thus increasing tentativeness of conclusions. At a time of the order of 10^{13} sec after the universe began, when the temperature was $\sim$0.3 eV, the photons that now compose the cosmic microwave radiation background last appreciably interacted with matter, which then "recombined" into transparent neutral atoms of hydrogen, helium, and lithium and began to form the large-scale structures familiar to us: stars, galaxies, and clusters of galaxies. At a time of the order of 10^{-2} sec after the beginning, when the temperature was $\sim$10 MeV, the free neutrons and some of the free protons underwent the primordial synthesis that formed the nuclei of these atoms, and the cosmic background neutrinos ceased having significant interactions with matter. At a time of the order of 10^{-5} sec after the beginning and a temperature $\sim$300 MeV, quarks became "confined" to form the hadrons as we now know them. At a time of the order of 10^{-12} sec after the beginning, the temperature was $\sim 10^3$ GeV, the present limit of terrestrial accelerators. The distinction between electromagnetic and weak interactions was not significant before then. The reconstruction of earlier history is of necessity much more tentative. The spontaneous symmetry breaking of the grand unification of electroweak and strong interactions is thought to have occurred at a time of roughly 10^{-34} sec and a temperature of the order of 10^{14} GeV. During the "inflation" preceding this epoch the baryon asymmetry of the universe may have been generated by fluctuations from thermal equilibrium in GUTs, and magnetic monopoles may have been produced by symmetry breaking. Any attempt to analyze events at substantially earlier times must address the unfinished program of constructing a quantum theory that unifies gravitation with the strong and electroweak interactions.

In the absence of complete understanding, the time and temperature scales of quantum gravitational effects can be estimated by dimensional analysis applied to the fundamental constants that must appear in any such theory. These are the quantum of action, the Newtonian gravitational constant, the speed of light, and the Boltzmann constant. The results, a time of the order of 10^{-44} sec and a temperature of the order of 10^{19} GeV, delineate conditions so near the classically predicted singularity of infinite density and curvature at the "big bang" that the very concept of a deterministic geometry of space–time breaks down. Violent fluctuations of space–time should generate particles in a manner analogous to that which occurs in the vicinity of a collapsing black hole, as first studied by Stephen Hawking in 1974. Such processes could conceivably be the source of all existing matter, and the possible removal of an infinite-density singularity at time zero would make it scientifically meaningful to ask what the universe was doing before the "big bang."

Not only the existence and structure of matter but even the topology and dimensionality of space–time become properties to be derived rather than postulated in the quantum gravity era. In 1957, John Wheeler suggested that space–time need not necessarily be simply connected at the Planck scale (of the order of 10^{-35} m) but could have a violently fluctuating topology. If so, its description in terms of a smooth continuum would not be appropriate and would have to be replaced by some other mathematical model. The first viable unification of electromagnetism and gravitation, proposed by Theodor Kaluza in 1921 and independently by Oskar Klein in 1926, used a five-dimensional space–time with an additional "compact" spatial dimension subject to constraints that reduced their model to a sterile fusion of Maxwell's and Einstein's field equations. Removing these constraints allows Kaluza–Klein spaces to be used for alternate formulations of gauge field theories. Models with a total of 11 dimensions are currently being actively explored in relation to supersymmetry theories, which seek to provide a unification of bosons and fermions and all interactions among them. Another approach, string field theory, involves replacing the pointlike particles of conventional quantum field theory with fundamental objects with extent in one spatial dimension. It has been demonstrated that topology-changing processes can be explicitly realized in a Kaluza–Klein superstring theory, encouraging the hope that this could be the long-sought basis for a theory of everything. The full implications of such studies for particle physics and for cosmology are not yet clear.

VI. THOUGHTS ON SOURCES OF FUTURE PROGRESS

A. Emerging Observational Technologies

Since the pioneering research of Galileo, advances in telescope capabilities have been the source of more and richer data to constrain cosmological speculation. The current generation of astronomers has seen photographic techniques increasingly augmented by electronic image intensifiers. Improvements in photon detectors, such as charge-coupled devices, are beginning to approach their limits, so that achieving substantial gains will involve increasing the aperture in the next generation of optical and infrared instruments. The twin Keck telescopes, which can be used as an optical interferometer of 85-m baseline, are the largest general-purpose telescopes on earth. A consortium of nations have built Gemini, a matched pair of 8-m telescopes, one in Hawaii and one in Chile. Arizona astronomers have built the Large Binocular Telescope, which will carry two 8.4-m mirrors in a single mounting. The European Southern Observatory has completed the first of four 8.2-m telescopes that will eventually observe as a single Very Large Telescope from the Andes in Chile. Even larger telescopes are planned, and the use of adaptive optics will endow modern telescopes with "seeing" much better than that possible in the past.

As more powerful telescopes looking farther into outer space observe signals from earlier in the history of the universe, higher energy accelerators probing farther into inner space measure particle behavior under conditions simulating earlier times during the "big bang." The installation of superconducting magnets in the tevatron at the Fermi National Accelerator Laboratory at Batavia, Illinois, enabled it to produce protons with an energy of 10^3 GeV in 1984. In 1986, it made available a total energy of 2×10^3 GeV by accommodating countercirculating beams of protons and antiprotons that are made to collide. Since the pioneering effort toward the detection of gravitational waves by Joseph Weber in the late 1960s, astrophysicists have eagerly anticipated the maturing of this technology to open a new window on the universe. Long experience with the Hulse–Taylor binary pulsar is strong evidence that gravitational waves do indeed exist and have the properties predicted by Einstein's general theory of relativity. The development of large dedicated facilities such as the laser interferometer gravitational-wave observatory (LIGO) at last promises to soon move gravitational wave astronomy from a curiosity with isolated applications to a tool for the exploration of any cosmic environment involving strong gravitational fields. As usual in the opening of a new area of science, the unexpected discoveries will surely be the most exciting. Continued experiments at energies of several thousand GeV will lead to important new insights into the structure of matter on a scale of less than 10^{-19} m.

B. Concepts and Mathematical Tools

The search for a viable extension of GUTs to a theory of everything (TOE), which would include quantum gravity as well as the strong and electroweak interactions, is an active area of particle theory research. The energies at which such unification was achieved in nature are presumably even higher than those for GUTs. Thus, no data from accelerators, even in the most optimistic projections of foreseeable future technology, can serve to constrain speculation as well as does information from cosmology. The currently fashionable attempts to derive a TOE are based on "supersymmetry," which is the idea that the fundamental Lagrangian contains equal numbers of Bose and Fermi fields and that they can be transformed into each other by a supersymmetry. This immediately doubles the particle spectrum, associating with each particle thus far observed (or predicted by GUTs) a "superpartner" of opposite quantum statistics. There is at present no experimental evidence for the existence of any of these superpartners, inviting doubt as to the necessity of the supersymmetry assumption. However, supersymmetric theories have the potential to address one otherwise unanswered issue of cosmology: Why is the cosmological constant so small, perhaps precisely zero? Supersymmetric theories are the only known quantum field theories that are sensitive to the vacuum energy level. This appears to imply that a derivation of the cosmological constant from first principles should be possible within a supersymmetric theory, but the problem remains unsolved.

As the number of degrees of freedom being considered in field theories increases, increased computing speed and power become more important on working out the consequences of various proposed models. Some new hardware architectures such as concurrent processing appear to be a means of achieving performance beyond the limits of any existing machines but require the further development of software exploiting their distinctive features to achieve their full potential. Progress in discrete mathematics is of benefit to both computer science and pure mathematics. In the past, fundamental insights have often been derived from mathematical analysis without the benefit of "number crunching," and there is no reason to expect that this process has come to an end. It is, of course, impossible to predict what new closed-form solution of a recalcitrant problem may be discovered tomorrow, or what impact such a discovery may have.

In the last analysis, any attempt to predict the direction of progress in cosmology further than the very near future seems futile. By the nature of the questions that cosmology seeks to answer, the scope of potentially relevant concepts and information is limitless. It is entirely possible that within a decade carefully reasoned thought, outrageous unexpected data, or some combination of the two may overthrow some of today's cherished "knowledge." Aware of the questions still unanswered and of the possibility that some of the right questions have not yet been asked, we can only hope that future discoveries, anticipated or unforeseen, will result in ever greater insights into the structure, history, and destiny of the universe.

SEE ALSO THE FOLLOWING ARTICLES

Celestial Mechanics • Chaos • Cosmic Inflation • Dark Matter in the Universe • Galactic Structure

AND EVOLUTION • PARTICLE PHYSICS, ELEMENTARY • QUASARS • RELATIVITY, GENERAL • RELATIVITY, SPECIAL • STELLAR STRUCTURE AND EVOLUTION • UNIFIED FIELD THEORIES

BIBLIOGRAPHY

Auborg, E., Montmerle, T., Paul, J., and Paul, P., eds. (2000). *Texas Symp. on Relativistic Astrophys. and Cosmology, Nuclear Physics B (Proc. Suppl.)* 80.

Bothun, G. (1998). "Modern Cosmological Observations and Problems," Taylor & Francis, London.

Gribbon, J., and Rees, M. (1989). "Cosmic Coincidences," Bantam Books, New York.

Kolb, E., Turner, M., Lindley, D., Olive, K., and Seckel, D., eds. (1986). "Inner Space/Outer Space," Univ. of Chicago Press, Chicago.

Misner, C., Thorne, K., and Wheeler, J. (1973). "Gravitation," Freeman, New York.

Peebles, P. J. E. (1980). "The Large-Scale Structure of the Universe," Princeton Univ. Press, Princeton, NJ.

Isham, C., Penrose, R., and Sciama, D., eds. (1981). "Quantum Gravity II," Oxford Univ. Press, London.

Rowan-Robinson, M. (1985). "The Cosmological Distance Ladder," Freeman, New York.

Gamma-Ray Astronomy

J. Gregory Stacy
Louisiana State University and Southern University

W. Thomas Vestrand
Los Alamos National Laboratory

GLOSSARY

Accretion disk A flattened, circulating disk of material drawn in and heated to high temperatures under the influence of the intense gravitational field associated with a black hole or other compact object (such as a neutron star or white dwarf).

Active galactic nuclei (AGN) A collective term for active galaxies whose emission is observed to come predominantly from the central nuclear region of the galaxy. Of these, blazars form a subclass that is observed to emit gamma radiation. It is likely that the viewing angle toward these latter objects is directed along jets of relativistic material ejected from the nucleus of the galaxy by supermassive black holes.

Bremsstrahlung The "braking" radiation given off by free electrons that are deflected (i.e., accelerated) in the electric fields of charged particles and the nuclei of atoms.

Cherenkov light Radiation produced by a charged particle whose velocity is greater than the velocity of light in the medium through which it travels. Cherenkov light is strongly directed along the line of travel of the particle.

Compton scattering The dominant process by which a medium-energy gamma ray interacts with matter by scattering and transferring a part of its energy to an electron. "Inverse" Compton scattering refers to the same process, but where a lower-energy photon is scattered to higher energy after interaction with a relativistic electron.

Cosmic rays High-energy charged particles, such as electrons, protons, alpha-particles (helium nuclei), and heavier nuclei that propagate through interstellar space.

Diffuse emission Radiation that is extended in angular size on the sky, such as the gamma-ray emission arising from the decay of radioactive nuclei dispersed throughout the interstellar medium. A diffuse source is distinguished from a pointlike or point source of emission that is not resolvable into further individual components given the limited angular resolution of a telescope.

Electromagnetic cascades A phenomenon that occurs in the upper atmosphere of the Earth when a very high-energy gamma ray interacts by the pair-production process, followed by further interactions resulting in extensive air showers (or EAS) of particles and photons. The relativistic cascade particles emit optical Cherenkov light that is observable from the ground. Electromagnetic cascades are distinguished from the nucleonic (or hadronic) cascades produced by high-energy cosmic-ray particles in the upper atmosphere.

Electron volt (eV, and keV, MeV, GeV, TeV, PeV) The electron volt (eV) is a fundamental unit of energy commonly used in high-energy astrophysics. It is defined to be the energy acquired by an electron when accelerated through a potential difference of one volt. One eV is equal to 1.602×10^{-19} J or 1.602×10^{-12} ergs.

Gamma rays The highest-energy form of electromagnetic radiation, above the X-ray portion of the spectrum. Gamma rays have energies measured in millions of electron volts and higher.

Pair production and annihilation The process by which the most energetic gamma rays (of MeV energies and above) interact with matter, producing an electron–positron pair (the positron, with positive charge, is the antiparticle to the electron). The inverse process is pair annihilation, in which an electron and positron mutually annihilate and produce a pair of high-energy gamma rays.

Parsec (pc, and kpc, Mpc, Gpc) A unit of distance commonly used in astronomy (an abbreviation for "parallax-arcsecond"). One parsec is equal to 3.086×10^{16} m, or 3.26 light-years.

Point source A source of emission that is not further resolvable into individual components given the limited angular resolution of a telescope. In gamma-ray astronomy angular resolution is relatively poor compared to other branches of astronomy. Thus, in some instances a gamma-ray point source may in fact consist of a number of individual sources whose summed emission is measured by the gamma-ray telescope.

Supernova An endpoint of stellar evolution for the most massive stars, an explosion triggered by the gravitational collapse of the stellar core following the exhaustion of fuel for nuclear burning. The collapsed stellar core, depending on its final mass, can become either a black hole or a neutron star (and some of the latter may be observable as pulsars).

Synchrotron radiation The emission produced by charged particles as they spiral (i.e., accelerate) around magnetic fields.

THE GAMMA-RAY regime constitutes one of the last regions of the electromagnetic spectrum to be opened to detailed astrophysical investigation. Only within the past decade has the field of gamma-ray astronomy become firmly established as a productive and dynamic discipline of modern observational astrophysics. This has been largely due to the successful operation during the 1990s of the Compton Gamma Ray Observatory, whose telescopes carried out the first comprehensive surveys of the sky at gamma-ray energies (as shown in Fig. 1). Gamma rays form the highest-energy portion of the electromagnetic (E-M) spectrum with individual photon energies extending from millions of electron volts (MeV) to values in excess of 10^{16} eV (optical photons in contrast carry energies of only a few electron volts). Observations of these highest-energy photons provide the means of investigating the largest transfers of energy occurring in the Universe and offer the key to understanding a host of challenging cosmic phenomena occurring in a wide variety of astrophysical settings. The environments in and around gamma-ray sources are among the most extreme to be found in the Universe, permitting the testing of models and hypotheses regarding high-energy phenomena under conditions impossible to achieve on the Earth. Further, the Universe is essentially transparent to the propagation of gamma radiation, and since, like all electromagnetic radiation, gamma rays are electrically neutral they are not deviated from their trajectories under the influence of magnetic fields. Gamma rays arriving at the Earth therefore serve as direct messengers from high-energy celestial sources within our own Milky Way galaxy and beyond, extending to the most distant reaches and earliest epochs of the cosmos. Gamma-ray astronomy quite literally provides a new window into space that extends our view out to the edge of the observable Universe.

I. INTRODUCTION AND HISTORICAL OVERVIEW

A. Fundamental Concepts and Terminology

To place the discipline of gamma-ray astronomy in context, we review some basic concepts and nomenclature. The electromagnetic spectrum encompasses the entire range of radiation from the radio through gamma rays, and includes the subregions of radio, infrared (IR), optical (or visible), ultraviolet (UV), X-rays, and gamma rays, in order of increasing energy or frequency of radiation. The fundamental relation between photon energy, frequency, and wavelength is the well-known expression due to Planck,

$$E = h\nu = hc/\lambda,$$

where E is the photon energy, ν the frequency, λ the wavelength of the radiation, c the speed of light ($=3 \times 10^8$ m/s), and h is Planck's constant ($=6.626 \times 10^{-34}$ J s $= 4.135 \times 10^{-15}$ eV s). The speed of an electromagnetic wave in vacuum is the speed of light, thus the product of

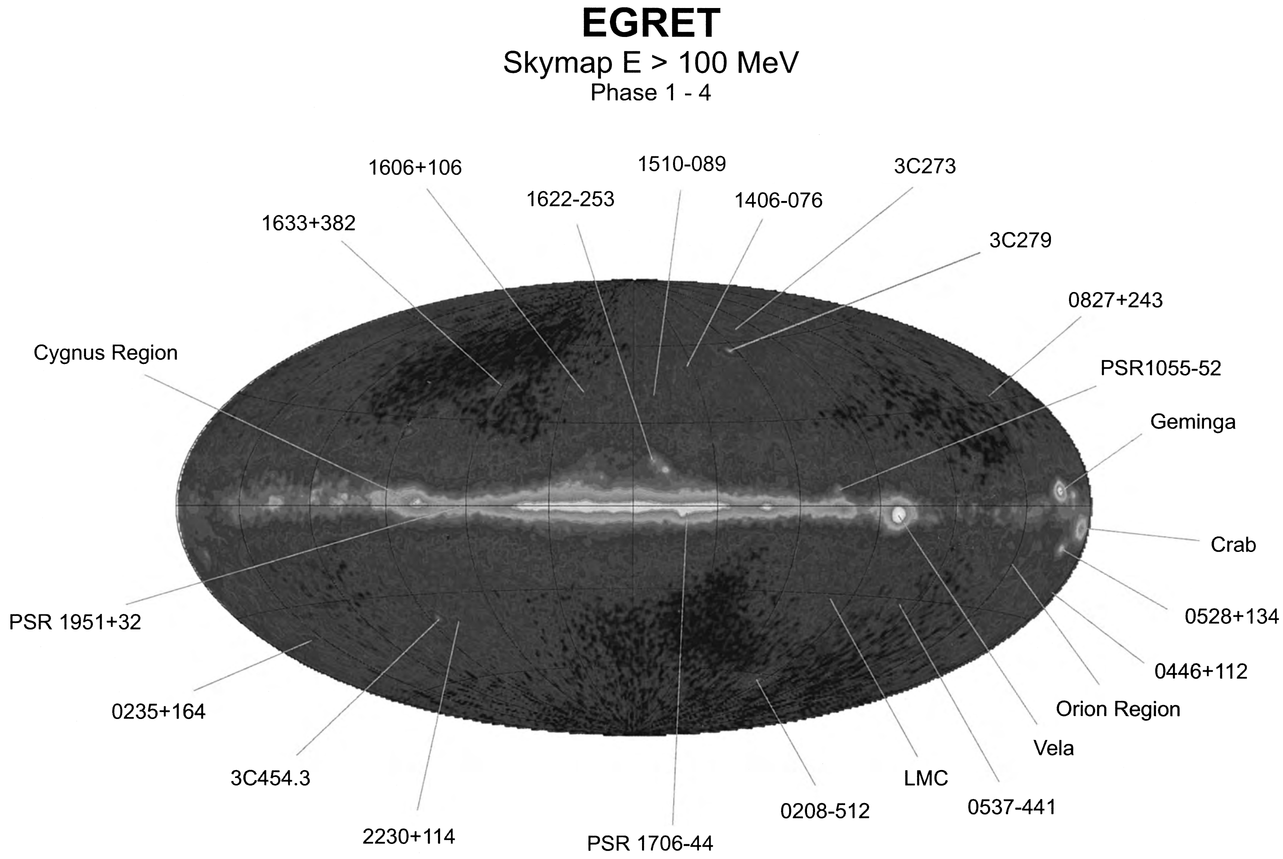


FIGURE 1 A map of the gamma-ray sky obtained with the high-energy EGRET telescope aboard the Compton Gamma Ray Observatory. The map is an all-sky Aitoff equal-area projection in Galactic coordinates with the direction of the Galactic Center at the middle of the map. The bright horizontal band is predominantly diffuse gamma-ray emission from the disk of the Milky Way. Prominent gamma-ray sources are indicated and labeled around the periphery of the figure. (Courtesy NASA, the Max Planck Institute, and the CGRO EGRET Instrument Team.)

frequency and wavelength equals c (or, $\nu\lambda = c$). At energies below or near the optical, sub-bands of the E-M spectrum are traditionally referred to either in terms of their wave properties (in the radio, for example, there are the meter, centimeter, millimeter, submillimeter, and microwave bands), or in terms of their relation to the central optical band (the *near* versus *far* infrared bands, for example, of shorter and longer wavelength, respectively, and the *extreme* ultraviolet beyond the optical and UV). Once in the X-ray band, however, one enters the realm of high-energy astrophysics and tends to abandon the wavelength/frequency nomenclature in favor of particle and energy terminology that is better suited to the description of photon interactions at these energies. Thus X-rays are generally referred to in broad terms as either *soft* or *hard* in energy content (with a loosely defined boundary in the keV energy range). Similarly, gamma rays themselves have been subdivided into soft and hard regimes, or otherwise characterized as of low, medium, or high gamma-ray energy. Given the very broad range of gamma-ray energies that are now observable (spanning more than 10 orders of magnitude in energy) one now usually refers to gamma rays by their energy designation alone, as either MeV, GeV, TeV, or even PeV, gamma rays, where standard order-of-magnitude prefixes apply (for the above, 10^6, 10^9, 10^{12}, and 10^{15} eV, respectively).

Much of the radiation with which we are familiar in everyday life is of *thermal* origin, arising by definition from matter in thermal equilibrium. In an ideal atomic gas in thermal equilibrium, for example, the upward versus downward transitions of bound electrons between energy levels in individual atoms are in close balance due to the exchange of energy between particles via collisions and the absorption and emission of radiation. The velocities of particles in an ideal thermal gas follow the well-known Maxwellian distribution, and the collective continuous spectrum of the radiating particles is described by the familiar Planck black-body radiation curve with its characteristic temperature-dependent profile and maximum.

Gamma rays, in contrast to other forms of electromagnetic radiation, are most often of *nonthermal* origin, arising usually from interactions involving high-energy, relativistic particles in an ionized plasma whose constituents are not in thermal equilibrium with their surroundings. A *relativistic* particle is one whose kinetic energy is comparable to or exceeds its rest-mass energy given by Einstein's famous relation, $E = mc^2$. Thus,

$$\begin{aligned}\text{total particle energy} &= (\text{rest-mass} + \text{kinetic})\text{energy}\\ &= \gamma mc^2,\end{aligned}$$

where $\gamma = (1 - v^2/c^2)^{-1/2}$ is the relativistic Lorentz factor for a particle traveling at velocity v. Interactions involving relativistic particles are properly treated within the framework of relativity. For sufficiently low velocities the familiar *nonrelativistic* case ($v/c \ll 1, \gamma \to 1$) is re-obtained in which the relations of classical Newtonian physics apply. In the standard relativistic regime the particle velocity approaches c, $v/c \to 1$ and $\gamma \geq 1$ (for example, a particle traveling at 90% of the speed of light will have $v/c = 0.9$ and $\gamma \cong 2.3$), whereas in the most extreme *ultra-relativistic* case ($v/c \sim 1, \gamma \gg 1$) the total energy of the particle is dominated by its kinetic energy. Photons of sufficiently high energy in cosmic sources can be diminished in intensity via the photon–photon pair-production process ($\gamma\gamma \to e^+e^-$) which is most likely to occur just above the reaction threshold when the product of the two photon energies is equal to the product of the electron–positron rest–mass energies, $E_{\gamma 1}E_{\gamma 2} \sim 2(m_ec^2)^2 \sim 0.52$ (MeV)^2. In astrophysical sources, then, the energy density of gamma rays may be sufficiently high to prevent their escape due to their greater likelihood of producing pairs. Such media are said to have a high pair-production opacity.

Given the relativistic nature of the major gamma-ray production mechanisms we adopt in this review the energy corresponding to the rest–mass energy of the electron, $m_ec^2 \sim 511$ keV ~ 0.511 MeV, as a natural reference energy defining the lower boundary of the gamma-ray regime. (We note that, historically, among physicists of a certain age, gamma rays were defined simply as the radiation resulting from nuclear transitions, independent of the energies involved, in recognition of the predominant nuclear origins of gamma radiation. We adopt here the more current view of a specific energy regime.)

Gamma-ray astronomy is closely linked to several other branches of modern observational astrophysics. A number of these subdisciplines of astronomy are described elsewhere in these volumes. Gamma-rays, by virtue of their high energy, also play a particularly key role in *broadband multiwavelength astronomy*, by which is meant the study of a celestial object or phenomenon over as broad a portion of the electromagnetic spectrum as possible.

B. The Production of Cosmic Gamma Rays

A wide variety of production mechanisms give rise to gamma radiation, resulting in either continuum or spectral-line emission. Gamma rays most often result from high-energy collisions between nuclei, particles, and other photons, or from the interactions of charged particles with magnetic fields. Line emission can arise from the deexcitation of nuclear states, from radioactive decay, or from matter–antimatter annihilation. The primary production mechanisms of interest to gamma-ray astronomy are summarized in the following.

1. Particle–Nucleon Interactions

High-energy nuclear collisions frequently yield charged and neutral mesons as unstable reaction products. Charged pions (π^+ and π^-) decay into positive and negative muons that decay in turn into relativistic electrons and positrons. Neutral pions (π^0) decay almost immediately ($t_{1/2} \sim 10^{-16}$ s) into two gamma rays of total energy equal to approximately 68 MeV in the rest frame of the decaying meson. The resulting gamma-ray spectrum depends on the distribution of particle energies of the original emitted pions, and is generally a broad continuum centered and peaked at $\mathrm{E}_\gamma \sim m_\pi c^2/2 \sim 68$ MeV. Nucleon–nucleon interactions play an important role in the production of diffuse high-energy gamma rays in the disk of the Galaxy following the collision of high-energy cosmic rays (primarily protons) with the nuclei of the atoms and molecules of the interstellar gas.

Collisions between high-energy particles and ambient matter can also result in the copious production of numerous other secondary particles, including neutrons. High-energy neutrons are capable of exciting nuclei in secondary collisions, leading to gamma ray line emission (see following). Further, neutrons can be slowed in the interacting medium to thermal energies whereupon they can be quickly captured by nuclei, again giving rise to gamma-ray lines. Neutron processes play an important role in solar flares, and in the production of gamma rays on planetary surfaces after cosmic-ray bombardment.

2. Nuclear Gamma-Ray Lines

Nuclear deexcitation following energetic collisions or radioactive decay gives rise to spectral line radiation whose specific energies are characteristic of the emitting nuclides. Nuclear gamma-ray line radiation extends up to ~9 MeV in energy for the most commonly abundant elements and likely interaction processes. The intensities and ratios of observed gamma-ray lines can provide detailed information on elemental composition and relative abundances (for example, in solar flares).

3. Relativistic Electron Interactions

Relativistic electrons can interact with charged particles via the bremsstrahlung process, with photons through Compton scattering, and with magnetic fields by emitting synchrotron radiation. These processes dominate many of the energy regimes in the field of high-energy astrophysics. They give rise to continuum gamma radiation, whose spectral characteristics can be used to deduce the physical conditions at the astrophysical source.

(a) Bremsstrahlung. Bremsstrahlung (or "braking radiation") is the radiation given off by free electrons that are deflected (i.e., accelerated) in the electric fields of charged particles and the nuclei of atoms. Thermal bremsstrahlung is the emission given off by an ionized gas of plasma in thermal equilibrium at a particular temperature, where the distribution of electron velocities follows the well-known Maxwellian distribution. Relativistic electrons, whose distribution of energies often follows a power-law shape in astrophysical settings, give rise to relativistic bremsstrahlung radiation that is also of power-law shape with the same spectral index as the emitting electrons.

(b) Compton scattering. Another major source of cosmic gamma radiation is the Compton scattering of lower-energy photons to gamma-ray energies by relativistic electrons. This process is often referred to as "inverse" Compton scattering since it is the low-energy photon that gains energy from the high-energy electron, in contrast to the more standard view of the Compton mechanism. In the ultrarelativistic case, it can be shown that the energy of the photons scattered by high-energy electrons is $E \sim \gamma^2 \mathrm{E_e}$ (in the Thomson limit when the energy of the photon in the center-of-momentum frame of reference is much less than $m_e c^2$), where γ is the relativistic Lorentz factor of the electrons. For relativistic electrons with $\gamma \sim$100–1000, as observed in many astrophysical sources, this implies that low-energy photons can be up-scattered to very high energies indeed, well into the gamma-ray regime.

(c) Synchrotron radiation. Synchrotron emission results when an electron gyrates around a magnetic field. For electrons of sufficiently high energy, or for magnetic fields of sufficiently high strength, high-energy photon emission readily results. Again, for a power-law distribution of electron energies, a power-law synchrotron emission spectrum follows. The relation between the observed intensity (I) of the synchrotron radiation as a function of frequency (ν), the magnetic field strength (B), and the power-law index (p) of the electron particle distribution is given by $I(\nu) \propto B^{(p+1)/2} \nu^{-(p-1)/2}$.

Collectively, these emission processes represent the primary energy-loss (or "cooling") mechanisms for relativistic electrons in astrophysical sources, the other major process being "ionization" losses via particle collisions. Observations of high-energy emission from celestial sources that can be decomposed into synchrotron, bremsstrahlung, and Compton components from characteristic spectral signatures therefore provides a wealth of information on the physical conditions within the emitting regions (such as particle densities, and the strengths of radiation and magnetic fields).

4. Electron–Positron Annihilation

A free electron and its antiparticle, the positron, may interact to produce annihilation radiation yielding two gamma rays ($e^+e^- \rightarrow \gamma\gamma$). The total energy of the two photons in the center-of-momentum frame of reference is equal to the combined rest–mass energy of the electron–positron pair, $2m_e c^2 \sim 1.022$ MeV. (Three-photon annihilation can also occur for free electrons and positrons, but is much less likely.) If an electron and positron are essentially at rest upon annihilation then two gamma rays of equal energy (0.511 MeV) are produced. In the more general astrophysical case, however, one or both particles are at relativistic velocities, and a more complicated emergent gamma-ray spectrum usually results.

An electron and positron of sufficiently low energy (typically thermal, $\leq$5 eV) may combine to briefly form a hydrogen-like state of matter referred to as *positronium*. Positronium almost immediately self-annihilates yielding either a two- or three-photon decay into gamma rays ($\tau_{2\gamma} \sim 10^{-10}$ s, $\tau_{3\gamma} \sim 10^{-7}$ s).

C. A Brief History of Gamma-Ray Astronomy

It was quickly recognized at the dawn of the nuclear age that the potential existed for the detection of celestial gamma rays from high-energy sources in the cosmos. In the early 1950s discussions were already underway on the likelihood of gamma-ray production via cosmic-ray interactions in interstellar space (cosmic rays are high-energy, relativistic particles and nuclei of celestial origin). In now-classic papers Burbridge, Burbridge, Fowler, and Hoyle, in 1957, laid out the principles governing the synthesis of heavy elements in stellar nuclear burning and during explosive nucleosynthesis in supernovae, and Morrison in 1958 similarly described many of the fundamental mechanisms and sources for the production of cosmic gamma rays.

Through the 1960s a number of balloon and early spacecraft observations (e.g., the Ranger spacecraft missions to the Moon) provided intriguing but inconclusive evidence for the existence of cosmic gamma-rays. The first positive detection of celestial gamma radiation was made with an instrument aboard the third Orbiting Solar Observatory (OSO-3), whose investigators reported in 1972 the detection of gamma rays from the Galactic disk, with a peak intensity observed toward the Galactic Center. In the early 1970s several high-altitude balloon experiments were also beginning to report positive results, including detection of the Crab pulsar and of diffuse gamma radiation from the disk and central region of the Galaxy. Nuclear gamma-ray lines from the Sun were detected from large solar flares on August 4 and 7, 1972, with a spectrometer aboard OSO-7. The first reported detection of likely positron annihilation radiation (at 0.511 MeV) from the direction of the Galactic Center was based on balloon measurements from 1971, and later confirmed by other investigators with a detector of higher spectral resolution in 1977. As described elsewhere in this review, gamma-ray spectrometers carried to the Moon by both U.S. and Russian spacecraft in the late 1960s and 1970s provided extensive orbital and in situ measurements of the elemental composition of the lunar surface. Similar experiments in the 1970s were carried to Mars aboard the U.S. Viking landers, and to Venus on the Russian Venera landers. Most intriguing was the announcement in 1973 of the discovery of the mysterious cosmic gamma-ray bursts with instruments aboard the Vela series of nuclear surveillance satellites (launched originally to verify compliance with the 1963 Nuclear Test Ban Treaty).

A major advance in the field of gamma-ray astronomy came with the launch in 1972 of the second Small Astronomy Satellite (SAS-2). Over its 7-month lifetime SAS-2 carried out a survey of high-energy gamma-ray emission (>50 MeV) from the Galactic plane, and provided a first measure of the extragalactic diffuse gamma-ray background. This pioneering mission was followed with the launch of the COS-B gamma-ray satellite by the European Space Agency in 1975. Over its 7-year lifetime COS-B greatly extended our knowledge of the gamma-ray sky, providing detailed maps of the diffuse gamma radiation arising from the Galactic plane, as well as cataloging a number of point sources of high-energy gamma rays, including the first detected extragalactic source, the quasar 3C 273.

Two gamma-ray instruments were carried into space in the late 1970s as part of NASA's High Energy Astronomical Observatory (HEAO) series of satellites. HEAO-1 conducted a survey of the sky from 10 keV to 10 MeV in energy, and identified a number of active galaxies and characterized their spectra in the 10- to 100-keV range. The HEAO-3 experiment discovered the first nonsolar nuclear gamma-ray line of celestial origin, the 1.809-MeV spectral line emitted by radioactive ^{26}Al that is produced in massive stars and is a tracer of recent star formation in the Galaxy. In 1980, NASA launched the Solar Maximum Mission (SMM) satellite which carried a gamma-ray spectrometer among its suite of instruments. Over its extended 10-year lifetime SMM provided a wealth of new information on gamma-ray processes occurring during flares on the Sun, and also made fundamental contributions to nonsolar gamma-ray astronomy. These included the discovery of greater-than-MeV emission from gamma-ray bursts, confirmation of the diffuse ^{26}Al emission detected by HEAO-3, and further observation of the positron annihilation radiation coming from the central region of the Galaxy. Particularly notable was the SMM detection of radioactive ^{56}Co line emission from the Type II supernova SN 1987A in the Small Magellanic Cloud, providing a long-awaited first direct measure of explosive nucleosynthesis in supernovae. The French coded-aperture SIGMA telescope was carried into space aboard the Soviet GRANAT satellite in 1989. Among other observations, this lower-energy (up to $\sim$1.3 MeV) instrument identified a number of black-hole candidate sources in the central region of the Galaxy based on the observed spectral and temporal behavior.

The realm of ground-based gamma-ray astronomy, where different observational challenges present themselves, begins at photon energies of $\sim$50 GeV (sometimes termed the very-high-energy, or VHE, gamma-ray regime). Gamma-rays approaching TeV energies become increasingly rare in number, and cannot be well sampled by existing spacecraft-borne instrumentation. Further, the absorbing medium of the Earth's atmosphere precludes their direct observation from the ground. Their presence, however, can be inferred indirectly from the electromagnetic cascades of electrons and positrons that they produce upon interaction in the upper atmosphere. (These cascades are also referred to as extensive air showers, or EAS.) The relativistic cascade particles emit Cherenkov light over a wide area that can be detected with optical telescopes on the ground. A particular difficulty, however, is distinguishing such photon-induced events from the nucleonic (or hadronic) cascades produced by high-energy cosmic-ray particles in the atmosphere, which exhibit very similar observable effects. Atmospheric Cherenkov imaging telescopes were originally proposed in the 1970s, but only in the 1990s did the techniques and instrumentation become sufficiently developed to achieve breakthrough detections of several high-energy gamma-ray sources. In the late 1980s the Whipple Observatory first detected TeV emission from the Crab pulsar and nebula, and this has been followed in recent years with detections by several groups of TeV emission from a small number of both Galactic and extragalactic sources (outlined in later sections).

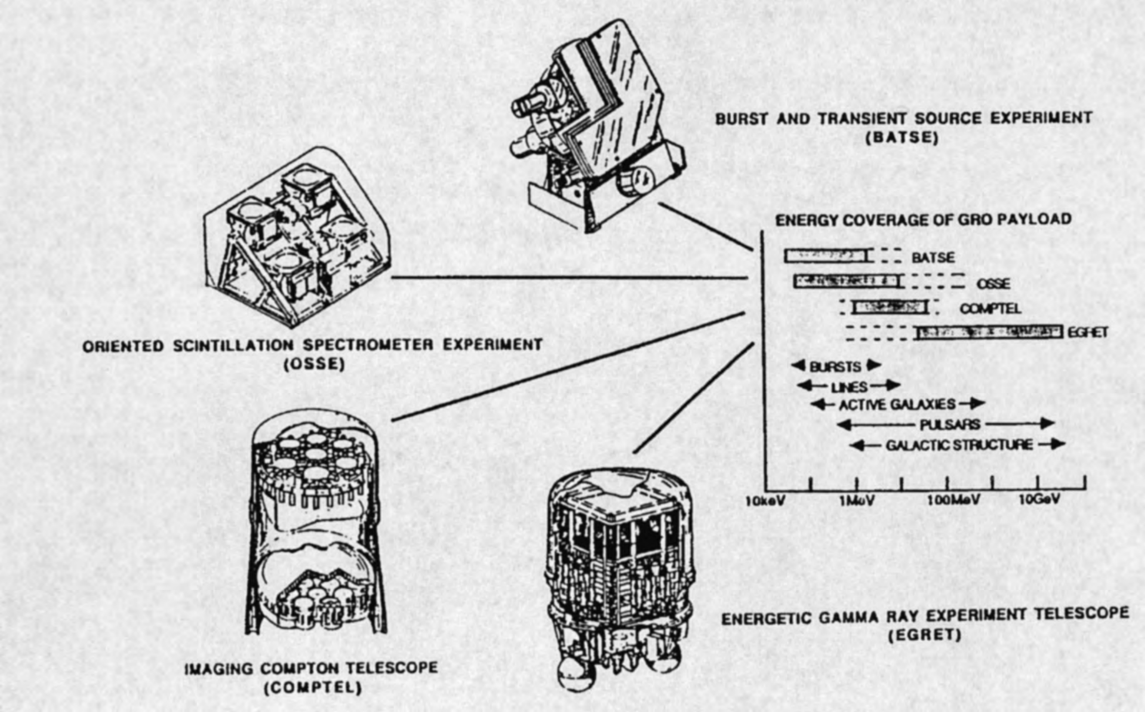

BURST AND TRANSIENT SOURCE EXPERIMENT
(BATSE)
ENERGY COVERAGE OF GRO PAYLOAD
BATSE
OSSE
COMPTEL
EGRET
BURSTS
LINES
ACTIVE GALAXIES
PULSARS
GALACTIC STRUCTURE
10keV
1MeV
100MeV
10GeV
ORIENTED SCINTILLATION SPECTROMETER EXPERIMENT
(OSSE)
IMAGING COMPTON TELESCOPE
(COMPTEL)
ENERGETIC GAMMA RAY EXPERIMENT TELESCOPE
(EGRET)

FIGURE 2 Schematic of the four gamma-ray telescopes aboard the Compton Gamma Ray Observatory and their corresponding overlapping energy ranges. Also indicated are classes of prominent gamma-ray sources by energy band. (Courtesy NASA.)

D. The Compton Gamma-Ray Observatory

The great potential of gamma-ray astronomy as a viable, productive branch of observational astrophysics was not fully realized until the launch of the Compton Gamma Ray Observatory (or CGRO) in April 1991. The Compton Observatory was the second of NASA's four planned Great Observatory missions that were designed to study the sky from space in different key regions of the electromagnetic spectrum. The first of these was the Hubble Space Telescope (launched in 1990), the second the CGRO (launched in 1991), and the third the Chandra X-ray Observatory (launched in 1999). At the time of writing, the Space Infrared Telescope Facility (SIRTF) is awaiting an anticipated launch in 2002. The CGRO was named in honor of the American physicist Arthur Holly Compton (1892–1962), whose pioneering investigations into the scattering of X-rays and gamma rays by charged particles earned him the Nobel prize for physics in 1927. After more than 9 years of successful operation, the CGRO mission was terminated in June 2000, when the spacecraft was deorbited for safety reasons.

The CGRO carried four separate, complementary gamma-ray telescopes with overlapping energy ranges, each designed for specific scientific objectives and developed by an international collaboration of scientists. The combined energy coverage of the CGRO detectors extended over 6 orders of magnitude from $\sim$30 keV to 30 GeV (see Fig. 2). The key characteristics of each of the four CGRO instruments are summarized in the following.

The *Burst and Transient Source Experiment* (*BATSE*) was an all-sky monitor consisting of eight separate detectors mounted on the corners of the main platform of the CGRO spacecraft. Its primary objective was to detect and measure rapid brightness variations in gamma-ray bursts and solar flares down to microsecond time scales over the energy range from 30 keV to 1.9 MeV. BATSE continually monitored the sky for transient phenomena, searching for variable emission from both known and new sources.

The *Oriented Scintillation Spectroscopy Experiment* (*OSSE*) was designed to carry out pointed spectral observations of gamma-ray sources in the range from 0.05 to 10 MeV, with capability above 10 MeV for solar gamma-ray and neutron observations. The four OSSE detectors were collimated scintillators (with a $4^\circ \times 11^\circ$ field of view) that were movable over a single axis, allowing a rapid response to targets of opportunity such as solar flares,

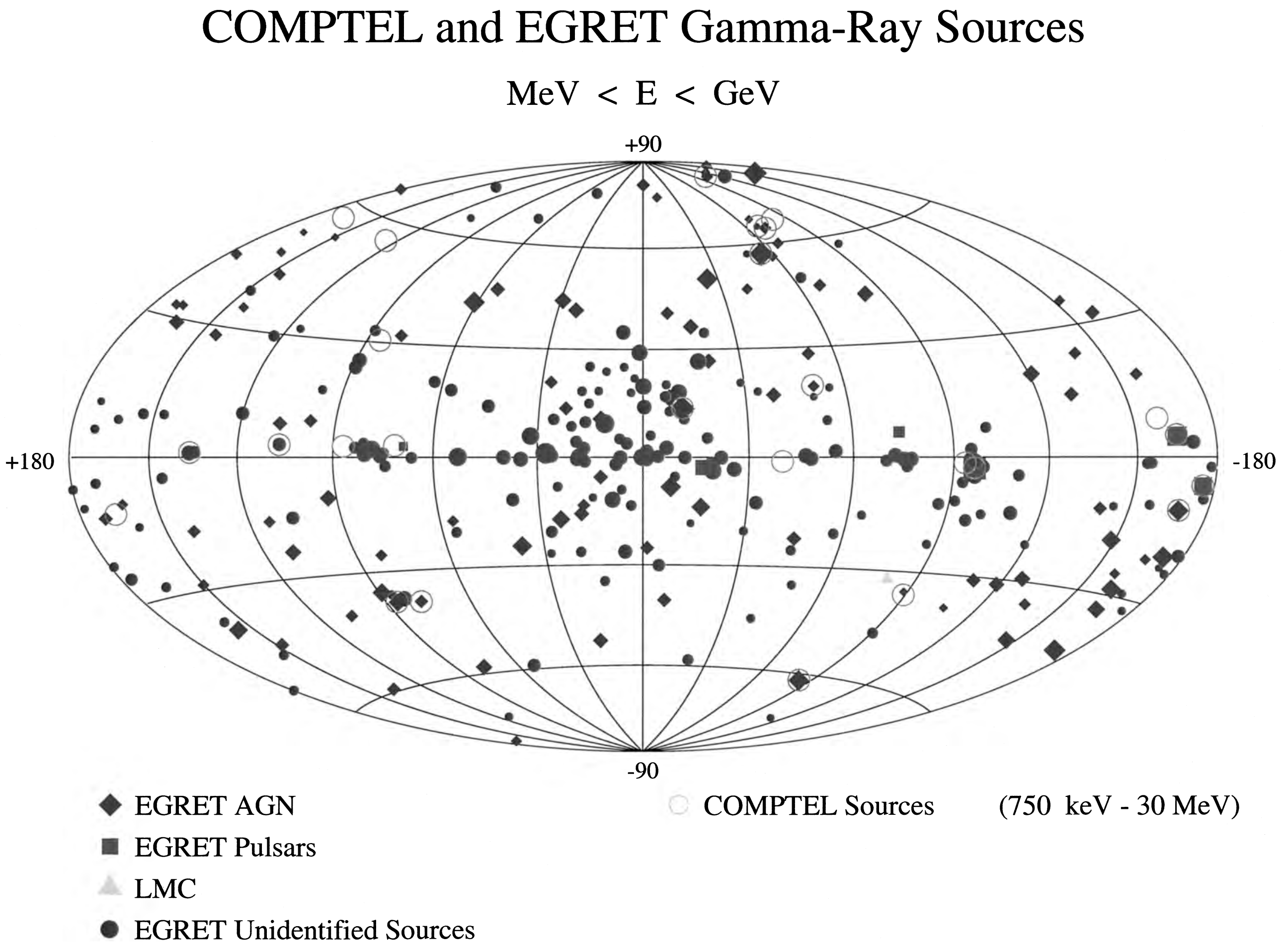

COMPTEL and EGRET Gamma-Ray Sources
MeV < E < GeV
+90
+180
-180
-90
◆ EGRET AGN
■ EGRET Pulsars
▲ LMC
● EGRET Unidentified Sources
○ COMPTEL Sources (750 keV - 30 MeV)

FIGURE 3 An all-sky map of gamma-ray sources detected with the COMPTEL and EGRET telescopes aboard the Compton Gamma Ray Observatory. Classes of sources are indicated by symbol, with increasing symbol size representing higher source intensity. The map is an Aitoff equal-area projection in Galactic coordinates (as in Fig. 1). (Courtesy NASA, the Max Planck Institute, and the CGRO COMPTEL and EGRET Instrument Teams).

transient X-ray sources, and other explosive astrophysical phenomena.

The *Imaging Compton Telescope* (*COMPTEL*) detected gamma-rays by means of a double-scatter technique whereby an incident gamma photon Compton scattered once in an upper detector module, and then was totally absorbed in a lower detector module. The COMPTEL instrument was sensitive over the energy range from approximately 0.75 to 30 MeV, and was also capable of detecting neutrons from solar flares. With its large field of view (∼1 steradian) COMPTEL carried out the first survey of the gamma-ray sky at MeV energies.

The *Energetic Gamma Ray Experiment Telescope* (*EGRET*) covered the broadest energy range of the CGRO instruments, from ∼20 MeV to 30 GeV. These high-energy photons interact primarily via the pair-production process, and the EGRET spark chamber was designed to detect the electron–positron pairs produced by high-energy gamma rays. EGRET also had a relatively wide field of view (∼0.6 sr), good angular resolution, and very low background.

The coaligned COMPTEL and EGRET instruments operated as wide-field imaging telescopes and together carried out a comprehensive survey of the gamma-ray sky from MeV to GeV energies (see Figs. 1 and 3). Taken together the four CGRO telescopes represented a major improvement in sensitivity, energy coverage, and spectral and angular resolution, compared to previous generations of gamma-ray instruments. It can be said without exaggeration that observations carried out with the CGRO have completely revolutionized our view of the high-energy Universe. The bulk of the scientific results discussed in the sections to follow are based on observations obtained with the four CGRO telescopes.

II. SOURCES OF COSMIC GAMMA RAYS

Our view of the gamma-ray sky has changed dramatically in recent years, and has been particularly influenced by the results obtained with the instruments aboard the Compton Gamma Ray Observatory. The energetic and variable cosmos revealed by gamma-ray telescopes stands in marked contrast to the quiescent night sky viewed in visible light on a placid summer evening. The Universe in the light of gamma rays is a dynamic, diverse, and constantly changing place.

A. The Sun

The Sun is a powerful site for the acceleration of energetic particles. The source of energy for particle acceleration is believed to be the tangled magnetic field in the solar atmosphere. However, our understanding of both the properties of the accelerated particles and the nature of their acceleration during the explosive release of energy in a solar flare is still emerging. Gamma-ray measurements have proved to be an essential tool for studying particle acceleration during flares.

The rich energy spectrum of gamma-ray emission from solar flares is quite complex and shows the signatures of many radiation processes. Below ∼1 MeV the observed emission is dominated by a strong line at 0.511 MeV from positron annihilation and a smooth continuum of bremsstrahlung radiation from mildly relativistic electrons. In the energy band from 1 MeV to 10 MeV the emission results predominantly from the deexcitation of nuclear levels following the bombardment of nuclei in the solar atmosphere by energetic particles. This nuclear deexcitation emission is composed of four components: (1) promptly emitted narrow lines from the excitation of a heavy atmospheric nucleus by an energetic proton or alpha particle, (2) broad lines from the excitation of an accelerated heavy ion by collision with an atmospheric hydrogen or helium nucleus, (3) delayed line emission such as the strong line at 2.22 MeV from the capture of secondary neutrons by atmospheric hydrogen to form deuterium, and (4) a quasi-continuum produced by the blending of lines from high-level transitions excited in both the accelerated and target nuclei. Above 10 MeV the emission is dominated by two mechanisms: bremsstrahlung from both ultra-relativistic primary electrons and secondary electrons/positrons from meson decay, and gamma rays from the direct decay of neutral pions. The complexity of the gamma-ray spectra of flares provides many diagnostics for probing the properties of flare-accelerated electrons and ions (see Fig. 4).

Measurements of the bremsstrahlung continuum during solar flares indicate that relativistic electrons are a common product of energy release in flares. The gamma-ray bremsstrahlung generated by relativistic electrons indicates that the yield in relativistic electrons scales roughly with the total energy released in thermal X-rays by the flare. Further, increasingly sensitive searches for gamma-ray bremsstrahlung over the last 2 decades have found no evidence of a flare-size threshold for relativistic electron acceleration. The gamma-ray evidence therefore suggests that relativistic electron acceleration is a property of all flares. The gamma-ray observations also show that the relative amount of high-energy bremsstrahlung increases as the position of the flare approaches the solar limb. Since high-energy bremsstrahlung is directed more strongly along the electron's velocity vector than bremsstrahlung at lower energies, the limb brightening can be explained by a distribution of emitting electrons that increases in directions away from the surface normal at the flare site. The nature of this electron distribution is regulated by the complex magnetic field structure in flaring regions. Future techniques that can measure the angular distribution of gamma-ray bremsstrahlung will allow us to explore the nature of relativistic electron transport in flaring regions.

Nuclear deexcitation emission during flares indicates that energetic ion acceleration is also a common property of solar flares. Gamma rays from nuclear deexcitations were first detected from two giant flares that occurred in August 1972. The enormous size of those flares and the fact that they were the only ones detected during that solar cycle led to an initial suspicion that ion acceleration might only occur when the flare energy surpasses a relatively high threshold. Sensitive detectors aboard the Solar Maximum Mission (SMM) satellite and the Compton Observatory, however, showed that nuclear line emission is present even in relatively small flares. While the relative importance of accelerated ions and electrons is observed to vary by approximately an order of magnitude, existing measurements are consistent with the hypothesis that both components are accelerated in all solar flares.

The temporal structure of variations in gamma-ray flux during flares can be quite rich. Gamma-ray flares can range from a single spiked pulse of 10-s duration to a complex series of pulses with total duration of more than 1000 s. Typically flares are composed of two or more pulses. An interesting property of the pulse structure is that the time of peak intensity is often energy dependent. When this energy dependence is present, the peak at higher energies tends to lag the peak intensity at lower energies by as much as 45 s. At one time, these delays were interpreted as reflecting the timescale needed for particle acceleration during flares. However, we now know that there are many flares where the peaks at X-ray through gamma-ray energies show time coincidence to better than 2 s and that,

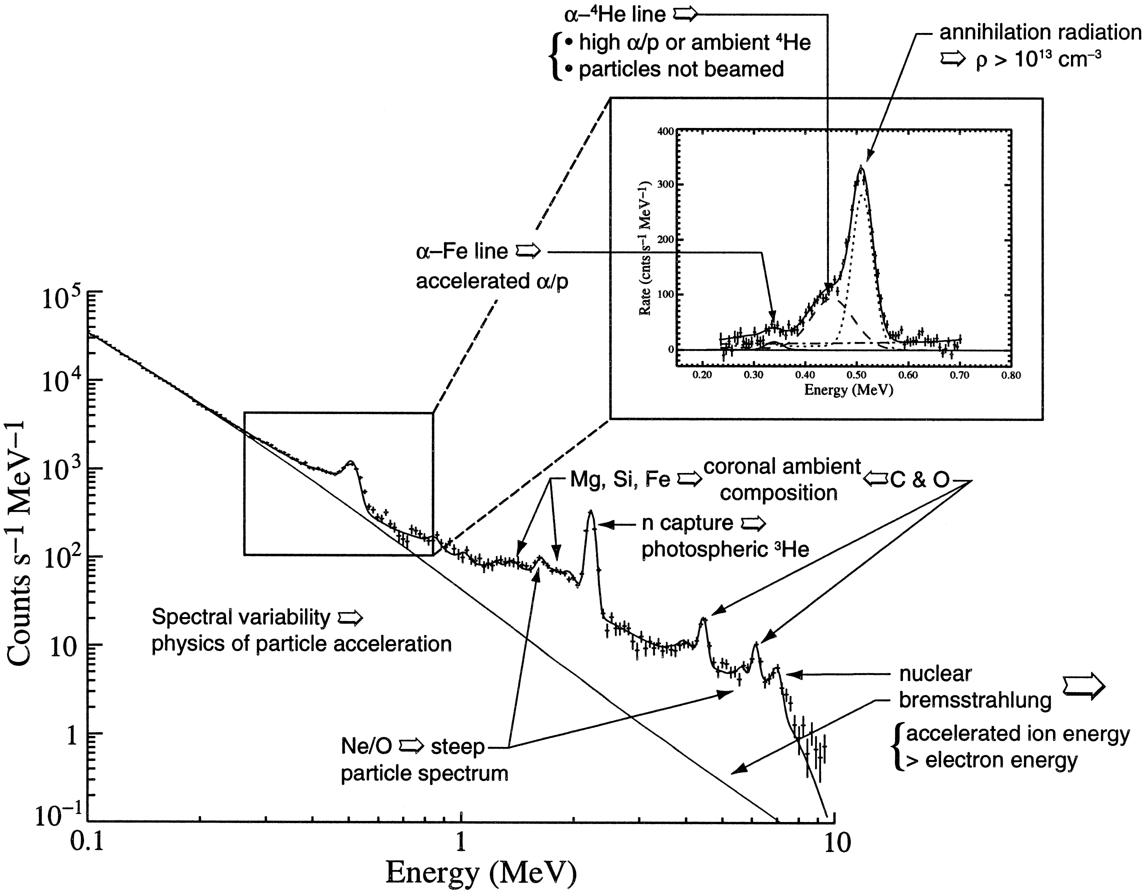


FIGURE 4 The gamma-ray spectrum of the June 4, 1991, solar flare obtained with the OSSE instrument aboard the Compton Gamma Ray Observatory. Prominent gamma-ray emission lines are identified, along with the primary processes responsible for both the observed line and continuum emission. (Courtesy NASA and the CGRO OSSE Instrument Team (G. H. Share and R. J. Murphy.))

even when significant peak delays are present, the pulse starting times are simultaneous to within 2 s. Those observations show that both electrons and nuclei can be rapidly accelerated to relativistic energies within seconds during solar flares. We now believe that the delays are largely generated by propagation and interaction effects as particles move from a low-density acceleration region high in the solar atmosphere to a higher-density interaction region deeper in the atmosphere where the high-energy emission is generated.

B. The Solar System

The question of the origin and evolution of the solar system is one of the most fundamental in astronomy. It bears directly on such related issues as stellar evolution, the formation of planetary systems, and on the existence of life itself. Gamma-ray observations from spacecraft, either via remote sensing from orbit or through *in situ* measurements from landers, contribute directly to the testing of evolutionary models of solar-system formation. Specifically, they provide a means of directly determining the elemental chemical composition of planetary surfaces, thus providing clues important to reconstructing the geochemical history of the solar system. Related complementary observational techniques include X-ray fluorescence measurements, and the detection of albedo neutrons and charged particles from planetary surfaces.

The gamma-ray observations relevant to planetary studies are spectroscopic in nature, aimed at identifying specific key elements present in planetary surfaces via their characteristic emission energies. The abundances of elements with different condensation temperatures and geochemical behavior relate directly to the origin and evolution of planetary bodies. For example, the K/U ratio provides a measure of the remelting of primordial condensates, while the K/Th ratio indicates the relative abundance of volatile to refractory elements.

Gamma-ray spectral lines originate in nuclear processes, either from radioactive decay or from nuclear deexcitation following particle collisions. Natural radioactivity results from the decay of the primordial radioactive elements ^{40}K, ^{138}La, ^{176}Lu, and those in the uranium and thorium decay sequences. Collisions of primary Galactic cosmic rays (of which ~90% are protons) with planetary material can give rise to numerous interactions and secondary particles, leading to gamma rays. Extensive model calculations of cosmic ray-induced gamma-ray emission from planetary bodies have been carried out and can be readily compared to the available observations.

Since the 1960s measurements of X-rays, gamma rays, alpha particles, and neutrons from the Moon, Mars, and Venus have been undertaken successfully with a variety of instruments aboard both U.S. and Russian spacecraft. The two U.S. Viking landers on Mars, for example, carried out X-ray fluorescence measurements of the Martian surface, while the Russian Venera 8, 9, and 10 spacecraft measured the natural radioactivities of potassium, uranium, and thorium at three landing sites on Venus.

Until very recently the most detailed and extensive remote-sensing observations of a planetary body were carried out during the Apollo 15 and 16 flights to the Moon (1971, 1972) when instruments aboard the orbiting command modules mapped approximately 20% of the lunar surface in X-rays, gamma rays, and alpha particles. The Apollo missions were also unique in that a detailed comparison could be made between the results of the remote mapping and follow-up compositional analysis of actual returned samples of lunar material.

The Apollo measurements clearly demonstrated that the Moon's crust is chemically differentiated, with a pronounced distinction between maria and highland regions, with the maria primarily basaltic in nature. On a more localized scale, material around craters tends to exhibit a significant chemical contrast relative to surrounding regions, suggesting an excavation of material from the subsurface due to impacts from asteroids and comets. Distribution patterns seem to favor an impact rather than a volcanic dispersal. The observed K/Th ratio has provided a measure of the volatile-to-refractory material variation over the lunar surface, which is found to be consistently lower than the terrestrial value, reflecting a global depletion of volatiles on the Moon compared to the Earth.

More recently the U.S. Lunar Prospector mission successfully obtained (1998–1999) global maps of the lunar surface using gamma-ray and neutron spectrometers. The results have generally confirmed the earlier Apollo findings. As a follow-up to the Apollo measurements, there is a particular interest in determining the distribution of "KREEP"-rich material on the Moon. KREEP refers to an unusual mixture of elements containing potassium (K), rare-earth elements (REE), and phosphorous (P) that is believed to have formed at the lunar crust-mantle boundary as the final product of the initial differentiation of the Moon. Understanding the composition and distribution of KREEP-rich material is thus considered key to reconstructing the evolution of the lunar crust. The Lunar Prospector data have demonstrated that KREEP-rich rocks tend to be found on the rims and boundaries of major lunar impact basins where there are surmised to have been exposed, dredged up, and dispersed as a result of these cataclysmic impact events. The most intriguing of the recent lunar neutron observations point to the possible presence of subsurface water ice at the lunar poles (whose existence was first indicated by radar measurements carried out with the Clementine spacecraft in orbit around the Moon in 1994).

In 2001, in an engineering tour de force, the Shoemaker-NEAR (Near Earth Asteroid Rendezvous) spacecraft completed its successful year-long orbital study of the asteroid 433 Eros with a spectacular unplanned landing on the asteroid's surface, transforming its on-board X-ray and gamma-ray spectrometers from remote-sensing to *in situ* instruments. Initial analysis of the NEAR spectrometer data suggests that Eros may remain in an undifferentiated state, unaltered by melting, constituting some of the most primitive material in the solar system yet studied. Also in 2001, the NASA's Mars Odyssey spacecraft is scheduled for launch, carrying a gamma-ray spectrometer among its suite of instruments. If the goals of these missions are fully realized complete global maps of elemental distribution for the Moon, Eros, and Mars, representing three bodies with distinct evolutionary histories, will be available for detailed comparative studies.

C. Galactic Sources

As anyone who has gazed at the night sky from a dark location knows, the distribution of visible stars is not random. Rather, stars tend to cluster in a bright band called the Milky Way that delineates the plane of the flat spiral galaxy in which we live. Like the stars, there is also a population of gamma-ray sources that cluster in the plane of the Galaxy and that are believed to reside within the Milky Way. However, our ability to associate them with visible counterparts in the crowded Galactic plane is hampered by the fact that the angular resolution of the best gamma-ray telescopes is still a hundred times coarser than even small backyard optical telescopes. As a consequence, many Galactic "point" gamma-ray sources have multiple counterpart candidates and, even worse, in many directions in the Galactic plane we know that the gamma emission from several "point" sources is actually blended together and gamma-ray telescopes are "source confused." Nevertheless, by studying the temporal variations of the gamma-ray emission and correlating it with intensity variations measured by higher-resolution X-ray, optical, or radio telescopes, we have been able to identify with certainty several classes of Galactic gamma-ray sources.

1. Isolated Pulsars

Among the best clocks known, natural or man-made, are the spinning, magnetized neutron stars called pulsars. Pulsars emit short bursts of electromagnetic radiation at intervals from one every few seconds to thousands of times a second with a regularity that exceeds that of watches of the highest precision. In some cases this electromagnetic pulse extends across the entire spectrum from radio to gamma-ray wavelengths, thereby allowing us to unambiguously identify these Galactic gamma-ray sources.

The best-known gamma-ray emitting pulsar is the so-called Crab pulsar. It is embedded in and is the source of power for the famous Crab nebula in the constellation Taurus, the first object (M1) listed in the renowned catalog of diffuse, nebular objects compiled by the French astronomer Charles Messier in the late 18th century. From medieval Chinese records describing the temporary appearance of a "guest star" near the star we now call Zeta Tauri, we know that the Crab pulsar was the product of a supernova explosion that occurred in the year 1054 AD. Every 33 ms, the Crab pulsar emits a pair of radiation pulses that are detectable from the radio band all the way up to TeV gamma-ray energies. While pulsar physics still has many open questions, it is generally agreed that the emission from isolated pulsars is generated by energetic particles that are accelerated by electric fields induced by the spinning magnetic field of a rapidly rotating, magnetized, neutron star. The energy reservoir that ultimately powers all the observed emission is therefore the rotational energy of the rapidly spinning pulsar.

At least six other isolated gamma-ray pulsars are currently known to exist and from that small sample a few general patterns are apparent. First, for all known isolated gamma-ray pulsars the gamma emission represents the largest observable fraction of the total power emitted by the pulsar. As a consequence, observational study of the gamma rays provides important diagnostics on the overall efficiency for particle acceleration and interactions in the extreme pulsar environment. Second, the gamma-ray visibility increases with the spin-down luminosity (or, the ratio of the magnetic field strength to the square of the spin period). Finally, all of the isolated gamma-ray pulsars appear to be unvarying point sources when their emission is averaged over the spin period.

Perhaps the most remarkable object in the sample of known gamma-ray-emitting pulsars is Geminga. For nearly 20 years this source, which is the second brightest source in the gamma-ray sky, was a puzzle because it did not appear to emit radiation in any other energy band. The unusual name, Geminga, was coined by Italian astronomers and is derived both from the source's location in the constellation Gemini and from a play on words, geminga meaning "is not there" in the Milanese dialect. The breakthrough in our understanding of Geminga occurred in 1991 when pulsating soft X-ray emission with a period of 0.237 s (=237 ms) was detected from the direction of Geminga by the ROSAT X-ray satellite. Identification of Geminga as pulsar was therefore clinched when a phase analysis revealed that the gamma-ray emission was also modulated at the same 0.237-s period (see Fig. 5). A particularly interesting property of Geminga is that, unlike other rotation-powered pulsars, it is not detectable as a radio pulsar even though it is thought to be only 100 parsecs from the Sun. Since our census of the population of pulsars is based on radio observations, it is possible that Geminga is just the nearest member of a large population of previously unknown pulsars that are only visible at X-ray and gamma-ray energies. Many of the steady, unidentified gamma-ray sources could therefore be Geminga-type pulsars with still unknown spin periods.

2. Accreting Pulsars

Not all pulsars are isolated. Some reside in binary stellar systems and some of these pulsars are X-ray sources that are powered by the gravitational energy released when gas from the companion star is accreted onto the neutron star. Those accreting pulsars are also likely sources of gamma-ray emission. Indeed, several groups using ground-based Cherenkov telescopes have reported the detections of TeV gamma-ray emission from systems containing an accreting pulsar. Unfortunately, most of the reported detections have been of low statistical significance and/or not confirmed with subsequent more sensitive observations. If real, they indicate that the gamma-ray emission from accreting pulsars is sporadic. On theoretical grounds, such behavior is plausible because accretion flows are often unstable and shocks in the flow could efficiently accelerate gamma-ray emitting particles.

Support for the idea that accreting pulsars sporadically emit gamma-ray emission was also found with the EGRET telescope aboard the Compton Gamma Ray Observatory. In October 1994 a week-long outburst of GeV gamma-ray emission was detected from the direction of the massive X-ray binary system Centaurus X-3. During the outburst, the accreting pulsar in Cen X-3 underwent an interval of rapid spin-down. Phase analysis of the gamma-ray emission showed evidence for spin modulation in step with the rapidly drifting X-ray period.

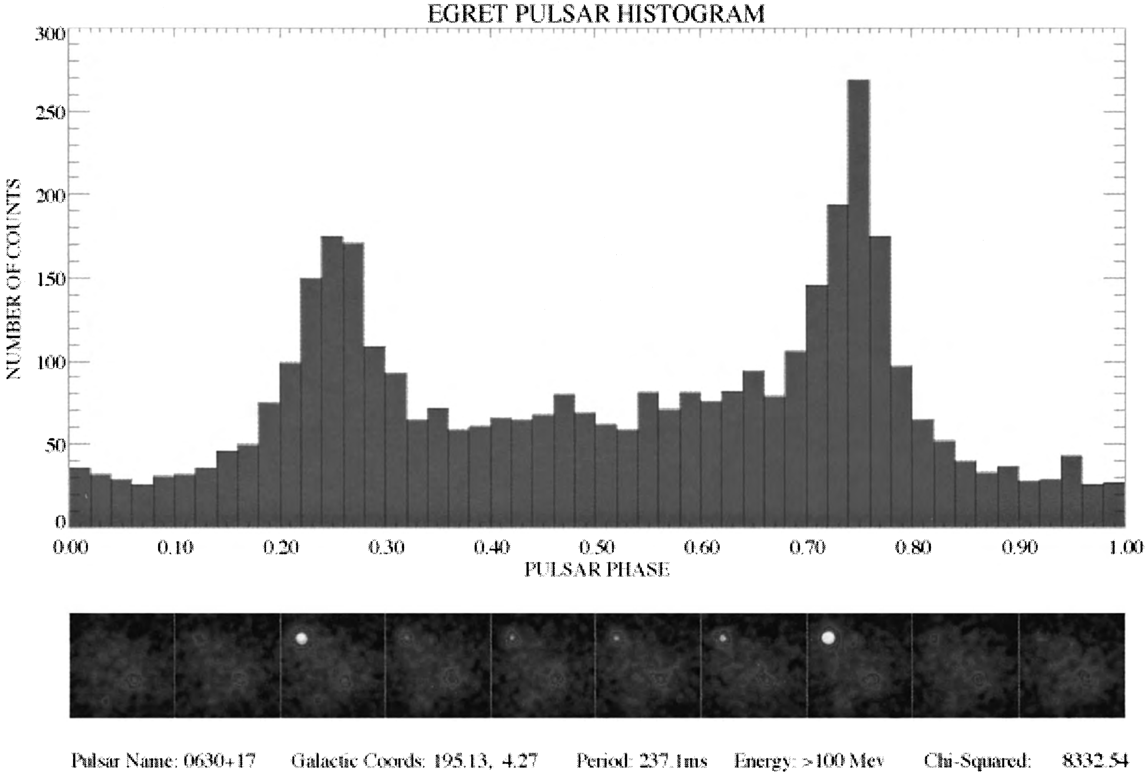


FIGURE 5 Pulsed gamma-ray emission from the Geminga pulsar. *Above*: The light curve of the gamma-ray emission from Geminga, with the gamma rays binned according to the pulsar period of 237 milliseconds. *Below*: Gamma-ray images of the region of the sky containing the Geminga pulsar, showing the brightening of the source (in the upper left of the image plots) in phase with the pulsed emission. The fainter object at the lower right is the Crab pulsar (with a different pulse period of 33 ms) which appears as a steady source in this representation. (Courtesy NASA and the CGRO EGRET Instrument Team (P. Sreekumar)).

3. Black-Hole Binary Systems

Some X-ray binary systems are believed to contain stellar-mass black holes that are embedded in disks of inwardly spiraling accreted gas. These accretion disks are known sources of both soft and hard X-ray emission and, in some cases, are also sources of soft gamma-ray emission. Perhaps the best-known example is Cygnus X-1, the brightest X-ray source in the constellation Cygnus. The X-ray luminosity of this high-mass X-ray binary (in which the normal companion is an O or B star many times the mass of the Sun) is variable and is correlated with different high-energy spectral "states." These states of activity are composed of different admixtures of two principal components: a soft thermal blackbody-like component and a hard nonthermal power-law component. It is this power-law component that can, at times, extend up to gamma-ray energies of at least an MeV.

The origin of the MeV gamma ray emission from Cygnus X-1 is not well understood. Most models assume that the gamma rays are X-rays that were Compton scattered by mildly relativistic electrons. The precise origin of those energetic electrons is still unclear. Some modelers have speculated that reconnecting magnetic fields or shocks in disk outflows accelerate the electrons. However, a particularly attractive idea is that the scattering electrons arise naturally in the convergent accretion flow from the innermost stable orbit of the accretion disk. If the accretion rate is high, Compton emission from the bulk flow could generate the observed gamma ray emission.

Finally, Galactic sources that generate intense outbursts of X-ray emission that can persist for months before fading are termed X-ray novae. They are known to occur in low-mass binary systems in which the normal companion star is of approximately solar mass in close orbit around a compact object. Such binary systems can undergo recurrent outbursts that for a brief period can make them the brightest high-energy sources in the sky. Optical observations of these low-mass X-ray binary systems during their quiescent states indicate that the compact object is typically more massive than 3.0 solar masses and is therefore most likely a black hole. Some outbursts of X-ray novae are also accompanied by low-energy gamma-ray emission. At least one such nova outburst, that of Nova Muscae in 1991 observed with the SIGMA instrument, displayed a variable positron annihilation line. Later observations with telescopes aboard the Compton Observatory of Nova Persei 1992 confirmed that X-ray novae can give rise to gamma-ray emission extending up to photon energies of at least 2 MeV.

D. Gamma-Ray Lines of Galactic and Extragalactic Origin

One of the great triumphs of modern astrophysics has been an increasingly detailed understanding of the nuclear reaction processes that govern the production of energy in stars and that determine the course of stellar evolution. It is now clear that nucleosynthesis, or the production of elements from the primordial building blocks of hydrogen and helium, occurs almost exclusively in the cores of stars, or in the cataclysmic explosive events, supernovae and novae, that mark the end of a star's lifetime. The study of gamma-ray spectral lines provides one of the few direct means of verifying the predictions of the various models of stellar nucleosynthesis, and of the explosive events that disperse this material back into the interstellar medium, out of which new generations of stars are formed.

Gamma-ray lines result predominantly from nuclear processes, either from the decay of radioactive nuclides and the deexcitation of excited nuclei (see Table I), or from the collisions of high-energy particles. The advantages of gamma-ray line spectroscopy are manifest. Spectral line transitions occur at specific characteristic energies that provide immediate identification of the isotopic species that produced them. The comparison of line strengths or intensities between different elements and isotopes can be translated into isotopic abundances, densities, and temperatures in the emitting region. Similarly, the presence of broad, narrow, or Doppler-shifted lines provides a measure of gas motions and velocities, all of which can be interpreted in terms of specific models of production.

TABLE I Primary Radioactive Nuclear Decay Lines from Nucleosynthesis

Decay process	Mean halflife	Line energies (MeV)
^{56}Ni → ^{56}Co → ^{56}Fe	111 d	0.511, 0.847, 1.238
^{57}Co → ^{57}Fe	272 d	0.014, 0.122
^{22}Na → ^{22}Ne	2.6 y	0.511, 1.275
^{44}Ti → ^{44}Sc → ^{44}Ca	~60 y	0.068, 0.078, 0.511, 1.157
^{26}Al → ^{26}Mg	7×10^5 y	0.511, 1.809
^{60}Fe → ^{60}Co → ^{60}Ni	2×10^6 y	0.059, 1.173, 1.332

1. Nucleosynthesis in Stars and Supernovae

Gamma-ray line radiation in the Galaxy results primarily from two broad categories of production: either *steady-state thermal* or *explosive nucleosynthesis*. In the former case, stars in hydrostatic equilibrium throughout the bulk of their lives generate energy via thermonuclear fusion reactions in their high-temperature cores. Depending on the stellar mass (which determines core temperature), these fusion reactions may progress over time through successive stages of nuclear burning, leading to a buildup of different layers of nuclear reaction products at the center of the star. For low- and intermediate-mass stars such as the Sun, nuclear burning ceases with the fusion of helium into carbon. For the most massive stars (10–100 times the mass of the Sun), however, with much higher core temperatures, nuclear burning leads ultimately to an iron core surrounded by layers of silicon, magnesium, neon, oxygen, and carbon, along with remnant amounts of helium and hydrogen in the outer envelope of the star. Toward the end of its life, as it runs out of nuclear fuel, a star becomes increasingly unstable, and the heavier elements at the center of the star are brought to the surface via mixing and convective processes, and are ultimately dispersed back into the interstellar medium through high-speed stellar winds, flares, outbursts, and variable pulsations of the outer envelope and atmosphere of the star. Mixed in with all of the reaction products are long-lived radioisotopes that act as tracers of the various stages of nuclear burning.

Elements beyond iron in the periodic table are formed in the course of supernovae explosions that result either from the cataclysmic collapse of a white dwarf into a neutron star due to the accretion of matter from a binary companion (a Type I supernova), or from the catastrophic core-collapse of the most massive stars at the end of their lives when nuclear fuel is exhausted (a Type II supernovae). Subcategories of each type also exist. In the former case, the entire star is disrupted, liberating approximately 10^{51} ergs in the explosion and about 0.5 to 1.0 solar masses in synthesized radioactive material, while in the latter class of event slighter higher energies may be liberated (up to $\sim 10^{53}$ ergs) but thick layers of ejecta partially mask for a time the 0.1 solar masses of radioactive material synthesized in the explosion. Supernovae are among the most violent and luminous events known in the Universe, with ejected material attaining speeds of thousands of kilometers per second. A supernova at peak light may completely outshine its host galaxy (containing billions of stars), and its light curve decays at the exponential rates characteristic of the primary radioactive species produced in the explosion.

The gamma-ray lines of primary observational interest associated with nucleosynthesis are of MeV energies, and are listed in Table I. The radioisotopes ^{56}Ni, ^{57}Ni, ^{44}Ti, and ^{26}Al are particularly important since they span a range of half-lives, from days to millions of years, thus together providing a measure of both the "prompt" emission from individual events, as well as of the "delayed" or integrated cumulative emission dispersed throughout the interstellar medium of the Galaxy arising from generations of star formation. The lines originally predicted to be the most luminous from individual events are the ^{56}Ni and ^{56}Co lines from Type Ia supernovae. The long-anticipated breakthrough in gamma-ray line detection, however, occurred for the now-famous SN 1987A, a type II supernova which occurred at a distance of 55 kpc in our neighbor galaxy, the Large Magellanic Cloud (the LMC). The gamma-ray spectrometer aboard the Solar Maximum Mission (SMM) satellite detected and studied the ^{56}Co lines from this event. The gamma-ray lines were detected earlier after the event than predicted, implying that the inner layers of material containing the radioactive ^{56}Co were more thoroughly mixed than expected, or indicating perhaps that the ejecta were clumpier, allowing clearer lines of sight to the inner regions of the exploding star through which the gamma rays could escape. After the launch of the Compton Observatory in 1991, several years after the event, the OSSE instrument detected the longer-lived ^{57}Co line ($t_{1/2} \sim 272$ days) at 122 keV in energy. Its intensity implied a ratio of synthesized ^{57}Ni to ^{56}Ni (from ^{56}Fe and ^{57}Fe, respectively) of about 1.5 times the solar value, providing constraints on models of the progenitor star's evolution. In other supernova observations, there were tantalizing hints of gamma-ray detections with COMPTEL of the 0.847- and 1.239-MeV lines of ^{56}Co from the Type Ia supernova SN 1991T in the galaxy NGC 4527 at a distance of about 13–17 Mpc, at the limit of this telescope's range of detectability.

Of particular interest in gamma-ray line astronomy is the detection of ^{44}Ti ($t_{1/2} \sim 60$ years) at 1.157 MeV from young, distant obscured supernovae in the Galaxy. About two to three supernovae per century are predicted to occur in the Milky Way, but most have remained undetected because they are believed to remain hidden behind intervening clouds of interstellar gas and dust in the spiral arms. Gamma rays, in contrast to optical light, easily penetrate the interstellar material and provide a means of detecting directly these interesting objects and confirming predictions regarding star formation rates in the Galaxy. Further, ^{44}Ti is formed in the deepest layers of the supernova ejecta, and its predicted line strength is sensitive to the details of the explosion models and to the likelihood of fall-back onto the collapsed, compact stellar core. The COMPTEL instrument aboard the CGRO has reported the detection of ^{44}Ti from the Cas A supernova remnant (a relatively close remnant about 3 kpc distant, believed to have exploded about 250 years ago) and from the Vela region of the Galaxy. The next generation of gamma-ray spectrometer and imager aboard the INTEGRAL spacecraft will provide critical confirmation of these first detections and should also be able to map out in detail the spatial distribution of the emitting radioisotope, and, from a determination of the exact shape of the gamma-ray line profiles, provide a measure of the symmetry of the initial explosion and subsequent expansion of the ejected material.

The longer-lived radioisotope ^{26}Al with a decay line at 1.809 MeV ($t_{1/2} \sim 710{,}000$ years) traces the sites of massive star formation and nucleosynthesis in the Galaxy over the past million years. This interstellar line was originally detected by the HEAO-3 and SMM spacecraft, followed by a number of balloon instruments. In a long-awaited result, the first all-sky map in the light of this radioisotope was produced by the COMPTEL instrument team following extensive observations and analysis. Clearly evident is the disk of the inner Galaxy, with enhancements of emission in particular regions (see Fig. 6). These tend to coincide in direction to spiral arms in the Galaxy, where recent star formation and supernova activity are most likely to occur (e.g., the Cygnus, Vela, and Carina regions). Further measurements of the spectral line of ^{26}Al with high-resolution spectrometers and imagers are expected to yield detailed identification of specific sources for this emission.

2. Electron–Positron Annihilation Radiation

A gamma-ray line at 0.511 MeV results from the mutual annihilation of an electron and a positron, a particle-antiparticle pair. A number of radioactive decay chains (see Table I) result in the emission of a positron as a decay product, which will annihilate upon first encounter with an electron. Also of astrophysical importance is the production of electrons and positrons via the photon–photon pair-creation process. Such pair plasmas are found in the vicinity of compact objects, such as neutron stars and black holes, that are associated with heated accretion disks and relativistic flows and jets, within which particle acceleration is known to occur. Thus, relatively narrow lines of 0.511-MeV annihilation radiation are expected to arise in the interstellar medium through the decay of dispersed, nucleosynthetic radionuclides, while broadened, Doppler-shifted, and possibly time-variable lines may occur in the high-energy and dense environments associated with compact objects.

Direct annihilation of an electron–positron pair leads to the emission of two photons. If the particles are at

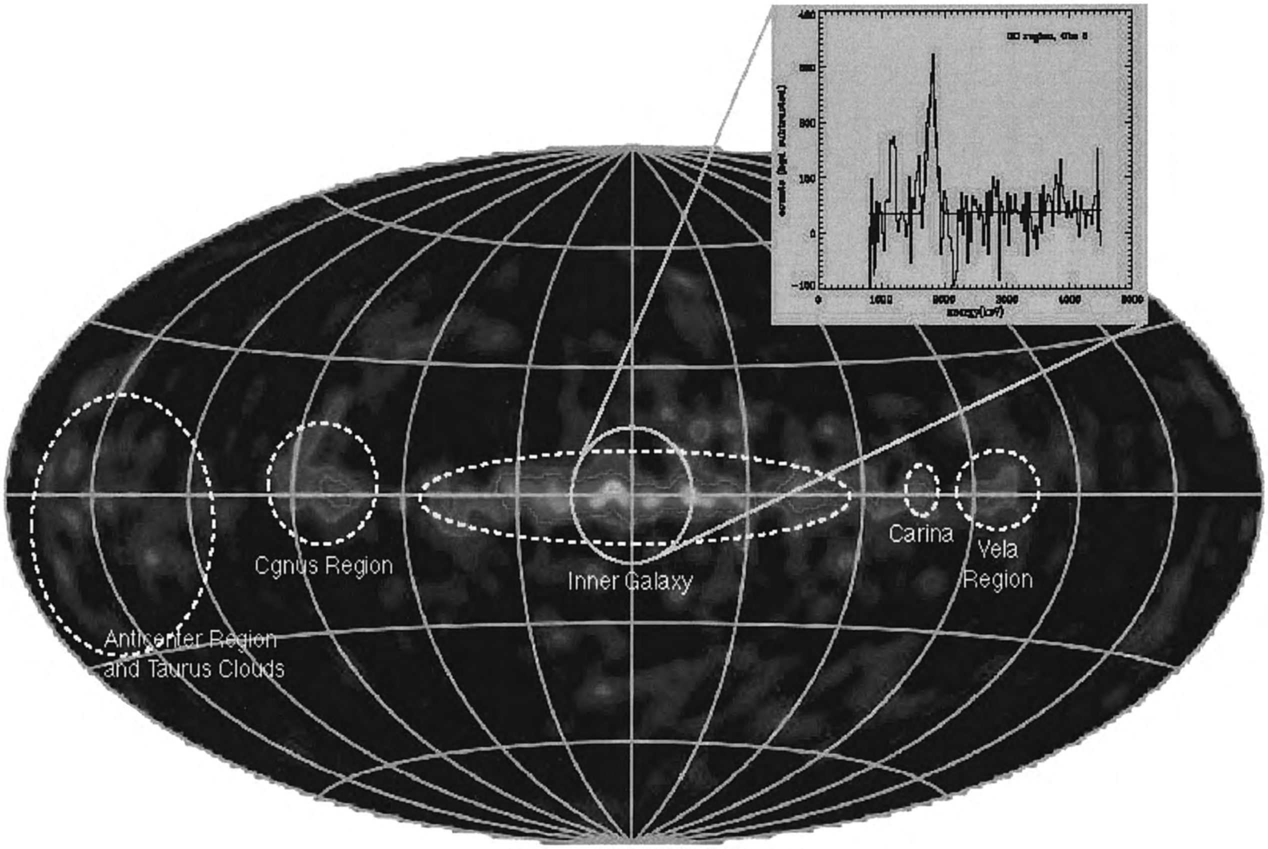


FIGURE 6 An all-sky map of the gamma-ray line emission at 1.809 MeV due to the decay of radioactive ^{26}Al in the Galaxy, obtained with the COMPTEL instrument aboard the Compton Gamma Ray Observatory. The map is an Aitoff equal-area projection in Galactic coordinates (as in Fig. 1). The gamma-ray line from ^{26}Al traces the sites of massive star formation and nucleosynthesis in the Milky Way. Regions with enhanced emission are identified. *Inset*: A gamma-ray spectrum showing the emission line from ^{26}Al at 1.809 MeV observed in the inner Galaxy. (Courtesy NASA, the Max Planck Institute, and the CGRO COMPTEL Instrument Team.)

rest two gamma rays of equal energy, 0.511 MeV, are produced, while in-flight collisions will result in a range of possible photon energies (whose sum must equal the kinetic plus the rest–mass energies of the two particles, $2m_ec^2$). If the velocities of the two particles are small, a bound state consisting of a positron and an electron is possible, the *positronium* atom. Positronium decays after a very short lifetime ($<10^{-7}$ s) via one of two channels. Two-photon decay results again in the emission of two gamma photons of energy 0.511 MeV, while three-photon decay leads to a continuum of emission below the main 0.511-MeV spectral peak. The observed *positronium fraction*, or continuum-to-line ratio, provides a unique diagnostic measure of the physical conditions of the emitting region.

Prior to the launch of the CGRO, from the 1970s on, a number of balloon and early satellite instruments detected the presence of apparently variable annihilation radiation from the direction of the Galactic Center. This led to the widespread supposition that the observed emission arose from two types of sources: a steady, dispersed diffuse component of nucleosynthetic origin in the disk of the Galaxy, and a time-variable point source (or sources) near the Galactic Center. Extensive observations of the 0.511-MeV line from the inner Galaxy with the CGRO OSSE instrument (see Fig. 7), however, have revealed a somewhat different picture: a central bulge, along with diffuse emission in the Galactic plane, and an apparent extension of emission above the Galactic disk in the direction of the Galactic Center. No temporal variability of the annihilation radiation is evident in the OSSE observations. The annihilation emission observed by OSSE can be explained entirely in terms of radioactive isotopes from supernovae and similar sources. The apparent enhancement of 0.511-MeV radiation above the Galactic plane has been variously interpreted as a "positron fountain" arising from an asymmetric outflow of positrons from the region of the Galactic Center following a period of enhanced star

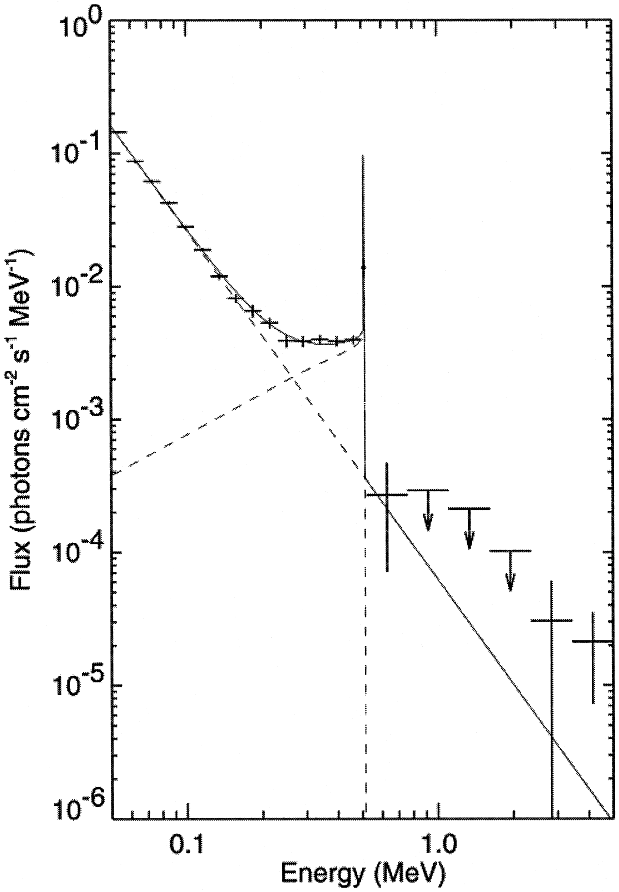


FIGURE 7 The gamma-ray spectrum of the positron annihilation line at 0.511 MeV and associated positronium continuum toward the region of the Galactic Center, obtained with the OSSE instrument aboard the Compton Gamma Ray Observatory. The data were fit with a model consisting of a narrow annihilation line, a positronium component, and an underlying power-law continuum, indicated by the dashed lines. (Adapted from W. R. Purcell *et al.*, "OSSE mapping of galactic 511 keV positron annihilation line emission," *Astrophys. J.* **491,** 725–748, Copyright 1997, reproduced with permission of the AAS.)

formation and supernovae activity, or the result of jet activity from one or more black-hole sources in or near the center of the Galaxy, or even the result of a single cataclysmic gamma-ray burst-like event occurring near the Galactic Center about a million years ago.

E. Diffuse Galactic Gamma-Ray Emission

The most conspicuous feature in the gamma-ray sky is the diffuse continuum radiation arising from the Galaxy itself, originating in the interstellar clouds of gas and dust that reside within the spiral arms, disk, and bulge of the Milky Way. A narrow band of diffuse gamma-ray emission is observed along the Galactic plane, with enhancements toward the inner Galaxy and in directions that are tangent to the spiral arms (where the column density of interstellar material is greatest along the line of sight) and with hot spots of intensity in particular directions that may be due to unresolved point sources of emission (see Fig. 1). First detected by the OSO 3 satellite, this diffuse radiation was also observed and mapped at medium resolution by the SAS 2 and COS B satellites. The two wide-field gamma-ray telescopes aboard the CGRO, COMPTEL and EGRET, were specifically designed to carry out as one of their primary scientific objectives a complete mapping of the entire sky in gamma rays. A key result of this all-sky survey was the first complete map of the Galactic diffuse radiation from approximately 1 MeV to 30 GeV in energy. The OSSE instrument aboard the CGRO also derived a measure of the Galactic diffuse radiation toward the center of the Galaxy up to an energy of approximately 10 MeV.

The Galactic diffuse emission arises from the interaction of energetic cosmic-ray particles, predominantly electrons and protons, with ambient material and low-energy photons in the interstellar medium. The primary interaction mechanisms are (1) nucleon–nucleon collisions of cosmic-ray protons that result in the creation of neutral pions that decay into gamma-rays, (2) bremsstrahlung from high-energy electrons, and (3) Compton up-scattering by cosmic-ray electrons of the low-energy photons (IR, optical, and UV) that comprise the interstellar radiation field (or ISRF). Each of these gamma-ray production mechanisms is dominant over a particular energy range. Below about 100 MeV the bremsstrahlung component is important, while above that energy the decay of neutral pions from nucleon–nucleon collisions is key. The relative importance of the Compton component is dependent on the spectral index (or "hardness") of the cosmic-ray electron population, and on the details of the ISRF, neither of which is well determined, hence the added interest in studies of the diffuse gamma-ray emission to help fix these contributing physical quantities. Finally, there is also an isotropic component of the diffuse gamma radiation that is believed to be extragalactic in origin (described in the next section).

Since the interstellar medium is essentially transparent to the propagation of gamma rays, the observed gamma-ray intensity in a particular direction represents the total, cumulative emission from all particle interactions and sources along that line of sight. Thus, the diffuse gamma radiation provides a simultaneous measure of both the cosmic-ray and matter distribution throughout the Galaxy. The challenge is to disentangle the cosmic rays (the "projectile" particles) from the interstellar matter (the "target" particles) in the observed gamma-ray signal. The study of the Galactic diffuse gamma radiation is therefore intimately related to Galactic radio astronomy. The distribution of matter in the Galaxy is derived primarily from radio surveys, in particular, those at 21-cm wavelength that map the distribution of atomic hydrogen, and the millimeter CO surveys that serve as tracers of molecular hydrogen. A major advantage of the spectral-line radio observations is that Doppler shifts in the observed line emission can be interpreted in terms of a kinematic model of differential Galactic rotation, which allows a determination of the distance to the emitting gas based on its measured radial velocity. Continuum radio surveys of the synchrotron emission from electrons interacting with Galactic magnetic fields also provide important observational constraints on the distribution and properties of the cosmic-ray electron population. In turn, studies of the diffuse gamma-ray emission complement the radio observations in that they serve to constrain a critical parameter in the molecular radio surveys, namely, the CO-to-H_2 conversion factor (the so-called "X-value," for which the EGRET-determined average over the whole Galaxy is $(1.56 \pm 0.05) \times 10^{20}$ H-molecules cm^{-2} $(\text{K km s}^{-1})^{-1}$).

The diffuse gamma-ray studies, when combined with the radio data, provide the best measure to date of the distribution of high-energy cosmic rays within the Galaxy. Cosmic rays presumably arise following particle acceleration via shocks in supernovae and their remnants in the interstellar medium. As charged particles, however, their trajectories are influenced by magnetic fields and thus their exact sites of origin and acceleration, as well as their composition, modes of propagation, and overall lifetime in the Galaxy are still not well understood. This situation is most severe for cosmic-ray electrons, since these less-massive particles "cool" rapidly due to energy losses via synchrotron, bremsstrahlung, and Compton emission, and consequently have relatively little time to diffuse far from their place of origin in the Galaxy to be detected.

A number of approaches have been followed to model the diffuse gamma-ray observations. They can be characterized as either *parametric* models that are fit to the data to study intensity and spectral variations as a function of Galactic radius, *dynamic balance* models that seek to balance the gravitational attraction of interstellar matter against the expansive pressures due to cosmic rays, matter and magnetic fields, and *cosmic-ray propagation* models that mimic the propagation of cosmic rays through the Galaxy by including such processes as diffusion, convection, reacceleration, fragmentation, the production of secondary particles, and the generation of gamma-ray and synchrotron radiation, as constrained by all available observations.

Analysis of the results obtained with the COMPTEL and EGRET instruments aboard the Compton Observatory suggest that the inverse Compton process may be a more important contributor to the Galactic diffuse emission than previously thought. The bremsstrahlung component of the medium-energy (1- to 50-MeV) diffuse emission, though comparatively small, remains of interest in that it can be combined with surveys of the Galactic radio synchrotron emission (which traces electrons with energies in the range 100 MeV to 10 GeV) to derive the shape of the cosmic-ray electron spectrum. Historically, the electron spectrum below ~10 GeV has been difficult to determine since these particles are excluded from direct detection near the Earth due to the periodic modulation of the solar wind. At higher energies in the EGRET sensitive range (30 MeV to 30 GeV) the observed gamma-ray spectrum is softer in the direction of the outer Galaxy compared to the inner Galaxy, suggestive of a corresponding change in the spectrum of cosmic-ray nuclei with Galactic radius. This could be explained if cosmic rays are accelerated preferentially in the inner Galaxy and then propagate to the outer Galaxy, or if the high-energy cosmic rays are less well-confined in the outer Galaxy. While overall agreement between model predictions and the gamma-ray data is quite good, one surprising finding is an excess in the diffuse emission observed above ~1 GeV (see Fig. 8).

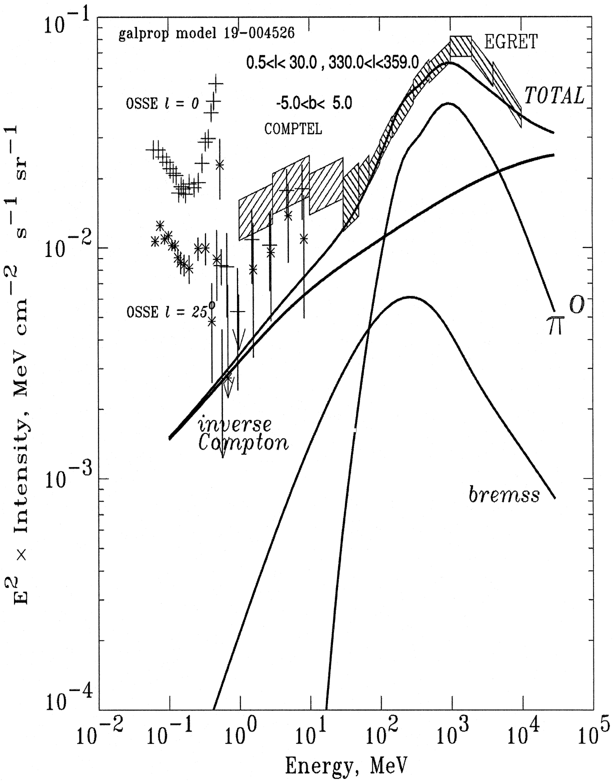


FIGURE 8 Spectrum of the diffuse gamma-ray emission from the inner Galaxy compared with calculations based on cosmic-ray propagation models. The data were obtained with the OSSE, COMPTEL, and EGRET instruments aboard the Compton Gamma Ray Observatory. Curves show the calculated contributions to the observed gamma-ray emission due to inverse Compton, bremsstrahlung, and π^0-decay processes, and the summed total. (Adapted from Strong, A. W., Moskalenko, I. V., and Reimer, O. (2000). "Diffuse continuum gamma rays from the Galaxy," *Astrophys. J.* **537,** 763–784, Copyright 2000, reproduced with permission of the AAS.)

This has been the subject of some debate and may reflect uncertainties in the neutral pion production function used in the model calculations, or perhaps is due to variations in the cosmic-ray spectrum with Galactic radius. Another possible explanation is enhanced Compton emission from a harder interstellar electron spectrum that may also give rise to an electron/inverse-Compton gamma-ray halo surrounding the disk of the Galaxy. Finally, at the very low end of the gamma-ray regime, it has been suggested that a population of unresolved Galactic sources may be responsible for the upturn in emission observed at low gamma-ray energies.

Future space missions (such as GLAST) will be able to confirm any spectral variations in the diffuse emission on sufficiently small angular scales (a few degrees) to permit a serious test of the various production models. Further observations with the next generation of ground-based air-Cherenkov GeV and TeV telescopes (where to date only tantalizing upper limits have been obtained) should also provide important confirmation of the GeV excess observed by EGRET.

F. Active Galaxies

Active galaxies, though they comprise only a small percentage of all known galaxies, rank among the most energetic and exotic objects in the Universe. They are exceptionally luminous (up to 10,000 times brighter than normal galaxies), and their emission is typically broadband and nonthermal in nature, spanning the entire electromagnetic spectrum. Their luminosity is concentrated in the central nucleus, which completely outshines the rest of the galaxy by factors of 100 or more, hence the alternate designation of "active galactic nuclei" (or AGNs). The emission from AGNs is highly variable and fluctuates rapidly enough ($\leq$day) to indicate that the source, or central engine, that drives the active behavior must occupy an extremely small region at the very unresolved core of the galaxy. AGNs are also associated with collimated bipolar jets and lobes of relativistic material emanating from the nucleus of the galaxy, within which ejected blobs of plasma can appear to travel at superluminal velocities, in apparent violation of the laws of physics (though this latter effect is now known to be a consequence of highly relativistic motion viewed along the direction of travel). For decades, as discoveries mounted, sub-categories of AGN proliferated, usually derived from some observational characteristic by which a particular class was originally identified. Thus, AGNs are categorized as radio-loud or radio-quiet, flat-spectrum or steep-spectrum sources, core-dominant or lobe-dominant in their emission, with broad spectral lines (indicative of high-velocity gas motions), narrow lines, highly polarized lines, or *no* lines, Seyfert galaxies of type I or II, Fanaroff-Riley galaxies of type I or II, and, at the most energetic extreme, the quasars, optically violent variables, and BL Lacertae objects (or BL Lacs) which collectively make up the class known as blazars (with luminosities in excess of 10^{48} ergs s^{-1}).

With the successful launch and operation of the Compton Gamma Ray Observatory (CGRO), a remarkable new class of active galaxy was discovered, the gamma-ray blazars. The high-energy EGRET experiment aboard the CGRO detected over 90 definite or probable gamma-ray AGNs (see Figs. 1 and 3), all but a few associated with active galaxies of the blazar class. These discoveries comprise the single largest category of gamma-ray source detected with EGRET and are all the more noteworthy in that prior to the launch of the CGRO only one such gamma-ray AGN was known, the quasar 3C 273 (discovered with the COS-B satellite). Gamma-ray blazars exhibit strong flaring behavior, and during active periods their gamma-ray luminosity can exceed that in other wavebands by factors of 100 or more. The gamma-ray blazars also span a large range in redshift ($z \sim 0.03$ to 2.3) and thus serve as cosmological probes of the intervening intergalactic medium. The discovery and study of the gamma-ray blazars has greatly enhanced our understanding of the blazar phenomenon in general, and has also provided strong confirmation of the relativistic jet model for active galaxies, since energy considerations require that the observed gamma-rays must be beamed in our direction via relativistic Doppler boosting.

In recent years, it has become clear that the vast majority of AGNs can be explained in terms of a single "unified model" for active galaxies. According to this concept the various subclasses of AGN can be explained as arising from geometric effects due to different viewing angles of the observer with respect to the central source. In the now-standard picture, the central engine of an AGN is a supermassive black hole (of $\sim 10^6$–10^{10} solar masses) that is powered by the accretion of surrounding infalling material. It is the partial conversion via accretion of the huge reservoir of gravitational potential energy into thermal energy due to the presence of the black hole that drives the energetic behavior of AGNs. In the standard scenario the central supermassive black hole is surrounded by an accretion disk and a thicker outer obscuring torus of material in the equatorial plane of the rotating black hole (see Fig. 9). Highly collimated bipolar jets of relativistic plasma are formed and ejected along the rotational axis, and intervening clouds of material of varying velocity, subject to bombardment from accelerated particles and radiation from the accretion disk and jets, give rise to the broad and narrow spectral lines observed. Predominantly thermal emission, extending up to X-rays, arises from the heated accretion disk and surrounding torus, while broadband nonthermal synchrotron and Compton emission, including gamma radiation, is given off by the relativistic

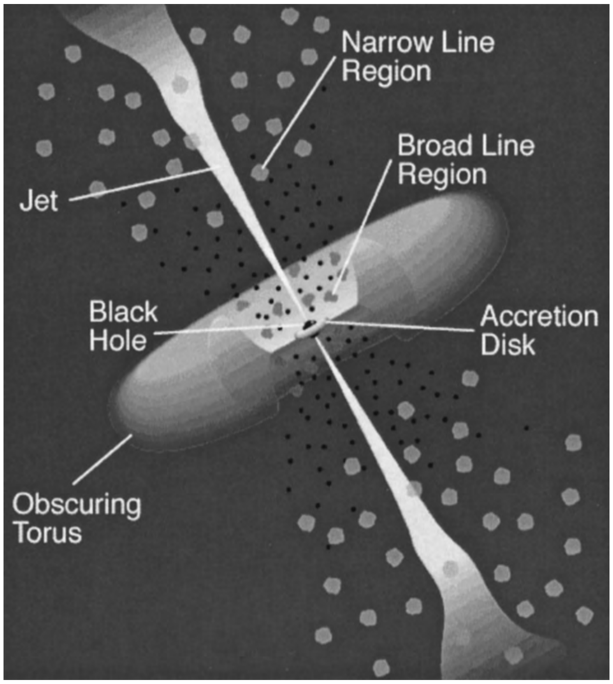


FIGURE 9 Schematic diagram of the central region of an active galaxy illustrating the main components of the "unified model" for active galactic nuclei (AGN). According to this model the various subclasses of AGN can be interpreted as arising from the observer's viewing angle with respect to the central source (a supermassive black hole) and the surrounding accretion disk, torus, intervening clouds, and relativistic jets. Gamma-ray blazars are presumed to arise when the observer's viewing angle is very closely aligned with the relativistic jets. (Adapted from Urry, C. M., and Padovani, P. (1995). "Unified schemes for radio-loud active galactic nuclei," *Publication of the Astronomical Society of the Pacific* **107,** 803–845, Copyright PASP 1995, reproduced with permission of the authors.)

particles in the jets. The appearance of an active galaxy thus depends critically on one's observing angle with respect to the central source and the surrounding accretion disk, torus, and relativistic jets. The most extreme variable behavior will be noted when the observer's viewing angle is very closely aligned to one of the jets, when relativistic effects reach their maximum.

The rapid variability observed in gamma-rays ($\leq$ days) indicates that the high-energy emission arises from regions very near the central engine of the galaxy, where jet formation and the acceleration of relativistic plasma occurs, phenomena that are not yet well understood. The gamma rays therefore constitute a new direct probe of the inner-jet region, heretofore unobservable, even by the techniques of very long baseline radio interferometry. Relativistic beaming plays a critical role in the explanation of the luminous time-variable gamma-ray emission from blazars, allowing rapid variations in the observed high-energy emission, without the penalty of severe attenuation of the radiation due to the pair-production opacity. Shocks and instabilities in the bulk flow of relativistic particles can lead to variable nonthermal emission over a wide range of energies. Given the broadband nature of the emission from blazars it has become increasingly clear that coordinated multiwavelength observations of flares from blazars offers the prospect of determining the physical structure and properties of the inner-jet region. A key ingredient of such multiwavelength campaigns is the ability to measure time delays, between wavebands, of the brightness variations occurring during a flare. The relative order and delay of these frequency-dependent variations differ according to the predictions of the various models.

The emission mechanism most often employed to model gamma-ray production in AGNs is the Compton scattering of lower-energy photons by high-energy electrons. The scattered "seed" photons are typically either synchrotron photons generated within the jet by the electrons themselves (the so-called Synchrotron Self-Compton, or SSC, model) or photons propagated directly or scattered into the jet from the accretion disk surrounding the central engine (labeled "external" Compton scattering). The broadband spectra predicted by these models present a characteristic double-peaked appearance (see Fig. 10). The low-energy peak, broadly centered in the millimeter radio to UV bands, is representative of the low-energy photons (usually of synchrotron origin) that are then Compton-scattered to higher gamma-ray energies via the relativistic electrons. Of particular interest in this regard are the recent measurements of TeV gamma rays from relatively nearby BL Lac objects (most notably MRK 421 and MRK 501, both at $z \sim 0.03$) with ground-based Cherenkov detectors. These detections have revealed high-energy emission extending up to 30 TeV, combined with the most rapid variability (flux-doubling times of less than 15 min!) observed to date at gamma-ray energies. Further confirming TeV measurements of gamma-ray AGN are eagerly anticipated to place more precise observational constraints on the models described above.

Finally, it is important to note the utility of gamma-ray blazars as cosmological probes of the extragalactic diffuse background radiation, sometimes referred to as the extragalactic background light (or EBL). The propagation of high-energy gamma rays through intergalactic space is fundamentally limited by the pair-production opacity of the intervening medium. That is, the likelihood of interaction of the gamma radiation with the lower-energy photons that make up the universal background radiation. Though the cosmic microwave background radiation, relic of the Big Bang, has been well measured, background

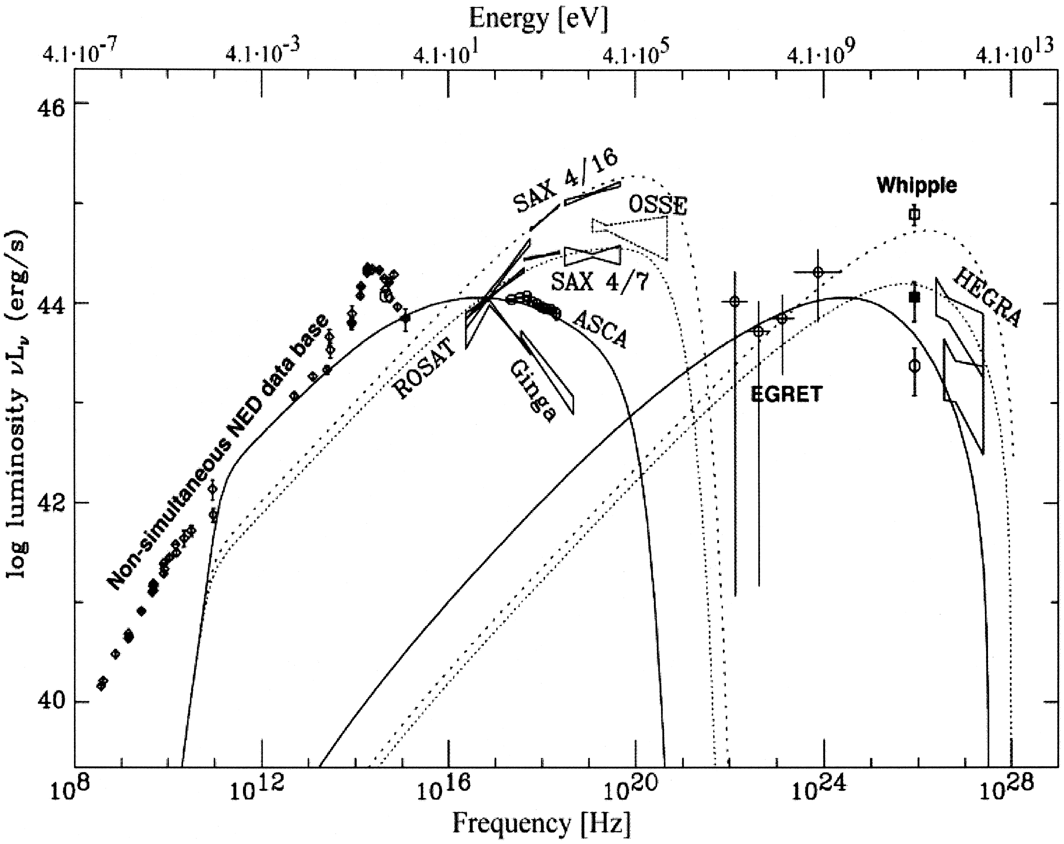


FIGURE 10 Broadband multiwavelength spectrum of the gamma-ray blazar MRK 501 from radio to TeV gamma-ray energies. The broadband spectrum shows the double-peaked structure characteristic of gamma-ray blazars in which lower energy photons are Compton scattered to gamma-ray energies by relativistic electrons. (Adapted from Kataoka, J. *et al*. (1999). "High-energy emission from the TeV blazar Markarian 501 during multiwavelength observations in 1996," *Astrophys. J.* **514,** 138–147, Copyright 1999, reproduced with permission of the AAS.)

radiation fields at other wavelengths are not nearly as well known, and yet are of considerable interest since they provide information about earlier epochs in the history of the Universe. TeV gamma rays, for example, are most likely to interact with infrared and optical photons. The optical/IR background radiation fields therefore limit the distance to which TeV gamma-ray sources can be detected. Observations of gamma-ray blazars as a function of redshift, then, provide a means of estimating the intensity of the EBL at the lower optical and IR wavelengths that provide information on star and galaxy formation in the early Universe.

G. Gamma-Ray Bursts

As their name implies, cosmic gamma-ray bursts (GRBs) are intense bursts of gamma radiation, lasting from fractions of a second to minutes, which emit the bulk of their energy in the gamma-ray regime (above ∼0.1 MeV). Unpredictable in occurrence, these transient events form one of the most long-standing and challenging puzzles in modern astrophysics, dating back to their accidental discovery over thirty years ago with the Vela series of nuclear-testing surveillance satellites. The Burst and Transient Source Experiment (BATSE) aboard the CGRO was specifically designed to serve as an all-sky monitor to detect these mysterious events. Over its nine-year lifetime BATSE detected a total of 2704 GRBs, many times the number recorded previously, and amassed a formidable collection of data on their properties. During their brief appearance GRBs are the brightest objects in the sky, outshining all other gamma-ray sources combined. Indeed, gamma-ray bursts may be the most distant and explosive events (with energies greater than 10^{53} ergs) ever observed in Nature. GRBs occur at random intervals (∼1/day); they do not seem to repeat (implying likely destruction of the source); and they are isotropic in their distribution, coming from every direction in the sky. Each burst is different in its temporal structure, which can vary dramatically in duration and complexity from burst to burst (see Fig. 11). Further, the rapid variability of the emission (on the order of

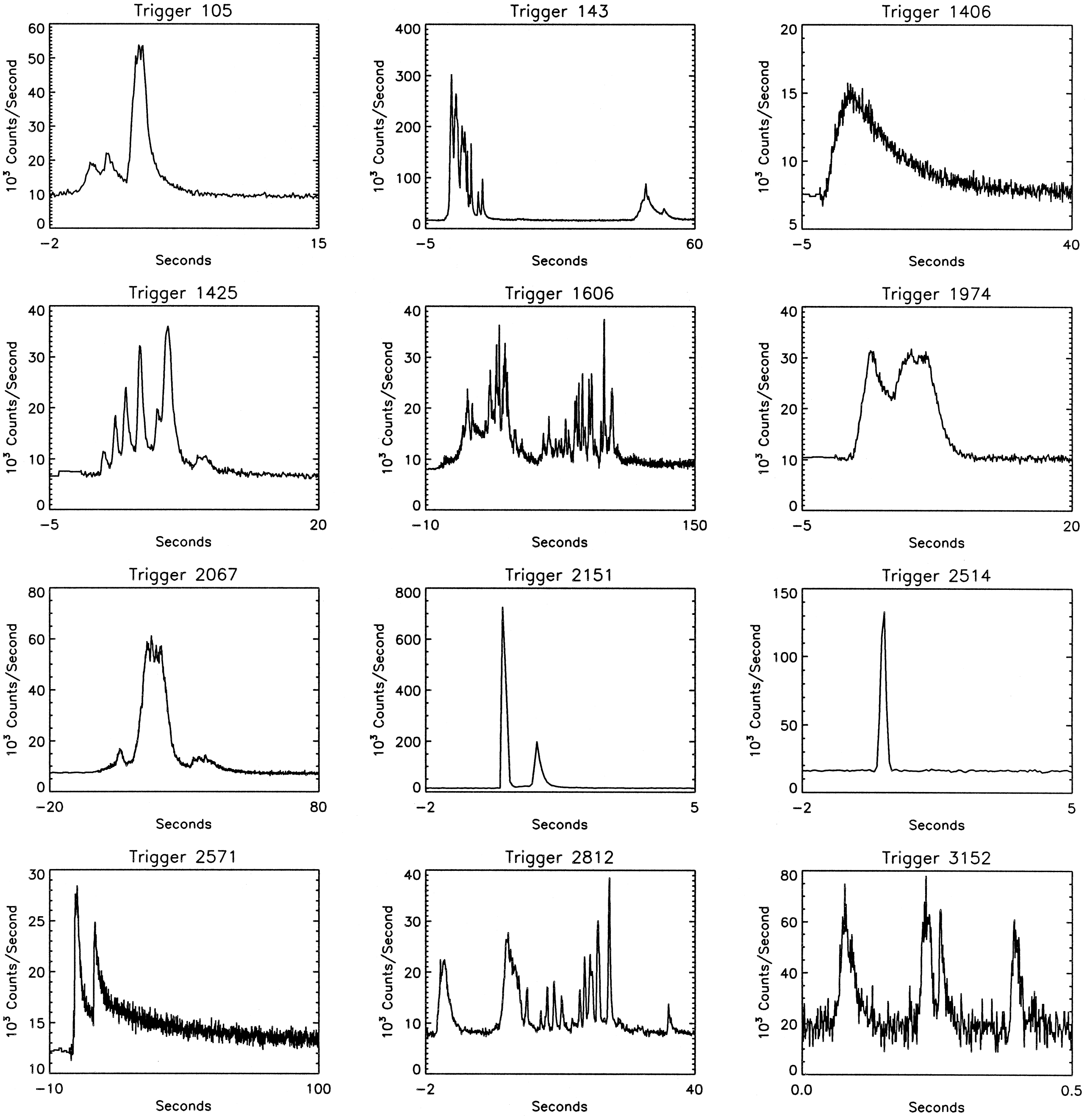


FIGURE 11 Time profiles for a sample of gamma-ray bursts (GRBs), obtained with the BATSE instrument aboard the Compton Gamma Ray Observatory. These profiles illustrate the rich diversity in temporal structure, intensities, and durations for gamma-ray bursts, no two of which are exactly alike in all respects. (Courtesy NASA and the CGRO BATSE Instrument Team.)

milliseconds) measured during a given burst implies that the observed radiation arises from an extremely compact source, requiring the relativistic expansion of the emitting particles to avoid the photon–photon pair-creation opacity that would otherwise quench the observed gamma radiation. This rapid expansion of emitting material could take the form of a relativistic fireball resulting from an initial explosion or from collimated beams of relativistic jets emanating from a central source.

For many years the mystery of the origin of gamma-ray bursts was compounded due to their lack of detection in any frequency band below the hard X-rays, a fact difficult to reconcile with such an apparently catastrophic release of energy. Further complicating the situation was that few, if any, observational constraints could be placed on the distances to the sources of bursts, nor could GRBs be associated with any known class of object. Consequently, the distance scale to bursts was studied indirectly through the angular distribution on the sky and the intensity distribution of the overall burst ensemble. The observed deficit in the number of weak bursts detected with BATSE (compared to the number expected for a uniform, homogeneous distribution of burst sources in flat three-dimensional Euclidean space) implied, for example, that we were seeing the far "edge" of the burst source population. Still, this allowed for burst sources to be either "local" to our own Galaxy, or "cosmological" at the edge of the observable Universe, leading to endless variety in proposed burst models and widespread debate and controversy within the field of burst studies.

It had long been recognized that the key to unraveling the GRB mystery was the identification of burst counterparts at other wavelengths. The short duration of gamma-ray bursts, however, combined with the unpredictability of their occurrence anywhere in the sky, and the relatively poor location-determination capabilities of gamma-ray instruments (compared to detectors operating in other wavebands) severely limited the effectiveness of coordinated follow-up searches for GRB counterparts. Despite repeated and valiant attempts at many wavelengths to detect either the remnant afterglow of a burst event, or of a quiescent counterpart within a gamma-ray burst error box, none was detected for many frustrating years. An observational breakthrough occurred in early 1997, however, shortly after the launch of the Italian-Dutch BeppoSAX X-ray satellite. On a number of occasions BeppoSAX observed with one of its Wide-Field Cameras (WFCs) the unmistakable X-ray afterglow associated with a gamma-ray burst. The occurrence of a burst was simultaneously registered with the cesium–iodide (CsI) detectors that served primarily as shields for the satellite's main X-ray telescope, but were also configured to operate as a separate gamma-ray burst monitor. Following a suspected burst event the spacecraft was reoriented to allow its high-resolution X-ray instruments to observe the same field, confirm the fading X-ray emission, and to exactly fix its point of origin (see Fig. 12). After rapid communication of precise coordinates, followup observations were immediately initiated by observatories around the world, resulting in the first detections of fading afterglow emission, lasting from hours to weeks, from a gamma-ray burst. In a few instances, further observations of the fading optical source revealed the presence of faint, extended underlying emission, suggestive of very distant host galaxies. Measurement of a host galaxy's redshift has provided the first incontrovertible evidence that GRBs occur at cosmological distances. In the 3 years following the initial detection of a burst counterpart, over thirty GRBs have been rapidly localized by BeppoSAX, with over a dozen events yielding counterpart detections in other wavebands, including X-ray, optical, infrared, millimeter, and radio, as well as redshift measurements to likely host galaxies (of characteristic redshift $z \sim 1$). These results electrified investigators in the field of gamma-ray burst research, and have completely revolutionized the study of GRBs.

The detection of both prompt and afterglow counterpart emission has permitted a serious revaluation of the many theories put forward to explain the origin of GRBs. In particular favor is the "relativistic fireball" model that posits an enormous, instantaneous energy release within a small volume. In this scenario a relativistic fireball consisting of an electron-positron pair plasma with a Lorentz factor of 100–1000 propagates outward following the initial energetic event. The temporal structure observed in the gamma-ray burst itself results from the collision of shocks with somewhat different Lorentz factors within the relativistic outflow (so-called "internal" shocks). Prompt burst emission (e.g., the dramatic optical flare seen by the robotic ROTSE telescope during GRB 990123) is attributed to "reverse" shocks traveling backward through the dense ejecta. The longer-term afterglow emission arises from forward-moving "external" shocks plowing through the surrounding medium. The relativistic particles accelerated in these shocks radiate both synchrotron and Compton emission. The afterglow spectrum observed hours to weeks following a burst typically consists of a series of synchrotron power laws whose breaks are dependent on the energy distribution in the population of radiating nonthermal electrons. With time, the observed synchrotron spectrum shifts to lower energies as a result of the cooling expansion and the decreasing Lorentz factor of the bulk flow. The delayed onset and characteristic smooth decay observed for the counterpart emission at X-ray, optical, and radio wavelengths can be naturally explained as arising from the expanding pair plasma that becomes progressively more optically thin to lower-frequency radiation during the later stages of the cooling fireball. The fireball model agrees remarkably well with the observations, though the exact details of the broadband emission depend critically on the physical properties of the interstellar or intergalactic environment into which the fireball expands.

A fundamental question is the nature of the energetic event that sparks the detonation of the fireball itself. The enormous energies ($>10^{53}$ ergs) implied by the cosmological distances inferred from the most recent observations of GRBs greatly restricts the number of possibilities. Only the mergers of the components of evolved binary systems

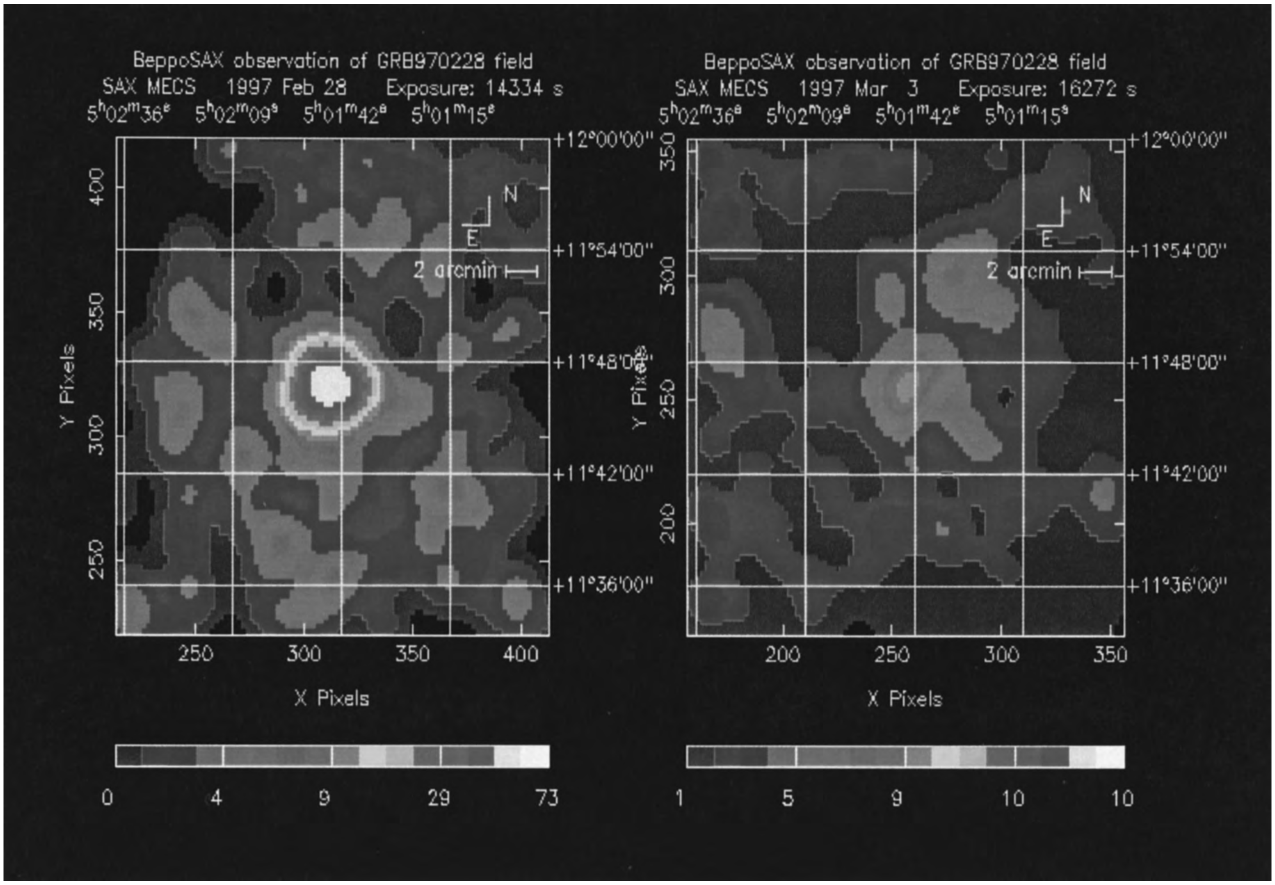

BeppoSAX observation of GRB970228 field
SAX MECS 1997 Feb 28 Exposure: 14334 s
5h02m36s 5h02m09s 5h01m42s 5h01m15s
+12°00'00"
+11°54'00"
+11°48'00"
+11°42'00"
+11°36'00"
N
E
2 arcmin
Y Pixels
400
350
300
250
X Pixels
250
300
350
400
0
4
9
29
73
BeppoSAX observation of GRB970228 field
SAX MECS 1997 Mar 3 Exposure: 16272 s
5h02m36s 5h02m09s 5h01m42s 5h01m15s
+12°00'00"
+11°54'00"
+11°48'00"
+11°42'00"
+11°36'00"
N
E
2 arcmin
Y Pixels
350
300
250
200
X Pixels
200
250
300
350
1
5
9
10
10

FIGURE 12 The discovery images of the first X-ray afterglow detected from a gamma-ray burst (GRB 970228), observed with the BeppoSAX satellite. The left panel shows bright X-ray emission from the location of the GRB shortly after burst occurrence on February 28, 1997, while the right panel shows the fading X-ray afterglow seen several days later on March 3, 1997. (Adapted from Costa, E. *et al*. (1997). "Discovery of an X-ray afterglow associated with the gamma-ray burst of 28 February 1997," *Nature* **387,** 783–785, Copyright 1997, by permission.)

containing pairs of compact objects (neutron stars or black holes), or the collapse of the most massive stars at the ends of their lifetimes (as described in the "hypernova" and "collapsar" models), appear to meet the gigantic energy requirements. The recent detection of iron-line emission at X-ray wavelengths in GRB afterglows with instruments aboard BeppoSAX and the Chandra Observatory favors the latter class of models, since the quantity of iron estimated from the observations seems most likely to have originated in an evolved massive star that exploded in a supernova-like event.

Virtually all models for the ultimate energy source of GRBs involve an endpoint of stellar evolution, particularly of the most massive stars. Thus it has been proposed that the burst rate must be proportional to the overall cosmic star formation rate. This view is supported by the fact that the typical redshifts ($z \sim 1$) associated with GRB host galaxies correspond to an epoch of early active star formation in the Universe. Burst counterparts also tend to be found in the outer regions of blue galaxies undergoing recent star formation, or in irregular galaxies that may have undergone recent collisions or mergers, promoting a burst of star-forming activity at an early epoch.

H. The Extragalactic Diffuse Gamma-Ray Emission

Studies of the Galactic diffuse gamma-ray emission revealed the presence of an isotropic component of diffuse emission that is now considered to be extragalactic in origin. The existence of extragalactic diffuse gamma rays was first demonstrated with the SAS-2 satellite and confirmed on a large scale with the all-sky mapping carried out with the COMPTEL and EGRET instruments aboard the CGRO. Sometimes referred to as the cosmic diffuse gamma-ray (CDG) background, this radiation has been measured from ~1 MeV to ~100 GeV in energy. Since it is by definition the constant, isotropic "background" radiation against which all other sources are measured, it is not characterized by any spatial or temporal signature.

The determination of the extragalactic diffuse gamma-ray background presents a particular observational challenge. This stems from the fact that the basic procedure for determining it involves the subtraction from the accumulated gamma-ray signal of all *other* known sources of emission, including the contribution of point sources, the Galactic diffuse emission, and the effects of the instrumental background. In other words, the extragalactic diffuse emission is assumed to be what is "left over" after all other sources of emission are accounted for and removed from the observations. Thus, by necessity it requires a profound understanding of the instrumental response, and a detailed knowledge of the sources and structure of the gamma-ray sky. Consequently, it is usually the last major result obtained with a gamma-ray telescope, often requiring years of observation, analysis, and study.

In this regard, the COMPTEL measurements of the extragalactic diffuse emission represent a major achievement. The MeV energy band over which COMPTEL was sensitive has long been recognized as a notoriously difficult one in which to operate, due to the high levels of instrumental background to which experiments are susceptible. As the first MeV telescope launched into space for extended observations, COMPTEL was in a unique position to provide a definitive measurement of the CDG at medium gamma-ray energies. The COMPTEL results were particularly anticipated in light of a series of earlier, shorter measurements that had indicated the existence of a puzzling excess, or "MeV bump," in the diffuse spectrum between ∼2 and 9 MeV that could not be readily explained and which had provoked widespread discussion as to its possible origin. Initial COMPTEL results were derived from observations of the Virgo region, well removed from the Galactic plane. In contrast with the earlier observations, the COMPTEL measurements provided *no* hint of the MeV bump previously reported. Later confirming observations were obtained in the direction of the South Galactic Pole, and these also did not indicate any excess emission, or any temporal variation or anisotropy in the diffuse gamma radiation. Most significantly, then, COMPTEL disproved the earlier reports of an MeV excess, and provided a first definitive measurement of the cosmic diffuse radiation between ∼1 and 30 MeV. At 30 MeV the COMPTEL results join smoothly with an extrapolation of the EGRET measurements at higher energies (see Fig. 13).

The EGRET results were obtained following the analysis of data from 36 distinct regions of the sky well removed from the contaminating influence of the Galactic disk and bulge. From 30 MeV to 100 GeV the extragalactic diffuse emission is well described by a strikingly smooth power-law photon spectrum of index-2.1, in remarkable agreement with the average spectrum derived from the large pool (>90) of gamma-ray blazars detected with EGRET over its lifetime. This immediately suggests that the extragalactic emission at EGRET energies most likely arises from numerous unresolved sources of the blazar class.

The origin of the extragalactic diffuse emission has been the subject of much theoretical speculation for many years. Theories of truly diffuse processes include matter–antimatter annihilation in a baryon-symmetric cosmology, the evaporation of primordial black holes, supermassive black holes ($\sim 10^6$ solar masses) that collapsed at high redshifts ($z \sim 100$) at very early epochs, and the annihilation of exotic supersymmetric particles. Unfortunately, many of these hypotheses cannot be tested realistically at the sensitivity of the current observations. The general consensus at present is that the extragalactic diffuse radiation most likely results from the superposition of a number of classes of unresolved point sources. The CDG, since it represents the integrated emission from sources extending back to the earliest cosmological epochs, provides an important observational constraint for theoretical models describing source evolution with time.

In the COMPTEL energy range around ∼1 MeV it has been proposed that a significant fraction of the diffuse emission could arise from gamma-ray lines due to the decay of radionuclides produced in the course of supernovae of types Ia and II in distant, unresolved galaxies. Other possible sources at low MeV energies include active galaxies such as those of the Seyfert I and II class. At the higher gamma-ray energies observed with COMPTEL and EGRET, blazars provide the most likely explanation of the observed diffuse emission. Of note in the EGRET band is the fact that the diffuse spectrum clearly extends up to 100 GeV without significant deviation, implying that the quiescent emission from gamma-ray blazars, assuming that they are responsible for the observed emission, must necessarily extend up to this energy as well. Consequently, the relativistic particles that give rise to gamma-ray emission in blazars must extend to even higher energies. It is anticipated that the next generation of gamma-ray telescopes with higher resolution and sensitivity will begin to detect and map out the individual sources responsible for the extragalactic diffuse gamma-ray emission.

I. Unidentified Gamma-Ray Sources

When any region of the electromagnetic spectrum is opened to new investigation, or detectors suddenly achieve a marked improvement in sensitivity, an exciting era of discovery inevitably ensues. This is particularly true in the gamma-ray regime, where the launch of each new satellite mission has resulted in the discovery of previously

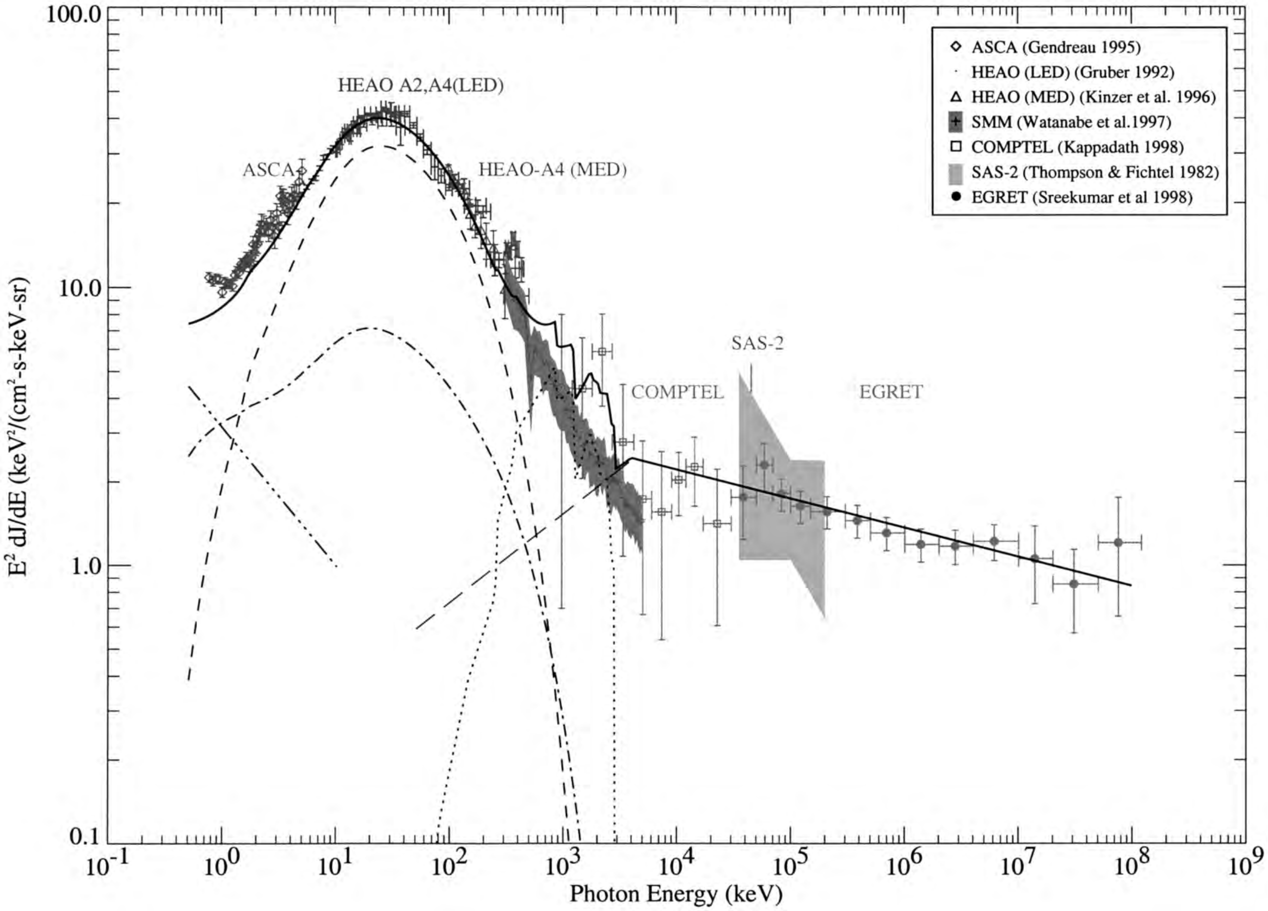


FIGURE 13 Broadband multiwavelength spectrum of the extragalactic diffuse emission from X-ray to gamma-ray energies combining the observations of several high-energy satellite missions, most recently those of the COMPTEL and EGRET telescopes aboard the Compton Gamma Ray Observatory. The dashed and dotted lines indicate the estimated contribution of several classes of unresolved point sources to the observed emission. (Adapted from Sreekumar, P. *et al*. (1998). "EGRET observations of the extragalactic gamma-ray emission," *Astrophys. J.* **494,** 523–534, Copyright 1998, reproduced with permission of the AAS.)

unknown sources of gamma radiation. A large fraction of the gamma-ray sources discovered to date, however, remain unidentified. Indeed, one of the great continuing mysteries of gamma-ray astronomy is the nature of the unidentified sources, which in the Third EGRET Catalog outnumber all other known gamma-ray sources combined (∼170 of the 271 cataloged 3EG sources are unidentified, see Fig. 3). These unidentified sources have no obvious counterpart at other wavelengths, and cannot be clearly associated with any other known class of object based on their high-energy emission properties.

The unidentified gamma-ray sources can be broadly categorized by their location in relation to the plane of the Galaxy, and by their spectral and temporal properties. The sources at high Galactic latitudes, well removed from the disk of the Galaxy, are most likely to be extragalactic objects, presumably galaxies. This supposition is supported by the variable emission often observed from these objects, characteristic of flaring active galaxies. The positional error boxes of these high-latitude unidentified sources, however, do not contain the bright, radio-loud galaxies of the blazar type that are typically associated with extragalactic gamma-ray sources. This suggests that some previously unrecognized class of radio-quiet or active galaxy can give rise to gamma radiation. The detection of gamma rays from the nearby radio galaxy Centaurus A lends support to this hypothesis.

The vast majority of the unidentified gamma-ray sources (>120) appear to belong to a Galactic population. Many of these are likely to be obscured objects whose exact properties are masked by the Galactic diffuse radiation, or cannot be accurately distinguished from those of other nearby sources in the crowded disk of the Milky Way. Gamma rays from the unidentified sources are presumed to arise from the same objects and processes that generate gamma radiation elsewhere in the Galaxy. The objects and phenomena most closely linked to the production of high-energy gamma rays in the Milky Way include (1) molecular clouds within the spiral arms and disk (where high-energy cosmic rays interact to produce gamma-rays); (2) supernova remnants (whose shock waves are a presumed site of particle acceleration and cosmic-ray production, leading to gamma-ray emission in their vicinity); (3) flares, winds, and outflows (within which gamma-rays can be produced) from massive and evolved stars; (4) relativistic jets and accretion disks associated with compact binary systems containing neutron stars or black holes; and (5) radio-quiet pulsars (such as Geminga) in whose intense magnetic fields particles are likely to be accelerated, leading to gamma-ray production. Efforts have been made to study the statistical properties of the unidentified sources and to correlate them with the possible source classes outlined above, though with limited success to date.

It has been noted that the unidentified Galactic sources can be separated into two apparently distinct populations: brighter sources appear confined more closely to the Galactic disk, while fainter sources are found to lie at medium Galactic latitudes (greater than about 5°). The bright sources in the disk are interpreted as more distant luminous objects, while the fainter sources may be much closer and more local to the Sun. Several investigators have proposed that these weaker, medium-latitude sources may be associated with Gould's Belt in the Galaxy. Known for well over a century from optical observations, Gould's Belt is a ring of bright massive O and B stars that define a great circle on the sky inclined by about 20° to the Galactic plane, tilted toward negative Galactic latitudes in the direction of Orion, and toward positive latitudes in Ophiucus. Observations over the years at optical, radio and other wavelengths have established that a slowly expanding ring of material (with an expansion age of about 30 million years) is interacting with and compressing the ambient interstellar gas along the periphery of Gould's Belt. This expansion has likely contributed to periods of enhanced star formation along the boundary of the Belt. The ring of expanding material is presumed to have resulted from a supernova explosion, thus explaining the origin of the local "superbubble" or evacuated region in the interstellar medium within which the Sun currently resides. In the context of this scenario, the most prominent grouping of weak, unidentified gamma-ray sources at medium latitudes lie on the boundary of Gould's Belt that is closest to the Sun, at a distance of about 100–400 pc in the direction of Ophiucus. The exact physical nature of individual sources remains to be determined, but is expected to be found among the list of candidates cited above.

A definitive determination and classification of the unidentified gamma-ray sources discovered to date must await the launch of the next generation of gamma-ray telescopes whose development is currently underway. While investigators are confident that the mystery of the unidentified sources will ultimately be resolved, there still remains the exciting prospect that some of the unidentified gamma-ray sources may in fact represent truly new types of high-energy cosmic sources previously unknown to science. The possibility that such fundamental discoveries remain to be made provides heightened motivation for the continued study of these mysterious objects.

III. THE CHALLENGE OF OBSERVATION AND TECHNIQUES OF DETECTION

As an observational discipline gamma-ray astronomy has always been extremely challenging. The very low fluxes of cosmic gamma-rays, combined with their high energy, necessitate the construction of large complex detectors. Since cosmic gamma rays are severely attenuated by the Earth's atmosphere, the experimental apparatus must be lifted above as much of this absorbing medium as possible by high-altitude scientific balloon or placed in low-Earth orbit via satellite. Either of these possibilities places severe constraints on the size and weight of a gamma-ray experiment. Further, since the detector must be designed to conduct observations quasi-autonomously, and since sophisticated analytical techniques must be employed to accurately identify true cosmic gamma-ray events from the multitude of background interactions that can masquerade as such, the successful operation of a gamma-ray telescope presents unique challenges to the experimental astrophysicist.

A. The Interaction of Gamma Rays with Matter

To observe a cosmic gamma ray, one must first devise a means to stop it (or at least "to slow it down" in some measurable way) within a detecting medium. In the gamma-ray regime, the primary interaction mechanisms of photons with matter are the photoelectric effect, the Compton effect, and pair production. Extensive analyses are available of these fundamental interaction processes. Here, we only briefly review their basic physical characteristics.

1. The Photoelectric Effect

Photoelectric absorption occurs when an incident photon is completely absorbed in an atomic collision with practically all of its energy transferred to an atomic electron, which is ejected. For photoionization to occur, the photon must necessarily have enough energy to overcome the binding energy of the liberated electron. A small, usually minimal, fraction of the initial photon energy is converted into the recoil of the atomic nucleus, required for the conservation of momentum. As a consequence of this conservation requirement, it is the most tightly bound innermost (K-shell) electrons that have the greatest probability of interacting with photons of energies approaching those of gamma rays. The cross section for photoelectric absorption has a strong dependence on atomic number (proportional to Z^5) and an inverse dependence on photon frequency (proportional to $\nu^{-7/2}$). This is valid for photon energies somewhat removed from the so-called "absorption edges," those energies at which strong rises in the absorption cross section occur as photon energies approach the binding energies of individual electron shells.

2. The Compton Effect

As mentioned previously, the Compton scattering process consists of interactions between photons and free electrons. In "standard" Compton scattering, an incident high-energy photon loses a fraction of its energy to an electron, while in the "inverse" Compton process it is the photon that gains energy at the expense of the electron. By applying the standard criteria of conservation of momentum and energy, it is a straightforward exercise to derive the Compton wavelength shift for the scattered photon,

$$\Delta\lambda = \lambda' - \lambda = h/m_e c(1 - \cos\theta) = \lambda_c(1 - \cos\theta),$$

where $\lambda_c = h/m_e c = 2.426 \times 10^{-10}$ cm is referred to as the Compton wavelength. Related expressions are easily obtained for the resultant electron and photon energies. The cross section for Compton scattering by an electron is given by the Klein–Nishina formula, and declines with increasing photon energy. There is a geometric dependence to the scattering probability, and at very high energies the cross section rapidly becomes strongly peaked in the forward direction.

3. Pair Production

In the pair production process, an incident gamma ray of sufficiently high energy is annihilated in the Coulomb field of a nearby charged particle, resulting in the creation of an electron–positron pair. As with the photoelectric effect, conservation of momentum requires the presence of a third body which takes up the balance of the momentum in the form of a recoil. For a recoil nucleus of mass much greater than the mass of the electron, the threshold energy for pair production is simply, $E_{\text{th}} \sim 2m_e c^2 = 1.022$ MeV. The total cross section for pair production rises rapidly above threshold and approaches at strongly relativistic energies a limit that is roughly independent of energy.

4. Attenuation and Absorption of Gamma Radiation

The relative importance of the interaction mechanisms described above as a function of photon energy is represented graphically in Fig. 14. The cumulative effect of these primary interaction processes is usually expressed in terms of attenuation coefficients. The probability that an individual photon will traverse a given amount of absorbing material without interaction is simply the product of the probabilities of survival for each type of possible interaction. Given a collimated beam of monoenergetic gamma rays of initial intensity I_0, the intensity observed after passage through a thickness x of absorber is given by

$$I(x) = I_0 e^{-\mu x},$$

where $\mu = \kappa_{\text{PE}} + \kappa_{\text{C}} + \kappa_{\text{PP}}$ is the total linear attenuation coefficient (units: cm^{-1}), and κ_{PE}, κ_{C}, and κ_{PP} are the attenuation coefficients for the photoelectric, Compton and pair-production processes, respectively. The linear attenuation coefficient is related to the cross section σ by an expression of the form $\kappa = \sigma N$ cm^{-1}, where N is the number density of atoms in the target medium. Of more fundamental usefulness in terms of detector design is the *mass attenuation coefficient*, which is simply the linear coefficient divided by the density of the given material (units: cm^2/g), permitting straightforward comparison of the attenuation properties of various types of materials. Numerous tables and graphs of mass attenuation coefficients as a function of energy for the most commonly used detector materials are readily available.

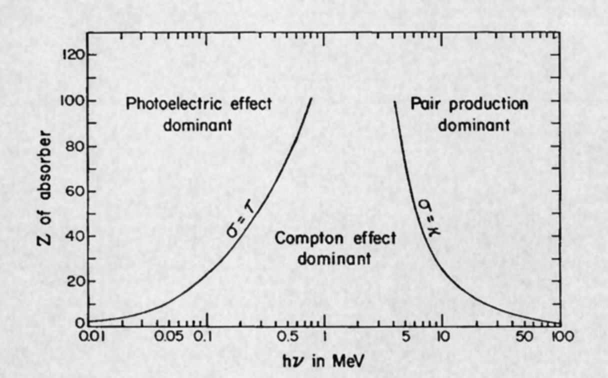


FIGURE 14 The relative importance of the major processes by which gamma rays interact with matter as a function of photon energy for elements of varying atomic number Z. (Adapted from Evans, R. D. (1995). "The Atomic Nucleus," Copyright 1955, McGraw-Hill, by permission.)

Some common detector materials include plastic organic scintillators (i.e., hydrocarbon compounds) typically used in charged-particle anticoincidence shielding, inorganic crystal scintillators such as sodium iodide (NaI), cesium iodide (CsI), and bismuth germanate oxide (BGO), and solid-state semiconductors such as germanium (Ge), and cadmium zinc telluride (CdZnTe) often used in high-resolution spectroscopic applications. As their name implies, scintillators are transparent materials that scintillate in visible light when high-energy photons or particles interact within them. The scintillation light, proportional to the amount of energy deposited, is collected by attached photomultiplier tubes and converted to an electronic pulse for further signal processing and recording. In semiconductor detectors interacting gamma rays deposit energy in the detector material, creating electron-hole pairs that are collected following application of an electric field to the material. Semiconductors are far superior spectroscopically to scintillators, due to their relatively low threshold for electron-hole creation ($\sim$3 eV for Ge). CdZnTe has the advantage that is a room-temperature semiconductor, while germanium, due to its narrow band gap, must be operated at cryogenic temperatures for optimal performance. The technology associated with the detection of cosmic gamma rays is very often derived from experimental techniques developed originally for application in the fields of high-energy and nuclear physics.

B. Sources of Instrumental Background

Gamma-ray telescopes, due to their physical size and mass, are particularly susceptible to background radiation and particle interactions. Below a few MeV in energy, there is a high background due to continuum emission and nuclear decay lines, arising primarily from cosmic rays interactions in and around the detector. Above a few MeV the background is somewhat less severe, but the photon fluxes are much weaker, still presenting a sizable signal-to-noise problem. Gamma-ray telescopes of necessity must combine high sensitivity with superior background-suppression capability. The major sources of experimental background in gamma-ray detectors are summarized below.

1. Atmospheric Gamma Rays

Gamma rays are produced in copious quantities in the upper atmosphere of the Earth as a consequence of cosmic-ray interactions. Balloon experiments rely typically on the "growth curve" technique to estimate the contribution of atmospheric gamma rays to the observed event count rate. In this method, the total count rate of the detector is determined as a function of the residual atmosphere remaining above the balloon-borne instrument as it rises to float altitude. Since the downward vertical atmospheric gamma-ray flux is assumed to be zero at the top of the atmosphere, all remaining event counts are assumed to be truly cosmic in nature (or locally produced within the experiment itself, see following). Both Monte Carlo calculations and semiempirical models are employed to test the reliability of such measurements.

2. Intrinsic Radioactivity

An ubiquitous source of background gamma radiation within a detector is the decay of naturally occurring radioisotopes contained within the structure of the instrument itself. The two most common naturally occurring background lines result from the presence of the long-lived isotopes ^{40}K and ^{232}Th in many detectors and structural materials. ^{40}K, with a half-life of 1.26×10^9 years, undergoes decay to ^{40}Ar giving rise to a gamma-ray line at 1.46 MeV, while a 2.62 MeV gamma ray results from an excited state of ^{208}Pb, a daughter product of ^{232}Th ($t_{1/2} = 1.4 \times 10^{10}$ years).

3. Cosmic-Ray and Secondary Charged-Particle Interactions

Background counts can arise in a gamma-ray telescope due to high-energy cosmic-ray collisions (of protons predominantly) with material in or around the detector, often accompanied by secondary particle production. Fortunately, active anticoincidence shielding has proven to be an effective means of screening out direct charged particle interactions within a detector. Of particular concern, however, are high energy ($E \sim 100$ MeV) proton collisions that result in radioactive spallation products which decay with longer characteristic lifetimes, yielding secondary "activation" gamma rays that can interact within the detector long after the original charged-particle event has been vetoed. The gradual build-up of activation species within detectors in low-Earth orbit, for example, must be monitored with care given the long-term exposure of such instruments to high-energy particles from space or trapped in the Earth's radiation belts. Extensive tables have been compiled detailing the dominant induced radioactive states with corresponding lifetimes and energies for common detector materials.

4. Neutron-Induced Background

Atmospheric and albedo neutrons can interact within detectors at balloon altitudes and in low-Earth orbit. Target nuclei can emit secondary particles or gamma rays that contribute to the background counting rate of the instrument. Three basic types of neutron interactions must be taken into account.

(a) Elastic scattering. Neutrons with energies below the first excited level of target nuclei are elastically scattered without excitation of the recoil nucleus. The amount of energy transferred to the recoil nucleus for the intermediate and heavy elements typical of gamma-ray detectors is relatively small, implying that elastic scattering is unlikely to generate a significant background in most cases. Over a series of such collisions, however, a neutron may lose sufficient energy to permit thermal neutron capture, giving rise to secondary gamma rays as outlined in (c) below.

(b) Inelastic scattering. The inelastic scattering of energetic "fast" neutrons (in interactions of the type $(n, n'\gamma)$ or $(n, x\gamma)$, where the "x" particle is a proton or alpha particle) can lead to significant background counts in a gamma-ray detector. The secondary gamma radiation results from the deexcitation of residual product nuclei. Such gamma-ray emission is usually prompt due to the relatively short lifetimes of low-lying nuclear excited states, though delayed activation gamma rays may result if the product nuclide is also radioactive. An important example of activation gamma rays resulting from fast neutron interactions is one involving aluminum, a commonly used structural material. The reaction $^{27}\mathrm{Al}(n, \alpha)^{24}\mathrm{Na}$ yields *two* gamma rays of energies 1.37 and 2.75 MeV following the radioactive decay via cascade of ^{24}Na ($t_{1/2} = 14.96$ h).

(c) Thermal neutron capture. Secondary gamma radiation can also result from the capture of thermal neutrons by nuclei within a detector. In this type of interaction, the secondary gamma emission is prompt and of energy comparable to the binding energy of the nucleus. A prominent example of a neutron-capture background line occurs for organic scintillator detectors, where the capture of thermal neutrons by the proton nuclei of hydrogen atoms leads to the production of 2.22 MeV photons from the deuterium recoil product.

C. Background Rejection Techniques

In order to optimize the cosmic gamma-ray signal measured with a detector, a number of background-suppression techniques are usually employed. The most common are outlined in the following.

1. Passive Shielding

One of the most fundamental techniques for background rejection is the use of passive high-Z materials (usually lead) to shield the central detector elements from unwanted photons and particles. Typically, several radiation lengths of the chosen material are employed to sharply attenuate the background flux. While the concept is simple in principle, weight constraints tend to limit the total amount of material that can be utilized. Additionally, since it is a passive technique, quantitative information as to the effectiveness of the shield during the course of actual measurements is often difficult to obtain.

2. Active Shielding

Active shielding, as the name implies, provides additional information over the passive approach. The active shield, commonly a plastic organic scintillator, has the advantage of being able to trigger a signal upon passage of a charged particle. Thus, active shielding is most commonly used in an anticoincidence mode. In other words, for a detector placed within charged-particle anticoincidence shielding, only those events which trigger the main detector element and *not* the surrounding active shield will be registered as valid events.

3. Pulse-Shape Discrimination (PSD)

One limitation of most active shields is that they are not efficiently triggered by the passage of neutral particles. Since neutron-induced reactions contribute a significant background to gamma-ray telescopes, it is imperative to devise a scheme for eliminating the neutron component of the detector signal. This neutron rejection is most efficiently achieved by means of pulse-shape discrimination (or PSD). With this technique, one exploits the fact that neutrons which interact in a scintillator crystal give rise to optical pulses that have a fundamentally different time profile compared to signals resulting from photon interactions. Electronic means can then be employed to separate one from the other.

4. Time-of-Flight (TOF) Discrimination

Finally, in telescopes containing multiple detectors one may take advantage of the physical separation between the detector elements to veto certain background events. By precisely timing interactions as they occur in two detectors, one may discriminate between upward-moving (i.e., atmospheric) and downward-moving (celestial) events. By defining a time-of-flight window for allowable events, one may also eliminate "slowly moving" downward particle transitions which could not have resulted from speed-of-light photon interactions. The rejection capability of a TOF system is, of course, limited by the detector-separation distance and the temporal resolution attainable with the available electronics.

TABLE II Gamma-Ray Telescopes by Energy Band

Name	Energy band	Energy (eV)	Telescope type	Instruments
Low energy	keV to MeV	10^5–10^6	Collimated scintillator, Coded-aperture telescopes	HEAO, SMM, GRANAT/SIGMA, CGRO/BATSE, HETE-2, Swift, CGRO/OSSE, INTEGRAL, HESSI
Medium energy	MeV	10^6	Compton telescopes	CGRO/COMPTEL
High energy	MeV to GeV	10^6–10^9	Pair-production telescopes	SAS-2, COS-B, CGRO/EGRET, GLAST
Very high energy (VHE)	GeV to TeV	10^9–10^{12}	Ground-based optical Cherenkov	Whipple, CAT, CANGAROO, HEGRA, Durham, MILAGRO, STACEE, CELESTE, MAGIC, VERITAS
Ultra high energy (UHE)	TeV to PeV	10^{12}–10^{15}	Ground-based optical Cherenkov and particle detection	CASA-MIA, CYGNUS, MILAGRO, HEGRA

D. Types of Gamma-Ray Telescopes

In the descriptions given in the following, refer to Table II for a listing of types of gamma-ray telescopes by energy band.

1. Low-Energy (keV to MeV)

In the low-energy portion of the gamma-ray spectrum photons interact primarily via the photoelectric and Compton processes. Historically, gamma-ray detectors sensitive to this energy range have been collimated, well-type instruments in which the primary detector (often an inorganic scintillator such as NaI or CsI) is surrounded by an active or passive shielding structure. The field of view of the telescope is determined typically by the opening angle of the shields or by passive collimators of some high-Z material, such as lead. The OSSE instrument aboard the CGRO is an example of such a configuration.

Another experimental approach used in the hard X-ray/low gamma-ray regime is that of coded-aperture imaging. The coded-aperture telescope operates essentially as a “multi-pinhole” camera. The telescope employs an absorbing mask to cast a shadow pattern of incident radiation on a position-sensitive detector plane below. Sky images and the positions of sources within the field of view are determined after applying standard deconvolution techniques to the data. To minimize artifacts in the reconstruction process the pattern of the coded mask is usually of a type referred to as a “uniformly redundant array” (or URA) of elements. The angular resolution and source-location accuracy of a coded-aperture telescope ($\sim 10'$) is superior to that of more traditional gamma-ray instruments, and is determined by three factors: the size of the individual mask elements, the mask-detector separation distance, and the spatial resolution attainable in the determination of photon-interaction locations in the central detector. While coded-aperture telescopes have excellent angular resolution, there is a practical upper limit to their sensitive energy range, determined by the thickness of the mask required to fully attenuate the gamma rays of interest. The French SIGMA experiment aboard the Russian GRANAT spacecraft was a coded-aperture telescope, as are both the imaging and spectroscopic instruments aboard the INTEGRAL spacecraft (see Fig. 15).

2. Medium-Energy (MeV)

The medium-energy gamma-ray regime extends from approximately 1 to 30 MeV and is one of the most difficult in which to observe. The background over this energy range is particularly severe, due to nuclear excitation and decay lines and continuum radiation resulting from particle interactions in and around the detector. The Compton process is the most likely interaction mechanism for incident gamma photons in this energy range, and double-scatter Compton telescopes were developed to exploit this fact.

A Compton telescope consists of two separate detector assemblies which register in coincidence the Compton scattering of an incident cosmic gamma ray in an upper detector array followed by the subsequent total absorption of the scattered photon in a lower detector. Hence, *two* simultaneous interactions (for which interaction positions and energies are recorded) are required for the registration of a valid event. The energy and direction of an incident cosmic photon are determined after event reconstruction. Assuming total absorption of the scattered gamma ray, a straightforward application of the Compton scattering formula can be used to determine the Compton scattering angle from the upper to lower detectors of the telescope. The scattering angle and recorded interaction positions within the detectors define an event circle on the sky on which the source of the incident gamma ray must lie. The position of the gamma-ray source therefore is not unambiguously determined from a single event. Rather, numerous events must be recorded, and event circles computed, whose

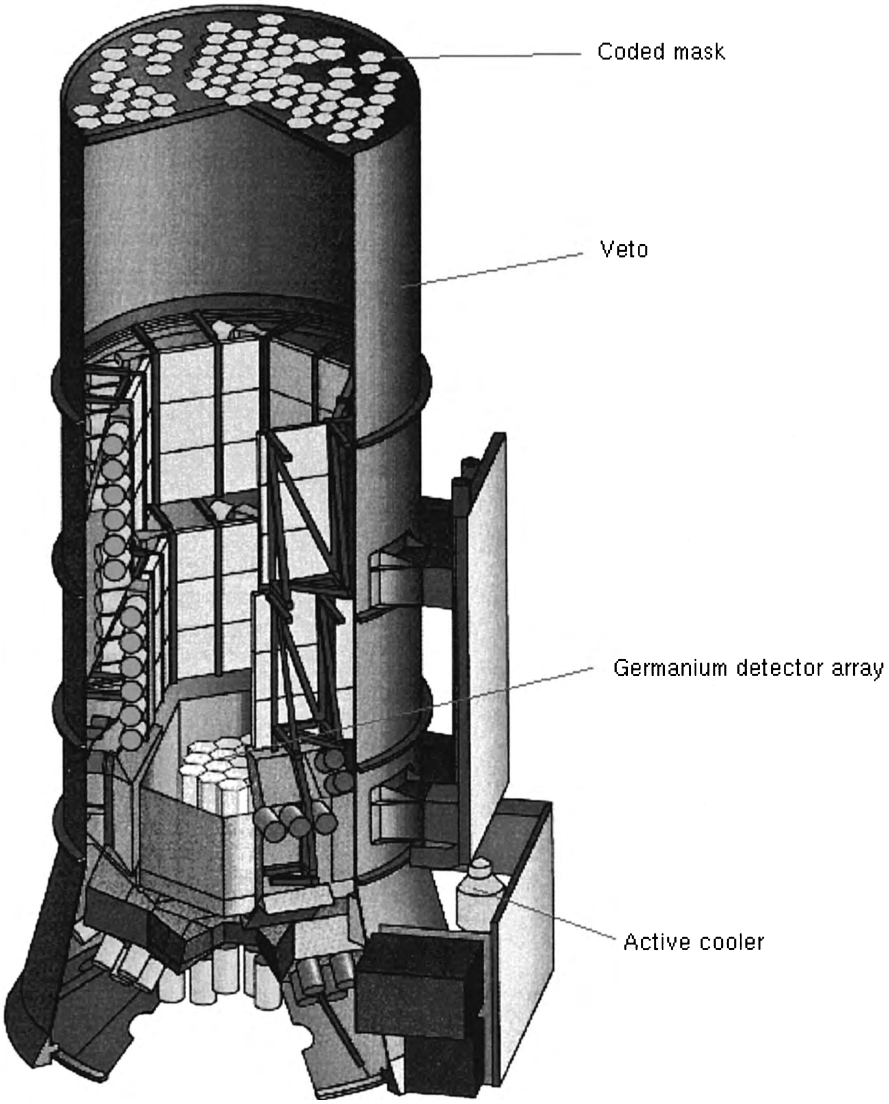


FIGURE 15 Schematic of the INTEGRAL SPI spectrometer showing the main components of this coded-aperture instrument. (Courtesy INTEGRAL SPI Instrument Team.)

common point of intersection reveals the true source location. The partial, rather than total, absorption of incident gamma rays, combined with other measurement errors, introduces uncertainties in source identification. In practice, detailed modeling of the instrumental response to incident radiation is required to accurately identify and characterize cosmic sources.

The COMPTEL instrument aboard the CGRO was a Compton telescope (see Fig. 16). The next generation of Compton telescopes currently under development will seek to remove the scattering angle ambiguity inherent in the traditional design by also recording the energy and direction of the recoil electron in the upper detector from the initial Compton scatter. This will greatly enhance instrument performance and aid in reducing the instrumental background.

3. High Energy (MeV to GeV)

Above approximately 30 MeV in energy, gamma rays interact predominantly via the pair-production process. The traditional method of detecting such photons is through the use of high-voltage spark-chamber grids, interleaved with sheets of a high-density converter material (such as tungsten). In this technique an incident cosmic gamma ray is first converted into an electron–positron pair. As the pair traverses the layers of the gas-filled spark chamber the trajectories of the charged particles are detected via the

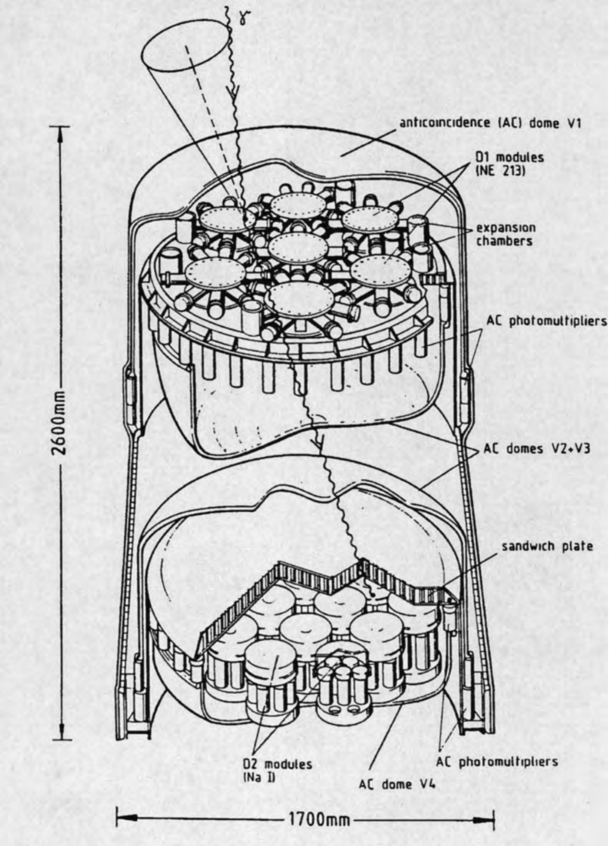


FIGURE 16 Schematic of the Imaging Compton Telescope (COMPTEL) on board the Compton Gamma Ray Observatory illustrating the double-scatter technique employed in Compton telescopes. (Courtesy NASA, the Max Planck Institute, and the CGRO COMPTEL Instrument Team.)

sparks they produce, the positions of which are recorded. A large scintillator below the spark chamber is typically used as a calorimeter to absorb any remaining particle energy. The diverging tracks recorded for the electron and positron are used to determine the incident direction of the original gamma ray. Spark chambers are normally filled with a gas (usually some mixture of neon) that acts as a spark-quenching agent, and which slowly degrades with use. The gas is thus a consumable that must be periodically flushed from the spark chamber and replenished for best performance.

The SAS-2, COS-B, and EGRET experiments all employed the same basic spark-chamber design, albeit on different scales of size and complexity (see Fig. 17). The next generation of pair-production gamma-ray telescope will be the Gamma-Ray Large-Area Space Telescope (GLAST) in which the spark-chamber grids will be replaced by semiconductor silicon strip detectors. GLAST is designed to be sensitive to gamma photons up to ∼300 GeV in energy.

4. Very High Energy (GeV to TeV, and above)

Gamma rays in the energy range above ∼30 GeV have not been well sampled from space to date, given their low fluxes, and ground-based techniques have only recently matured to the point where efficient observations are feasible. As outlined previously, cosmic gamma rays are absorbed in the Earth's atmosphere and are not directly observable from the ground. Photons of energy greater than ∼30 GeV are detectable in principle, however, from the extensive air showers they produce upon interaction in the upper atmosphere. A particular signature of such interactions is the visible pulse of Cherenkov light emitted collectively by the relativistic electrons and positrons produced in these electromagnetic cascades. Ground-based optical telescopes are used to detect this Cherenkov light (see Fig. 18), from which energies and directions of incident cosmic gamma rays can be inferred.

To date the atmospheric Cherenkov technique has been employed with greatest success for photon energies above ∼300 GeV by several groups. The U.S.-based Whipple collaboration (USA–UK–Ireland) has pioneered many of the techniques now commonly applied to TeV

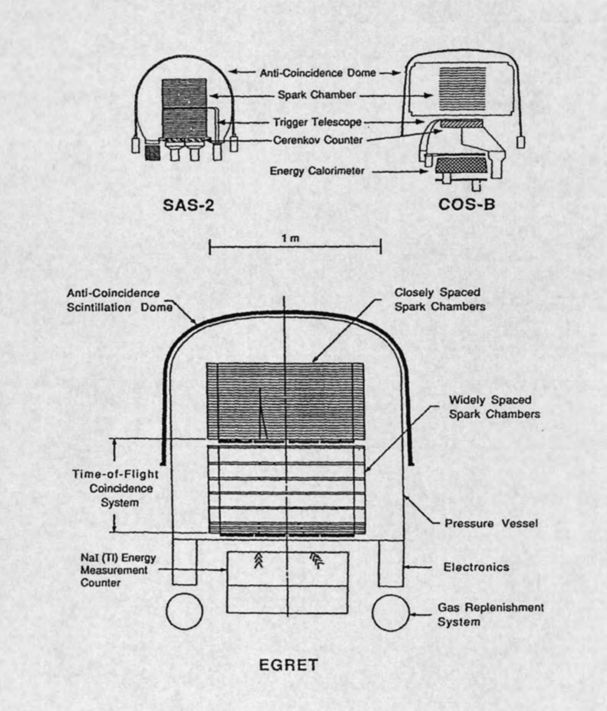


FIGURE 17 Schematic of the EGRET instrument on board the Compton Gamma Ray Observatory illustrating the spark-chamber design characteristic of pair-production telescopes. For comparison, schematics of the earlier SAS-2 and COS-B instruments are shown to scale. (Courtesy NASA and the CGRO EGRET Instrument Team.)

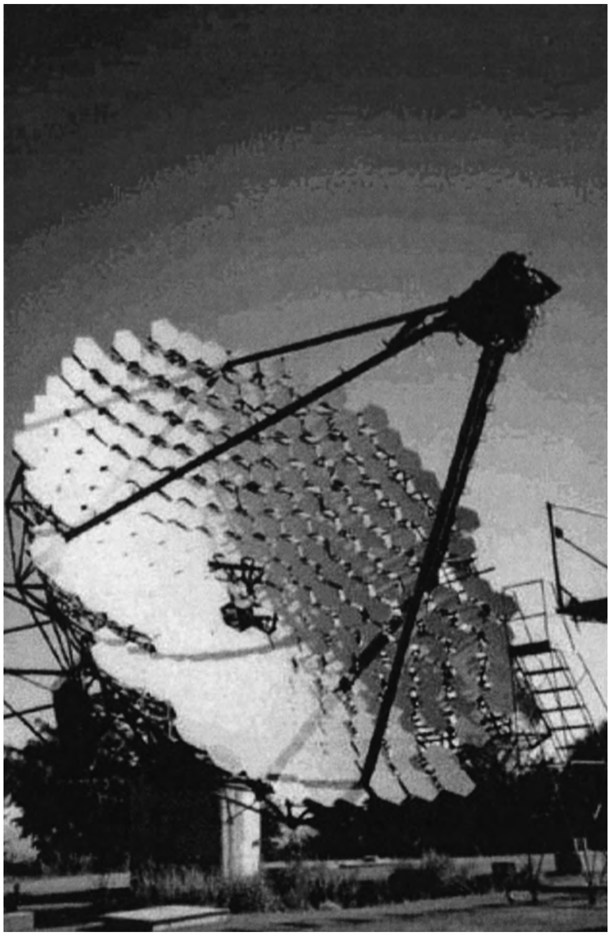


FIGURE 18 Photograph of the Whipple Observatory 10-m imaging atmospheric Cherenkov telescope. (Courtesy of the Whipple Collaboration.)

measurements. Other collaborations actively operating air-Cherenkov telescopes include CAT (in France), the CANGAROO (Japan–Australia) and Durham, UK groups (with telescopes in Australia), and HEGRA (Germany–Armenia–Spain, in the Canary Islands). Among the several TeV sources detected in recent years are the Crab and Vela pulsars, and AGNs such as MRK 421 and MRK 501.

The slightly lower energy domain from approximately 30 to 300 GeV still remains largely unexplored. A lower detection threshold requires a much larger collecting area for the atmospheric Cherenkov light. Two groups, STACEE and CELESTE, are in the process of converting large-area heliostats from experimental solar-power stations to this purpose. A German-Spanish project (MAGIC) is also underway to build a 17-m air-Cherenkov telescope.

At even higher gamma-ray energies, in the PeV range (10^{15} eV, the “ultra-high energy,” or UHE, regime), particles associated with the photon-induced electromagnetic cascades have sufficient energy to reach the ground, where arrays of particle detectors can be used to detect them. Very large collecting areas are required, however, given the extremely low gamma fluxes expected. Early attempts to detect photons of PeV energy with particle-detector arrays included the CASA-MIA and CYGNUS efforts, among others. Ongoing attempts are being carried out by the HEGRA and MILAGRO collaborations. Early reports of positive PeV detections in the 1980s proved to be optimistic and have been unconfirmed by later more sensitive measurements.

IV. FUTURE MISSIONS AND PROSPECTS

The development of any new telescope has as its primary goal an increase in sensitivity, combined with improved angular and spectral resolution. In the gamma-ray regime this translates invariably into improved determination of photon-interaction positions and energy-depositions within the detecting medium. More accurate determination of the properties of interacting gamma-rays leads directly to a reduced background rate, since true celestial events are less likely to be confused with background interactions. Virtually every gamma-ray telescope under current development seeks to improve these interaction measurements by exploiting new detector technologies. Spatial and energy resolution within detector materials are greatly enhanced, for example, with the use of newly developed semiconductor strip and pixel detectors (such as silicon, germanium, and CdZnTe). The continuing challenge is to fabricate such sensitive small-scale devices in sufficiently large and reliable quantities to be incorporated into new large-area instrumentation, at costs that can be reasonably afforded. Another common characteristic of high-energy telescopes is the large number of data signals that must be processed and recorded in multi-channel detector systems. Increased use of custom Application-Specific Integrated Circuit (ASICs) designs employing Very Large Scale Integration (VLSI) techniques is imperative for the efficient operation of high-energy instruments. Fortunately, computational speed and data-storage capacities continue to rise at a steady pace, and experimentalists are quick to exploit these new capabilities in their instrument designs.

At the time of writing (2001), a number of gamma-ray missions are scheduled for launch in the near future (see Table II). Key among these is the International Gamma-Ray Astrophysics Laboratory (INTEGRAL), a mission of the European Space Agency (ESA) with participation from Russia and NASA. INTEGRAL is to be launched in 2002 and will be dedicated to high-resolution spectroscopy ($E/\Delta E \sim 500$) and imaging ($\sim 12'$ FWHM) over the energy range from 15 keV to 10 MeV. INTEGRAL carries two gamma-ray instruments, the SPI spectrometer and the IBIS imager, both operated as coded-aperture telescopes for accurate source identification. The SPI employs high-purity germanium detectors, while the IBIS uses two detector planes, a front layer of CdTe elements and a second layer composed of CsI pixels. In recognition of the need for broadband coverage INTEGRAL also carries two coded-aperture X-ray monitors (JEM-X), as well as an optical monitoring camera (the OMC). The primary scientific objective of the INTEGRAL instruments is to carry out high-resolution spectroscopic studies of sources over the nuclear-line region of the spectrum.

The Gamma-Ray Large Area Space Telescope (GLAST), scheduled for launch by NASA in 2005, will be the follow-on mission to the highly successful CGRO EGRET experiment. The sensitivity of GLAST, from 20 MeV to 300 GeV, will extend well beyond the EGRET range, providing much-need coverage in the poorly observed GeV region of the spectrum. A more modern particle-tracking technology (silicon strip detectors) will be employed in GLAST in place of the spark-chambers grids used in earlier pair-production telescopes. GLAST will have a large field of view ($\sim$2 sr) and achieve a factor of 30 improvement in flux sensitivity and a factor of 10 improvement in point-source-location capability compared to EGRET. GLAST will also carry a gamma-ray burst monitor.

Missions designed specifically for gamma-ray burst studies include HETE-2 and Swift. The High-Energy Transient Experiment-2 (HETE-2) was launched in 2000, and became operational in early 2001. This satellite carries three science instruments: a near-omnidirectional gamma-ray spectrometer, a wide-field X-ray monitor, and a set of soft X-ray cameras. A major goal of the HETE-2 mission is the rapid identification and accurate localization of gamma-ray bursts, whose coordinates will be relayed within seconds to ground-based observatories for deep counterpart searches. The recently selected Swift mission (scheduled for launch in 2003) will also carry out multiwavelength studies of gamma-ray bursts, in the manner of BeppoSAX and HETE-2. Like its avian namesake Swift will “feed on the fly” by rapidly localizing gamma-ray bursts to $\sim$1–4$'$ precision, and transmitting coordinates to the ground within $\sim$15 s for follow-up counterpart searches. Swift can also be rapidly reoriented to carry out observations with its X-ray and ultraviolet/optical telescopes that will be used to study afterglow properties, fix positions to arcsecond levels, and determine distances via redshift spectral measurements.

The High-Energy Solar Spectroscopic Imager (HESSI) is a NASA-funded mission to study the characteristics of particle acceleration in solar flares via the X-ray and gamma-ray emission produced in these energetic events. HESSI, scheduled for launch in 2001 at the peak of the solar cycle, will carry out high-resolution spectroscopic measurements of nuclear lines and underlying bremsstrahlung continuum over the energy range from 3 keV to 20 MeV with a set of cooled high-purity germanium detectors. HESSI will carry out Fourier-transform imaging of the full Sun at $\sim$2$''$–36$''$ resolution over its sensitive range by using rotating modulating collimators. Since HESSI is unshielded it can also carry out other nonsolar observations, including measurement of the Galactic diffuse lines due to radioactive ^{26}Al (at 1.809 MeV) and positron annihilation (at 0.511 MeV).

In the area of planetary studies, NASA’s Mars Odyssey mission is also scheduled for launch in 2001. Among its suite of instruments are a gamma-ray spectrometer and two neutron detectors. These will be used to fully map the Martian surface and determine its elemental composition. The neutron and gamma-ray measurements in combination will also be used to obtain an estimate of the water content of the Martian near-surface.

Other gamma-ray experiments and missions have been identified as a high-priority by the Gamma-Ray Astronomy Program Working Group, an advisory panel to NASA composed of scientists from the high-energy community. Among their recommendations for future development is an advanced Compton telescope employing the latest detector technologies for application in the MeV region of the spectrum.

High-altitude scientific ballooning has long served as a test-bed for new instrumentation. Gamma-ray telescopes require long exposures, due to comparatively low source fluxes and high instrumental backgrounds, while the duration of a typical balloon flight, unfortunately, can often be rather limited (a few days at most). To counter this drawback NASA has recently initiated the Ultra-Long Duration Balloon (ULDB) project whose planned 100-day around-the-world balloon flights will greatly extend the time aloft for scientific instruments. The ULDB program will provide much-needed opportunities for longer-exposure balloon flights, as well as an attractive low-cost alternative to full-scale space missions.

Among the collaborations actively engaged in ground-based air-Cherenkov studies of TeV gamma rays there are also a number of efforts underway to upgrade existing facilities, primarily through an increase in optical collecting area. Perhaps the most ambitious are those of the VERITAS collaboration, with a planned array of seven 10-m telescopes in the USA, the German-French-Italian HESS group with 4 to 16 12-m class telescopes to be built in Namibia, the German-Spanish MAGIC project with a telescope of 17-m aperture, and the Japanese SuperCANGAROO array of four 10-m telescopes in Australia. In a related effort, the MILAGRO collaboration is constructing a water-Cherenkov detector with a wide field of view in New Mexico in the USA for TeV measurements. As a covered light-tight detector MILAGRO has the added advantage that it can remain operational for 24 h a day.

V. CONCLUDING REMARKS

The future of gamma-ray astronomy looks very bright indeed, as the next generation of gamma-ray telescopes and missions stands poised to extend our knowledge of the high-energy sky. Over the past decade gamma-ray astronomy has become firmly established as a productive and dynamic discipline of modern observational astrophysics. With the completion of comprehensive surveys with the telescopes aboard the Compton Gamma Ray Observatory the pioneering phase of gamma-ray exploration has now been achieved and the promise of gamma-ray astronomy confirmed and realized. New questions have been raised and remain to be addressed, and ongoing technological advances will certainly continue to spur the development of future successor missions to build upon the present findings. The results obtained with the Compton Observatory and other spacecraft have revolutionized our view of the high-energy Universe, and have underlined the fundamental importance of the gamma-ray region of the spectrum to a fuller and more complete understanding of the high-energy processes that dominate the cosmos.

ACKNOWLEDGMENT

This review is dedicated to the memory of Frank J. Kerr, one of the founding fathers of the field of Galactic radio astronomy, and an early advisor to both of the authors. His gentlemanly manner and professional example will be long remembered by those who knew him. He would be delighted and amazed at the many recent discoveries in gamma-ray astronomy, and fully appreciative of their relevance to much of his own work.

SEE ALSO THE FOLLOWING ARTICLES

Cosmic Radiation • Gravitational Wave Astronomy • Infrared Astronomy • Neutrino Astronomy • Pulsars • Solar Physics • Supernovae • Ultraviolet Space Astronomy • X-Ray Astronomy

BIBLIOGRAPHY

Catanese, M., and Weekes, T. C. (1999). Very high gamma-ray astronomy. *Publications of the Astronomical Society of the Pacific* **111,** 1193–1222.

Diehl, R., and Timmes, F. X. (1998). Gamma-ray line emission from radioactive isotopes in stars and galaxies. *Publications of the Astronomical Society of the Pacific* **110,** 637–659.

Fichtel, C. E., and Trombka, J. I. (1997). "Gamma-Ray Astrophysics: New Insights into the Universe, Second Edition," NASA Reference Publication 1386. NASA, Greenbelt, MD.

Fishman, G. J., and Meegan, C. A. (1995). Gamma-ray bursts. *Annu. Rev. Astr. Astrophys.* **33,** 415–458.

Gehrels, N., and Paul, J. (1998). The new gamma-ray astronomy. *Physics Today* **51,** 26–32.

Hoffman, C. M., Sinnis, C., Fleury, P., and Punch, M. (1999). Gamma-ray astronomy at high energies. *Rev. Mod. Phys.* **71,** 897–936.

Longair, M. S. (1997). "High Energy Astrophysics, Volume 1: Particles, Photons and their Detection," Second Edition, Cambridge University Press, Cambridge.

Longair, M. S. (1994). "High Energy Astrophysics, Volume 2: Stars, the Galaxy and the Interstellar Medium," Second Edition. Cambridge University Press, Cambridge.

van Paradijs, J., Kouveliotou, C., and Wijers, R. A. M. J. (2000). Gamma ray burst afterglows. *Annu. Rev. Astr. Astrophys*. **38,** 379–425.

Strong, K. T., Saba, J. L. R., Haisch, B. M., and Schmelz, J. T. (eds.). (1998). "The Many Faces of the Sun: A Summary of the Results from NASA's Solar Maximum Mission," Springer-Verlag, Berlin.

Urry, C. M., and Padovani, P. (1995). Unified schemes for radio-loud active galactic nuclei. *Publications of the Astronomical Society of the Pacific* **107,** 803–845.

Gravitational Wave Astronomy

Patrick R. Brady
Jolien D. E. Creighton
University of Wisconsin–M:Lwauxee

GLOSSARY

Characteristic strain The gravitational wave strain of a signal at some characteristic frequency times the square root of the number of cycles over which the signal is observed near that frequency.

Detector sensitivity Characteristic strain due to gravitational waves that would exceed instrumental strain noise in a detector.

Gravitational wave strain The fractional change in the distance between two free-falling bodies produced by a gravitational wave as it passes the bodies.

Laser interferometric detector A laser interferometer which measures the relative lengths of two orthogonal (or nearly orthogonal) cavities. When a gravitational wave impinges on the system, the relative change in cavity length causes a change in the interference pattern registered by the interferometer.

Resonant-mass detector A massive cylinder of aluminium, or nobium crystal, suspended in vacuum and mechanically isolated from its surroundings. When a gravitational wave impinges on the cylinder, the relative accelerations excite the cylinder's natural modes of oscillation.

Tidal distortion The deformation of a massive body by the gravitational field produced by other bodies.

GRAVITATIONAL WAVE ASTRONOMY is the study of our universe using gravitational radiation emitted by astrophysical sources. Experimental efforts to directly observe gravitational waves have a history spanning at least 40 years. The construction and commissioning of long-baseline interferometric detectors is a landmark for the field. Scientists believe that these instruments will make the first (unambiguous) detection of gravitational waves, and will become tools for routine astronomical observations.

I. INTRODUCTION TO GRAVITATIONAL RADIATION

Gravitational radiation, like electromagnetic radiation (radio, infrared, light, and X-rays), transports energy via propagating field fluctuations, or waves. Where electromagnetic radiation involves fluctuations of the electromagnetic field, gravitational radiation involves fluctuations of the gravitational field. Gravitation and electromagnetism are both long-range forces which exhibit radiative behavior. Two significant differences are (i) the electromagnetic force is stronger than the gravitational force, and (ii) the electric charge can be positive or negative, while the gravitational charge—the mass of an object—has only one sign.

Electromagnetic radiation originates from individual charged particles which undergo rapid acceleration. Astronomical bodies tend to be electrically neutral due to the strength of the electromagnetic force and the repulsion between like charges. Consequently, the brightest sources emit an incoherent superposition of high-frequency radiation from ionized material. Because electromagnetic radiation is readily absorbed and scattered, observed radiation comes only from the outermost layers of the source. By observing electromagnetic radiation, astronomers gain information about the thermodynamic state of the radiator and its constituent materials.

In contrast, strong gravitational waves originate from the bulk motion of large masses. Large accumulations of mass are possible because the gravitational charge (mass) has only one sign and the force between charges is attractive. Gravitational radiation tends to be strongest at low frequencies since significant coherent changes in motion occur on macroscopic scales. The information carried by the waves describes the dynamics of the source rather than its thermodynamic state. Moreover, gravitational waves couple weakly to matter, so there is almost no absorption or scattering of the gravitational radiation—the universe is almost entirely transparent to gravitational waves. This weak coupling also makes it difficult to directly observe gravitational waves.

General relativity is the simplest of modern theories of gravitation; it attributes the gravitational field generated by massive bodies to the curvature of spacetime. More precisely, the curvature of spacetime is generated by the mass, energy, and stress contained in matter. The familiar notion of gravitational acceleration arises from the assertion that a free-falling test particle (i.e., a particle that does not contribute to the gravitational field) follows a shortest path in spacetime. Tidal effects—the distortion of massive bodies by the gravitational field produced by other bodies—are manifestations of curvature; a pair of free-falling test particles which start out moving parallel may move closer together or further apart due to a tidal field. Tidal fields and gravitational acceleration are also well known in Newtonian gravity theory. In general relativity, however, the gravitational field propagates at finite speed.

Gravitational waves are propagating fluctuations in the spacetime curvature. These fluctuations produce an oscillating tidal field which changes the distance between nearby free-falling bodies. The fractional change in the distance between two free-falling bodies initially separated by a distance ℓ is the strain

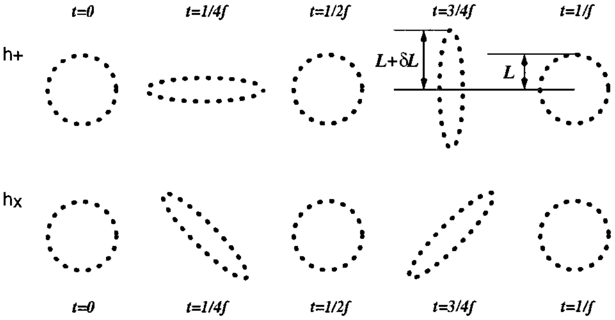


FIGURE 1 The two polarizations of gravitational waves.

$$h = \Delta\ell/\ell$$

produced by the gravitational wave as it passes the two bodies. In general relativity, there are two polarizations of gravitational waves, h_+ and $h_\times$, which cause displacements in independent directions (see Fig. 1). The strain h on a particular pair of objects is a linear superposition of these two polarizations.

According to Einstein's equations, accelerating masses generate gravitational radiation. The strongest sources consist of large accumulations of dense matter. By analogy with electromagnetic radiation, it is both illuminating and useful to describe radiating matter distributions in terms of their changing multipole moments. The monopole moment is the total mass of the accumulation. It is conserved and thus cannot change. Similarly, the gravito-electric and gravito-magnetic dipole moments are the momentum and angular momentum of the distribution. They are also conserved, so there is no dipole radiation. (This contrasts with electromagnetism, in which the charge of an object is independent of its inertia. The absence of dipole radiation in general relativity is related to the equivalence of gravitational and inertial mass.) Consequently, gravitational waves are radiated if the source has a changing quadrupole (or higher) moment.

For weakly gravitating sources composed of matter moving slowly compared to the speed of light, the amplitude of the gravitational wave strain h is proportional to the second time derivative (retarded, to account for the finite speed of propagation of the gravitational wave) of the quadrupole moment of the system. The strain, due to a passing gravitational wave, on two free-falling bodies located a distance D from the source is

$$h \approx \frac{2G}{c^4 D} \frac{d^2 Q}{dt^2}\bigg|_{\text{retarded}}$$

where Q is the projection of the quadrupole moment tensor first transverse to the direction of propagation of the wave and then along the direction separating the two free-falling bodies. Because $Q \sim mx^2$ (see below), the second time derivative of Q is essentially the kinetic energy of the source attributable to nonsymmetric motion of the particles in the source, $E_{\rm NS}$. To produce significant amounts of gravitational radiation, the nonsymmetric mass–energy of the source M must be compressed into a small region. In general relativity, this region must be bigger than a black hole of the same mass, i.e., $R = 2GM/c^2$ where G is Newton's gravitational constant and c is the speed of light. For a stellar mass source ($M \sim M_\odot$) at astrophysical distances, the strain at gravitational wave observatories can be estimated as

$$\begin{aligned} h &\sim \frac{G(E_{\rm NS}/c^2)}{c^2 D} \\ &\sim 10^{-20}\left(\frac{E_{\rm NS}/c^2}{M_\odot}\right)\left(\frac{\rm Mpc}{D}\right). \end{aligned}$$

That is, to observe waves from a system with a solar-mass worth of energy in nonsymmetric motion at the megaparsec scale (the distance to the nearest galaxy), astronomers need to measure strains on the order of 10^{-20}. This sensitivity is just now becoming possible for broad-band detectors.

Until the 1960s there was considerable debate whether gravitational waves were observable at all. Complications of general relativity made it difficult to separate physical effects from artifacts associated with the mathematical treatment of the theory. The most convincing theoretical arguments in favor of gravitational waves as observable physical quantities are based on energy content of the waves. It is well known that work is done on bodies subjected to time-varying tidal fields. For example, work is done to raise ocean tides by the changing tidal field of the moon and sun as the Earth rotates. Gravitational waves are produced by accelerating masses that produce time-varying tidal fields. Since the propagating tidal field can do work, gravitational waves must carry the energy required to do the work. Consequently, bodies that produce dynamic tidal fields lose energy in the gravitational waves they produce. The amount of energy carried by gravitational waves through an element of area dA in an element of time dt scales as the rate of change of the two polarizations of the gravitational wave strain squared:

$$dE = \frac{c^3}{16\pi G}\left[\left(\frac{\partial h_+}{\partial t}\right)^2 + \left(\frac{\partial h_\times}{\partial t}\right)^2\right] dA\, dt.$$

Within the quadrupole approximation, the rate of energy loss from a source is therefore given by

$$\frac{dE}{dt} = \frac{G}{5c^5}\sum_{i,j=1}^{3}\left(\frac{d^3 Q_{ij}}{dt^3}\right)^2$$

where Q_{ij} is the quadrupole moment tensor given by

$$Q_{ij} = q_{ij} - \sum_{k=1}^{3} q_{kk}$$

where

$$q_{ij} = \sum_{\rm particles} m x_i x_j.$$

Here m is the mass of a given particle in the source and $\{x_i\}$ is the set of components describing its position relative to the (irrelevant) origin of the coordinates.

II. OBSERVATION TECHNIQUES

The methods used to observe gravitational radiation suggest that four frequency bands be identified. The bands are:

- Extremely Low Frequency (ELF) band: 10^{-18}–10^{-15} Hz. The canonical source in this band is gravitational waves from the big bang. These waves form a cosmological stochastic background much like the famous cosmic microwave background discovered by Penzias and Wilson. Limits on the cosmological stochastic background gravitational waves are provided by observations of the cosmic microwave background.
- Very Low Frequency (VLF) band: 10^{-9}–10^{-4} Hz. Pulsar timing provides a method to detect these gravitational waves. For example, timing measurements of the Hulse–Taylor binary pulsar provide indirect evidence for gravitational waves emitted by binary neutron star systems. Promising sources for direct detection by radar ranging include waves from a variety of cosmological defects originating in the early universe.
- Low Frequency (LF) band: 10^{-4}–10^0 Hz. Gravitational wave detectors in space, e.g., the Laser Interferometer Space Antenna (LISA), and radar ranging of spacecraft can observe in this band. Close white dwarf binaries are so common in our Galaxy that they are expected to produce a stochastic background of gravitational waves below $\sim 10^{-3}$ Hz, while gravitational waves from super-massive black holes (10^5–10^9 $M_\odot$) will be the strongest point sources.
- High Frequency (HF) band: 10^0–10^4 Hz. Ground-based interferometers and resonant-mass detectors operate in this band. The most promising sources of gravitational waves are inspiral and coalescence of compact stellar-mass black holes and

neutron stars (1–10^3 $M_\odot$). Other sources include distorted pulsars and supernovae.

The experimental techniques used (or planned) to observe gravitational waves are described below in order from high frequency to low frequency.

A. Resonant Mass Detectors

Massive cylinders of aluminum, or niobium crystal, are suspended in vacuum and mechanically isolated from their surroundings. When a gravitational wave impinges on the cylinder, the relative accelerations excite the cylinder's natural modes of oscillation. These excitations are measured using transducers attached to the cylindrical bar. Detectors of this type are typically most sensitive to gravitational waves in the HF band traveling orthogonal to the cylinder's axis. The first instruments constructed with the express purpose of observing gravitational waves were resonant bar detectors. Future resonant mass detectors may be spheres or truncated icosahedra which would have omnidirectional sensitivity.

Gravitational waves deposit energy into resonant mass detectors when they excite their normal modes. The absorption cross section of the bar increases with the mass of the bar and with the square of the speed of sound in the bar's material. Thus, it is desirable to construct high-mass resonant bars with a high speed of sound. Accurate measurement of the normal mode excitations is hampered by noise, due to Brownian motion, in the bar and the transducer. Modern bar detectors are cryogenically cooled to liquid helium temperatures, or lower, to reduce this noise source. Bar detectors are also limited by the electronic noise in the amplifier. The strain-noise power spectral density (see below) of a bar detector is

$$S(f_0) = \frac{\pi}{2}\frac{kT}{Mv_s^2 Q f_0}$$

where f_0 is the frequency of the excited mode, Q is the quality factor (the number of radians of oscillation in the mode it takes to damp by a factor of e) of the mode, v_s is the speed of sound in the bar, T is its temperature, and k is Boltzmann's constant. This sensitivity is achieved in a narrow frequency band, $\Delta f \sim 1$ Hz for modern detectors, around the mode frequency.

Modern resonant bar detectors have mode frequencies of $f_0 \sim 900$ Hz with bandwidths of $\Delta f \sim 1$ Hz, quality

TABLE I Gravitational Wave Astronomy Frequency Bands

Frequency band	Sources and methods of observation
Extremely Low Frequency (ELF) 10^{-18}–10^{-15} Hz	Stochastic gravitational waves from the early universe. Can be detected as anisotropies of the Cosmic Microwave Background Radiation (CMBR). Waves from the big bang may be parametrically amplified by inflation. They may be produced by cosmic strings or domain walls during phase transitions as the universe cooled.
Very Low Frequency (VLF) 10^{-9}–10^{-4} Hz	(a) Gravitational waves from widely separated binary systems. Can be detected using radio timing observations of binaries containing millisecond pulsars, e.g., PSR 1913+16. (b) Stochastic background of gravitational waves. Has been constrained using pulsar timing data.
Low Frequency (LF) 10^{-4}–10^{0} Hz	Space-based laser interferometers provide the best method to observe in this frequency band. It is anticipated that the Laser Interferometric Space Antenna (LISA) should launch circa 2010. (a) Inspiral waves from binary systems comprised of white dwarfs, neutron stars, or stellar-mass black holes in our Galaxy. At the lower frequencies in this band, There are so many white dwarf binary systems that the signals will be confused; these systems will produce a stochastic background of gravitational waves at lower frequencies in this band. (b) Gravitational waves from supermassive ($M \sim 10^5$–10^9 $M_\odot$) black holes disturbed by infalling stars and other debris. (c) Cosmological stochastic background of gravitational waves; has been constrained by radar ranging of interplanetary spacecraft.
High Frequency (HF) 10^{0}–10^{4} Hz	Ground-based interferometers and resonant-mass detectors provide the best method to observe in this frequency band. Resonant mass detectors have been operating since the early 1960s; prototype interferometric detectors have been operating since the 1970s. The new generation of kilometer-scale interferometric detectors will be in continuous operation from 2002 onward. (a) Waves from the inspiral and merger of binary systems comprised of neutron stars or black holes (with mass $M < 1000 M_\odot$). (b) The birth of neutron stars and supernova explosions. Rotation-induced asymmetries, convection, or unstable modes in the nascent neutron star (e.g., bar mode instabilities and r-modes) will generate gravitational waves in this band. (c) Wobbling or distorted pulsars will produce nearly monochromatic continuous waves. (d) Cosmological stochastic background of gravitational waves.

factors of $Q \sim 10^6$, masses of $M \sim 2000$ kg, and temperatures $T \sim 1$ K. The rms strain sensitivity of these detectors is $h_{\rm rms} = [f_0 S(f_0)]^{1/2} \sim 10^{-20}$. There are five currently operating cryogenic resonant bar detectors: EXPLORER (located at CERN), ALLEGRO (located in Baton Rouge, USA), NIOBE (located in Perth, Australia), NAUTILUS (located in Frascati, Italy), and AURIGA (located in Legnaro, Italy).

B. Ground-Based Interferometers

Interferometric gravitational wave detectors measure the relative lengths of two orthogonal (or nearly orthogonal) cavities. The simplest interferometric detector is based on the Michelson interferometer. In this configuration, a beam of light is split at a beam splitter and enters the two cavities (see Fig. 2). The light traverses the cavities and reflects off the mirrors at the ends of the cavities. The light beams from the two cavities interfere at the beam splitter (after traversing the cavities). A photodiode is used to detect changes in the interference pattern when the cavities change length. The interferometer is normally configured so that no light hits the photodiode if the two cavities have identical length; the photodiode is trained on the dark port of the beam splitter. When a gravitational wave passes the detector, the relative lengths of the two cavities change as the end mirrors are affected by the tidal field. This causes light to fall on the photodiode, and thus provides a read-out of the gravitational wave signal.

Modern interferometric detectors often use modified optical topologies which provide greater sensitivity to gravitational waves. Instead of a simple Michelson interferometer, a Fabry–Perot cavity is created within each of the arms to store the laser light. The quality of the cavity determines the storage time for the light and the corresponding effective increase in the length of the cavities. Since readout is sensitive to the relative difference in length of the two arms, this provides additional sensitivity if the light is stored for less than one gravitational wave cycle.

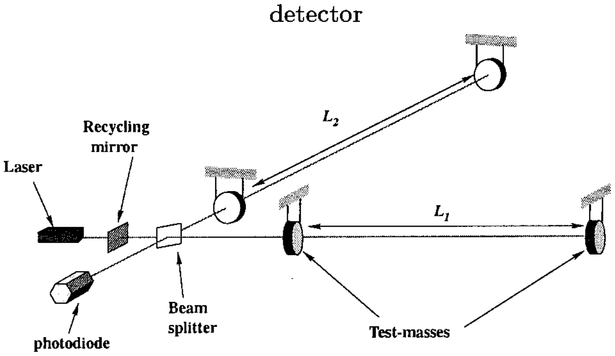


FIGURE 2 Schematic of an interferometric gravitational wave detector.

There are three dominant sources of noise in ground-based interferometric detectors. At low frequencies, seismic noise creates a "wall" at ~10 Hz below which the ground motion is no longer isolated by the suspension. At intermediate frequencies, the thermal noise of the test masses and their suspensions is dominant. At high frequencies, the combination of shot (photon counting) noise, and, for cavities with a high light-power, radiation pressure noise are dominant. The sensitivity at high frequencies can be improved using a power-recycling mirror to reflect light coming from the beam splitter back into the cavities. This increases the light power stored in each arm, thus reducing photon shot noise. Further improvement may be achieved using signal-recycling—a mirror is added to the dark port of the beam splitter and the signal is fed back into the interferometer.

Currently operating kilometer-scale interferometers include the Laser Interferometer Gravitational-wave Observatories (LIGO) with sites located in Hanford, Washington, and Livingston, Louisiana; the GEO-600 interferometer in Hannover, Germany; and the TAMA-300 interferometer in Tokyo, Japan. In addition, the VIRGO observatory is under construction near Pisa, Italy.

C. Space-Based Interferometers

Seismic and gravitational field gradients limit the sensitivity of terrestrial detectors in the LF band. Very long baseline ($\sim 10^6$ km) interferometers in space would be sensitive to gravitational waves in this band. A joint ESA/NASA mission for a space-based gravitational wave detector called Laser Interferometer Space Antenna (LISA) is proposed and under consideration for launch around 2010. The mission would place three identical spacecraft in a heliocentric orbit about 20° behind the Earth. The spacecrafts would each house a laser and two proof masses. These masses would be allowed to orbit freely inside the spacecraft—they would not be attached in any way, rather they would be protected from buffeting by the solar wind and radiation pressure. The spacecraft would use ion propulsion drives to follow the proof masses in drag-free orbits. The positions of the proof masses with respect to each other would be monitored using heterodyne interferometry; this provides the readout sensitive to the gravitational wave signal.

D. Pulsar Timing

Millisecond pulsars are excellent clocks, comparable in accuracy to the best atomic clocks on earth. When a gravitational wave is present, either at the pulsar or at the earth,

TABLE II Physical Data for Modern Cryogenic Resonant Bar Detectors

Detector	ALLEGRO	AURIGA	EXPLORER	NAUTILUS	NIOBE
Material	Aluminum	Aluminum	Aluminum	Aluminum	Niobium
Length (m)	3.0	2.9	3.0	3.0	2.75
Mass M (kg)	2296	2230	2270	2260	1500
Mode frequencies f_0 (Hz)	895, 902	912, 930	905, 921	908, 924	694, 713
Quality factor Q	2×10^6	3×10^6	1.5×10^6	5×10^5	2×10^7
Temperature T (K)	4.2	0.2	2.6	0.1	5
Sensitivity h_{rms}	3×10^{-20}	6×10^{-21}	2×10^{-20}	6×10^{-21}	2×10^{-20}
Latitude N	$30^\circ 27' 45''$	$45^\circ 21' 12''$	$46^\circ 27'$	$41^\circ 49' 26''$	$-31^\circ 56'$
Longitude E	$-91^\circ 10' 44''$	$11^\circ 56' 54''$	$6^\circ 12'$	$12^\circ 40' 21''$	$115^\circ 49'$

there will be a dilation of time due to the gravitational wave; this time dilation can be measured as the clocks become unsynchronized. Pulsar timing data puts the best current observational limits on the stochastic background of gravitational waves with periods on the order of the total observation time (several years), i.e., in the very low frequency band.

E. Radar Ranging

Tracking of interplanetary spacecraft affords the opportunity to use radar ranging over very long baselines for gravitational wave detection. A highly stable radio signal is sent to the spacecraft which transponds it back to an earth-based receiver. The presence of a gravitational wave would create a Doppler signature $\delta f/f$ proportional to the wave-strain in the radio signal. Gravitational waves with periods on the order of the round-trip time ($\sim 10^{-3}$ Hz for solar system distance scales) can be detected if their amplitudes exceed $h \sim 10^{-15}$. This detection method is limited by dispersion of the radio waves in the interplanetary plasma and by water vapor in the earth's atmosphere.

III. DATA ANALYSIS METHODS

Despite the many observational methods available to search for gravitational wave signals, a direct observation remains elusive to this day. With the advent of long-baseline, ground-based interferometers, a new era in gravitational physics has arrived. These sensitive instruments will likely transform the field of gravitational wave detection into the new field of gravitational wave astronomy. With the improved sensitivity and the planned duty cycle, i.e., continuous operation of multiple detectors around the world, comes a new challenge: data analysis. The anticipated sources of gravitational waves will be close to the limit of detectability, so it is necessary to use optimal data analysis tools. Since routine astronomical observations are likely to be made with the interferometric or resonant bar detectors, attention is restricted to these instruments below.

The output from gravitational wave detectors is generally time series data that provide a direct measure of the strain on the detector. Extraction of the gravitational wave signal is frustrated by noise which limits the sensitivity of the instrument. If the calibrated strain signal from a given detector is $h(t)$, the detector sensitivity can be expressed in terms of the mean square instrumental noise per unit frequency. Specifically, the noise power spectral density of the instrumental strain noise is

$$S(f) = \frac{2}{T}\langle |\tilde{h}_{\text{noise}}(f)|^2 \rangle$$

where $\langle \cdot \rangle$ represents an average over an ensemble of noise realizations and

$$\tilde{h}_{\text{noise}}(f) = \int_0^T h_{\text{noise}}(t) e^{2\pi i f t} dt$$

is the Fourier transform of the strain noise $h_{\text{noise}}(t)$ over a time interval T. The power spectral density provides a measure of instrumental noise as a function of frequency. As indicated above, each detector has a frequency band of optimal sensitivity which determines the type of source it may observe.

Optimal detection of signals in noise relies on accumulation of all the signal power into a single number. For periodic signals, this is achieved by taking the Fourier transform of $h(t)$ and examining the power at the signal's frequency. In general, the strength of a signal is described by its characteristic amplitude h_{char}, which represents the intrinsic amplitude of the signal at some characteristic frequency times the square-root of the number of cycles over which the signal is observed near that frequency (see below for precise definitions). For comparison with signal strength, detector sensitivity is better expressed in terms of the root mean square dimensionless strain per logarithmic frequency interval

$$h_{\text{rms}} = \sqrt{f S(f)}.$$

TABLE III Physical Data for Modern Ground-Based Interferometric Detectors

	LIGO					
Detector	**LHO 2 km**	**LHO 4 km**	**LLO 4 km**	**GEO-600**	**TAMA-300**	**VIRGO**
Arm length(s) (m)	2000	4000		600	300	3000
Frequency f_{min} (Hz)	125	125		200	300	100
Bandwidth Δf (Hz)	200	200		300	300	800
Strain sensitivity $h_{\text{rms}}(f_{\text{min}})$	6.5×10^{-22}	3.5×10^{-22}		8×10^{-22}	3×10^{-21}	6×10^{-22}
Elevation (m)	142.554		−6.574	114.425	90	51.884
Latitude N	46°27′18″.528		30°33′46″.4196	52°14′42″.528	35°40′35″.6	43°37′53″.0921
Longitude E	240°35′32″.4343		269°13′32″.7346	9°48′25″.894	139°32′9″.8	10°30′16″.1878
X arm orientation N of E	125°.9994		197°.7165	115°.9431	180°	70°.5674
Y arm orientation N of E	215°.9994		287°.7165	21°.6117	270°	160°.5674

A. Detection of Burst Sources

1. Interferometers

Since interferometers have broadband sensitivity, one can monitor the phase evolution of a gravitational wave over many cycles. This provides a method of discriminating the signal from noise, and increases the detectability of a gravitational wave signal.

a. Known waveform. When the waveform of the source is known, the method of matched filtering is optimal in the sense that the deduced probability of a signal being present depends upon the collected data only through the matched filter. The matched filter is computed by correlating the observed interferometer output $h(t)$ with the known gravitational wave signal $h_{\text{gw}}(t)$, weighted inversely by the noise power spectrum:

$$x(t) = 4\text{Re}\int_0^\infty \frac{\tilde{h}^*(f)\tilde{h}_{\text{gw}}(f)}{S(f)} e^{-2\pi i f t} df.$$

If the detector noise originates from stationary, Gaussian random processes and the instrument is calibrated with zero-offset, this matched filter output is a zero-mean Gaussian random variable with variance equal to the power signal-to-noise ratio of the known signal:

$$\sigma^2 = 4\int_0^\infty \frac{|\tilde{h}_{\text{gw}}(f)|^2}{S(f)}\, df = 4\int_0^\infty \left[\frac{h_{\text{char}}(f)}{h_{\text{rms}}(f)}\right]^2 d\ln f$$

where the characteristic amplitude of a signal is

$$h_{\text{char}}(f) = |f\tilde{h}_{\text{gw}}(f)| \approx h\sqrt{n}$$

if it spends roughly n cycles with amplitude h in a frequency band $\Delta f \sim f$ about the frequency f. In order for a signal to be detectable by matched filtering, it must have an amplitude signal-to-noise ratio, σ, somewhat greater than unity. In this way, comparison between h_{char} for a particular source and h_{rms} for a particular detector provides an estimate of detectability of the source.

b. Unknown waveform. The class of sources for which the gravitational wave signal can be accurately determined in advance is very small. Other techniques are needed when the precise phase evolution of the expected gravitational wave burst is not known. For signals with known frequency support Δf and known time duration Δt, the signal can be effectively detected by examining the amount of power for durations Δt in the detector output that has been band-limited to the band of interest, again weighted by the noise power spectral density in the band:

$$\varepsilon(t) = \int_t^{t+\Delta t} \left| \int_f^{f+\Delta f} \frac{\tilde{h}(f)}{S(f)} e^{-2\pi i f t} df \right|^2 dt.$$

(This method is optimal when only the band and duration of the signal are known.) Because this method lacks the discriminating power of known phase evolution that the matched filter provides, it is somewhat less powerful at distinguishing a signal from noise; consequently, a signal needs to have a greater amplitude to be detected. Typically, the signal has to be stronger (in amplitude) by a factor of $(\Delta t \times \Delta f)^{1/4}$ in order to have the same probability of detection.

There are other methods of detecting "unknown" signals when different information about them is known. For example, if it is known that the signal is nearly monochromatic at any instant but sweeps from low to high frequencies with time, it can readily be detected by searching for tracks in a time-frequency plot.

2. Bars

In contrast to interferometric detectors, resonant bar detectors have optimal sensitivity only in a narrow frequency band Δf about a resonant frequency f_0. Gravitational wave bursts with time duration $\tau < 1/\Delta f$ act like impulsive hammer blows at the resonant frequency with amplitude $h_{\text{char}} = |f_0\tilde{h}_{\text{gw}}(f_0)|$. Consequently, significantly less information about the waveform is needed to achieve optimal sensitivity to gravitational wave bursts in bar detectors. A burst is detected by searching for excess power deposited in the bar by the impulsive burst within the sensitive band of the bar.

B. Detection of Continuous Wave Sources

Sources in this category emit gravitational waves for times long compared to the expected observation time. Two circumstances must be distinguished: (i) sources whose gravitational waveform can be inferred by electromagnetic observations and (ii) sources which can only be detected by their gravitational wave signals.

Rapidly spinning neutron stars are the main source of continuous waves in the HF band accessible to earth-based interferometers and resonant bar detectors. When the neutron star can be observed using radio (or other) telescopes, the expected gravitational waveform can be inferred (up to small uncertainties) from observations of the spin period. In this case, the optimal data analysis strategy is matched filtering. The implementation may be slightly different than for burst sources, but the idea is the same. Successful detection of waves from these sources will rely on direct interaction between the radio astronomers and gravitational astronomers.

Generally, continuous-wave signals are expected to be nearly periodic and very weak. When a target source cannot be identified using electromagnetic observations, the optimal strategy involves coherent accumulation of signal-to-noise using Fourier transforms of long stretches of data (months to years). The power signal-to-noise ratio at frequency f_0 is given by

$$\sigma^2 = \frac{2|\tilde{h}(f_0)|^2}{TS(f_0)}$$

where $\tilde{h}(f)$ is the Fourier transform of T seconds of calibrated data. By analogy with burst sources, the characteristic amplitude of a continuous wave signal is defined as $h_{\text{char}} = h_0\sqrt{(f_0 T)}$ where h_0 is the rms strain amplitude of the signal, f_0 is the frequency of the signal. Here the observation time T is the total integration time for truly continuous wave signals, or the duration of quasi-monochromatic bursts.

Earth-motion-induced Doppler shifts, and intrinsic frequency evolution of the sources, will reduce the narrow-band signal-to-noise by spreading power across many frequency bins; therefore, it is necessary to correct for these effects before performing the Fourier transform. The corrections can be implemented by a parametrized model in which one searches over a discrete set of points in the parameter space of corrections. The signal-to-noise ratio required for a confident detection depends on the number of points which must be considered in the search.

C. Detection of a Stochastic Background

A stochastic background of gravitational waves creates correlations between (otherwise) independent gravitational wave detector outputs. The correlation of the stochastic background of gravitational waves between two sites is

$$\langle \tilde{h}^*_{\text{gw}\,1}(f)\tilde{h}_{\text{gw}\,2}(f)\rangle = \frac{3H_0^2}{20\pi^2}\frac{T}{f^3}\Omega_{\text{gw}}(f)\gamma(f)$$

where $H_0 \approx 65\,\text{km s}^{-1}\,\text{Mpc}^{-1}$ is Hubble's constant, T is the length of the observation, and $\Omega_{\text{gw}}(f)$ is the measure of the stochastic background of gravitational waves that is convenient for describing cosmological sources (see below). The overlap reduction function $\gamma(f)$ describes how sensitive the two sites are at detecting the stochastic background. Several factors are important in determining the overlap reduction function: (i) the relative alignment of the two detectors (if the two detectors are aligned, they are sensitive to the same polarization of gravitational waves) and (ii) the relative separation of the two detectors (widely separated detectors are most sensitive to low-frequency stochastic background waves that have wavelength greater than the separation). The overlap reduction function approaches a constant at low frequencies. This constant is unity for aligned detectors but less than unity if the detectors are misaligned. The overlap reduction function decreases and becomes an oscillatory function at higher frequencies.

To detect the stochastic background, one simply cross correlates the detector output at the two sites, with an appropriate weighting factor to suppress those frequencies at which the detector combination is not sensitive:

$$2\text{Re}\int_0^\infty \frac{3H_0^2}{20\pi^2}\frac{\gamma(f)\Omega_{\text{gw}}(f)}{f^3 S_1(f)S_2(f)}\tilde{h}_1^*(f)\tilde{h}_2(f)\,df.$$

This method is tuned to detect a stochastic background of gravitational waves with an anticipated spectrum $\Omega_{\text{gw}}(f)$.

If the stochastic background is significantly above the known instrumental noise, it is possible to measure it using the autocorrelation of the noise in a single detector rather than the cross correlation between detectors. The relevant detection method is the same as above but where $\tilde{h}_1(f)$ is the same as $\tilde{h}_2(f)$, $S_1(f)$ is the same as $S_2(f)$, and $\gamma(f) = 1$. In this case, it makes sense to define the characteristic strain

$$h_{\text{char}}(f) = (f_{\text{H}}/f)[\Omega_{\text{gw}}(f)]^{1/2}$$

with $f_{\text{H}} = H_0\sqrt{(3/20\pi^2)} \approx 3 \times 10^{-19}$ Hz; then the autocorrelation statistic is the same as the statistic used to detect an unknown burst (above). Even if the cross correlation statistic is used, this quantity is a reasonable characterization of the strength of the background for detection purposes if the detectors used are sufficiently similar.

IV. ASTROPHYSICAL SOURCES

Since the direct observation of gravitational waves remains elusive (but see Outlook below), candidate sources can only be identified by a combination of theoretical calculations and observations of electromagnetic phenomena which might reasonably be powered by gravitational potential energy. In this section, the likely sources are listed along with estimates of their detectability using current technology.

A. Supernovae and Neutron Star Formation

Supernova explosions are among the most violent and spectacular events in the electromagnetic universe and are believed to be the birthplace of neutron stars and black holes. They occur a few times per century in our Galaxy, and once every few weeks in the Virgo cluster of galaxies. The physics of supernovae is under active investigation. Research suggests several viable scenarios which end in a supernova.

Type-II supernovae, distinguished by strong hydrogen lines, occur when the core of a massive super-giant star ($M > 10M_\odot$) collapses at the end of silicon burning. The core, which is composed of iron nuclei, cannot undergo exothermic nuclear reactions; it is supported against gravitational collapse by the degeneracy pressure of its electrons. The onset of the collapse occurs when the star runs out of silicon to burn. The outer layers collapse and crush the core. The iron nuclei fragment in endothermic reactions, and the electrons are captured by protons to form neutrons; both of these processes reduce the ability of the core to support the star, and the collapse is accelerated. Eventually, the core collapses to the point at which neutron degeneracy pressure is sufficient to support the core. At this stage, the outer layers of the in-falling core bounce off the neutron-degenerate inner core and form a shock wave. A second shock wave is generated by an expanding bubble of neutrinos, which are being convected from the center of the star outward. These shocks expand through the outer layers, heating them as they go, and eventually produce the supernova explosion.

Type-I supernovae have weak or no hydrogen lines. Evidence suggests that these events occur when a white dwarf, or other small dense star, accretes material from a companion in a binary system. Eventually, the star cannot support the additional material, and collapse is initiated. Type-Ia supernovae are thought to be explosions of helium white dwarf stars that accrete enough matter to detonate their thermonuclear fuel. Other type-I supernovae may be caused by the accretion-induced collapse of white dwarfs containing heavier metals, by the collapse of Wolf–Rayet stars, which are the cores of very large stars ($M > 50M_\odot$) whose strong stellar winds have blown off all of their mantle, or by core collapse in helium stars with $M \sim 5M_\odot$ whose outer hydrogen envelope has been stripped by a companion star.

The star which seeds a supernova explosion is known as a progenitor. For gravitational astronomers, bulk properties of the progenitor distinguish possible scenarios for gravitational wave emission. If the supernova progenitor is slowly rotating, the collapse will be nearly spherical, and only weak gravitational waves are expected. (Even in this case, a few cycles of strong gravitational waves may be produced by convective turnover in the hot core.) The larger the rotation rate of the progenitor, the less symmetric will be the collapse. High rotation rates can trigger secular instabilities in the core.

1. Axisymmetric Supernovae

If the core of the supernova progenitor is slowly rotating, the core collapse will be nearly spherical. The gravitational collapse, though violent, produces only a modest change in the quadrupole moment of the star; such supernovae emit weak bursts of gravitational waves. Numerical simulations suggest characteristic strains of $h_{\text{char}} \sim 10^{-21}$ at distances of 10 kpc in frequency bands of ~100 Hz between 100 Hz and 1 kHz.

2. Nonaxisymmetric Supernovae

When the collapsing stellar core has significant rotation, it will flatten to form a rapidly rotating disk (a few hundred kilometers in radius) around a nascent neutron star. At large rotation rates, this disk may become unstable. Matter can clump together into an elongated bar which tumbles end-over-end; this bar-mode grows secularly under dissipation, which can originate with fluid viscosity or gravitational wave emission. If a sizable rotating bar forms, it can generate significant amounts of gravitational radiation. Numerical estimates suggest $h_{\text{char}} \sim 10^{-20}$ at 10 kpc in the frequency band 500–1000 Hz. The gravitational radiation may provide the mechanism by which the disk loses angular momentum during the collapse.

Bar formation may also occur in the proto-neutron star when it is ~20 km in radius. In this case, the growth of the bar is caused by a hydrodynamic instability. Typical characteristic strain amplitudes of $h_{\text{char}} \sim 10^{-19}$ can occur at frequencies of ~100 Hz for sources within our Galaxy ($D \sim 10$ kpc).

3. Nascent Neutron Star Boiling and Neutrino Emission

The newborn neutron star will start its life at a very high temperature and central density. The dominant cooling mechanism in the mantle of the neutron star will be the emission of neutrinos generated during weak nuclear processes within the star. The outer layers of the neutron star will be transparent to neutrino emission, but the neutron star core will initially be opaque. Within the opaque core, fluid will boil up from the center to the point where the neutron star becomes transparent to neutrino emission (the "neutrinosphere"), it will cool, and then settle back to the center. This "boiling" of the neutron star core will generate gravitational radiation. In addition, a neutrino bubble is created around the proto-neutron star, which is responsible for generating the shock that powers the supernova explosion. The characteristic amplitude of gravitational waves due to these convective instabilities is $h_{\text{char}} \sim 10^{-22}$ for events within our Galaxy (within a few tens of kpc).

4. Unstable Modes of Nascent Neutron Stars

Hot, rapidly rotating, newborn neutron stars may be subject to unstable modes of fluid oscillation that will grow, driven by gravitational radiation, on a time scale much shorter than the viscous time scales that would damp these modes in colder neutron stars. The mechanism by which gravitational radiation can cause a mode of fluid oscillation to grow is known as the Chandrasekhar–Friedman–Schutz (CFS) instability. This instability arises when a fluid mode, rotating in a retrograde motion with respect to the neutron star, is dragged forward (prograde) with respect to the inertial frame due to the rotation of the neutron star. In the inertial frame, the gravitational field sees the fluid oscillation as having positive angular momentum; consequently, the gravitational radiation produced by the oscillation carries away positive angular momentum. With respect to the neutron star, however, the oscillation possesses negative angular momentum, so the gravitational radiation reaction causes the oscillation to lose even more angular momentum, which results in the growth of the amplitude of the oscillation.

There are several types of neutron star modes that may exhibit the CFS instability, but the most significant among them are the r-modes. R-modes are described by circular flow patterns with (almost) no radial component; the restoring force for these cycles of fluid motion is the Coreolis force. The r-modes are very similar (and derive their name from) Rossby waves, which are observed in the earth's oceans. The r-modes have the property of always rotating retrograde with respect to the neutron star, but are dragged prograde in the inertial frame due to the neutron star's rotation, and are thus always subject to the CFS instability. The r-modes also radiate copious amounts of gravitational radiation (they possess a significant gravitomagnetic quadrupole moment), and the growth time scale is significantly shorter than viscous time scales in young, hot, and rapidly rotating neutron stars.

It is believed that gravitational radiation from the r-modes in nascent neutron stars is the mechanism by which newborn neutron stars, which could be rotating at near their breakup speed (~1 kHz), lose most of their angular momentum. The result is the slowly spinning neutron stars that are observed. R-modes could produce nearly monochromatic gravitational radiation with a characteristic strain as large as $h_{\text{char}} \sim 10^{-22}$ at frequencies of ~1 kHz and distances of ~10 Mpc, lasting for several tens of seconds.

B. Continuous Waves from Neutron Stars

The most likely sources of quasi-periodic gravitational waves in the frequency bands of terrestrial detectors are rapidly rotating neutron stars. A rotating neutron star will radiate gravitational waves if its mass distribution (or mass-current distribution) is not symmetric about its rotation axis. A neutron star with nonzero quadrupole moment which rotates about a principle axis produces gravitational waves at a frequency equal to twice its rotation frequency. Equally strong gravitational waves can be emitted at other frequencies when the rotation axis is not aligned with a principal axis of the source. If the star precesses, the gravitational waves will be produced at three frequencies: the rotation frequency, and the rotation frequency plus and minus the precession frequency.

1. Isolated Pulsars

Rapidly rotating neutron stars (pulsars) tend to be axisymmetric; however, they must break this symmetry in order to radiate gravitationally. Several mechanisms may lead to deformations of the star, or to precession of its rotation axis, and hence to gravitational wave emission. The characteristic amplitude of gravitational waves from neutron stars at 10 kpc distance scales as

$$h_{\text{char}} \sim 10^{-19} \frac{I}{10^{45}\ \text{g cm}^2} \frac{\epsilon}{10^{-5}} \left(\frac{f}{1\ \text{kHz}}\right)^{5/2} \left(\frac{T}{10^7\ \text{s}}\right)^{1/2}$$

where I is the moment of inertia of the star, f is the gravitational wave frequency, ϵ is a measure of the deviation from axisymmetry, and T is the observation time.

Neutron stars are thought to form in supernova explosions. The outer layers of the star crystallize as the newborn neutron star cools by neutrino emission. Estimates, based on the expected breaking strain of the crystal lattice, suggest that anisotropic stresses, which build up as the pulsar loses rotational energy, could lead to $\epsilon \lesssim 10^{-5}$; the exact value depends on the breaking strain of the neutron

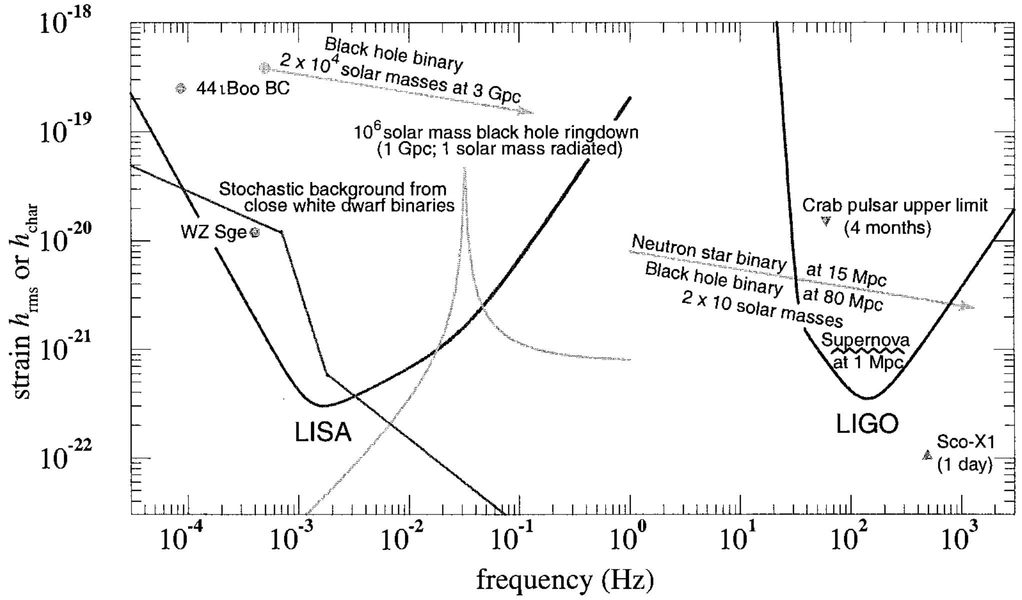


FIGURE 3 Sources of gravitational radiation and detector sensitivities.

star crust as well as the neutron star's "geological history," and could be several orders of magnitude smaller. Nonetheless, this upper limit makes neutron stars a potentially interesting source for kilometer-scale interferometers (see Fig. 3).

Large magnetic fields trapped inside the superfluid interior of a neutron star may also induce deformations of the star. Numerical simulations suggest that this effect is extremely small for standard neutron star models ($\epsilon \lesssim 10^{-9}$).

Another plausible mechanism for the emission of gravitational radiation in very rapidly spinning stars is accretion-driven asymmetries. The principal axes of the moment of inertia can be driven away from the rotational axes by accretion from a companion star. Accretion can produce relatively strong radiation, since the amplitude is related to the accretion rate rather than to structural effects in the star.

2. Low-Mass X-Ray Binaries

A low-mass X-ray binary (LMXB) is a neutron star orbiting around a stellar companion from which it accretes matter. The accretion process deposits both energy and angular momentum onto the neutron star. The energy is radiated away as X-rays, while the angular momentum spins the star up. It has been suggested that the accretion could create nonaxisymmetric temperature gradients in the star, resulting in a substantial mass quadrupole and gravitational wave emission. The star spins up until the gravitational waves are strong enough to radiate away the angular momentum at the same rate as it is accreting; estimates suggest that the equilibrium occurs at a gravitational wave frequency ~500 Hz. The characteristic gravitational wave amplitudes from these sources would be

$$h_{\rm char} \gtrsim 2 \times 10^{-23}\left(\frac{R}{10\ \rm km}\right)^{3/4}\left(\frac{M}{1.4M_\odot}\right)^{-1/4} \times \left(\frac{F}{10^{-8}\ \rm erg\ cm^{-2}\ s^{-1}}\right)^{1/2}\left(\frac{T}{1\ \rm day}\right)^{1/2},$$

where R and M are the radius and mass of the neutron star, and F is the observed X-ray flux at the earth.

The amplitude of the gravitational waves from these sources make them excellent candidates for targeted searches using interferometric detectors. If the source is an X-ray binary pulsar—an accreting neutron star whose rotation is observable in radio waves—then one can apply the exact phase correction deduced from the radio timing data to optimally detect the gravitational waves. (In this process, one must assume a relationship between the gravitational wave and radio pulsation frequencies.) Unfortunately, radio pulsations have not been detected from the rapidly rotating neutron stars in all LMXBs (i.e., neutron stars which rotate hundreds of times a second). In the absence of direct radio observations, estimates of the neutron star rotation rates are obtained from high-frequency periodic, or quasi-periodic, oscillations in the X-ray output during Type I X-ray bursts. But this does not provide precise timing data for a coherent phase correction. To detect gravitational waves from these sources, one must search over the parameter space of Doppler modulations due to the neutron star orbit around its companion, and fluctuations in the gravitational wave frequency due to variable accretion rates. The Doppler effects of the gravitational wave detector's motion can be computed exactly, because the sky position of the source is known.

C. Massive Compact Binaries

The end-over-end tumble of binary star systems is an excellent source of gravitational waves in both the LF and HF bands. The gravitational wave frequency increases with the total mass of the system, and is inversely proportional to the separation of the binary elements. Thus, compact binaries composed of neutron stars and/or stellar-mass black holes radiate at the highest frequencies since the elements can get very close together without merging.

As a binary system emits gravitational waves, it looses energy. The loss of energy is compensated by decay of the orbit so that the binary elements orbit with a smaller separation and at a higher frequency. As energy is drained from the system, the binary elements spiral together and eventually collide. The orbital evolution can be calculated to high accuracy using post-Newtonian formalism which is valid until the very late stages of inspiral for compact objects. (Post-Newtonian formalism is a mathematical scheme which captures Newtonian gravitational effects and slow-motion corrections introduced by general relativity.) The gravitational wave luminosity depends on the masses of the two bodies, m_1 and m_2, their separation (semi-major axis), a, and the eccentricity of the elliptical orbit, e:

$$\frac{dE}{dt} = -\frac{32}{5}\frac{G^4\mu^2M^3}{c^5a^5(1-e^2)^{7/2}}\left(1+\frac{73}{24}\,e^2+\frac{37}{96}\,e^4\right)$$

where $M = m_1 + m_2$ is the total mass of the system and $\mu = m_1m_2/M$ is the reduced mass. The time it takes the binary elements to coalesce due to gravitational wave emission is the decay time of the orbit. For a nearly circular orbit, with initial separation given by a_0, the decay time is

$$\tau = \frac{5}{256}\frac{c^5a_0^4}{G^3\mu M^2}.$$

A pair of neutron stars, with initial separation 500 km, would spiral together in a matter of minutes, whereas a neutron star would take more than a decade to spiral into a super-massive black hole ($10^6 M_\odot$) from 5 times the radius of the black hole. The decay time is shorter for eccentric orbits.

In the LF band, the typical decay time for a binary will be longer than any reasonable observation time, and the gravitational wave signal can be considered to be a continuous wave signal. The characteristic amplitude (for nearly circular orbits) is

$$h_{\text{char}}(f) \approx 1\times10^{-17}\,(4\eta)\left(\frac{M}{10^6M_\odot}\right)^{5/3}\left(\frac{D}{3\text{ Gpc}}\right)^{-1}$$
$$\times\left(\frac{f}{10^{-4}\text{ Hz}}\right)^{7/6}\left(\frac{T}{4\text{ months}}\right)^{1/2},$$

where T is the observation time, D is the distance to the neutron stars, M is the total mass of the binary, and $\eta = \mu/M$ is the mass fraction ($\eta = 1/4$ for equal mass companions, or less if one companion is less massive). Neutron star and stellar-mass black holes are sufficiently compact that they will enter the HF band before merging. In the HF band, the inspiral waveform will "chirp" up in frequency in a relatively short period of time, and the signal will be detectable as a burst rather than a continuous wave. In this case, the characteristic strain is given by

$$h_{\text{char}}(f) \approx 5\times10^{-20}\,\sqrt{4\eta}\left(\frac{M}{M_\odot}\right)^{5/6}$$
$$\times\left(\frac{D}{1\text{ Mpc}}\right)^{-1}\left(\frac{f}{1\text{ Hz}}\right)^{-1/6}.$$

1. Binary Neutron Stars

Compact binary systems with pairs of neutron stars have been observed using radio telescopes. The famous millisecond pulsar PSR 1913+16 discovered in radio data by Hulse and Taylor orbits with a (radio quiet) neutron star companion. Emission of gravitational waves from this system has been confirmed by measuring the decay of the orbit, and the loss of energy agrees with the predictions of general relativity to high precision.

There are a total of three known binary neutron star systems in our Galaxy whose orbits are decaying fast enough that they will collide within the age of the Universe. This suggests that there are many extra-galactic binary neutron star systems that are coalescing even now. Current estimates indicate that the rate of binary neutron star coalescences is in the range between 10^{-8} and 5×10^{-5} events per year per Mpc^3. The expected sensitivity of the future ground-based gravitational wave observatories should make several events per year accessible to gravitational wave astronomers.

During the final few minutes before coalescence, the gravitational wave from a neutron star binary sweeps up in amplitude and frequency ("chirps") through the HF band. When the binary system reaches a frequency of ~1 kHz, the orbit will become unstable either due to the tidal interaction between the two stars or because of a dynamical instability of orbital motion in general relativity. At this stage, the details of the merger may depend on the internal properties and spins of the two neutron stars.

2. Binaries With Black Holes

Black holes in binary systems spiral together, as they emit gravitational waves, just as binary neutron stars do. The gravitational waves from, and the dynamics of, the coalescence will be quite different, however.

Neutron stars in orbit around stellar-mass black holes will emit strong gravitational wave signals in the HF band. Moreover, if the black hole is rapidly rotating, the gravitomagnetic spin-orbit coupling may cause measurable precession of the binary orbit. Also, if a neutron star orbits close to a rapidly spinning but relatively small black hole, tidal disruption of the neutron star may imprint its signature on the gravitational waves emitted from the system.

Unlike a neutron star, the mass of a black hole has no upper bound. Indeed, black holes are known to exist with masses of $\sim 10\ M_\odot$ (e.g., in Cyg X-1), $\sim 10^6\ M_\odot$ (e.g., at the center of our Galaxy), $\sim 10^9\ M_\odot$ (e.g., in M87 at the center of the Virgo cluster of galaxies), and there is recent evidence that black holes with masses of $\sim 1000\ M_\odot$ also exist. Since the frequency at coalescence of the binary system decreases with increasing total mass, more massive binary systems ($M > 10^5\ M_\odot$) will emit gravitational waves primarily in the LF band. The merger of two supermassive black holes will be the strongest emitters of gravitational radiation in the LF band, and could easily be detected to cosmological distances (Gpc) with a space-based interferometer such as LISA.

Binary black hole systems with masses of tens of solar masses should be among the most interesting sources in the HF band. For these systems, the late stages of inspiral and merger will be observable to earth-based interferometric detectors; a direct observation of these waves would provide great insight into strong-field interactions of general relativity. At present, theorists can accurately predict only the early phase of the evolution during which the binary elements have speeds significantly less than the speed of light. The late inspiral and merger phases of these binary systems require the full nonlinearity of Einstein's equations to be addressed; the black holes rapidly spiral together in a whirl of spacetime curvature, and finally merge into a single distorted black hole. Numerical relativity remains the best tool to make theoretical predictions about the dynamics of this phase. Finally, the distortions of the event horizon are rapidly dissipated as the black hole emits gravitational waves; this is known as black-hole ringdown since the waves are predominantly emitted at a single frequency with a decaying amplitude determined by the mass and angular momentum of the black hole.

The merger and ringdown of the black hole will produce a burst of gravitational radiation that is in the HF band for stellar-mass black holes, or in the LF band for supermassive black holes. Typically, the black holes will be rotating close to their maximum angular momentum, in which case the frequency of the ringdown radiation then scales inversely with the mass of the black hole

$$f_{\text{ringdown}} \approx 32\ \text{kHz}\ \left(\frac{M}{M_\odot}\right)^{-1}$$

and the characteristic strain produced by the ringdown burst can be related to the fraction ϵ of the total mass-energy of the black hole that is radiated (a canonical value is 1%):

$$h_{\text{char}} \approx 2 \times 10^{-21}\left(\frac{\epsilon}{1\%}\right)^{1/2}\left(\frac{M}{M_\odot}\right)\left(\frac{D}{\text{Mpc}}\right)^{-1}.$$

The inspiral of compact stellar-mass objects into supermassive (10^5–$10^9 M_\odot$) black holes can be readily calculated by treating the small stellar-mass object as a perturbation to the black hole geometry. If the supermassive black hole is rapidly rotating, the motion of the inspiralling object can be quite complex, with epochs of whirling near the black hole followed by zooms in which it is thrown far away from the black hole. The resulting gravitational waves will be in the LF band and should be easily detected by space-based gravitational wave observatories.

On the speculative side, there may be a population of substellar mass black holes ($M \sim 0.5 M_\odot$) in the Galactic halo. If they exist, these black holes could comprise a substantial portion of the MACroscopic Halo Objects (MACHOs) that have been observed; the population of black hole MACHOs could be as high as 10^{12} if all MACHOs are black holes. If this is the case, then event rates of coalescence of black hole MACHO binaries could be as large as 10^{-5} events per year in our Galaxy, or several per year in the Virgo cluster of galaxies. Once again, detection of the gravitational waves from such a binary would be a dramatic discovery with substantial implications for Galactic dynamics.

3. Binaries with White Dwarfs and Dwarf Stars

Gravitational waves from binary systems with white dwarfs, or other dwarf stars, are generally in the LF band accessible to space-based detectors. Since dwarf stars are considerably larger than neutron stars and stellar-mass black holes, the orbital separation is greater and the wave frequency lower. For example, close white dwarf binary systems and cataclysmic variables (binary systems involving a low-mass red star and a white dwarf) will be detectable at mHz frequencies. In fact, since such systems are so common in our Galaxy, the superposition of the numerous unresolved (for some finite observation time) signals will become hopelessly confused and will produce an effective stochastic background of gravitational waves. Only the strongest emitters will be detectable above this stochastic background, though the background itself will provide valuable information about the distribution and density of the binary systems in our Galaxy. The gravitational waves from sources like WZ Saggitae (a cataclysmic variable) and 44ι Boötes BC (a contact binary of dwarf stars) may be directly detectable because of their known location and orbital periods.

D. Cosmological Sources

A variety of phenomena can produce a stochastic background of gravitational waves arising from the early universe. The conventional measure of the spectrum as a function of frequency f of this stochastic background is given by the function

$$\Omega_{\rm gw}(f) = \frac{1}{\rho_{\rm c}} \frac{d\rho_{\rm gw}}{d \ln f}$$

where $\rho_{\rm c} = 3c^2 H_0^2/(8\pi G)$ is the critical energy density (energy per unit volume) required to make the universe spatially flat ($H_0 \approx 65\,{\rm kms}^{-1}\,{\rm Mpc}^{-1} = 2 \times 10^{-18}\,{\rm s}^{-1}$ is Hubble's constant), and $\rho_{\rm gw}$ is the energy density contained in gravitational waves.

Current knowledge of the physics of the early universe is highly speculative, as are the sources of the stochastic background described below. However, gravitational wave astronomy may provide the only method of directly observing the conditions in the universe as early as $\sim 10^{-22}$ sec after the big bang (in the HF band) since the universe was opaque to electromagnetic radiation until $\sim 10^5$ years after the big bang. The signature of processes which occurred between these times may be imprinted on the spectrum of the stochastic background of gravitational radiation.

1. Microwave Background Radiation Measurements

The measured anisotropy of the Cosmic Microwave Background Radiation (CMBR) places strong constraints on the gravitational wave background from the early universe in the extremely-low-frequency band.

The CMBR was produced in the early universe when the temperature of the universe dropped below $\sim$3000 K and the plasma of protons and electrons combined to form atomic hydrogen. At that time, known as the time of last scattering, the universe became (essentially) transparent to electromagnetic radiation. Observations of the CMBR show that it is highly uniform, with only very small temperature fluctuations. Inflationary models of the very early universe explain this uniformity. However, even if the universe had a perfectly uniform temperature at the time of last scattering, gravitational waves produce observed temperature fluctuations via the Sachs–Wolfe effect. The presence of an extremely-low-frequency gravitational wave causes a difference, from point to point in the sky, in the gravitational potential through which the CMBR must travel to reach the earth. This difference in the gravitational potential induces a change in the apparent temperature of the CMBR. The observed degree of homogeneity of the temperature of the CMBR thus provides an upper limit on the gravitational wave spectrum $\Omega_{\rm gw}(f) < 10^{-10}(H_0/f)^2(h_{\rm char}(f) < 10^{-4}(f_{\rm H}/f)^2)$ for $H_0 < f < 30H_0$.

Future measurements of the polarization of the CMBR with the Microwave Anisotropy Probe (MAP) and the Planck Surveyor satellites will be able to distinguish between scalar perturbations, which are caused by density perturbations, and tensor perturbations, which are caused by gravitational waves, and thus provide a method of directly observing cosmological gravitational waves.

2. Pulsar Timing Limits

The regularity of the electromagnetic pulses from pulsars rivals the best atomic clocks. By measuring the timing residual, which is the amount of time drift between the pulsar "clock" and earth-based atomic clocks, after correcting for Doppler effects due to the relative motion of the earth and the pulsar, a limit on the stochastic background of gravitational waves between the earth and the pulsar can be set: Gravitational waves, with periods comparable to the observation time, cause a slight time dilation which would be in the very-low-frequency band. Measurements of several stable millisecond pulsars for about a decade have given a bound $\Omega_{\rm gw} < 10^{-8}(h_{\rm char} < 10^{-15})$ at $f = 10^{-8}$ Hz. (This bound is on any stochastic background of gravitational waves, not just those of cosmological origin.)

3. Limits Based on Nucleosynthesis

Comparison of measured primordial element abundances to those predicted by proposed cosmological models provides a sensitive test of the model. A similar test can be used to constrain the stochastic background of gravitational waves. The total amount of energy in gravitational waves will affect the expansion rate of the universe during the period of nucleosynthesis, and is thus tightly constrained by the observed abundances. Only those frequencies of the stochastic background radiation with frequencies greater than the Hubble expansion rate at the time of nucleosynthesis are so constrained. The resulting bound is

$$\int \Omega_{\rm gw} d \ln f < 10^{-5}$$

where the integral includes frequencies greater than 10^{-9} Hz.

4. Inflation Models

Inflationary cosmological models invoke an epoch of rapid expansion (power-law or exponential) of the universe at early times to solve various problems in cosmology such as the flatness problem (why the universe is so close to being spatially flat) and the horizon problem (why the CMBR is so homogeneous). Inflation is often predicted to have taken place when the universe had cooled to energy densities associated with physics at the Grand Unification Theory (GUT) scale, approximately 10^{-35}sec after the big bang, when the strong (color) force becomes distinct and quarks form.

If such an inflationary period did exist in the early universe, it would have created a stochastic background of gravitational waves by parametric amplification of the primordial gravitational wave spectrum produced in the big bang. The initial spectrum of gravitational waves produced by the big bang, which is usually taken to be the ground-state quantum vacuum fluctuations of the gravitational field, is dramatically red-shifted during a rapid inflation, but the amplitude of the waves grows so as to maintain nearly the same amount of energy in the gravitational wave background. Typical models of inflation involve expansions of the universe by a factor of 10^{27}, and thus will produce a dramatic increase in the amplitude of the primordial stochastic background of gravitational waves. Predictions based on an exponentially growing inflationary period triggered at GUT-scale energy densities yield a flat spectrum of stochastic gravitational waves:

$$\Omega_{\mathrm{gw}} = (1 + z_{\mathrm{eq}})^{-1}\frac{16}{9}\frac{\rho_{\mathrm{gut}}}{\rho_{\mathrm{pl}}} \approx 10^{-16}$$

where $\rho_{\mathrm{gut}}/\rho_{\mathrm{pl}} \approx 10^{-12}$ is the ratio of the GUT-scale energy density of the universe during inflation to the Planck-scale energy-density $\rho_{\mathrm{pl}} = \hbar^{-1}G^{-2}c^{7}$ at the big bang, and $z_{\mathrm{eq}} \approx 6000$ is the red-shift of the universe when it changed from being radiation dominated to matter dominated. The characteristic amplitude of this stochastic background is $h_{\mathrm{char}} \approx 10^{-8}(f_{\mathrm{H}}/f)$ for $f \gg f_{\mathrm{H}}$.

The stochastic background spectrum in this scenario is flat over many decades of frequency including the range from the ELF band to the HF band. The last two decades of the ELF band, 10^{-16}–10^{-18} Hz, however, contain waves that have “emerged” (have frequencies greater than Hubble’s constant) since the universe became matter dominated; for these waves, the stochastic background spectrum $\Omega_{\mathrm{gw}}(f) \sim f^{-2}$.

The spectrum of the stochastic background of gravitational waves depends on the details of the inflation model as well as on the assumptions about the initial state of gravitational waves produced by the big bang.

5. Gravitational Waves from Cosmic Strings

Cosmic strings may have formed during a phase transition as the strong force separated from the electroweak force when the universe was at the GUT-scale. They are one-dimensional defects, or strings, which oscillate rapidly and produce gravitational waves. Depending on how many (if any) cosmic strings are present in the universe, these oscillations may produce a sizable component of the stochastic background of gravitational waves. Most models of the spectrum produced by cosmic strings show that $\Omega_{\mathrm{gw}}(f)$ is relatively flat in the LF and HF bands. The spectrum typically increases in the very-low-frequency band, where pulsar timing data provides an upper limit on the cosmic string contribution to the stochastic background.

6. Bubbles from a Phase Transition

As the universe cools after the big bang, matter may undergo a phase transition and produce bubbles of material with a lower energy-density than the surrounding material. These bubbles expand relativistically as the latent heat of transition from the newly incorporated matter is converted into the kinetic energy of the bubble wall. The collisions of the expanding bubbles produce large amounts of gravitational radiation.

Unlike the other examples of cosmological sources of gravitational radiation, the spectrum of gravitational waves from colliding bubbles will be peaked at a characteristic frequency set by the time the phase transition took place. For example, for the standard-model electroweak phase transition, the peak frequency is $f_{\mathrm{peak}} \sim 10^{-3}$ Hz, in the LF band, and the spectrum of gravitational waves is expected to be $\Omega_{\mathrm{gw}}(f_{\mathrm{peak}}) \sim 10^{-22}(h_{\mathrm{char}} \sim 10^{-27})$ at this peak frequency. Though this level of stochastic background could not be observed, it is possible that there are other first-order phase transitions in the earlier universe, which could produce significantly larger contributions to the stochastic background of gravitational waves.

V. OUTLOOK

Efforts to directly observe gravitational waves have a history spanning at least 40 years; Joseph Weber constructed the first resonant mass detectors in the early 1960s. At the start of the 21st century, a new generation of interferometric detectors will have sufficient sensitivity to detect many anticipated astrophysical sources. The commissioning and scientific operation of these instruments marks the birth of gravitational wave astronomy.

In the short term, observations will bring information about source populations beyond that available using current astronomical techniques. Direct detections will provide information that is complementary to, but qualitatively different from, that provided by electromagnetic radiation. It also seems likely that novel and unexpected astrophysical sources of gravitational radiation will be discovered soon after the detectors begin operation. Efforts are already under way to improve the sensitivities of these instruments to the point where detection will be routine. As gravitational wave astronomy matures, it will provide unique tests of general relativity and strong-field gravity; these tests will become more powerful when the LF band is opened to observation using space-based antennas.

In the long term, gravitational wave observations will revolutionize both astronomy and gravitational physics. These observations will support some predictions discussed in this article, they will confute others, and they will bring surprises about the Universe in which we live.

ACKNOWLEDGMENTS

The authors are grateful to the National Science Foundation for continued support of their research into gravitational waves through Grants PHY-9970821 and PHY-0071028.

SEE ALSO THE FOLLOWING ARTICLES

Electromagnetics • Gamma-Ray Astronomy • Global Gravity Modeling • Gravitational Wave Detectors • Gravitational Wave Physics • Infrared Astronomy • Neutrino Astronomy • Pulsars • Radio Astronomy, Interferometry Supernovae • Ultraviolet Space Astronomy • X-Ray Astronomy

BIBLIOGRAPHY

Brown, J. D. (2000). Gravitational waves from the dynamical bar instability in a rapidly rotating star. *Physical Rev. D* **62,** 084024.

Hulse, R. A. (1994). Nobel lecture: The discovery of the binary pulsar. *Rev. Modern Physics* **66,** 699–710.

Lindblom, L., Tohline, J. E., and Vallisneri, M. (2001). Non-linear evolution of the r-modes in neutron stars. *Physical Rev. Lett.* **86,** 1152–1155.

Müller, E. (1997). Gravitational radiation from core-collapse supernovae. *Classical Quantum Gravity* **14,** 1455–1460.

Nakamura, T., Sasaki, M., Tanaka, T., and Thorne, K. S. (1997). Gravitational waves from coalescing black hole macho binaries. *Astrophy. J.* **487,** L139–L142.

Peters, P. C. (1964). Gravitational radiation and the motion of two point masses. *Physical Rev. B* **136,** 1224–1232.

Peters, P. C., and Mathews, J. (1963). Gravitational radiation from point masses in a Keplerian orbit. *Physical Rev.* **131,** 435–440.

Saulson, P. S. (1994). "Fundamentals of Interferometric Gravitational Wave Detectors." World Scientific, Singapore.

Schutz, B. F. (1999). Gravitational wave astronomy. *Classical Quantum Gravity* **16,** A131–A156.

Taylor, J. H. (1994). Nobel lecture: Binary pulsars and relativistic gravity. *Rev. Modern Physics* **66,** 711–719.

Thorne, K. S. (1994). "Black Holes and Time Warps: Einstein's Outrageous Legacy," W. W. Norton & Company, New York.

Thorne, K. S. (1997). Gravitational radiation: A new window onto the universe. *http://arXiv.org/abs/gr-qc/9704042*.

Gravitational Wave Detectors

Rosa Poggiani
Università di Pisa and Istituto Nazionale di Fisica Nucleare

GLOSSARY

Black hole Celestial body with such a high density that even light cannot escape from it.

Equation of state Description of the state of a system in terms of its thermodynamics variables pressure, volume, and temperature.

Event horizon Black hole boundary.

Fabry–Perot cavity Instrument in which a light beam is reflected several times between two transparent plates before transmission.

Interferometer Optical instrument that splits a light beam into two beams recombined to produce interference.

Michelson interferometer Interferometer that uses a beam splitter and a pair of mirrors to split and recombine the beams.

Neutron star High-density star consisting mainly of neutrons as a result of gravitational collapse.

Pulsar Rotating neutron star that emits periodic bursts at radio frequencies.

Space–time The four-dimension space which includes the three spatial coordinates and time as the fourth dimension.

Supernova Explosion of a massive star whose core collapses gravitationally.

Transducer Device converting one form of energy to another one, generally electromagnetic.

GRAVITATIONAL WAVES are ripples in space–time caused by the coherent accelerated motion of massive bodies. The potential gravitational wave sources span a wide range of cosmic environments: neutron stars, black holes, supernovae, and binary systems. This review focuses on the instruments currently under construction that are expected to perform the first direct detection of gravitational waves.

I. INTRODUCTION

A. Gravitational Waves in General Relativity

The existence of gravitational waves has been predicted by Einstein since 1916, but to date it remains the last prediction of general relativity that has not been experimentally verified. When large masses are accelerated, the resulting perturbation of the metric of space–time propagates at the speed of light as a ripple. A gravitational wave exerts a tide-like force between pairs of free masses or across solid bodies. The force causes a variation of the relative distances of free masses or of the solid faces. The wave amplitude is expressed by the dimensionless amplitude, or *strain*, h. The strain $h = \frac{\Delta L}{L}$ measures the fractional variation of the distance between two free masses initially at distance L. The gravitational waves have two polarization states rotated by 45°, labeled $+$ (*plus*) and $\times$ (*cross*). The effects of gravitational waves with the *plus* and *cross* polarizations on a ring of test masses are shown in Fig. 1.

Any gravitational wave detector uses masses that can move freely with respect to each other. Two main detection techniques are used:

- *Interferometric detection*: An interferometer is used to measure the change of the relative position between test masses equipped with mirrors; the gravitational wave produces a variation in the length of the interferometer arms and in the interference pattern.
- *Resonant detection*: This technique uses a mechanical resonator, such as an elastic solid body, whose resonance is excited by the gravitational wave.

The prominent characteristic of the gravitational waves is the extreme weakness of their interaction with matter. Gravitational waves are only marginally absorbed or scattered by matter between the source and the detector, opposite to what happens to electromagnetic waves. Moreover, the gravitational waves are preferably emitted by the bulk motion of their sources, typically objects with strong gravity, while astronomical electromagnetic waves are the incoherent superposition of the emission of individual electrons or atoms. Common electromagnetic sources (interstellar gas, stellar atmospheres) are not relevant gravitational wave emitters. On the other hand, gravitational sources are often not electromagnetically visible because they do not emit or because their radiation is absorbed. Thus, the gravitational wave information is complementary to the electromagnetic information. The weakness of the wave coupling to matter preserves the original information of the gravitational event, allowing researchers to investigate the interior of strong gravity sources (neutron stars, supernovae) or to gain information about the very beginning of the universe (down to $\sim 10^{-24}$ s age). The direct detection of gravitational waves holds the potential of a rich physics and astronomy wealth: It could open the field of gravitational wave astronomy to give a new picture of the universe that could be as surprising as the opening of the radio window after the optical one.

B. Indirect Evidence for Gravitational Waves

Indirect evidence for the existence of gravitational waves was provided by the Hulse and Taylor discovery of the

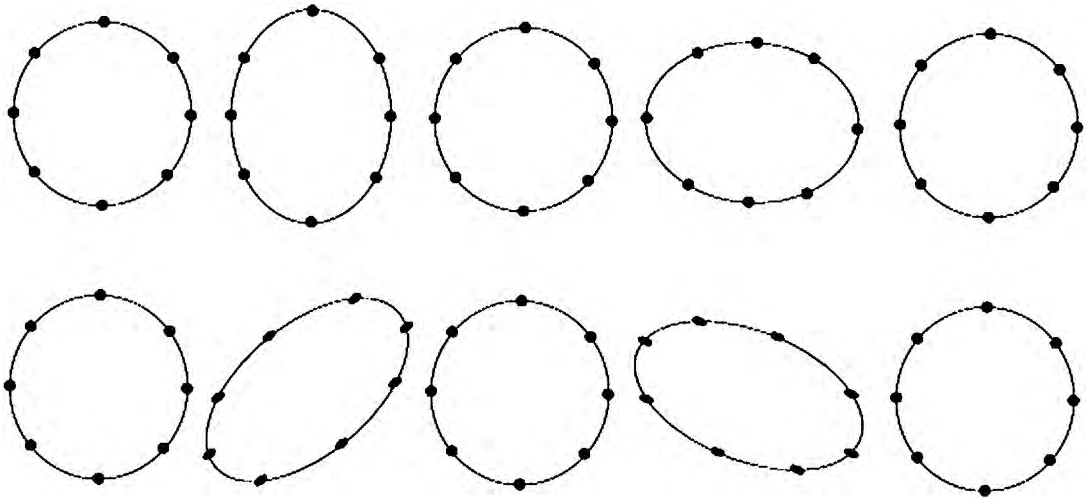


FIGURE 1 Effects of the *plus* (top row) and *cross* (bottom row) polarizations on a ring of masses during one cycle of the gravitational wave.

neutron star binary system PSR 1913+16 and successive observations over many years. The two bodies have masses $\sim$1.4 $M_\odot$ (where $M_\odot$ is the solar mass, $\sim 2 \times 10^{30}$ kg) and have an orbital period of 7.8 hr. The system is radiating energy that causes the two bodies to spiral toward each other speeding up the orbit by 3 mm/orbit. The observed increase in speed agrees with the general relativity prediction of gravitational wave emission to an accuracy better than 0.4%. Hulse and Taylor were awarded the Nobel Prize in 1993.

II. GRAVITATIONAL WAVE SOURCES

The emission of gravitational waves is caused by the quadrupole moment Q of the source, according to this relation:

$$h \sim \frac{G}{c^4}\frac{\ddot{Q}}{r}, \tag{1}$$

where $\ddot{Q}$ is the second derivative of Q, r the distance of the source, G the Newton gravitational constant (6.67×10^{-11} m^3 kg^{-1} sec^{-2}), and c the speed of light.

The radiated power is the gravitational wave luminosity:

$$L = \frac{G}{c^5}\dddot{Q}^2. \tag{2}$$

The quadrupole moment can be approximated by $Q \sim Ml^2$, where M is the source mass and l the scale of deviation from symmetry. An axisymmetric body does not radiate gravitational waves. The strongest gravitational radiators are nonspherical, with $\ddot{Q} \sim 2Mv^2 \sim 4E_\mathrm{K}^\mathrm{ns}$, where v is the velocity, E_K^ns is the nonspherical component of kinetic energy:

$$h \sim \frac{4G}{c^4 r}E_\mathrm{K}^\mathrm{ns} \sim \frac{2G}{c^4 r}Mv^2. \tag{3}$$

A binary system of two masses M at distance r_0 rotating with frequency f about their common center of mass exhibits the maximum variation of quadrupole moment since the whole kinetic energy is nonspherical. The emission occurs at twice the rotation frequency with this intensity:

$$h \sim \frac{4G}{c^4 r}Mr_0^2\omega^2. \tag{4}$$

Due to the tiny value of the factor $\frac{G}{c^4} \sim 8 \times 10^{-45} \frac{\mathrm{sec}^2}{\mathrm{kg\,m}}$, man-made gravitational waves are by far too weak to be detected. A hypothetical dumbbell made of two masses of 10^4 kg at the ends of a 10-m rod rotating at 10 Hz about the center of mass produces waves with an amplitude below $h \sim 10^{-40}$ in the wave zone. The candidate gravitational wave sources are very compact and heavy celestial bodies, like neutron stars and black holes. Bodies with strong gravity in binary systems radiate gravitational waves. For such bodies the radius is of the order of the Schwarzschild radius $r_\mathrm{s} = \frac{2GM}{c^2}$. The emitted wave has an amplitude of

$$h \sim \frac{1}{rr_0}r_{\mathrm{s}1}r_{\mathrm{s}2} \tag{5}$$

where $r_{\mathrm{s}1}, r_{\mathrm{s}2}$ are the Schwarzschild radii of the bodies. For a binary system of two neutron stars in the Virgo cluster of galaxies (at the distance of 15 Mpc, where 1 pc = 1 parsec = 3.1×10^{16} m), with $M \sim 1.4\,M_\odot$, $r_0 \sim 20$ km, and orbital rotation frequency of some hundreds of hertz, the resulting strain is $h \sim 10^{-21}$. This value is the initial goal of ground-based detectors.

Since a gravitational source of mass M cannot be smaller than its Schwarzschild radius, the emission frequency cannot exceed the reciprocal of the travel time of light along r_s:

$$f \le \frac{c^3}{4\pi GM}. \tag{6}$$

Having in mind compact celestial bodies with masses above the Chandrasekar limit of $\sim$1.4 $M_\odot$, the highest expected frequency is $\sim 10^4$ Hz.

The gravitational wave physics span a wide range of frequencies, which is traditionally divided into four regions:

1. *Extremely low frequency region*, 10^{-18}–10^{-15} Hz: The gravitational waves produce quadrupole anisotropies in the cosmic microwave background (CMB) radiation. The wave spectrum is described by the fraction of energy density $\Omega_g(f)$ (in a bandwidth f) needed to close the universe. From observations of the COBE satellite $\Omega_g \le 10^{-9}$ at 10^{-18} Hz.

2. *Very low frequency region*, 10^{-9}–10^{-7} Hz: The gravitational waves produce fluctuations in the arrival times of pulsar radio signals. From the timing of millisecond pulsars, $\Omega_g < \frac{6\times 10^{-8}}{H^2}$ at 4×10^{-9} Hz, where H is the Hubble constant in units of 100 km/sec.

3. *Low-frequency region*, 10^{-4}–1 Hz: The experimental approaches actually under investigation are Doppler tracking of spacecrafts from earth and laser interferometry in space (which is discussed in Section III.F). The lower frequency cutoff is determined by the fluctuations of solar radiation pressure. The potential sources in the low-frequency region are galactic binary stars, supermassive coalescing black hole binaries (masses from 10^2 $M_\odot$ to 10^8 $M_\odot$), neutron stars, and small black holes falling into massive black holes ($M \sim 10^6$ $M_\odot$).

4. *High-frequency region*, 1–10^4 Hz: This region includes ground-based interferometers and resonant detectors, which are discussed in detail in Sections III and IV. The lower frequency cutoff is given by the earth vibrations and by fluctuations of the gravity gradient. Several

sources are expected to emit at high frequency: rotating neutron stars, collapsing supernovae, neutron stars, and small coalescing black holes.

The gravitational wave observational window of the detectors described in this review covers the low- and high-frequency regions (10^{-4}–10^{4} Hz).

From the point of view of the signal they produce, the sources can be classified into three main types: burst sources, with very short duration, i.e., broadband in frequency; narrowband sources, which are periodic or quasi-periodic; and stochastic backgrounds. Catastrophic events, such as supernova explosions or coalescence of binary systems, produce burst signals. Rotating asymmetric neutron stars or binary systems far from coalescence are narrowband sources. The addition of several weak sources produces a stochastic background.

The characteristics of the most relevant sources are described in the next paragraphs. The last years have seen vigorous theoretical efforts directed at source modeling, mainly with the advance of the techniques of numerical relativity.

A. Supernovae

Supernovae have been the first historical target of gravitational wave detectors. The gravitational collapse of a star into a supernova leads to the birth of a neutron star. If the collapse of a type II supernova is not symmetrical, emission of gravitational waves is expected as a burst signal with a timescale on the order of milliseconds. The detailed features of the bursts are not yet very well known. The gravitational emission is related to the fraction of mass that is converted into gravitational waves. Assuming an efficiency of 0.1%, the current estimates predict one event every 30 years with $h \sim 10^{-18}$ within our galaxy (10 kpc) and a few events per year detecting events up to the Virgo cluster (15 Mpc), where several interesting sources are expected.

B. Pulsars

Pulsars are rotating neutron stars with typical radii of $\sim$10 km and masses of 1.4 $M_{\odot}$, with rotation periods ranging from fractions of hertz to hundreds hertz. Pulsars can emit gravitational waves if deviating from axial symmetry. The emission is continuous at twice the rotation frequency, with an amplitude of

$$h \sim \frac{8\pi^2 G}{c^4}\frac{If^2}{r}\epsilon, \tag{7}$$

where f is the rotation frequency, I the average moment of inertia ($\sim 10^{38}$ kg m^2), and ϵ the deviation parameter (ellipticity, estimated to be below 10^{-6}). The expected signal for known pulsar is below 10^{-26} but the signal can be integrated over long time intervals. To date, more than 1000 pulsars are known, out of the 10^9 predicted in the galaxy. The rotation frequency peaks at a few hertz, which demands a low-frequency sensitivity of detectors. Neutron stars could also be gravitational wave radiators for the first few years of life if born with a high rotation frequency.

C. Black Holes

Black holes should be quite common in the universe. There is indirect evidence that stellar mass black holes ($M \sim 10\, M_{\odot}$) are in X-ray binary systems and many galaxies host massive or supermassive (from $M \sim 10^6\ M_{\odot}$ to $M \sim 10^{10}\ M_{\odot}$) black holes at their centers. Despite the great variety of phenomena that can lead to their formation and history, unperturbed black holes are described by mass, charge, and angular momentum only. However, when objects are captured by a black hole, its event horizon vibrates emitting gravitational waves before coming back to equilibrium. The gravitational waves are a strong signature of the black hole physics. The natural emission frequency of a body of mass M and radius R is

$$f_n = \frac{1}{4\pi R}\sqrt{\frac{3GM}{R_{\mathrm{s}}}} \tag{8}$$

where R_{s} is the Schwarzschild radius of black hole. Thus stellar mass black holes oscillate in the kilohertz region, while more massive ones oscillate in the millihertz region or below.

D. Coalescing Binaries

Binary systems such as PSR 1913+16 emit gravitational waves. We have seen above the strain intensity [by Eq. (5)]; for a neutron star binary system in the Virgo cluster, the strain is $h \sim 10^{-21}$. Since the gravitational wave emission produces a gradual shrinking of orbit radius, the wave frequency increases as a chirp, until final coalescence. The lower the cutoff frequency of the detector, the longer the monitoring time of the waveform. The amplitude and the chirp rate depend on the *chirp mass* $M_{\mathrm{c}} = \mu^{3/5} M_{\mathrm{t}}^{2/5}$, where μ is the reduced mass and M_{t} is the total mass of the system. Measuring the chirp time allows for deducing the chirp mass, and the distance to the source can be determined from the amplitude.

The simplicity of the system makes this event the potentially clearest signature for gravitational waves. The coalescence of compact binary systems—neutron star/neutron star (NS/NS), neutron star/black hole (NS/BH), black hole/black hole (BH/BH)—can provide

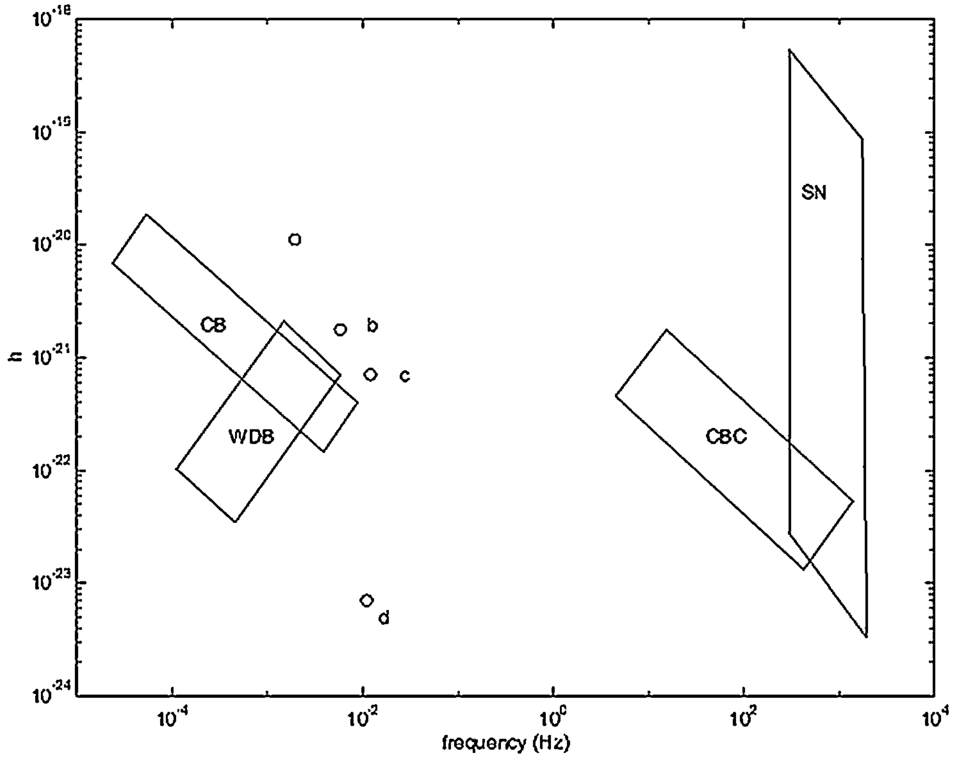


FIGURE 2 The intensity of some astrophysical sources. CB, compact binaries; WDB, white dwarf binaries; CBC, compact binary coalescence; SN, supernovae; *a*, coalescence of binary black holes with 10^6 $M_\odot$; *b*, black hole formations with 10^6 $M_\odot$; *c*, black hole binary with 10^6 $M_\odot$; *d*, black hole–black hole with 10^3 $M_\odot$.

information about several physics topics. The NS/NS coalescence can provide a probe of the nuclear equation of state and hopefully an explanation of the γ-ray burst phenomenon; BH/BH coalescence can provide a superb test of the general relativity theory in the strong gravity regime. The predicted event rate per galaxy is $\sim 10^{-5}$ yr^{-1} for NS/NS and $\sim 10^{-7}$ yr^{-1} for BH/BH coalescence. To have a few events per year, it is necessary to have detectors sensitive up to 200 Mpc (including 6×10^5 galaxies).

E. Stochastic Background

The stochastic background includes all unresolved sources and cosmological gravitational waves. Primordial gravitational waves are of special interest since they should have been generated $\sim 10^{-24}$ sec after the Big Bang. The gravitational waves could probe the universe at a much earlier life than any other measurement, such as the one of CMB.

The detectors under construction have a noise level of the same order of the signals mentioned above. The noise is usually described by the power spectrum $S_\mathrm{N}(f)$ or the spectral density $\tilde{N}(f) = \sqrt{S_\mathrm{N}(f)}$, measured in units of $\frac{\text{noise amplitude}}{\sqrt{\text{Hz}}}$. As an example, if a signal with $h \sim 10^{-21}$ is detected in a bandwidth of 1 kHz, then the spectral density is $\tilde{h} = 3 \times 10^{-23} \frac{1}{\sqrt{\text{Hz}}}$. All sources of noise in the detectors are assumed to be uncorrelated and thus add in quadrature.

The ability to detect signals is given by the *characteristic amplitude* $h_\mathrm{c} = h\sqrt{N_\mathrm{cycl}}$, where N_cycl is the number of cycles spent by the waveform close to the maximum amplitude, i.e., $N_\mathrm{cycl} \sim f^* t_\mathrm{o}$ with f^* typical frequency of the signal and t_o the observation time.

The intensity of the main astrophysical sources is shown in Fig. 2.

III. INTERFEROMETRIC DETECTORS

A. Basic Principles

The first suggestions to use electromagnetic radiation, namely, interferometers, to monitor the variation of relative distance between test masses were made in 1956 by Pirani and in 1962 by Gertsenshtein. The first tabletop prototype Michelson interferometer with an He–Ne laser and optical elements in vacuum was realized in 1970 by

Forward; it achieved a sensitivity of $h \sim 10^{-16}$ Hz$^{-1/2}$ at 1 kHz, comparable with the Weber bar interferometer (see below). In the early 1970s Weiss made the first detailed study of noise sources and the first design of a large-scale interferometer, with potentially a better sensitivity than bars. The study triggered the construction of several prototypes, at MIT (USA), at the University of Glasgow (UK), at the Max Planck Institut for Astrophysics in Garching (Germany), and at Caltech; the arm length ranged from 1 to 40 m. The prototypes allowed a deep understanding of the noise sources and the development of suitable techniques to improve the sensitivity, quickly achieving a sensitivity of $h \sim 10^{-18}$ Hz$^{-1/2}$ at 1 kHz. The success of the prototypes was a determinant for the submission of various proposals for full-scale interferometers to the funding agencies in Europe and in the United States. Several interferometers are now under construction or beginning to operate (see Section III.E).

Interferometric detectors monitor the relative position change between free masses using a laser beam. The basic configuration is the Michelson interferometer (Fig. 3). The laser beam is split into two beams traveling along distinct arms; the beams are reflected by mirrors at the end of each arm, then recombined. The interference signal, measured by a photodiode, is related to the relative change in the arm length due to the gravitational wave. The beam splitter and the mirrors are suspended as pendulums; with this arrangement they behave as free masses above the pendulum resonance frequency. Assuming that the gravitational wave is polarized along the arm directions, the arm lengths will change in antiphase, with a total change in the optical path:

$$\Delta L = hL, \tag{9}$$

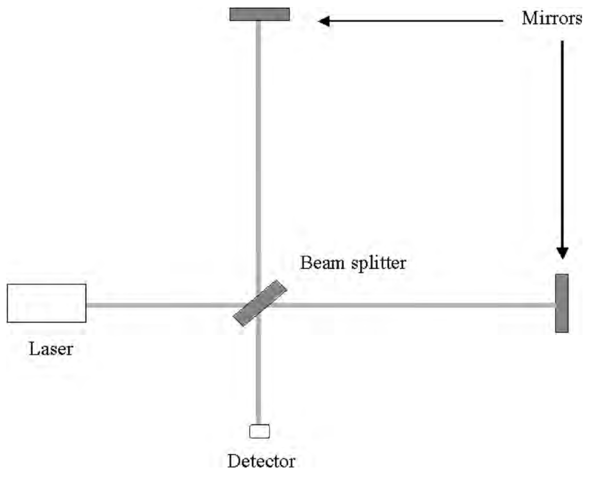


FIGURE 3 Scheme of an interferometric detector.

where L is the physical arm length. For an arm length of 3 km and a strain $h \sim 10^{-21}$, the change in the optical path is the very tiny value $\Delta L \sim 3 \times 10^{-18}$ m, which is $1/1000$ the diameter of a nucleus. We see that the sensitivity of the interferometer is proportional to the arm length. Practical considerations limit the physical arm lengths to a few kilometers. The optical path can be increased by folding the light up to the optimum value of one-half of the wavelength of the gravitational wave. This corresponds to 150 km for a 1-kHz wave, which gives an optical path variation of $\Delta L \sim 1.5 \times 10^{-16}$ m. Even with such a large optical path, the induced displacement is much smaller than the thermal vibrations of single atoms of test masses. However, the precision in the position of the test mass is defined by the averaged position of all atoms over the laser beam width, which is below the level of the displacement due to gravitational waves.

The first solution for folding the optical path is the Herriot delay line, with two concave mirrors having a radius of curvature such that the injected light exits from the entrance hole after N round-trips, which means after an optical path $2NL$. The number of reflections is limited by the reflection losses, thus the mirrors should have high optical quality over the whole surface. During the N trips the laser spots move over a circle with radius $N\omega_0$, where ω_0 is the minimum laser beam width; in typical working conditions, the mirrors should be 1 m in diameter, a number with a great impact on the scale and the cost of the vacuum system.

The second solution is the use of Fabry–Perot cavities, where the laser spots from all reflections coincide. A Fabry–Perot cavity consists of two plane mirrors at distance L. The power transmitted by the cavity is maximum at the resonance, when $kL = n\pi$, where $k = \frac{2\pi}{\lambda}$ is the wave vector. In this case, the power trapping in the cavity replaces the many round-trips of the laser beam. The phase of the light leaving from the cavity depends on the cavity length. If the cavity is working at resonance, the passage of a gravitational wave produces a phase shift near the resonance. The resonance characteristics of the Fabry–Perot cavity are described by the finesse F, which is related to the reflectivities r_1, r_2 of the mirrors:

$$F = \frac{\pi\sqrt{r_1 r_2}}{1 - r_1 r_2}\ . \tag{10}$$

A Fabry–Perot cavity with finesse F produces a phase shift $\frac{2F}{\pi}$ times the shift of a single trip. The Fabry–Perot mirrors must be large enough to allow the reflection of a single beam, thus with a diameter of a few tens of centimeters. In addition to the high reflectivity, the mirror substrates must be of very optical quality, since light enters the cavity through one of them. Since a single Fabry–Perot cavity is very sensitive to changes in cavity length and to changes

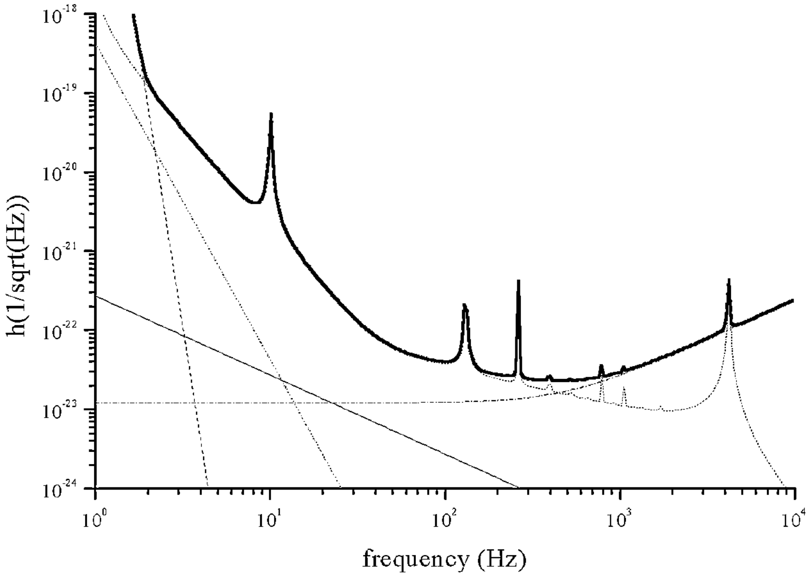


FIGURE 4 Sensitivity curve (thick solid line) for a 3-km baseline interferometer with 30-kg test masses. The sources of noise are outlined: standard quantum limit (thin solid line); seismic noise (long dashed line); gravity gradient noise (dotted line); thermal noise (short dashed line); shot noise (dash–dot line).

in laser frequency, there is a Fabry–Perot cavity in each arm. The laser is locked in frequency to the first cavity and the second cavity is locked to the wavelength of the laser. The locking signal of the second cavity gives the relative variation of the cavity length with respect to the first cavity.

The variation of the optical path corresponds to a phase shift:

$$\Delta\Phi = 4\pi \frac{\Delta L}{\lambda}, \tag{11}$$

where λ is the laser wavelength. The phase shift is measured operating the interferometer at a *dark fringe*, i.e., at a minimum of interference, to reduce the effect of the laser power fluctuations. The wave intensity is deduced from the error signal that must be applied to maintain the dark fringe when the test masses move because of the gravitational wave.

We see that the response of interferometers is inherently broadband. The beam pattern of the antenna is almost nondirectional. An interferometer is very different from a telescope, which can be pointed to a specific location in the sky. The position of the source in the sky is determined measuring the difference in the arrival time of signals in detectors at distant locations. The broad angular response of the interferometer, however, is an advantage for a survey of the sky.

B. Noise Sources

Noise in an interferometer arises from fundamental physics processes, such as shot noise and thermal noise, or from technical factors, such as seismic noise and laser fluctuation noise. The first category sets the intrinsic limits to the interferometer sensitivity. The various sources of noise and their contribution to the interferometer sensitivity are summarized in Fig. 4.

1. Photon Shot Noise and Radiation Pressure Noise

The shot noise is caused by the fluctuations in the number of photons detected at the photodiode. The shot noise has a spectral density of

$$\tilde{h}_{\mathrm{sn}} \approx \frac{1}{NL}\sqrt{\frac{\hbar c\lambda}{2\pi\eta P}}, \tag{12}$$

where P is the laser power and η the quantum efficiency of the photodiode ($\eta \sim 1$). The contribution of this noise is minimized by having a large laser power. All long baseline interferometers have chosen neodymium:YAG (Nd:Yag) lasers, with a wavelength of 1.064 μm and typical powers of 10–20 W. With an effective arm length of $NL \sim 150$ km, the power laser required to achieve $\tilde{h} = 3 \times 10^{-23} \frac{1}{\sqrt{\mathrm{Hz}}}$ is well above the current laser performance. The effective laser power is increased by using the *power recycling* technique. The optimal operation point of the interferometer is in *dark fringe*, with most of the laser light bounced back to the entrance. A mirror in front of the laser in a suitable position can reflect the light into the interferometer again, allowing a gain factor in the available power of up to 1000. The technique of power recycling and the high reflectivity needed demand mirrors with scattering and absorption losses below a few parts per million. The substrate material should have a very low expansion coefficient because of heating from the high-power laser.

The laser power cannot be arbitrarily high because the laser light induces mirror motion because of radiation pressure. The linear spectral density of the radiation pressure noise is

$$\tilde{h}_{\mathrm{rp}} \approx \frac{N}{LMf^2}\sqrt{\frac{2\hbar P}{\pi^3 c\lambda}}, \tag{13}$$

where f is the frequency and M is the mass of the optical element. Shot noise and radiation pressure noise contribute in quadrature as optical readout noise. There is an optimum power that minimizes this contribution ($P = \frac{\pi c\lambda M f^2}{2N^2}$), giving the *standard quantum limit* (SQL):

$$\tilde{h}_{\mathrm{SQL}} \approx \frac{1}{\pi f L}\sqrt{\frac{2\hbar}{M}}. \tag{14}$$

The quantum limit is below the sensitivity of the interferometers under construction but at the level of second-generation detectors. Thus the shot noise is the limiting noise at high frequencies for terrestrial detectors.

Despite the fact that the significant term for shot noise is the effective optical length *NL*, it is still important to have a large physical arm length L. In fact, the other sources of noise at low frequency are thermal noise and seismic noise, which are displacement noises causing the motion of the optical elements according to

$$h = \frac{2x}{L}, \tag{15}$$

where x is the displacement and L is the physical arm length. Thus the physical scale of the interferometer is mainly determined by displacement noises.

2. Seismic Noise

For ground-based interferometers, the ground motion (seismic noise) is several orders of magnitude above the expected path difference induced by the gravitational wave. The optical elements must be isolated with special isolation systems. The main goal of isolation systems is the reduction of horizontal motions along the laser direction. The seismic noise is described by the linear spectral density:

$$\tilde{x} = \frac{A}{f^2}\frac{\mathrm{m}}{\sqrt{\mathrm{Hz}}} \tag{16}$$

where the coefficient A ranges from 10^{-7} to 10^{-6} for different sites. An attenuation factor of 10^{10} at least is required at 10 Hz. The isolators use multistage pendulums. A pendulum suspension with resonant frequency f_{r} provides an isolation factor of $\frac{f_{\mathrm{r}}^2}{f^2}$ above f_{r}, i.e., a factor of 100 at 10 Hz for a 1-Hz pendulum. A cascade of N pendulums provides an attenuation of $(\frac{f_{\mathrm{r}}^2}{f^2})^N$ above f_{r}. The presence of unavoidable cross-couplings between different degrees of freedom, mainly the vertical one, demands additional isolation. The vertical cross-coupling is due to asymmetries of the suspension. However, hanging pendulums point toward the earth's center: there is an intrinsic lower limit given by the ratio of interferometer arm length to the earth radius ($\sim 10^{-3}$ for long baseline interferometers).

The isolation systems are designed in such a way that their mechanical resonances do not inject additional noise. Mechanical isolators are generally passive systems consisting of mass–spring systems. Historically, the first developed detectors were stacks of alternating layers of steel and rubbers. The stacks provide horizontal and vertical isolation at the same time. They generally have a large temperature coefficient and how to make them vacuum compatible is not obvious. Another solution is the use of cascaded multistage pendulums equipped with cantilever springs to achieve vertical isolation. The suspension materials should have low creep since the sudden release of energy could introduce additional noise. In active isolation techniques the relative motion of test masses and the suspension point are sensed and the signal is used to move the last one. The isolation system can include a preisolator stage with very low resonant frequency since the residual motion of test mass is on the order of a few microns: The locking of the laser becomes difficult since one needs large forces.

There is a lower limit to the achievable seismic attenuation because of direct coupling of seismic motion to test masses (gravity gradient noise) due to motion of the ground close to the interferometers, motion of close objects, etc. This noise is well below the sensitivity of the forthcoming interferometers, but could maybe preclude the achievement of SQL.

The isolation systems of most interferometers currently under construction are passive systems (with a small degree of active parts) designed to provide good attenuation down to tens of hartz or a few hertz at best. Thus seismic noise is the dominant noise source at low frequencies in terrestrial interferometers.

Whatever the choice of suspension systems, the optical elements of first-generation interferometers are suspended with one or two loops of wires as pendulums to behave as free masses.

3. Thermal Noise

As soon as seismic noise has been reduced, the dominant noise source at low frequency is thermal noise. This noise arises from the modes of the test masses and the last stage of their suspensions: pendulum mode, violin modes of the wires, and internal modes of the test masses. Each vibration mode can be described as a damped oscillator at temperature T that is excited by thermal noise to a mean energy $k_{\mathrm{B}}T$, where k_{B} is the Boltzmann constant. The contribution from the residual gas damping is made negligible by operating the interferometer in high vacuum. The main contribution to damping is coming from internal dissipation in the suspension material, or *anelasticity*. To describe oscillator damping, the usual elastic constant k is replaced by a complex elastic constant $k(1+i\phi(\omega))$, where $\phi(\omega)$ is the loss factor, which is a function of frequency. At the resonance frequency ω_{r} the loss factor is the reciprocal of the quality factor of the oscillator, i.e., $\phi = \frac{1}{Q}$. The power spectral density of the thermal noise of an oscillator of mass M and loss factor ϕ can be deduced from the fluctuation–dissipation theorem:

$$\tilde{x}^2(\omega) = \frac{4k_{\mathrm{B}}T\omega_{\mathrm{r}}^2\phi}{\omega M\left[\left(\omega_{\mathrm{r}}^2-\omega^2\right)^2+\omega_{\mathrm{r}}^4\phi^2\right]}. \tag{17}$$

At the resonant frequency the noise spectral density has a maximum; thus resonant frequencies should lie outside the bandwidth of the detector or be narrow enough to be filtered in the data analysis. Far from resonance, the spectral density of thermal noise is lower for low loss factors ϕ: The test masses and the last stage of isolation systems are made of low loss materials. There is experimental evidence that for most materials the loss factor is almost independent of frequency; thus below resonance $\tilde{x}^2(\omega) \propto \frac{1}{\omega}$ and above resonance $\tilde{x}^2(\omega) \propto \frac{1}{\omega^5}$.

The isolation systems contribute with several modes, mainly the pendulum mode of the test mass and the violin modes of the suspension wires. Since test masses are suspended as pendulums, the effective loss factor is

$$\phi_{\text{pendulum}} = \phi\frac{n}{2Mgl}\sqrt{ITY}, \tag{18}$$

where ϕ is the loss factor of wire material, n the number of wires, M the pendulum mass, l the pendulum length, I the moment of cross section of wire, T the wire tension, and Y Young's modulus. Typical values of ϕ are 10^{-3}–10^{-4} for metals and $\sim 10^{-8}$ for monolithic fused silica suspensions. Most long baseline interferometers will start with conventional wire suspensions.

The violin modes are the vibrations of the suspension wires, an harmonic series starting at about a few hundreds of hertz. The violin modes have a quality factor on the order of the pendulum quality factor and look like narrow peaks in the sensitivity band.

The normal modes of the test mass have a typical frequency on the order of $f_{\text{int}} = \frac{v_{\mathrm{s}}}{2d}$ (where v_{s} is the sound velocity): The frequencies are above the bandwidth of the detector. Due to the $\frac{1}{\omega}$ tail below resonance, materials with very low losses are used to reduce this contribution. The material should also be low absorption and low expansion because of shot noise requirements. The material currently used is fused silica, with a quality factor of $\sim 10^7$. The way test masses are suspended has an impact on the losses and must be carefully studied.

The choice of the quality factors of the suspension components is not trivial. From the point of view of thermal noise, the quality factor of the suspension modes should be high to reduce dissipation, but from the point of view of seismic noise it should be low to reduce noise amplification $\sqrt{Qf_{\mathrm{r}}}$ at the resonance. The strategy is to use relatively low quality factors for the suspensions stages and a high value for the last stage.

In typical working conditions, the thermal noise dominates the radiation pressure noise at low frequency.

4. Other Noise Sources

The sensitivity of forthcoming laser interferometers is limited by seismic noise at low frequencies (from a few hertz to tens of hertz according to the isolation strategy), by thermal noise up to some hundreds of hertz, and by shot noise above. Other sources of noise, however, can contribute to the total noise budget:

- The fluctuations of the refraction index of residual gas along the laser beam can induce noise. For this reason, all interferometers operate in a high vacuum ($\leq 10^{-8}$ mbar). High-vacuum operation also prevents test mass damping by friction. The vacuum should also be contamination free to preserve the optical properties of the optics. The long baseline laser interferometers will be the largest evacuated volumes on earth, which accounts for a large part of the construction costs.
- The light scattered by the optical elements can be reflected by the vacuum vessel, which is not isolated from seismic vibrations, and can recombine with the unscattered beam. The vacuum tubes are thus equipped with baffles to eliminate the reflected beam rays.
- The fluctuations in the laser frequency contribute to noise if the interferometer arm lengths are not equal. For a mismatch x in the arm length, the required frequency stability is

$$\frac{\Delta f}{f} \simeq h\frac{L}{x}. \tag{19}$$

The lasers are stabilized using high finesse reference cavities, typically a Fabry–Perot cavity in one arm.

- The laser power fluctuations contribute to noise if the interferometer is slightly offset from the dark fringe by an amount δL. The required power stability is

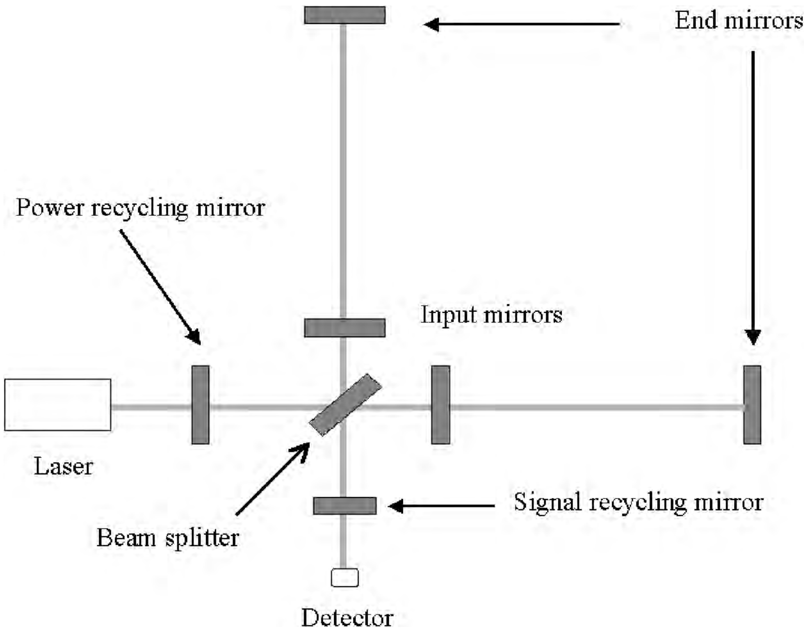


FIGURE 5 Schematics of an advanced interferometer using power recycling and signal recycling.

$$\frac{\Delta P}{P} \simeq h \frac{L}{\delta L}. \tag{20}$$

Active stabilization techniques are used.

- The laser beam geometry fluctuations (lateral and angular movements, wavefront or size variation) produce noise if the recombining beams are not perfectly aligned. The forthcoming interferometers will include long mode cleaning cavities (high finesse optical resonators) in front of the laser.

C. Technical Requirements

An interferometer needs both local and global controls. Local controls provide the damping of the normal modes of the suspension systems at low frequency. Global controls provide control of the interferometer arm length to a relative motion of $\sim 10^{-12}$ m. Since residual motion of test masses induced by seismic noise is still present, the interferometer output moves over many fringes, unless it is operated at a fixed point (dark fringe). Feedback techniques keep the interferometer fixed at the working point; a sensor reads how far the device is from the working point and an actuator brings it back there. The useful signal of the interferometer is no longer the output power, but the intensity of the feedback needed to keep the interferometer at the working point (locking). The position of test masses is corrected using coils acting on magnets on their surface. To reduce the effect of laser power fluctuations and $1/f$ noise in general, the measurement is performed in the megahertz region using modulation and demodulation techniques.

The basic interferometer comes in different variants to improve sensitivity. In a standard interferometer only a very small part of the light gets to the detection region; most goes back to the input. In the *power recycling* technique mentioned above, the light is reused using a partially transmitting mirror to add it coherently in phase with the laser light. The sensitivity of the interferometer can be increased in a narrow bandwidth using the *signal recycling* scheme. A partially reflecting mirror at the output of the interferometer makes the whole interferometer a resonating cavity. In the *resonant recycling* scheme, a mirror set is used in such a way that after a gravitational wave period the light in each arm is exchanged, building at the end a large phase shit. The scheme is useful for realizing enhanced sensitivity in a narrowband and useful to detect periodic signals. In the *dual recycling* scheme, both a power recycling mirror and a signal recycling mirror are used. A scheme of an advanced design interferometer using dual recycling is shown in Fig. 5.

D. Interferometric Detectors in Operation

The sensitivity of the prototypes with tens of meters of arm length are on the order of a few units of $10^{-19} \frac{\mathrm{m}}{\sqrt{\mathrm{Hz}}}$ above some hundreds of hertz. In addition to the invaluable contribution to the development of the techniques of interferometric detection, the prototype interferometers have been used to perform gravitational wave searches.

Among these first runs, the 100-hr coincident run between the 10-m Glasgow and the 30-m Garching prototypes has demonstrated the possibility of reliably and simultaneously operating such complex instruments for long times. The coincident run has set an upper limit of 4.9×10^{-16} on gravitational wave bursts. More recently, the 40-m prototype at Caltech has posed the upper limit of 0.5 hr^{-1} on the rate of coalescence of binary neutron stars in the galaxy.

E. Long Baseline Interferometers under Construction

The major planned ground-based interferometers are LIGO, VIRGO, GEO600, TAMA, and AIGO. The sensitivity is shown in the top part of Fig. 6.

The LIGO project is a Caltech/MIT collaboration consisting of two detectors, each with a 4-km arm length, at Hanford (Washington state) and Livingston (Louisiana), plus a 2-km detector sharing the vacuum system of the 4-km apparatus at Hanford. All detectors are Michelson interferometers equipped with Fabry–Perot cavities and use power recycling. The seismic attenuation system is a four-stage stack cascade. The detectors are expected be operative in 2002.

The VIRGO project is a French/Italian collaboration using a 3-km arm length detector close to Pisa, Italy. The Michelson interferometer is equipped with Fabry–Perot cavities. The seismic attenuation system is the combination of an inverted pendulum for preisolation and a five-stage pendulum cascade (*superattenuator*). The arrangement allows for an extension of the sensitivity band down to 4 Hz. The interferometer is actually under construction and is expected to be operative in 2002.

The GEO600 project is a British/German collaboration using a 600-m arm length detector, which is a pure Michelson interferometer. The seismic suspension system uses stacks in the first stages and multiple pendulums in the last stages. The detector is designed to use from the beginning several second-generation techniques for the optical configurations (signal recycling) and the suspensions (monolithic fused silica) to approach the sensitivity of the longer baseline detectors in some regions. The interferometer will be operational in 2001.

The TAMA300 project is a Japanese interferometer with 300-m arm length and Fabry–Perot cavities. The detector began operating in 1999 with a sensitivity of $\sim 3 \times 10^{-20} \frac{1}{\sqrt{\text{Hz}}}$ and has recently set an upper limit of 0.59 hr^{-1} on the rate of coalescence of close binary neutron stars.

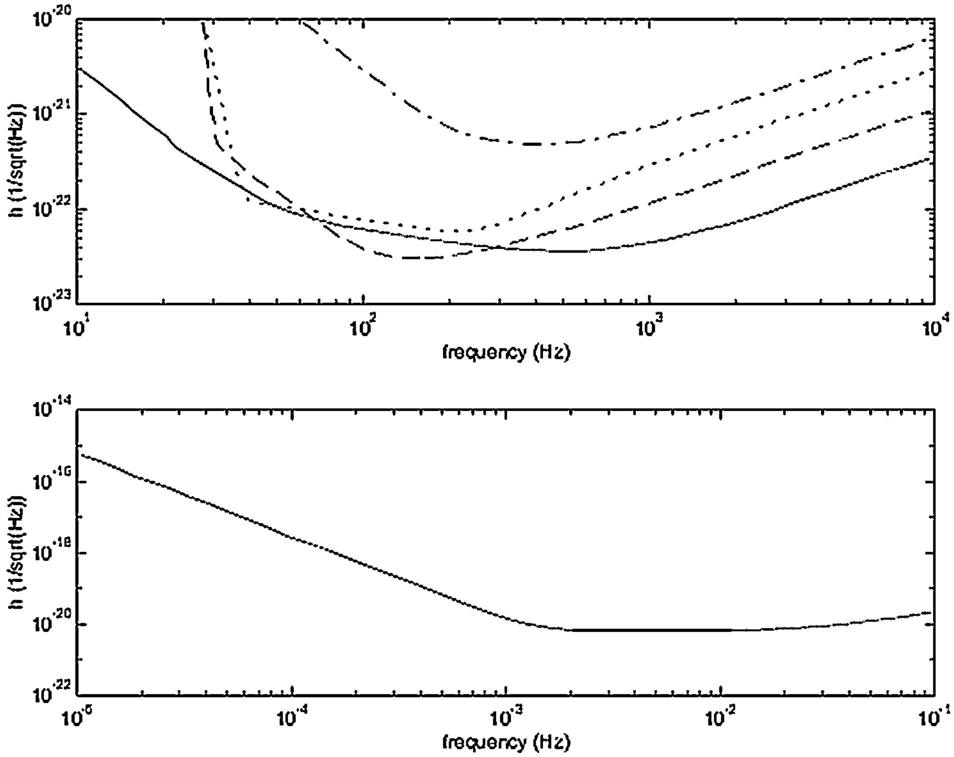


FIGURE 6 Strain noise sensitivity of the interferometric detectors. Top figure: ground-based interferometers VIRGO (solid line), LIGO (dashed line), GEO600 (dotted line), and TAMA (dot–dash line). Bottom figure: the space interferometer LISA.

The first stage of the AIGO detector in Australia was completed in 2000.

F. Very Long Baseline Interferometers in Space

Laser interferometry in space is immune form seismic noise and allows for investigation of the frequency range from 0.1 mHz to 1 Hz. The LISA interferometer is a Michelson made of three spacecraft 5×10^6 km distant from each other at the vertexes of an equilateral triangle. The third arm is used for polarization studies (and gives redundant information). The three spacecraft are in a heliocentric orbit lagging of 20° behind the earth. The relevant noises are a result of solar radiation pressure and solar wind, which demand accurate shielding of the test masses. The sensitivity is shown in the bottom part of Fig. 6.

IV. RESONANT DETECTORS

A. Basic Principles

Historically, resonant detectors were the first gravitational detectors, since the work by Weber during the 1960s. Weber searched gravitational waves from star collapse using room-temperature aluminum bars resonating at 1600 Hz. Tentative evidence for gravitational events was found using two detectors at a 1000-km distance in coincidence, but not confirmed by other detectors. In search of a higher sensitivity, in the 1970s the development of the first cryogenic detectors started all over the world: Stanford, Louisiana State University, University of Rome in Italy, and University of Western Australia. Several cryogenic antennas are now in operation.

A pair of masses attached to a spring is the simplest type of resonant gravitational wave detector. A gravitational wave with a frequency spectrum including the resonance frequency of the detector f_r induces an oscillation of the masses about their center of mass. The lowest resonant frequency of an elastic solid body such as a metal cylinder can be modeled as a pair of masses connected by a spring. The typical resonant detectors are cylindrical metallic bars of some meters length and some tons mass, oscillating in the fundamental longitudinal mode (resonant detectors are also named *bars*). The fundamental oscillation occurs at frequency $\omega_r = 2\pi f_r = \frac{\pi v_s}{L}$, where v_s is the sound velocity in the bar material and L the bar length. A bar is most sensitive to signals with a frequency very close to its resonant frequency. The actual bars are resonant at about 1 kHz, a frequency relevant for detection of supernova events. The efficiency of a bar is given by its cross section, the ratio of absorbed energy to the incoming energy:

$$\Sigma \sim \frac{8GM}{\pi c} \frac{v_s^2}{c^2}. \qquad (21)$$

The bar should be as massive as possible and made of a material with a high sound velocity. Similar to interferometers, resonant detectors are almost omnidirectional. A resonant detector needs some further stages to produce a signal (see Fig. 7). A transducer transforms the mechanical oscillation into an electric signal, which is then amplified. The transducer is a sensor that can also be modeled as a mass connected to a spring. The transducer frequency f_t is tuned to resonate at the same frequency f_r of the bar. The resonant detector and the transducer behave as two coupled oscillators in series. This system has two normal modes at the frequencies $f_\pm = f_0(1 \pm \frac{\sqrt{\mu}}{2})$, where $f_0 = f_r = f_t$, $\mu = \sqrt{\frac{m}{M}}$, where M, m are the masses of the bar and of the transducer, respectively. The ratio μ is low to achieve a large transducer motion. When the gravitational wave arrives, the bar initially experiences the maximum displacement; the transducer is almost quiet. Bar and transducer exchange energy through beatings of the normal modes, until the relative displacement is maximum. Finally the transducer vibrates with an amplitude $\sqrt{\frac{M}{m}}$ times larger than the original vibration of the bar. The bandwidth of a resonant detector is of the order of the distance between the normal modes $\Delta f = f_r\sqrt{\frac{m}{M}}$, of the order of a few hertz for current detectors. A much larger bandwidth would require close mass values for transducer and antenna, at the cost of a smaller transducer motion.

The efficiency of the transducer is characterized by the ratio β of the electrical energy stored in the transducer to the total mechanical energy stored in the detector. A transducer with high β value allows a larger bandwidth since it reaches equilibrium with the bar faster. Since $\omega_r t_m \sim \frac{1}{\beta}$, if $\beta \sim 1$ the bandwidth could be as large as the resonant frequency. Achieving such performance is difficult since it has been experimentally observed that a high β causes a

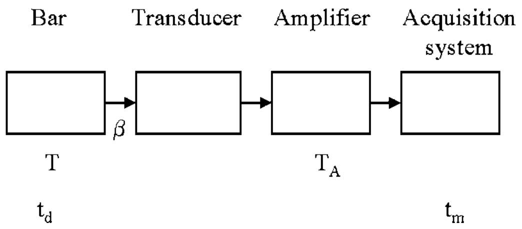


FIGURE 7 Schematics of a resonant detector. The bar, at temperature T and with damping time t_d, is coupled (with coupling β) to a transducer that converts the mechanical energy to electrical energy. The signal is then amplified and acquired with an integration time t_m.

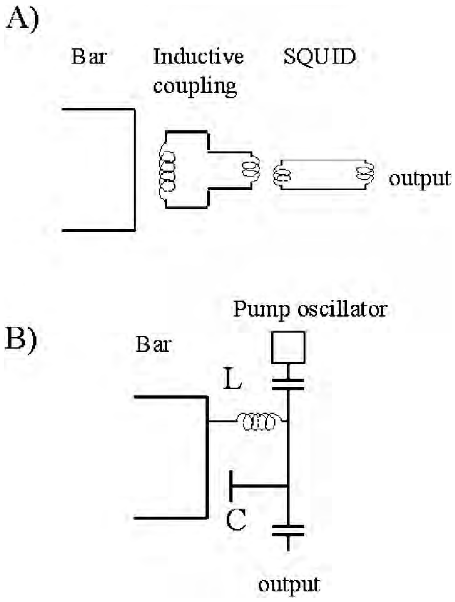


FIGURE 8 (A) Inductively coupled transducer; (B) parametric transducer using capacitive modulation.

degradation of the antenna damping time. The transducer can be either passive or parametric (Fig. 8). Passive transducers use capacitive or inductive sensing; the variation of the capacitance or inductance is proportional to the bar motion relative to the capacitor or inductor. Since there is no intrinsic gain, an amplifier is needed, usually a SQUID (superconducting quantum interference device). The limits of this technique are electrical losses in the circuits and the SQUID noise. Parametric transducers use a high-frequency resonator, whose frequency is modulated by the changing capacity or inductance. There is an intrinsic gain since they have an intrinsic pump oscillator. The coupling factor β is enhanced by the electrical quality factor of the resonator. Most detectors uses passive transducers with $\beta \leq 10^{-2}$.

B. Noise Sources

1. Thermal Noise

In interferometric detectors thermal noise mainly contributes far from resonances, since care is taken to shift most of them outside the detection band. In bars the resonance provides the detection mechanism. The resonance can also be excited from noise, including thermal noise. Thermal noise has a power spectral density $S_{\text{tn}} = k_B T \frac{M\omega}{Q}$, thus the expected rms amplitude is $x_{\text{tn}} = \sqrt{\frac{k_B T}{M\omega_r^2}}$. The recipe to reduce the thermal noise contribution is the choice of material with a high quality factor and low-temperature operation. The material of the bar should have a high quality factor to reduce thermal noise, but it should have density and high sound velocity to maximize the absorbed energy. The best choices are Nb, Al5056, and sapphire. Sapphire exhibits the highest quality factor (3×10^9), but it is not available on the scale of a few tons. The current operating bars are made of Nb or Al5056 and operate at cryogenic temperatures. For a typical bar with a mass of 1 ton and a length of 3 m, resonant at 1 kHz, the expected minimum detectable $h = \frac{x_{\text{tn}}}{L}$ is $\sim 10^{-16}$ for operation at room temperature and $\sim 10^{-17}$ for operation at liquid helium (4.2 K). In both cases, the noise is too high to allow the detection of the astrophysical sources discussed above. However, if the quality factor Q of the resonance is high, the oscillation induced by a burst event will be very slowly damped with a typical time $t_d = \frac{1}{\omega_r Q}$. If the measurement is performed with an integration time t_m shorter than t_d, the effective rms noise budget is diluted by the factor $\sim\sqrt{\frac{t_m}{t_d}}$. The damping time is maximized by having a resonant detector with a high quality factor Q. The dependence of the dissipation factor on the frequency is not relevant, since the bandwidth of the detector is very narrow. In this way, the effective noise temperature $T_e = T\frac{t_m}{t_d}$ can be much lower than the physical temperature T. The spectral density of thermal noise is

$$S_{\text{tn}}(\omega) = 2k_B T \frac{t_m}{t_d}. \qquad (22)$$

2. Electronic Noise

Any transducer modulates a source of electrical energy, acting in some way on an electrical circuit and producing either a voltage, current, or frequency output. A transducer can be modeled as a two-port device described by a 2×2 impedance matrix Z_{ij}. It has two inputs, force and velocity, and two outputs, current and voltage. The parameters Z_{12} (forward transconductance) and Z_{21} (reverse transconductance) are the transducer sensitivity and the back-action effect on the bar. In fact, any transducer can operate in reverse mode and a current noise at its output can apply a fluctuating force on the bar. Moreover, an amplifier is in series to the transducer to amplify the signal. The noise sources due to the transducer and amplifier are described using the voltage and current noise $V(\omega)$ and $I(\omega)$ at the entrance of the amplifier. The spectral density of the back-action noise is

$$S_{\text{ba}}(\omega) = \frac{|Z_{12}|^2}{2M} I(\omega) t_m. \qquad (23)$$

The spectral density of the series noise is

$$S_{\text{sn}}(\omega) = \frac{2M}{|Z_{12}|^2} \frac{V(\omega)}{t_m}. \qquad (24)$$

While the thermal noise and the back-action noise are proportional to t_{m}, the series noise is proportional to t_{m}^{-1}; there is an optimal measurement time.

The total noise budget must be compared to the standard quantum limit. Using the spread $\sigma_{\mathrm{p}} \sim M\omega_{\mathrm{r}}\sigma_x$ in the uncertainty principle, it is possible to deduce the minimum detectable strain:

$$h_{\min} \sim \frac{1}{L}\sqrt{\frac{\hbar}{M\omega_{\mathrm{r}}}}. \qquad (25)$$

For a typical bar with a mass of 1 ton and a length of 3 m, resonant at 1 kHz, $h_{\min} \sim 10^{-21}$, on the order of the relevant astrophysical signals. Having fixed the frequency of interest, there is not too much room to improve the limit since the sound velocity is not varying very much from material to material; thus the size of the bar is fixed as well.

3. Seismic Noise

For resonant detectors it is also necessary to reduce the seismic noise at the resonant frequency below the gravitational signal level. The vibration isolation systems are multistage suspensions that have normal mode frequencies below the bar frequency and internal modes frequencies above it. The isolation band extends from a few hundreds of hertz to a few kilohertz. A typical attenuation system provides a 10^{10} reduction at 1 kHz. The actual bars operate at cryogenic temperatures, using a room-temperature isolator to which the cryogenic system is suspended. The cryogenic section uses cable suspensions to hang the bar as a pendulum or cantilevers. Although parametric transducers are noncontacting, passive transducers require additional cabling, which must be damped to avoid direct injection of noise. Cryogenic antennas operating at tens of millikelvins temperature have additional problems since they need cooling by conduction, which can inject noise from the cryostat vibrations.

4. Disturbances from Cosmic Rays

High-energy cosmic rays incident on the bar produce heating and expansion of the bar. The main sources are high-energy hadrons and muons. There are a few significant events per day for bars operating at a few millikelvins and hundreds for bars at a few microkelvins. The resonant detector should be underground to minimize cosmic ray noise.

C. Resonant Detectors in Operation

A world array of five resonant detectors is in actual operation as the International Gravitational Event Collaboration, The detectors are Explorer at CERN, Auriga in Legnaro (Italy), Nautilus in Frascati (Italy), Allegro in Baton Rouge (Louisiana, USA), and Niobe in Perth (Australia). The characteristics of the currently operating detectors are shown in Table I. The bandwidth is on the order of 1 Hz for all detectors. The bars are aligned within a few degrees of each other to maximize the chance of coincident detection.

TABLE I Main Parameters of the Resonant Detectors Currently in Operation

	Allegro	Auriga	Explorer	Nautilus	Niobe
T (K)	4.2	0.25	2.6	0.1	5
Material	Al5056	Al5056	Al5056	Al5056	Nb
Length (m)	3.0	2.9	3.0	3.0	2.8
Mass (kg)	2296	2230	2270	2260	1500
Q (10^6)	2	3	1.5	0.6	20
Frequency (Hz)	900	900	900	900	700
h ($\times 10^{-19}$)	6	4	6	4	6

The resonant detector network has recently set upper limits on burst events using coincidence analysis from three and four event detectors (complete with numbers). The limit on monochromatic pulsar signals is $\sim 10^{-23}$.

V. GRAVITATIONAL WAVE ASTROPHYSICS

Gravitational signal detection involves the use of suitable data processing techniques for burst, continuous, and stochastic signals.

The observations with resonant detectors are narrowband. The optimum filter watches for changes in the amplitude or in the phase of the bar oscillation during the measurement interval t_{m}; thus the filter is not sensitive to slow drifts. The bandwidth of this filter is $\sim t_{\mathrm{m}}^{-1}$, much larger than the intrinsic resonance width $\frac{f_{\mathrm{r}}}{Q}$. Such a filter is sensitive to the magnitude and phase of the Fourier transform of the signal, but not to the details of the signal shape.

The output of interferometer is a time series where the genuine signal is hidden in the noise. Burst signals from supernova events can be identified by choosing a suitable threshold: in fact, assuming that the noise is Gaussian, the probability of a large deviation due to noise is very small (for example, 0.046 and 0.0027 for 2σ and 3σ). Very careful environmental monitoring is necessary; there is no way to measure noise distribution alone since gravitational waves cannot be shielded. The solution is to use several detectors at different locations and use a coincidence method; while the genuine gravitational signal can produce coincident signals, the noises should be uncorrelated. The rate of chance coincidence is the product of threshold crossing probability in each detector. The time delay between events for detectors at distance D^* (measured through the earth) is $\Delta t = \pm\frac{D^*}{c}$. The delay between European detectors and LIGO is ∼20 msec, while for the two LIGO detectors it is ∼6 msec. The coincidence window should be large enough to allow detection of real events (but at the cost of a larger rate of chance coincidences). Coincidences can also occur between detectors in the same location, like the two Hanford interferometers: The second detector can provide a veto signal, since a genuine gravitational event must be detected in both of them with a signal ratio scaling as their sensitivities.

For the burst signal of binary coalescences, it is possible to use a *matched filter* since the signal waveform is known. The output of the detector is multiplied by the template of the waveform and integrated: the presence of the signal will give a higher output than the presence of noise alone. The matched filter is very demanding from the point of view of computation; in fact, the expected signals depend on several parameters and the burst arrival time is not known. The evolution of the signal frequency during coalescence provides information about the masses of the components, while the intensity is related to the system distance from earth. The final part of coalescence is particularly interesting since it has not been computed yet.

For continuous signals, there is no need for templates since the waveform is a sinuosoid. The Fourier transform of monochromatic signals exhibits peaks at the signal frequencies, which are a clear signature of the signal since the probability of having a spectral peak with noise alone is very low. The coincidence of detection in different detectors is of great help. There is another strong signature: the relative motion of the earth and the source causes both amplitude and frequency modulations of the signal, which depend on the source position. However, the number of operations needed to reconstruct the source direction is computationally demanding.

Stochastic background can be detected by cross-correlating the signal of different detectors. The relative distance of the planned detectors is of the same order or smaller than the gravitational wave wavelength.

A. Gravitational Detectors Network

In the near future, a global network of interferometric and resonant detectors will be performing observations in coincidence. The coincident observations improve the significance of the observations and allow correlation with the measured intensity at each detector with the delay due to the finite propagation speed. The network operation will allow a precise determination of the source direction and of the wave polarization. It is estimated that with a time resolution of 0.1 msec and a network of detectors at a few thousands of kilometers from each other, the source position can be reconstructed at about (5 mrad)^2. The determination of position is important to check the consistency of data measured by different detectors.

B. Coincidence with Nongravitational Detectors

Gravitational wave astronomy will benefit from coincidence with observations in other astronomical bands. If the source position has been determined with good precision, there could be an optical or radio or γ- or X-ray counterpart. As an example, coincidence with the intense optical emission of a supernova should be observed in coincidence with a gravitational wave burst. The coincidence with neutrinos from a supernova collapse offers another opportunity. The observation of optical and neutrino emission together with gravitational emission allows researchers to check the equality of gravitational wave velocity propagation with the speed of light. On the other hand, assuming that gravitational waves propagate with velocity c, the measurement can be used to put limits on the neutrino mass.

The coalescence of a binary neutron star system in coincidence with a γ-ray burst should allow researchers to test the origin of this process. A coalescing binary system can also be used as a standard candle, up to hundreds of Mpc. In fact, the measurement of the chirp time allows for deduction of the chirp mass and prediction of the signal intensity, thus determining the distance of the binary system. If an optical counterpart of the source is found, its redshift can be determined by classical methods. In this way, the Hubble constant can be determined.

VI. FUTURE DEVELOPMENTS

Hopefully the first generation of gravitational detectors will allow the first direct detection of gravitational waves. To build the field of gravitational astronomy, further improvements in sensitivity are needed. A vigorous research and development program is being carried out for both interferometers and bars in attempts to approach the fundamental SQL.

A. Interferometric Detectors

The LIGO collaboration is planning to upgrade the initial detector in 2006 (LIGO II), improving the sensitivity by one order of magnitude at least. The research is being performed by the LIGO Scientific Collaboration, which includes groups from the United States, Europe, and Australia. On the European side, Virgo and GEO600

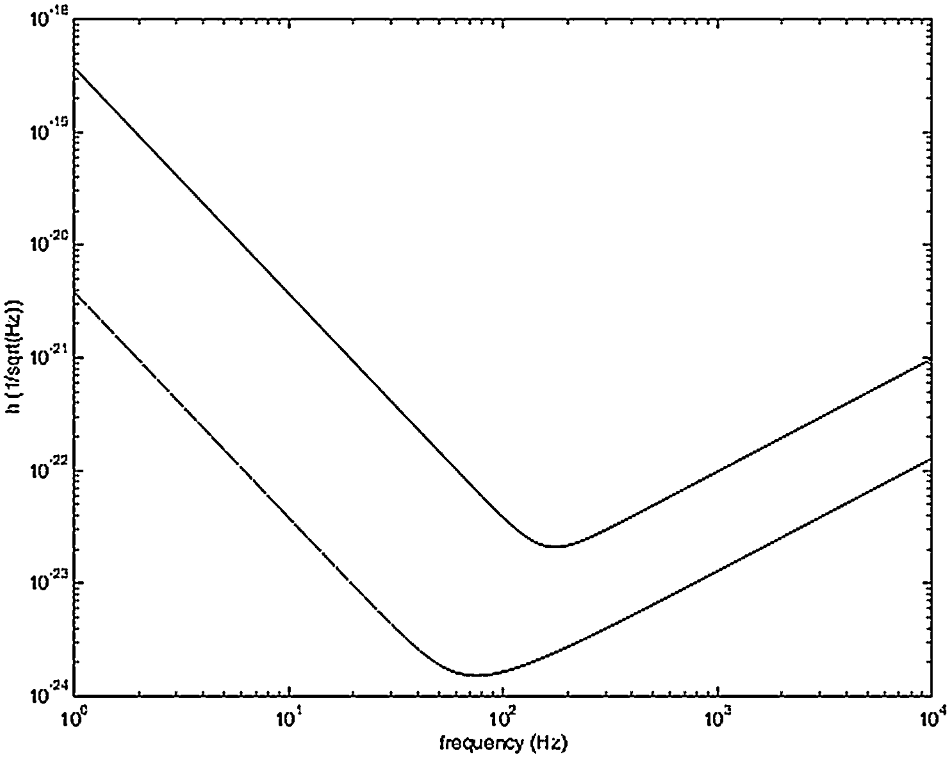


FIGURE 9 Sensitivity of the initial (dashed line) and advanced LIGO (solid line) detectors.

collaborations have considered the realization of the detector EURO around 2010. In Japan there is a project for a 3-km scale interferometer.

The research in sensitivity improvements covers all sources of noise. All collaborations are planning to have effective seismic isolation down to a few hertz. The thermal noise can be reduced, in principle, by working at cryogenic temperatures, which is foreseen in future Japanese detectors. However, the loss factor of various materials does not always improve at low temperature, and the test mass becomes more sensitive to heating from the laser beam. For this reason, other techniques are now being studied. The first improvement comes from using monolithic suspensions (fused silica fibers or ribbons) to hang test masses. Losses on the order of 10^{-8} have been achieved for several suspension modes. Another planned improvement is the replacement of fused silica test masses with sapphire test masses, where losses are one order of magnitude smaller. The improvements of sensitivity in the region where shot noise dominates are performed with several techniques. First of all, plans call for increaseing the input laser power from 10 to 100 W. Moreover, signal recycling will be used to improve sensitivity in a narrowband.

The possible improvements in sensitivity of the LIGO detector are shown in Fig. 9.

B. Resonant Detectors

The cross section of a bar is proportional to its mass, thus there are efforts to build a 100-ton spherical resonant detector, with high sensitivity and omnidirectional response. The sphere must be equipped with several transducers to detect the fundamental modes: the quadrupole modes and the monopole mode (not sensitive to gravitational waves). Efforts are also being devoted to developing more efficient electronics systems (transducers and amplifiers) to achieve better sensitivity. Another major improvement foreseen is the increase of the bandwidth from the current few hertz to several tens of hertz. In this way, coincident runs could allow the determination of source position.

VII. CONCLUSIONS

The forthcoming gravitational wave detectors have an initial sensitivity close to the expected level of the signals from astrophysical sources and should perform the first direct detection of waves. The research in gravitational physics involves several aspects of physics and technology, ranging from theoretical and numerical calculations to applied research in lasers and optics. Further developments in detector sensitivity should allow the birth of gravitational wave astronomy, to give a new view in the physics of neutron stars, black holes, and the primordial universe. The connection of the gravitational measurements with other astronomical observations should contribute to a more complete view of the universe.

SEE ALSO THE FOLLOWING ARTICLES

Binary Stars • Cosmology • Gravitational Wave Astronomy • Gravitational Wave Physics • Neutron Stars • Pulsars • Radio Astronomy, Interferometry • Relativity, General • Supernovae

BIBLIOGRAPHY

Blair, D. G., ed. (1991). “The Detection of Gravitational Waves,” Cambridge University Press, Cambridge.

Ju, L., Blair, D. G., and Zhao, C. (2000). “Detection of gravitational waves.” *Rep. Prog. Phys.* **63,** 1317.

Saulson, P. R. (1994). “Fundamentals of Interferometric Gravitational Wave Detectors,” World Scientific, Singapore.

Thorne, K. S. (1987). *In* “300 Years of Gravitation” (S. W. Hawking and W. Israel, eds.), Cambridge University Press, Cambridge.

Image Restoration, Maximum Entropy Method

R. J. Sault
Australia Telescope National Facility

GLOSSARY

Bayesian probability theory The approach to probability theory which views a probability as a measure of our uncertainty of knowledge rather than as relative frequency of occurrence.

Convolution A position-independent blurring operation on an image.

Deconvolution The process of inverting a convolution operation.

Lagrange multiplier A variable introduced into an optimization problem. It ensures the solution of the optimization obeys a particular equality constraint.

Modulation transfer function The Fourier transform of the point-spread function for a position-independent system.

Mosaicing The practice in radio astronomy of making an image of a large region of the sky from many observations of smaller regions.

Pixel "Picture element." The smallest element of a picture or image, when represented as an array of elements, as in a computer.

Point-spread function The response of an imaging system to a hypothetical point source. Point-spread functions can be position-dependent or position-independent.

Sidelobe The nonzero response of the point-spread function well away from the main response peak.

IMAGE RESTORATION is the process of estimating a "true" image, given a degraded (generally noisy and blurred) version of that image. Maximum entropy image restoration is a class of approaches to this problem. This takes as the solution the image that has a maximum entropy measure while still remaining consistent with the measured data. Unlike classical linear techniques, maximum entropy methods provide a framework where it is simple to include a wide variety of forms of data in the restoration process. Being a nonlinear technique, maximum entropy methods allow extrapolation and interpolation of unmeasured Fourier information. These properties can result in substantially better results than those possible with classical linear techniques. One of the main disadvantages of maximum entropy techniques is that the results tend to have significant biases.

I. IMAGE RESTORATION

A. Introduction to Imaging

Image restoration can be simply defined as the process of forming a high-quality estimate of an image from available data. The term *restoration* is generally used where the data are in the form of a degraded (blurred) image, whereas the term reconstruction tends to be used where the data are some more general measure of the image. Restoration and reconstruction are quite similar (indeed restoration is a subset of reconstruction), and although image restoration will solely be considered here, many of the general principles apply equally well to the reconstruction problem.

The need for image restoration arises because image acquisition systems are never perfect, regardless of whether the image is formed from a simple camera or from a multibillion dollar space telescope. Many imaging systems are a good deal more complex than simple lens systems, and so are the degradations that are present in their measured image. Whereas some imaging techniques involve nonlinear processes (e.g., the response of photographic film is related to the light intensity through a very nonlinear relationship), many are linear to good approximation. That is, the output of the imaging system is proportional to some intensity of the input. The response of such linear imaging systems is conveniently described in terms of the point-spread function. This is the response to a point source (this is the two-dimensional analog of the impulse response function in traditional time-series analysis). The point-spread function can generally be thought of as a blurring function. At the very least, any real imaging system has a finite resolution, and the point-spread function is that function which describes how an infinitely narrow source is blurred by that system. In general, the point-spread function will be affected by many other characteristics of the imaging system. In addition to this blurring, the output image will be corrupted by noise. Representing the image as a function of two variables x and y, the true image $i(x, y)$, is related to the measured image, $i_M(x, y)$, by an integral

$$i_M(x, y) = \iint i(x', y')b(x, y; x', y')\,dx'\,dy' + n(x, y).$$

Here $b(x, y; x', y')$ is the point-spread function of the imaging system and $n(x, y)$ is additive noise. This gives the general case where the point-spread function varies with position in the image. That is, the response of the system to a point source at (x', y') is $b(x, y; x', y')$. This integral recognizes that the true image can be thought of as a sum of infinitesimal point sources, and so the measured image will be the sum of the responses to these infinitesimal point sources.

Many imaging systems are *position-independent*. This means that the point-spread function is independent of the position of the point source in the image. The point-spread function at (x', y') is then of the form $b(x - x', y - y')$, and the integral relating true image to the measured image is a two-dimensional convolution relationship:

$$i_M(x, y) = \iint i(x', y')b(x-x', y-y')\,dx'\,dy' + n(x, y).$$

This convolution is often written as

$$i_M(x, y) = i(x, y) * b(x, y) + n(x, y).$$

Whereas image processing and image manipulation were once the domain of optical engineers, digital representation of images is now pervasive. Digitally, images are represented as a discrete array of picture elements or pixels, with each pixel taking on a value (or perhaps three values for color images). In this digital representation, the integrals above are replaced with sums, and the point-spread function is represented as an array of numbers. We will use $i(m, n)$ and $b(m, n)$ to represent these discrete-valued images.

Real systems often are approximately position-independent, at least over small regions or near the center of the region being imaged. One of the attractions of modeling an imaging system as being position-independent is that they are amenable to analysis using Fourier theory. The Fourier transform of the image $i(x, y)$ is defined

$$I(u, v) = \iint i(x, y)\exp(i2\pi(xu + yv)\,dx\,dy.$$

Similarly the inverse Fourier transform is

$$i(x, y) = \iint I(u, v)\exp(-i2\pi(xu + yv)\,du\,dv.$$

Note that there is no generally agreed convention on the sign in exponentials in the forward and inverse Fourier transform, other than the forward and inverse transformations have opposite signs. By analogy with time-domain signals, the (u, v) coordinates in the Fourier domain are called spatial frequencies. Fourier transforms of real functions will generally be complex-valued. A discrete version of the Fourier transform (for sampled data) also exists.

One of the more important properties of Fourier transforms (both continuous and discrete) is the convolution theorem: a convolution in the image domain transforms to a multiplication operation in the Fourier domain. If $I_M(u, v)$, $N(u, v)$, and $B(u, v)$ are the Fourier transforms of $i_M(x, y)$, $n(x, y)$, and $b(x, y)$, then

$$I_M(u, v) = I(u, v) \cdot B(u, v) + N(u, v).$$

$B(u, v)$, the Fourier transform of the point-spread function, is known as the modulation transfer function. As well as being amenable to Fourier analysis, position-independent systems are attractive from a computational point of view: for an image of size $N \times M$, the computational cost of simulating a position-dependent system is of order $(NM)^2$, whereas that for a position-independent system is of order $NM \log(NM)$. For images of realistic size, this difference in computational complexity often separates the computational feasibility and total infeasibility. The difference is purely because of the existence of the "Fast Fourier Transform" algorithm (devised by J. W. Cooley and J. W. Tukey in 1965). As the name implies, this algorithm allows a fast computation of the discrete Fourier transform, which in turn allows fast convolution algorithms to be developed. Because of this computational advantage, instances containing position-dependent point-spread functions are usually decomposed and approximated as multiple position-independent systems. For these reasons, we will largely restrict our attention to position-independent systems.

For an aberration-free lens system, the point-spread function will be the diffraction limit, which gives a response of the form

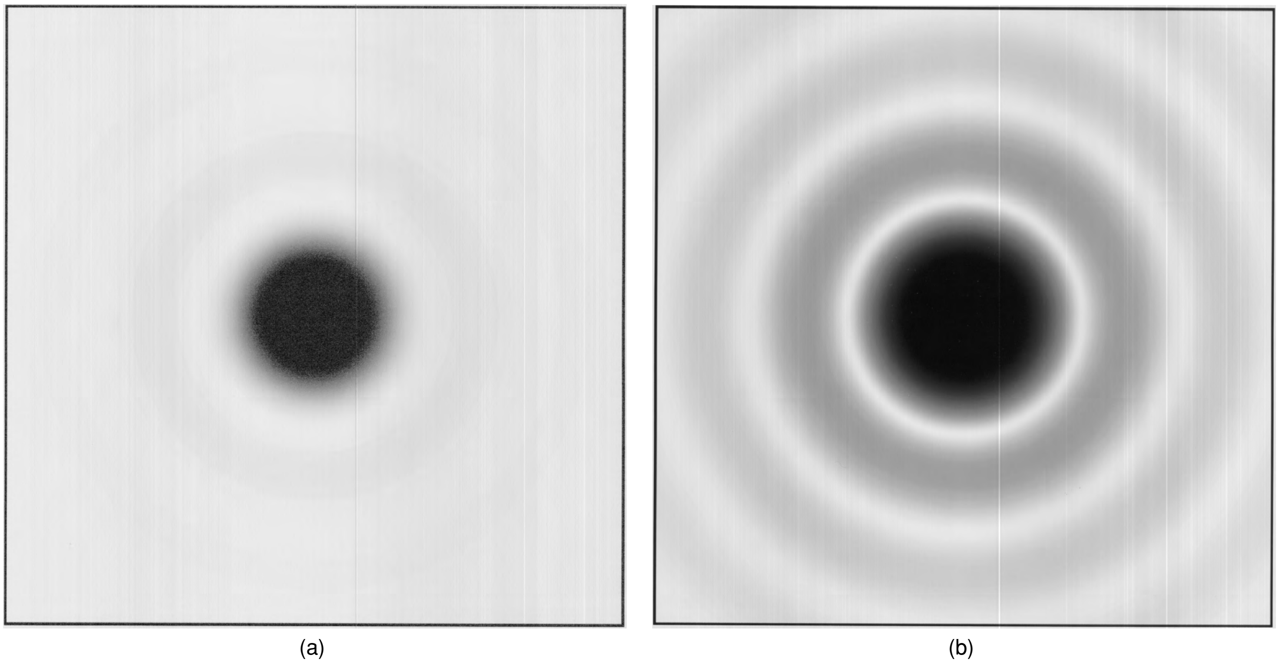
$$b(x, y) = \left(\frac{2\mathrm{J}_1(ar)}{ar}\right)^2.$$

Here $\mathrm{J}_1(r)$ is the first-order Bessel function, $r = \sqrt{x^2 + y^2}$, and a is a constant related to the lens size and wavelength of the light. This point-spread function is known as the Airy disk, after Sir George Airy who was the first to derive its form. The Airy disk is a comparatively clean point-spread function: most of the power in the response is confined to the main peak.

Many point-spread functions, however, are far less well-behaved. For example, Fig. 1 shows an Airy disk and the point-spread functions of particular observations of the Hubble Space Telescope and the Very Large Array (a radio telescope) are shown in Figs. 3 and 4.

B. Image Restoration

The technique of image restoration (or "deconvolution") can now be simply put: given the measured, noisy, image $i_M(x, y)$ and the point-spread function $b(x, y)$, estimate the true image $i(x, y)$.


(a)

(b)

FIGURE 1 The point-spread function of an aberration-free lens system. The left grayscale uses a linear transfer function, whereas the right uses a logarithmic transfer function to bring out low-level details.

The basic problem of image restoration is that it is fundamentally ill-posed: the blurring resulting from the point-spread function means that some information is lost. As a consequence, the restoration step is nonunique—many images, if not an infinite number, are typically consistent with the measured data. Using Fourier analysis, it is clear that image information is fundamentally lost at those spatial frequencies where the modulation transfer function falls to zero. Although it is clear that no spatial frequencies are measured beyond the diffraction limit, many imaging instruments fail to measure spatial frequencies within this limit. That is, the Fourier information is incomplete.

The ill-posed nature is even more apparent when noise is considered. If the modulation transfer function is such that the signal-to-noise ratio drops below unity, then the signal has, again, effectively been lost. Any image restoration technique must be robust to noise or errors in the data and the assumed point-spread function.

The challenge of image restoration is to make a good estimate of the true image despite noise and nonunique solutions.

C. "Classical" Image Restoration Algorithms

The "obvious" image restoration algorithm is the so-called inverse filter. Given that

$$I_M(u, v) = I(u, v) \cdot B(u, v) + N(u, v),$$

then an estimate of the true image is

$$\hat{I}(u, v) = I_M(u, v)/B(u, v).$$

There is a major flaw with the inverse filter which renders it useless: when $B(u, v)$ falls to near zero, the correction becomes large, and any noise present is substantially amplified. Even computer rounding error can be substantial. An alternative approach, which avoids this problem, is based on the approach of Wiener. This approach models the image and noise as stochastic processes, and asks the question: What re-weighting in the Fourier domain will produce the minimum mean squared error between the true image and our estimate of it? The Wiener solution has the form

$$\hat{I}(u, v) = I_M(u, v) \cdot \frac{B^*(u, v)}{|B(u, v)|^2 + R_N(u, v)/R_I(u, v)}.$$

Here the spectral density functions of the true image and noise are $R_I(u, v)$ and $R_N(u, v)$, respectively. Note that at spatial frequencies where the signal-to-noise is very high, the ratio $R_N(u, v)/R_I(u, v)$ approaches zero, and the Wiener filter reduces to the inverse filter. However, when the signal-to-noise ratio is very poor (i.e., $R_N(u, v)/R_I(u, v)$ is large), the estimated spatial frequencies approach zero. This is a moderately interesting property—whereas the inverse filter gives an unbiased estimate of the true image, the Wiener filter output is biased toward 0 when the signal-to-noise ratio is poor.

The Wiener filter is seldom used in real-image restoration applications for a simple but critical reason: it is a linear technique. With a linear technique, the spatial frequencies present in the true image estimate are those in the measured image multiplied by some factor. This is adequate if all spatial frequencies are reasonably sampled. However, in those cases where the spatial frequencies are unmeasured, the Wiener filter will simply estimate them as 0. Linear techniques, in general, cannot extrapolate or interpolate to estimate unmeasured spatial frequencies. Techniques that do not estimate these missing spatial frequencies will produce images with plainly visible defects. To produce a good restoration technique, some plausible values need to be estimated for these missing spatial frequencies. Consequently nonlinear restoration techniques are required.

II. BAYESIAN INFERENCE

A. Bayes Theorem

To provide a grounding in some of the theory around maximum entropy methods, we digress here to describe some of the important aspects of Bayesian probability theory. Although this is arguably the original form of probability theory, Bayesian theory fell into disrepute, however, interest in it has been re-awakened and re-invigorated in the last half century. Although the differences between conventional and Bayesian theory can appear subtle at times, they are based on quite different approaches. Conventional theory defines probability functions as normalized frequency distributions ultimately derived from a nominally infinite number of trials (or perhaps a physical principle that allows us to predict the outcome of an infinite number of trials). Bayesian probability, on the other hand, views probabilities as degrees of belief that are modified by information. A Bayesian probability is refined as more information becomes available. In the limit of an infinite number of trials, Bayesian and conventional probability converge. In the presence of limited information, Bayesian probabilities are often easily assigned when conventional probabilities cannot.

In Bayesian probability, Bayes theorem is an important tool used to incorporate information to refine the probability assigned to a hypothesis. Bayes theorem can be stated as a relationship between conditional probabilities:

$$\mathrm{P}(H \mid D) = \mathrm{P}(H) \cdot \frac{\mathrm{P}(D \mid H)}{\mathrm{P}(D)}.$$

Here H is a hypothesis and D is some data, and so $\mathrm{P}(H \mid D)$ is the probability of the hypothesis given the data. Bayes theorem relates this to the probability of the hypothesis before the data became available, $\mathrm{P}(H)$, the probability of the data given the hypothesis, $\mathrm{P}(D \mid H)$, and finally the probability of the data $\mathrm{P}(D)$. The term $\mathrm{P}(H)$ is usually called the prior probability, or simply the prior, as it is obviously the probability that is assigned before the data is available. As the probability $\mathrm{P}(D)$ is independent of the hypothesis, knowledge of this is not needed when comparing the relative probability of different hypotheses. In this sense, it is a normalization term that is often not important to an argument.

B. The Principle of Maximum Entropy

A major problem in Bayesian probability had been assigning the prior probability. Pioneering work by E. T. Jaynes has shown that there is a unique consistent way to assign prior probabilities, with this approach becoming known as the principle of maximum entropy. This principle states that the prior probabilities, $p_1, p_2, \dots, p_N$, that we associate with the mutually exclusive hypotheses $H_1, H_2, \dots, H_N$, should be those which maximize

$$S = -\sum_{i}^{N} p_i \log(p_i),$$

subject to this being consistent with whatever information is already available.

The concept of entropy is present in many disciplines. In statistical mechanics, Boltzmann introduced entropy as a measure of the number of microscopic ways that a given macroscopic state can be realized. A principle of nature is that it prefers systems that have maximum entropy. Shannon has also introduced entropy into communications theory, where entropy serves as a measure of information. The role of entropy in these fields is not disputed in the scientific community. The validity, however, of the Bayesian approach to probability theory and the principle of maximum entropy in this, remains controversial.

To illustrate Bayesian probability and the principle of maximum entropy, consider the roll of a die (or perhaps a computer program that generates a number between 1 and 6). What is the probability of the roll showing a 6? The probabilities of the hypotheses for rolling a 1, 2, through to 6 will sum to 1 (no other events are possible). Lacking any information, the Bayesian prior probability is that which maximizes the entropy. This leads to the intuitive result that all faces of the die have equal probability of being thrown (all have the probability of $\frac{1}{6}$). This is the crux of the difference between Bayesian and conventional probability—a Bayesian is comfortable in saying that the faces of the die have equal probability in coming up. This is despite lack of information supporting this conjecture, but, importantly, despite lack of information to refute it. Strict conventional probability would not be able to answer such a seemingly simple question as its probability function is derived from either a notional infinite number of throws of the die to derive a frequency distribution or some physical argument to assure us that the die was fair.

Note that whereas the equal probability agrees with everyday intuition, it does not necessarily represent the "truth." If it were a computer program, rather than a die, then it is far less obvious that the numbers would have equal probability. Furthermore, if the die was in a gambling den of dubious repute, then you might be quite right to doubt the fairness of the die—further tests of the die might be called for before you are willing to bet. In such a den, if a person with Bayesian beliefs was presented with the extra data that the mean value from the throw of the die was 5, then this extra information could be included in determining the prior probabilities. Maximizing the entropy subject to this information, the chances of throwing a 1 through 6 are given by probabilities 0.02, 0.04, 0.07, 0.13, 0.25, and 0.48, respectively. Again, Bayesian probability is able to give such an answer, whereas a conventional probability would still not be willing to conjecture. In this case, however, intuition is no guide as to whether this is a reasonable choice, and the reader might be more comfortable with the conventional choice of not making a choice.

It is fair to say that there is no universal agreement between those who champion either Bayesian or conventional probability theory. Indeed, as a bystander, one can find it interesting to watch the tussle. Conventional probability theory is very successful at modeling situations with large numbers of random events (e.g., the pressure in a vessel resulting from gas molecules colliding with the vessel walls). Indeed, statistical mechanics is based on conventional probability theory. On the other hand, Bayesian approaches seem to be more successful when the number of random events is smaller, and where there is insufficient information available.

C. Maximum Entropy Applied to Images

Bayesian probability theory shares some similarities with the problem of image restoration; they both are required to make some choice in the presence of insufficient data or information. It is not surprising, then, that Bayesian techniques have been applied in image restoration. Applying Bayesian principles, of the possible solutions to an image restoration problem (i.e., of all the images that are consistent with the data), we choose that image which maximizes the entropy.

Historically, however, there have been two different schools of thought as to how the entropy of an image should be measured, and there has been no resolution as to whether one, both, or neither of these schools are correct in the arguments that they champion.

The original school (resulting from the work of J. P. Burg) applied the Bayesian logic and the principle of maximum entropy rigorously to the problem of estimating a power spectrum of a signal, given discrete and limited measurements of the autocorrelation function of that signal. This work showed that the power spectrum, $R(\upsilon)$, which maximized the measure

$$S = \int \log(R(\upsilon))\,d\upsilon$$

was the preferred spectrum. The argument behind this considers the signal as a Gaussian random process, and so the Fourier components of the signal will also be Gaussian variables. The entropy associated with these components is proportional to the logarithm of the variance, and the variance is simply $R(\upsilon)$. Hence we should be maximizing the above entropy measure. Extending this argument to imaging, the electric field pattern of an electromagnetic wave can also be treated as a Gaussian random process, and a similar argument leads to the entropy measure of an image as

$$S = \sum_{n}^{N}\sum_{m}^{M} \log(i(m,n)).$$

An alternative school, introduced by B. R. Frieden, and which will be called the Gull and Daniell form (after two early champions of this measure), takes a different approach. It models an image as being composed of a large number of *luminance elements*. If there are K luminance elements that are to be laid down on an image with $N \times M$ pixels, what are the most likely images to be formed? If $p(m,n)$ is the probability of a luminance element landing on pixel (m,n), then the image expected from K luminance elements will be $i(m,n) = Kp(m,n)$. The entropy measure of the image is then

$$\begin{aligned} S &= -\sum_{n}^{N}\sum_{m}^{M} p(m,n)\log(p(m,n)) \\ &= -\frac{1}{K}\sum_{n}^{N}\sum_{m}^{M} i(m,n)\log(i(m,n)/K) \end{aligned}$$

Using this argument, the preferred image is the one which is consistent with any data, but which maximized this entropy. An amusing analogy sometimes used to illustrate this is to imagine an infinite troop of monkeys at work, with each monkey having K building blocks or luminance elements to make an image. The image most often made, which is consistent with the data, will be that which maximizes entropy.

As will be discussed below, the scaling by K in this measure of entropy is important. A more common and intuitive way to represent this scaling is

$$S = -\sum_{n}^{N}\sum_{m}^{M} i(m,n)\log\left(\frac{1}{e}\frac{i(m,n)}{d(m,n)}\right),$$

where $d(m,n)$ is called the *default image* and e is the base of natural logarithms. Differentiating S with respect to the pixel values will show that, in the absence of other information, the image that maximizes the entropy is simply $d(m,n)$. While a uniform (flat) default image would be used where there is no prior information on the true image, a nonuniform default can also be used. In this case, solution image will be biased toward this default.

The argument with luminance elements suggests analyzing the situation in terms of photons. The Bose formula for the entropy of n photons in a mode is

$$(1+n)\log(1+n) - n\log n.$$

In the limit of $n \gg 1$, which is the case at radio wavelengths, this reduces to $\log n$, whereas when $n \ll 1$, as at optical wavelenghts, this reduces to $-n\log n$. Unfortunately, the statistics of photons does not resolve the dispute in entropy measures.

Although the entropy measure used can appreciably affect the solutions when there is limited information, in practice there is usually a good quantity of data to constrain possible solutions. The data ultimately should (and generally do) drive the solutions.

III. MAXIMUM ENTROPY IMAGE RESTORATION IN PRACTICE

A. Data Constraints

Maximum entropy image restoration problems are usually formulated as optimization problems, with the pixel values being the variables of the optimization. The data are generally included in these optimizations by way of equality constraints. Equality constraints are attractive, as they can be easily included in an optimization problem by augmenting the problem with extra variables called Lagrange multipliers. For example, if we wish to maximize S subject to a data constraint, $D = 0$, then by introducing the Lagrange multiplier α, the augmented problem of maximizing

$$J = S - \alpha D$$

will produce the desired result. That this is so requires a little thought. However, note that at the maximum, D will be 0 and so does not affect the value of S achieved, and the derivative of J with respect to α is simply the data constraint, and so will be 0. The Lagrange multiplier, α, does have a well-defined value at the maximum, and can be thought of as a mediator between the competing requirements to maximize S and to enforce the data constraint D.

One possible set of data constraints in image restoration is

$$b(m,n) * i(m,n) = i_M(m,n).$$

This uses a data constraint for each pixel in the measured image, and requires the introduction of one Lagrange multiplier per pixel. The main problem with this is that it fails to recognize that the data are noisy. Consequently, exact agreement with the data should not be enforced. An alternative, the so-called χ^2 constraint, requires only statistical agreement. In particular,

$$\chi^2 = \sum_n^N \sum_m^M (b(m,n) * i(m,n) - i_M(m,n))^2 / \sigma^2$$

is constrained to equal its statistically expected value. Here σ^2 is the expected noise variance of each pixel. For $M \times N$ pixels, the expected value of χ^2 will be MN. Thus, to find the maximum entropy solution, for entropy measure S, we maximize

$$J = S - \alpha(\chi^2 - MN).$$

Many other data constraints are possible. For example, the *integrated flux* constraint is common, where the sum of all the pixels is constrained to equal some fixed value,

$$F = \sum_n^N \sum_m^M i(m,n).$$

Again, this constraint can be included in the maximization process using a Lagrange multiplier.

In addition to the data, there is often other physical information to constrain possible solutions. These include:

1. Image data are often positive-valued by definition. Indeed, the maximum entropy measures discussed above only make sense for positive-valued images, and so positivity is a natural consequence. Whereas this enforcement of positivity is often seen as a strength, for images that can be negative-valued, it is certainly a hindrance.
2. Often an image is known to be nonzero in only a small subregion.

Including as many constraints in the entropy maximization process is important—the more information present, the better conditioned the problem will become, and the resultant image will be more reliable.

B. Entropy Measures

Even though there is no general agreement on the "correct" measure of entropy for an image, this does not prevent the use of these measures. Indeed, many maximum entropy practitioners take a more pragmatic approach, and are unmoved by the philosophical debates between the different camps as to whether the Gull and Daniell or the Burg form is the appropriate entropy measure. In many instances, practitioners use maximum entropy because it works—information-theoretic arguments are not required. It is possible to take the argument a step further and choose any *entropy* function that is effective.

Maximum entropy, as applied to image restoration, can be thought of as a particular case of a more general technique known as *regularization*. One approach to ill-posed problems, such as image restoration, is to find solutions that are consistent with the data, but which possess other desirable features. Maximum entropy is but one of many possible desirable features. Another possibility is a measure such as

$$S = \sum_n^N \sum_m^M \Big(i(m,n) - \frac{1}{4}(i(m+1,n) + i(m-1,n) + i(m,n+1) + i(m,n-1)) \Big)^2.$$

This measures the roughness of an image (it is the sum of the square of second derivatives). By choosing a solution image which minimizes this roughness measure, but which remains consistent with the data, a "good" reconstruction can be expected. This regularizing function is typical of many that convolves the image estimate with some filter kernel, $k(m,n)$, and then sums the squares of the result. In the example above, the kernel was a simple derivative filter. When using the sum of squares of a filter as the regularizing function, and when the data constraint is the χ^2 approach, the solution has a form similar to the Wiener filter discussed previously. In particular,

$$\hat{I}(u,v) = I_M(u,v) \cdot \frac{B^*(u,v)}{|B(u,v)|^2 + K(u,v)/\alpha},$$

where $K(u,v)$ is the Fourier transform of the filter kernel and α is the Lagrange multiplier in the χ^2 constraint. As with the Wiener filter, a regularization function like this will result in a linear algorithm, and this will not extrapolate/interpolate any missing Fourier data. To estimate missing Fourier data, the regularization/entropy function must have first derivatives that are neither constant nor linear functions of the data.

Apart from the two information-theoretic entropy measures, other forms that have been suggested include

$$S = \sum_{n}^{N} \sum_{m}^{M} \sqrt{i(m,n)}$$

and

$$S = -\sum_{n}^{N} \sum_{m}^{M} \log(\cosh(\lambda i(m,n))$$

$$\cong -\lambda \sum_{n}^{N} \sum_{m}^{M} |i(m,n)| + \text{constant} \quad \text{where } \lambda i(m,n) \gg 1.$$

The latter measure is called the *maximum emptiness* criterion. When λ is chosen to be comparable to the reciprocal of the noise level, this will be approximately equivalent to optimizing the L_1 norm of the image.

C. Combining Data

One of the advantages of a nonlinear approach, such as maximum entropy, is that it can extrapolate and interpolate missing information in the Fourier domain. Almost a corollary of this is that all available data must be used in the restoration process simultaneously. To see the difference here from a linear approach, assume we have two sets of data: D_1 and D_2. If we use a linear approach, we will achieve identical results if we restore D_1 and D_2 separately and then combine the result, as compared against combining the two sets of data and then restoring. That is, it does not matter with a linear approach when the combining is done; it can be done before or after the linear restoration. With a nonlinear approach, this is not the case as quite different results will generally follow. From a Bayesian's viewpoint, this should be no surprise. A better result should be possible when more information is available. Nonlinear processing encourages so-called "joint" approaches, where all data is processed jointly. In this way, full advantage can be made of the relationships between the different data. We will see examples of this below.

D. Maximum Entropy Algorithms

In general, a maximum entropy problem can be expressed as the maximizing of an expression such as

$$J = S - \alpha_1 D_1 - \alpha_2 D_2 + \cdots.$$

Here S is the entropy measure, D_1, D_2, etc., are data equality constraints and α_1, α_2, etc., are Lagrange multipliers. One of the important differences in formulating a restoration problem in this way is that it does not focus on transforming the data into the solution. Rather, this approach aims at a more general formalism, which readily accommodates adding more data and constraints. As no direct inversion between the data and true image is called for, it is easy to include data that is related to the true image in quite an indirect fashion. The approach, which could be called data-oriented rather than procedural, is far more general than just image restoration. The maximum entropy approach has been used in a broad variety of ill-posed inverse problems.

A maximum entropy image restoration can be solved in a fashion similar to many traditional optimization problems—the derivatives of J can be used to drive algorithms such as steepest descent, Newton-Raphson, or conjugate gradients. The algorithms, like solutions to most nonlinear problems, are invariably iterative. The computationally expensive part of such algorithms is usually the implementing of convolutions (or similar operations). In practice, these are generally done using Fourier transforms. Because the number of variables in the maximization is typically large (it equals the number of pixels plus the number of Lagrange multipliers), there are often some approximations and simplifications required when compared with more traditional optimization problems.

E. General Properties

Somewhat paradoxically, general characteristics of maximum entropy solutions are more easily analyzed than solutions to a number of other image restoration algorithms. This is despite maximum entropy methods being more difficult to implement. This ability to characterize the solutions results from them all obeying a general set of properties—they maximize the entropy. On the other hand, image restoration techniques that are described procedurally (particularly those described by an iterative procedure) are difficult to analyze as their solution is defined by nothing more than the end point of a procedure. Here we consider a number of general properties of maximum entropy solutions.

Positivity and compression of pixel range—With the exception of the maximum emptiness criteria, all the entropy measures described above are only defined for positive-valued pixels. Clearly then, these measures (or rather their optimization algorithms) enforce positivity.

Figure 2 plots the various entropy measures and their derivatives, with the entropy for a pixel S equalling $\log p$, $-p \log p$, $\sqrt{p}$ and $-\log(\cosh p)$. Although the entropy measures at first glance look quite dissimilar, their similarity becomes more apparent when their derivatives are compared. Ignoring the data constraints, for a small change, Δp, the change in the entropy will be

$$\Delta S = \frac{dS}{dp} \Delta p.$$

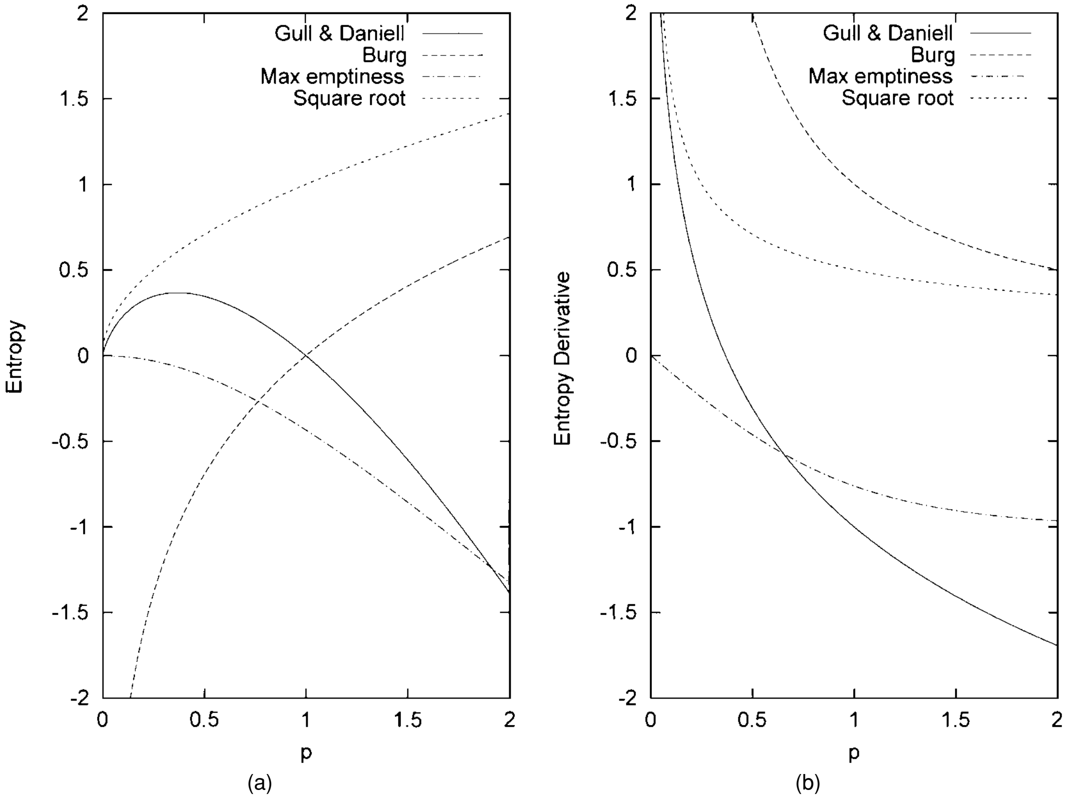


FIGURE 2 Plots of different maximum entropy functions and their derivatives. Shown are the Burg (log$p$), Gull and Daniell ($-p$ log p), maximum emptiness (−log(cosh p)), and square root ($\sqrt{p}$) measures.

The derivatives all have the property that, for a given change in Δp, there is a greater change in entropy for smaller rather than larger values of p (i.e., the second derivative of S is negative). For parts of an image that is mostly blank, this property tends to compress the pixels to small nonzero values, and to suppress ripples. This compressing effect becomes less effective for pixels with large values. It is this property that leads to maximum entropy's stabilizing influence on image restoration problems.

Flux-scale dependence—An interesting property of the Gull and Daniell measure is that, without the default image, it is flux-scale dependent. That is, the answer potentially depends on the units of the measurements. As a trivial example, consider an experiment where two lengths, x and y, are measured to be 1 and 2 inches, respectively, and there is a known uncertainty in each measurement of $\sigma^2 = \frac{1}{4}$. Formulating this as a maximum entropy problem with χ^2 constraint, we must maximize

$$J = -x \log x - y \log y + \alpha\left((x-1)^2 + (y-2)^2 - \frac{1}{2}\right).$$

This gives a result $(x, y) = (0.62, 1.40)$. On the other hand, using the same data and variance, apart from forming the problem in centimeters and centimeters squared, the result would be, after converting back to inches, $(x, y) = (0.69, 1.36)$.

None of the other forms of entropy described above suffers from this undesirable property*—it is a property of the Gull and Daniell measure only. The key to the difference here is that the other entropy measures described above have derivatives that are a power of $|p|$.

This flux-scale dependence of the Gull and Daniell measure underlines the effect the default image has on the final solution. Even when using a *flat* default, the final result will depend on the value of that flat default level. Indeed, this reveals a flaw in the luminance elements argument previously used to justify this entropy measure. It assumes that the number of luminance elements is known *a priori*, which is often not the case. Without knowing this *a priori*, the luminance element argument cannot be used.

Interestingly, if an integrated flux constraint is combined with the Gull and Daniell measure, the solution becomes scale-independent.

Bias—Maximum entropy solutions are invariably biased solutions. This means that the expected value of the

*Strictly, the maximum emptiness measure is also scale-dependent, but the dependence becomes negligible if $\lambda i(m, n) \gg 1$.

maximum entropy solutions, given an infinite number of different realizations of the possible noise, does not converge to the truth. Biased estimators in statistical techniques are by no means uncommon, there is always a trade-off between bias and stability, with the more stable estimators tending to be biased. Indeed, the Wiener filter, and those regularization techniques that share the Wiener form, are all biased estimators.

The Gull and Daniell measure will tend to bias the solution toward the default image. This generally means that peaks are underestimated, and regions of low value tend to be overestimated. For example, when using a χ^2 data constraint, some analysis shows that the solution image is approximately

$$\hat{i}(m, n) \cong d(m, n) \cdot \exp(\beta \cdot r(m, n)),$$

where β is a positive-valued constant, $d(m, n)$ is the default image as before, and $r(m, n)$ is the discrepancy between the measured and solution images:

$$r(m, n) = i_M(m, n) - b(m, n) * \hat{i}(m, n).$$

This shows that when the solution image deviates significantly from the default image, then the residuals must be large. This bias can become quite problematic when a good default image is not known. For example, a flat default will bias the solution to have flux across the image, even if the true image might be quite clumpy. This can lead to an artificial plateau in the image, where there is no real emission.

Super-resolution and variable resolution—One of the prime motivations in using a nonlinear restoration algorithm is its ability to interpolate *and* extrapolate unmeasured information in the Fourier plane. Thus, virtually by definition, nonlinear algorithms can generate solutions with a greater resolution than are present in the data. This is usually called super-resolution. Clearly, when there is a characteristic in an image that relies on extrapolation of unmeasured information, caution is needed in analyzing the results (extrapolating data is *never* as good as measuring the data in the first place).

Maximum entropy restorations also generally have *variable* resolution; the resolution can differ between several features in a single image. Features that are measured with good sensitivity will generally have better resolution than features with poor sensitivity. Because of variable resolution, and the dangers of interpreting super-resolved images, it is usual in some fields to convolve maximum entropy restorations back to the fundamental resolution of the data. In doing so, a convolving function that lacks some of the artifacts of the point-spread function is used.

IV. SOME APPLICATIONS

Maximum entropy image restoration is used in a number of fields, including photographic, medical, and astronomical imaging. As concrete examples of maximum entropy restoration, we will consider two specific applications from astronomy: the restoration of images from the Hubble Space Telescope and from radio interferometric imaging. In both these subfields, maximum entropy image restoration, using the Gull and Daniell measure, has found considerable favor.

A. Hubble Space Telescope

Image restoration is used much less in optical astronomy than one might first expect. This is because the careful design of optical instruments ensures that the point-spread function achieved is relatively artifact-free. Additionally, the earth's atmosphere is usually the main aberration in the observing system, and the point-spread function introduced by the atmosphere is often known only in a statistical sense as the atmosphere is constantly fluctuating. Unless the image can be sampled at a high rate (≈100 Hz), the high spatial frequencies in the image are completely lost. Regardless of the telescope size, the resolution achieved with a conventional system is approximately that of an ideal 20-cm aperture. Although optical observatories are specifically built on high mountains to be above much of the earth's atmosphere and its aberrating effect, diffraction-limited imaging can only be achieved by placing a telescope in space.* Such was the intention of the Hubble Space Telescope.

The Hubble is a 2.3-m aperture telescope in earth orbit that operates at infrared, optical, and ultraviolet wavelengths. Rather than producing diffraction-limited images, soon after its launch in 1990 it was discovered to have a significant spherical aberration in the main reflector of the telescope. This was traced back to an error in the manufacture and testing. Fortunately, the Hubble was designed to be serviced by the Space Shuttle, and in 1993 corrector optics were installed to eliminate the aberration. In the intervening years, however, much effort was placed in understanding the aberration and in deconvolving the resultant images.

The spherical aberration caused the point-spread function to cover a region of 5 arc seconds in diameter (an arc second is 1/3600th of a degree), with only 15% of the light falling in the central 0.1 arc second core. The telescope

*Increasingly, the techniques of adaptive optics are making this statement false. Adaptive optics works by measuring the aberrating effect of the atmosphere in real-time, and deforming the imaging system in such a way as to compensate for these aberrations.

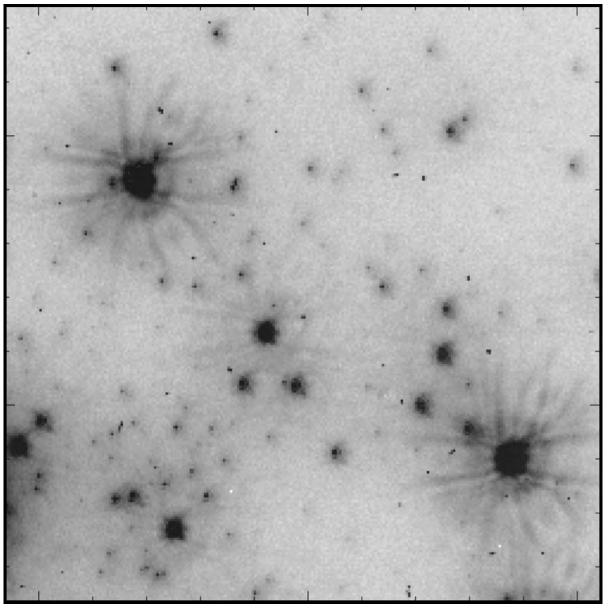


FIGURE 3 An image from the Hubble Space Telescope, showing the effects of spherical aberration. The point-spread function of the telescope is quite apparent around the two bright stars. The spots peppering the image result from cosmic ray hits or bad pixels in the detector.

design had intended that approximately 80% of the light should have fallen in this core. A significant extra complication was that the point-spread function varied with time, source position on an image, and observing wavelength. Time restrictions often prevented high-quality observations to determine the point-spred function. Attempts to model (and so predict) the point-spread function were never entirely satisfactory, and so direct measurement of this function (where possible) gave best results in image restoration processes. To compound the problems, images were often badly undersampled, and detector nonlinearities could be problematic (the Hubble had not been designed with image restoration in mind). Figure 3 shows an image from the Hubble. The halos and tentacle-like sidelobes around the two brightest stars are quite apparent.

The maximum entropy method was one of several techniques that were used to restore the Hubble images. However, all techniques were hampered by lack of complete knowledge of the point-spread function. Using a point-spread function that included errors (either because of noise in a measured point-spread function, or modeling imperfections in predicted point-spread function) further adds to the ill-posed nature of the restoration problem. Although the modulation transfer function tended not to be zero-valued (at least not as severely as in the radio interferometry case described below), this was of only small comfort when imaging weak objects, as the loss in senstivity over the design specification was quite substantial.

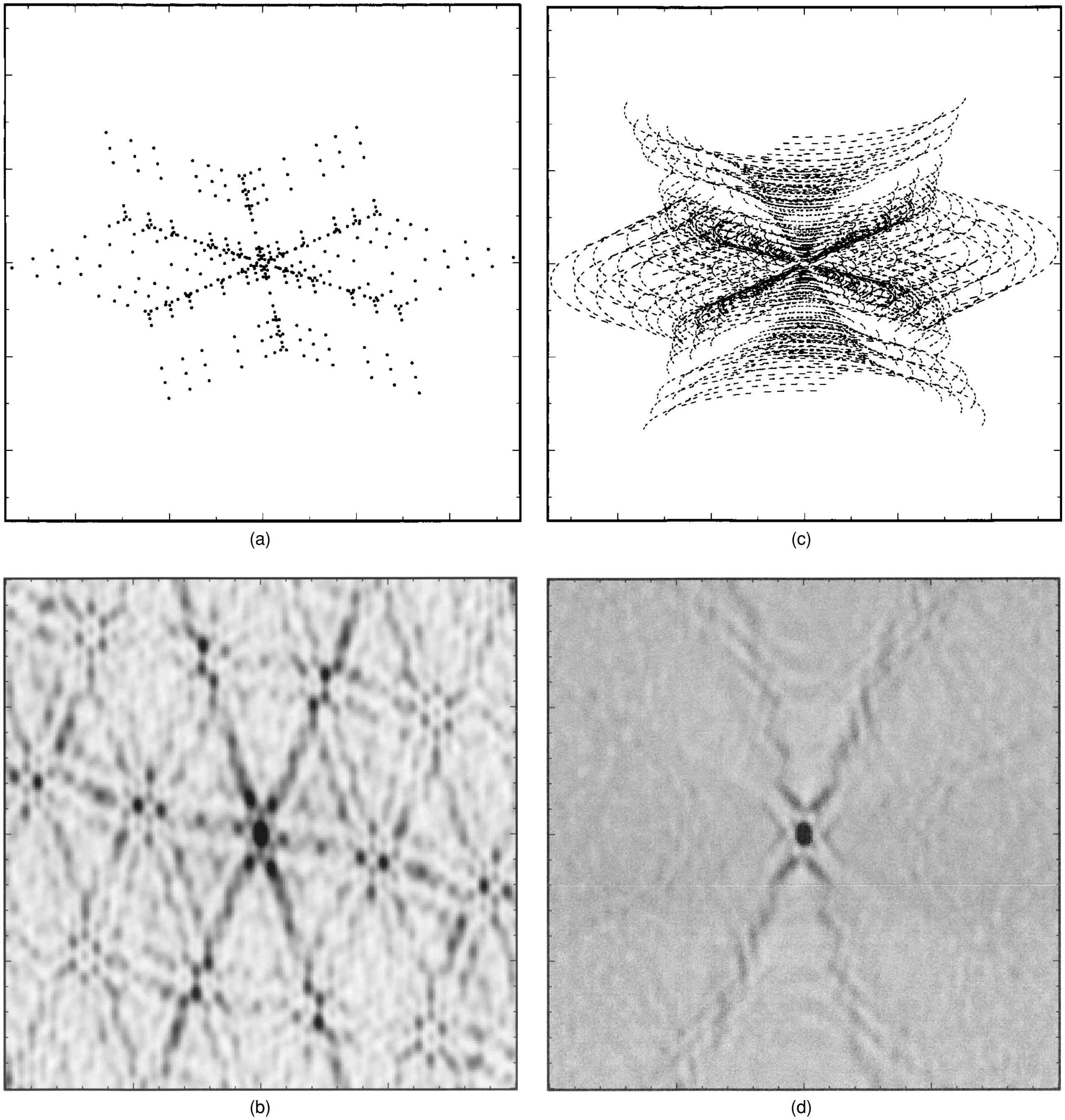
To restore large fields, the problem of the position-dependent point-spread function had to be addressed. Because of time constraints, the point-spread function could generally not be measured as a function of position. Instead, the restorations had to rely upon point-spread functions predicted by models. By segmenting the Hubble images into a large number of subimages, each subimage could then be deconvolved using a point-spread function that was assumed to be position-independent in this region.

An important issue in the restoration process is *photometric linearity*—the ability of the restoration technique to maintain a linear relationship between the brightness of a star and the response. Unfortunately, the biases in maximum entropy methods make photometric linearity a difficult proposition.

B. Radio Interferometric Image Restoration

There are two main forms of imaging used in radio astronomy: single dish telescopes, which form images by scanning a single antenna across the sky, and radio interferometric arrays. An interferometric array consists of a number of antennas, with the outputs of all antennas being brought back to a central site. The signal from each antenna is then cross correlated with all the other antennas. This cross correlation of each pair (which is a complex-valued quantity) is a sample of the Fourier transform of the sky intensity distribution. The Fourier coordinate of the sample is the projection onto the plane of the sky, at the source, of the position vector from one antenna to the other. As the earth rotates, this geometry changes, and so different Fourier samples are measured. For N antennas, there are potentially $N(N-1)/2$ different antenna pairs, and so this many different simultaneous measurements of the Fourier plane. By using many antennas, allowing for earth rotation and possibly using physical reconfiguration of the antenna array, many different measurements of the Fourier transform of the image can be collected. When complete, these data can be Fourier transformed to produce an image. If the sampling of the Fourier plane is complete, then the resultant image would be perfect. However, getting complete Fourier coverage is infeasible: the number of antennas clearly will be finite, and available observing time will be limited. In practice, no interferometric observation ever completely samples the Fourier plane, and the sampling can vary from quite good to very sparse. Two examples of Fourier plane sampling patterns and their corresponding point-spread functions are given in Fig. 4.

Given that an interferometer measures the data in the Fourier domain, the modulation transfer function will


(a)

(c)

(b)

(d)

FIGURE 4 The Fourier plane sampling pattern (top) and the resulant point-spread functions (bottom) for an observation with the Very large Array (near Socorro, New Mexico). The first sampling pattern corresponds to a very brief observation (a so-called snapshot). The second corresponds to a 6-hr observation, where the earth's rotation has been used to improve the sampling pattern in the Fourier plane.

simply have a value of 1 where a sample is measured, and a value of 0 otherwise.* Consequently, the modulation transfer function and the point-spread function for an interferometric observation are well defined. Radio interferometry produces a classical image restoration problem, where large parts of the modulation transfer function are zero-valued. Consequently, it is a very nonunique problem.

The point-spread functions in radio interferometry differ from those in many imaging applications. First, the

*Strictly, an extra weight is often applied to the Fourier samples before transforming them, and so the value of the modulation transfer function will not be 1, but the value of this weight. The weights are chosen in a way that is a compromise between good sensitivity and a clean point-spread function.

point-spread function is both positive- and negative-valued, and, secondly, there is typically a significant non-zero response to a point source across the entire image. The responses away from the main peak of the point-spread function are known as sidelobes (i.e., as distinct from the *main lobe* of the response). The sidelobes from a strong source, or the combined sidelobes from an extended source, make it impossible to analyze subtle features in the raw image (the raw image being that formed by simply transforming the Fourier data). Image restoration has become an integral part of radio interferometry. Rather descriptively, radio astronomers talk of the raw image as being "dirty," and the process of restoring (or deconvolving) this image as "cleaning."*

As with optical astronomy, bias of the maximum entropy solution can be problematic in blank regions of the sky. The bias tends to create an artificial plateau of emission. Although this is most noticeable in low-sensitivity images, the bias exists (but in a more subtle form) even when the sensitivity is very good. Deconvolving point sources is also problematic with maximum entropy algorithms, particularly when the point source lies on top of a plateau of emission.

A significant problem in interferometry is that it is not possible to measure the smallest spatial frequencies—these are inaccessible because the antennas physically cannot get closer than one diameter apart. Thus, the integrated flux of a source cannot be measured (it corresponds to a measurement at the origin of the Fourier plane), nor can large-scale structure. An alternative way of thinking of the same problem is to recognize that the field-of-view of an observation (or *pointing* of the telescope) is fixed by the antenna size. It is not possible to accurately measure structure with a scale size comparable or larger than this field-of-view.

Radio astronomers have solved this problem by using a technique known as *mosaicing*. Mosaicing is the practice of combining the data from many pointings of adjacent

*In interferometry, the term clean is used both as a generic term meaning to restore an image, and as a particular algorithm (introduced by Högbom) which implements such as restoration.

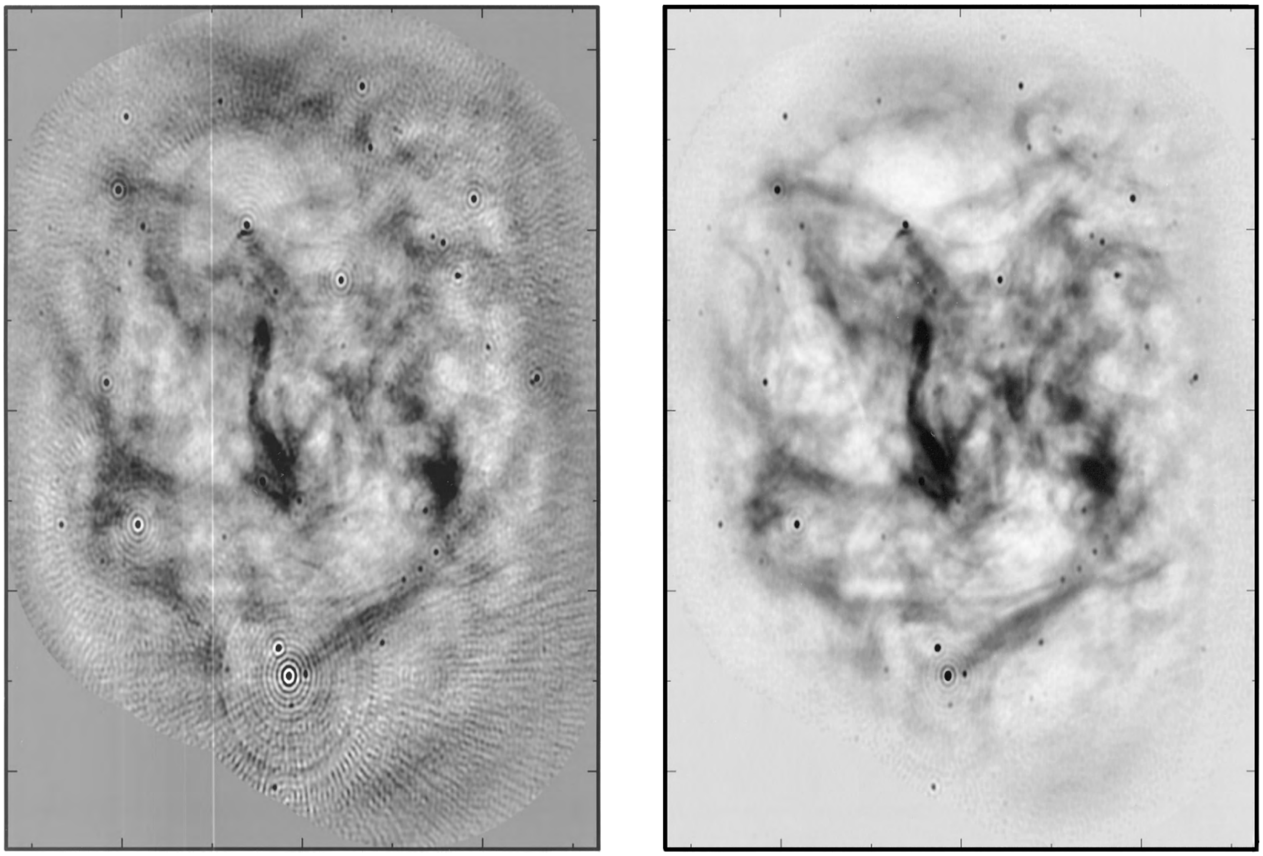


FIGURE 5 Images of the so-called Vela-X region formed by a radio interferometric observation at 20-cm wavelength. This image uses 35 pointings to form an image much bigger than the field-of-view of the telescope. The left and right panels are before and after restoration (using a Gull and Daniel maximum entropy algorithm). Single-dish data were included in the restoration process.

parts of the sky to form a single image. In principle, mosaicing allows spatial frequencies down to just above 0 to be recovered. Maximum entropy image restoration is easily generalized to handle a joint restoration of the mosaic data. The data from the different pointings simply constitute added constraints to be included in the entropy maximization process. Modeling of the differing fields-of-view and point-spread functions of the various pointings are readily handled.

Even with mosaicing, it is still not possible to measure the sample at the origin of the Fourier plane with an interferometer. The interferometer data must be augmented with single-dish observations (single-dish data measures all spatial frequencies up to the dish diameter). Again, using a maximum entropy approach, combining single-dish and interferometer data together is quite natural and straightforward—the single-dish data is yet more data constraints. Figure 5 gives an example of a mosaiced observation that includes single-dish data. Whereas the restoration shown in Fig. 5 is plainly far superior to the dirty image, one of the weaknesses is apparent in the strong point source to the south: maximum entropy has failed to completely remove the sidelobes from this source. The data does not rule out this source having rings around it, although physically such a structure is implausible.

V. CLOSING REMARKS

Nonlinear approaches to image restoration have proven to be far superior to classical linear approaches. The maximum entropy approach, in particular, has proven to be very flexible in allowing a wide and complex variety of data and constraints to be used in the restoration process. The maximum entropy method, however, is not without shortcomings. The biased nature of the solution, in particular, is problematic in a number of applications.

Ideally an image restoration technique will deliver an image that is consistent with available data and constraints (e.g., positivity), and which is free of obvious artifacts. Any technique that achieves this should be taken seriously, regardless of whether it is based on an ad hoc procedure or justified by a formalism such as maximum entropy. It is the data, ultimately, that must drive a restoration process. In analyzing the solution to any ill-posed problem, it is important to differentiate between those characteristics dicatated by the data, and those that are dependent on the solution technique. Any physically implausible feature that is not required by the data should be ignored.

SEE ALSO THE FOLLOWING ARTICLES

Image Processing • Imaging Optics • Imaging Through the Atmosphere • Optical Information Processing • Smart Pixels • Solid-State Imaging Devices • Statistics, Bayesian • Telescopes, Optical

BIBLIOGRAPHY

Taylor, G. B., Carilli, C. L., Perley, R. A., eds. (1999). "Synthesis Imaging in Radio Astronomy II," Astronomy Society of the Pacific Conference Series, Vol. 180.

Jansson, P. A., ed. (1996). "Deconvolution of Images and Spectra," Academic Press, San Diego.

Hanisch, R. J., and White, R. L., ed. (1994). The Restoration of HST Images and Spectra—II, Proceedings of a workshop held at the Space Telescope Science Institute (http://www.stsci.edu/stsci/meetings/irw).

Sivia, D. S. (1996). "Data Analysis: A Bayesian Tutorial," Oxford Science Publications, Oxford.

Wu, N. (1997). "The Maximum Entropy Method," Springer Series in Information Sciences, 32, Springer-Verlag, New York.

Infrared Astronomy

Rodger I. Thompson
University of Arizona

GLOSSARY

Blackbody An object that is a perfect absorber and emitter of electromagnetic radiation at all wavelengths.

Diffraction limit The smallest image possible for a telescope observing a point-like object such as a star. The larger the telescope the smaller the angular size of the image. Without adaptive optics, ground-based telescopes larger than about 12 inches cannot achieve diffraction-limited imaging due to distortions caused by the earth's atmosphere.

Dewar A container for very cold material or cryogens such as liquid nitrogen or helium. The dewar usually contains both the cryogen and the detector system.

Infrared bands Spectral regions, usually defined by atmospheric transmission, where the brightness of the infrared flux of an astronomical object is measured.

Jansky A unit of flux measurement equal to 10^{-26} watts m^{-2} Hz^{-1}.

Kelvin (K) A unit of temperature with increments equal to centigrade units but with 0° at absolute 0; 0°C is 273.15°K.

Luminosity The total amount of power emitted by an astronomical object. The luminosity of the sun is approximately 4×10^{33} ergs s^{-1}.

Magnitude A logarithmic system expressing the brightness of an astronomical object. The larger the magnitude, the fainter the object.

Spectral classification A system of classifying stars by their temperature. The range of spectral classifications goes as O, B, A, F, G, K, M, and recently added L. The hottest stars are O stars and the coolest L. Each temperature is also generally divided into subclasses, which depend on the luminosity of the star.

INFRARED ASTRONOMY encompasses astronomical observations in the spectral region from 1 μm to 1 mm, a factor of 1000 in wavelength. Optical astronomers often declare the region from 7000 Å to 1 μm as the near infrared; however, most infrared astronomers define the near infrared as the spectral region between 1 and 5 μm. The midinfrared includes the region between 5 and approximately 100 μm, and the region from 100 μm to 1 mm is declared the far infrared region. The region between about 400 μm to 1 mm is also often referred to as the submillimeter region by radio astronomers. The divisions historically derived from both atmospheric windows and detector types. In the modern world of space infrared astronomy, the boundaries between the definitions are quite blurred, with many detector types covering more than one infrared region. As we will see as this chapter progresses, observations in the near infrared region are predominantly of the intrinsic radiation of an astronomical object, whereas observations in the mid- and far infrared region are predominantly of reradiated emission, where the intrinsic radiation of an object has been absorbed by dust and re-emitted at infrared wavelengths.

The decade of the 1990s witnessed a phenomenal growth in the capabilities of infrared astronomy with the introduction of large-format detector arrays covering a large spectral band of the infrared region. At this time only the far infrared utilizes single detectors or small hand-crafted arrays. At the same time the sensitivity of the array detectors has increased, with each pixel of the current detector arrays being far more sensitive than any of the previous single detectors.

I. INFRARED BANDS

Ground-based infrared observations must be conducted in the wavelength regions where the earth's atmosphere is transparent to infrared radiation. These transmission bands generally define the common photometric bands used by infrared astronomers. Harold Johnson who also defined the ultraviolet, blue, visible (UBV) photometric system for visible wavelengths first defined the near infrared bands of J, K, and L. Table I gives the common infrared photometric bands in use today. Note that there are subtle differences in the exact definition of the bands.

TABLE I Common Infrared Photometric Bands

Band designation	Wavelength range
J	1.1–1.4
H	1.5–1.8
K	2.0–2.4
L	3.5–4.5
M	4.4–5.0
N	8.0–12.6
Q	16.5–23.0

Commonly used systems are the Johnson system, the Arizona system, and the Caltech system. For extremely accurate work it is important to know what system is being utilized. Some of these bands have alternative filters designated by a prime such as K′ that are designed to reduce the thermal emission inside the band by narrowing the spectral range to only the highest transmission regions. The longer wavelength filters that define the M, N, and Q region are often tuned to the expected spectral transmission for a particular observatory. Observatories at very high and dry locations may be able to use broader filter ranges than observatories at less favorable locations.

The intensity of the detected radiation is often expressed in magnitudes, units that are peculiar to astronomy. Magnitudes are a logarithmic unit. The flux of an object in a unit such as ergs cm^{-2} sec^{-1} is given by

$$F = C \times 10^{-mag/2}, \tag{1.1}$$

where C is a constant defined for each band and *mag* is the astronomical magnitude. The value of C varies with the magnitude system even for a given band, so all reported values should be checked to see which system is being used. As is evidenced by Eq. (1), the larger the magnitude the fainter the object. In one system of magnitudes the observed flux from the star Vega defines zero magnitude in each band. Another system of magnitudes called the Ab magnitude defines a flux of 3.6×10^{-20} ergs cm^{-2} sec^{-1}Hz^{-1} as zero magnitude. It is also common in infrared astronomy to express the flux in the more physical units of Janskys borrowed from radio astronomy. One Jansky is 10^{-26} watts m^{-2} hz^{-1}.

All astronomical objects emit infrared radiation. In fact, the peak of the integrated light from all of the stars in a normal galaxy is near a wavelength of 1 μm. The amount of infrared radiation emitted by a star, planet, or other astronomical object is determined by its temperature, the blackbody or Planck curve, and its emissivity at each wavelength, as given in Eq. (1.2).

$$F(\lambda) = \varepsilon(\lambda)B(\lambda), \tag{1.2}$$

In Eq. 1.2, $F(\lambda)$ is the flux as a function of wavelength λ, $\varepsilon(\lambda)$ is the emissivity and $B(\lambda, T)$ is the Planck function. Figure 1 shows the Planck function for two temperatures, 3000 K and 2500 K. 3000 K is a typical temperature of a late K star, the stellar type that generally dominates the spectrum of galaxies. Note that as the temperature decreases the peak wavelength shifts to longer wavelengths and the amount of emission decreases. It is important to note that the lower temperature Planck curve falls below the higher temperature curve at all wavelengths. The Planck curves of various temperatures are nested curves that do not cross each other.

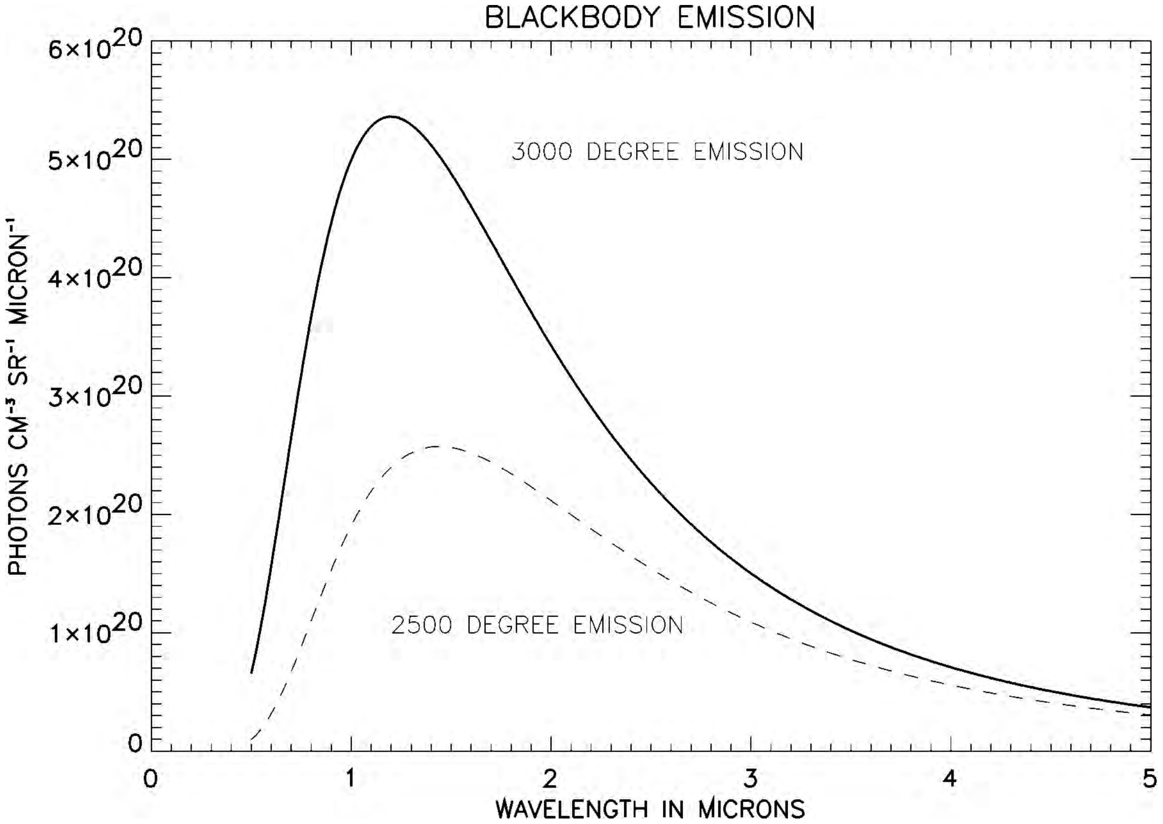


FIGURE 1 The blackbody spectrum at two different temperatures.

II. INFRARED OBSERVATIONS

There are some objects that radiate predominantly by infrared light and in fact are not detectable at the shorter optical wavelengths. Historically, infrared astronomy has concentrated on these types of objects, but recently near and midinfrared observations are increasingly studying objects that traditionally have been the study of optical astronomy.

A. Predominantly Infrared Objects

An object may radiate predominantly in the infrared for several reasons. It is in fact these reasons that make infrared astronomy essential for a complete understanding of an object's physics.

1. Objects Affected by Interstellar Dust

Interstellar dust is a ubiquitous component of our galaxy and most other galaxies. Extinction by interstellar dust is often called "reddening" since the dust is far more efficient at absorbing and scattering blue light than red or infrared light. We see this same phenomenon in terrestrial sunsets. The dust in our atmosphere removes the blue light much more efficiently, turning the sun into a red orb. In some cases the reddening can be so great as to make objects undetectable at optical wavelengths while remaining quite bright at infrared wavelengths. Primary examples of this are regions of star formation, which are generally obscured by large amounts of dust, and the center of our own and other galaxies. The central region of our galaxy suffers about 100 magnitudes of extinction at optical wavelengths but less than 10 magnitudes at the near infrared wavelength of 2.2 μm. We will discuss the fate of the absorbed radiation in section II.A.3.

2. Objects That Radiate Infrared Light

Objects that radiate mainly at infrared wavelengths may do so because of their low temperature by astronomical standards. Objects that fall in this category start with stars of spectral classifications of K or cooler extending down the newly designated spectral classification of L. The substellar classification of Brown Dwarf links stars generating energy by nucleosynthesis to planets such as the giant and terrestrial planets in our own solar system. The planets such as the earth radiate like blackbodies at their surface temperature modified by their emissivity, as described in Eq. (1.1). At an average temperature of approximately 300 K, the peak of the earth's radiation lies at the mid-infrared wavelength of 10 μm. The peak wavelength in micrometers of radiation is given simply by Wien's Law, where the temperature T is in Kelvins.

$$\lambda_{\text{max}} = 2898\ \mu\text{m} \tag{1.3}$$

3. Reradiated Light

Some objects are either surrounded by dust or are embedded in dense dust clouds. The dust may be directly associated with objects, such as the dust disks around young stellar objects or the ejected dust clouds in late-type giant or supergiant stars. Examples of embedded objects are clusters of newly formed stars in star formation regions, more evolved giant and supergiant stars such as infrared (OH/IR) stars or the central stars and perhaps central black hole of a galaxy. In either case the nearby dust is heated by the central object and reradiates the absorbed energy at mid- and far infrared wavelengths. In general, the temperature of the dust runs from a few 100s to a few 10s of degrees. The approximate maximum temperature of dust before it is destroyed is 1500 K. Equation (1.2) shows that even at this temperature the maximum emission wavelength is in the infrared range. The amount of reradiated emission is equal to the amount of light absorbed by the dust, as described in section II.A.1.

The InfraRed Astronomical Satellite (IRAS) described in section VI.B found that many galaxies have most of their starlight absorbed by dust and reradiated at mid- and far infrared wavelengths. Follow-up observations by the Infrared Space Observation (ISO) (section VI.D) measured the midinfrared spectra of many of these galaxies. Figure 2 shows the spectrum of a galaxy, M82, known to be producing stars at a greatly accelerated rate. Galaxies of this type are often designated as starburst galaxies.

The continuum emission in this spectrum is due to thermal emission from hot dust that has been heated by absorbing starlight. The sharp emission lines are due to radiation from various atoms and atomic ions. Atomic designations with square brackets such as [NeII] indicate the emission is due to a forbidden transition. A forbidden transition is a transition via the atom's electronic quadrapole moment or magnetic dipole moment rather than the electronic dipole moment. The broader emission features designated by PAH (polycylic aromatic hydrocarbons) are emission by dust components. One of the major discoveries with ISO spectroscopy has been that these PAH emission features

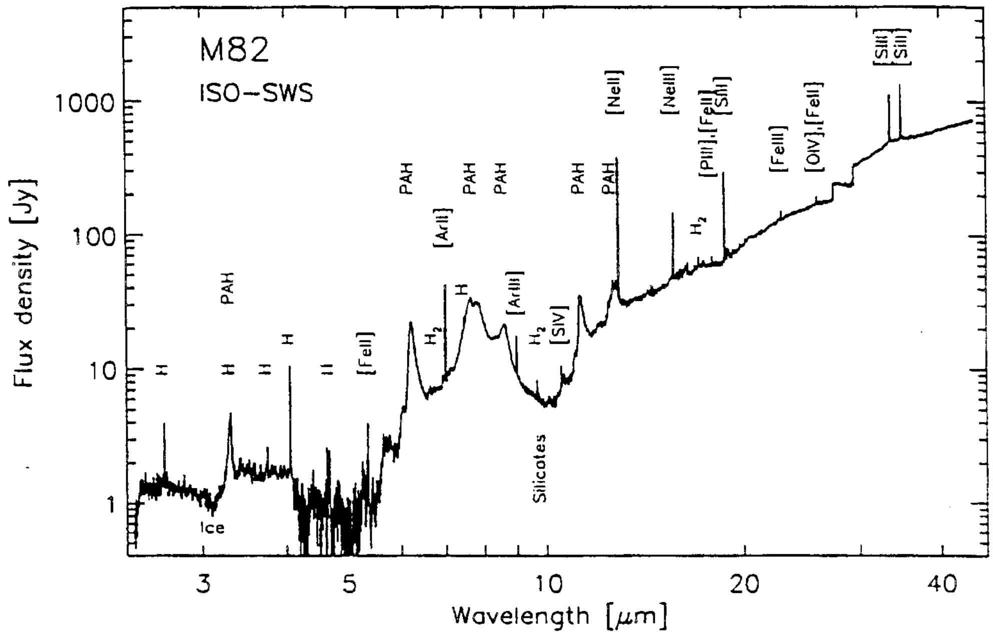


FIGURE 2 The infrared spectrum of M82 by D. Lutz et al. PAH, polycyclic aromatic hydrocarbons; ISO, Infrared Space Observatory.

Source: Lutz, D., *et al.*, "Probing starbursts with ISO mid-IR spectroscopy," *in* Extragalactic Astronomy in the Infrared, MPI fur extraterrestrische Physik.

appear to be ubiquitous in the spectra of galaxies with significant starburst activity. Also note the broad absorption feature near 10-μm designated silicates. This is a very commonly observed feature in the spectrum of dust-emitting objects. It indicates there is a solid-state silicon oxygen bond in the dust material. This is similar to the mineral olivine found on earth.

4. Objects That Are Red Shifted

Objects such as galaxies can have their observed radiation be predominantly at infrared wavelengths due to the redshift caused by the expansion of the universe. Before the advent of modern infrared array detectors, this was not a normally observed class of objects due to the lack of sensitivity in previous detectors. The redshift of light from distant objects is often described as a Doppler shift of light due to the high recession speed of distant objects. This is actually not a correct interpretation. The redshift is due to the expansion of the universe. When we observe very distant objects we see them as they were a long time ago. This time is simply the time it takes for the emitted light to travel from the object to us. In some cases the time is so long that we are seeing the universe when it was much smaller that it is now. If the galaxy or object emitted radiation with a wavelength λ at that early time, then the wavelength of the emitted radiation has expanded along with the universe to a longer or redder wavelength by the time we receive it. We call this expansion of wavelength the redshift and give it a quantitative name z defined by

$$\lambda_{observed} = (1+z)\lambda_{emitted}. \tag{1.4}$$

This simply states that the universe at the time of observation is $1+z$ times bigger than the universe was at the time the light was emitted. Since most galaxies are intrinsically red to begin with, very distant galaxies have most of their radiation in the infrared region due to the redshift. This makes the infrared spectral region very critical in the study of the early universe.

A second aspect of the universe makes the infrared spectral region important for high redshift observations. Hydrogen gas in a galaxy and in the intergalactic medium absorbs all of the starlight at wavelengths shorter than 912 Å. In fact, for high redshift objects the intergalactic gas absorbs almost all of the light shorter than the Lyman α line at 1215 Å. The absorption due to the intergalactic gas clouds is called the Lyman α forest. The absorption due to the Lyman α forest is so strong that for objects at a redshift of 4 or more no light with rest wavelengths shorter than 1215 Å reaches the earth. At a redshift of 7, the 1215 Å light is shifted to a wavelength of almost 1 μm. This means that for objects with redshifts greater than about 7, all of the light that reaches earth is in the infrared or longer spectral region and thus can only be observed at infrared and radio wavelengths.

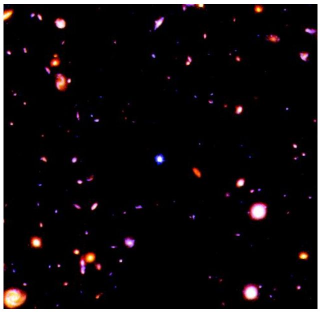


FIGURE 3 The near infrared view of part of the Northern Hubble Deep Field.
Source: R. Thompson, NASA.

At this time the majority of near infrared observations of very distant galaxies have been carried out either with large ground-based telescopes or with the NICMOS instrument on the Hubble Space Telescope (HST) (described in section IV.A). Figure 3 shows a region of sky observed very deeply with the near infrared camera and muliobject spectrometer (NICMOS) that contains the most distant objects observed to date. This is in a part of the sky called the Northern Hubble Deep Field that has been observed very deeply at optical wavelengths with HST as well. Since the NICMOS images were only in two bands centered on 1.1 and 1.6 μm, a third band was added, which is a band from previous Wide Field and Planetary Camera 2 observations at 0.606 μm. The 1.6-, 1.1-, and 0.606-μm images are combined as red, green, and blue to form a full-color picture. It is evident by looking at Fig. 2 that most galaxies are red, indicating more infrared than optical emission. Note that the only object in this image that is not a galaxy is the central bluish dot, which is an individual star in our own galaxy. The expected launch of the Space InfraRed Telescope Facility (SIRTF) (section IV.C) in 2001 should provide midinfrared observations of very distant objects.

Observations of distant objects at infrared wavelengths has been deemed so important that NASA's largest future project is dedicated to that mission. It is called the Next Generation Space Telescope (NGST) and is scheduled to be launched before 2010. A description of the current status of NGST is given in section IX.B.

5. Cosmic Background Radiation

Our universe started with an event called the Big Bang. At the moment of its creation the universe was an infinitesimally small point, called a singularity, that contained all of the matter and radiation. The radiation in the singularity was extremely hot and dense. Since that time the universe has expanded to its present size and the original radiation has cooled to a temperature of approximately 2.7°K. Actually, the radiation has redshifted to much longer wavelengths as explained in section (II.A.3). Radiation at this temperature peaks in the radio region, but it also has a significant far infrared component. The remnant big bang radiation is present throughout the universe and can be seen in all directions in the sky. It is called the cosmic background radiation and has been the subject of intense study recently. Much of this study has been stirred by the phenomenal success of the Cosmic Background Explorer (COBE) satellite (described in section IV.E). COBE operated over a large wavelength region spanning the submillimeter through the near infrared region. It first found the very subtle differences in temperature at various points in the sky that are the precursors of the density fluctuations that eventually grew into galaxies such as the one we live in. Current studies are attempting to detail these fluctuations to learn about the fundamental parameters of our universe. Figure 4 shows a map of the sky with fluctuations found by COBE. The higher temperature regions appear lighter in the map. This map contains about 1 year of data from the differential microwave radiometer (DMR) instrument on COBE.

Very recently there have been two balloon-borne instruments, Maxima and Boomerang, that have measured smaller pieces of the sky but at higher spatial resolution. These experiments have concentrated on how the spatial variations of the background match to predictions made from various models of the universe. They are consistent with a universe that is flat rather than curved and imply that most of the matter in the universe must be in a currently unobservable form, which has the designation dark matter.

The bottom of Fig. 4 shows another profound measurement by COBE. The far infrared absolute spectrophotometer (FIRAS) instrument on COBE measured the spectrum of the cosmic background radiation and found it to be an essentially perfect fit to a blackbody spectrum. The solid line in the figure is a theoretical perfect blackbody spectrum and the squares are the measurements. Any cosmological theory must be able to predict this remarkably perfect fit. These two measurements are among the most important observations ever made in infrared astronomy.

B. Normal Objects Studied in the Infrared

Infrared astronomy is no longer limited to those objects whose main emission is in the infrared spectral region. Infrared studies are now carried out on objects to determine characteristics that are not visible at other wavelengths, are easier to determine at infrared wavelengths, or are coupled with observations at other wavelengths to tell a more complete story about the object. This is particularly true in the near infrared spectral region, where many studies previously carried out predominantly in the optical spectral region are now performed in the near infrared. Below are just a very few examples of these types of observations.

1. Stellar Atmospheres

The infrared region offers some unique spectral features that yield significant information about the composition and conditions in stellar atmospheres. In particular, many molecules such as carbon monoxide (CO) have strong absorption features in the infrared due to vibration-rotation transitions. The higher energy electronic transitions will often lie at ultraviolet energies so that infrared observation is the only practical way to study these molecules with ground-based facilities. In cool stars there is also very little radiation at ultraviolet wavelengths.

2. Molecular Hydrogen

Near infrared spectra and images have been particularly fruitful in the observation of interstellar molecular hydrogen. The strongest line is a quadrapole transition at 2.12 μm first observed in a region of intense star formation. Although it is possible to excite the line through ultraviolet fluorescence, in most cases the excitation comes from shocks due to outflowing gas around young stars. Regions of high star formation, such as the Orion nebula where molecular hydrogen emission was first discovered, have extensive areas of intense emission. Molecular hydrogen emission has also been observed in other galaxies, particularly in galaxies with high star formation rates, which are often called starburst galaxies. Hydrogen molecules lack a dipole moment, so there are no spectral features in the visible wavelength region.

3. Stellar Radiation in Galaxies

Our optical view of galaxies, particularly of spiral galaxies, is often biased by the youngest, most luminous, and most massive stars. These stars so outshine their colleagues at optical wavelengths that the actual distribution of stellar masses is difficult to observe. It is also the case that the central regions or nuclei of spiral galaxies and very active galaxies called AGNs are obscured by dust, as are many of the very active star forming regions. Extension of our observations into the infrared gives us a much better view of the actual distribution of stars in the galaxy. Mid- and far infrared studies also measure the amount of optical and ultraviolet light that is absorbed by dust in the galaxy by observing the reradiated light described in

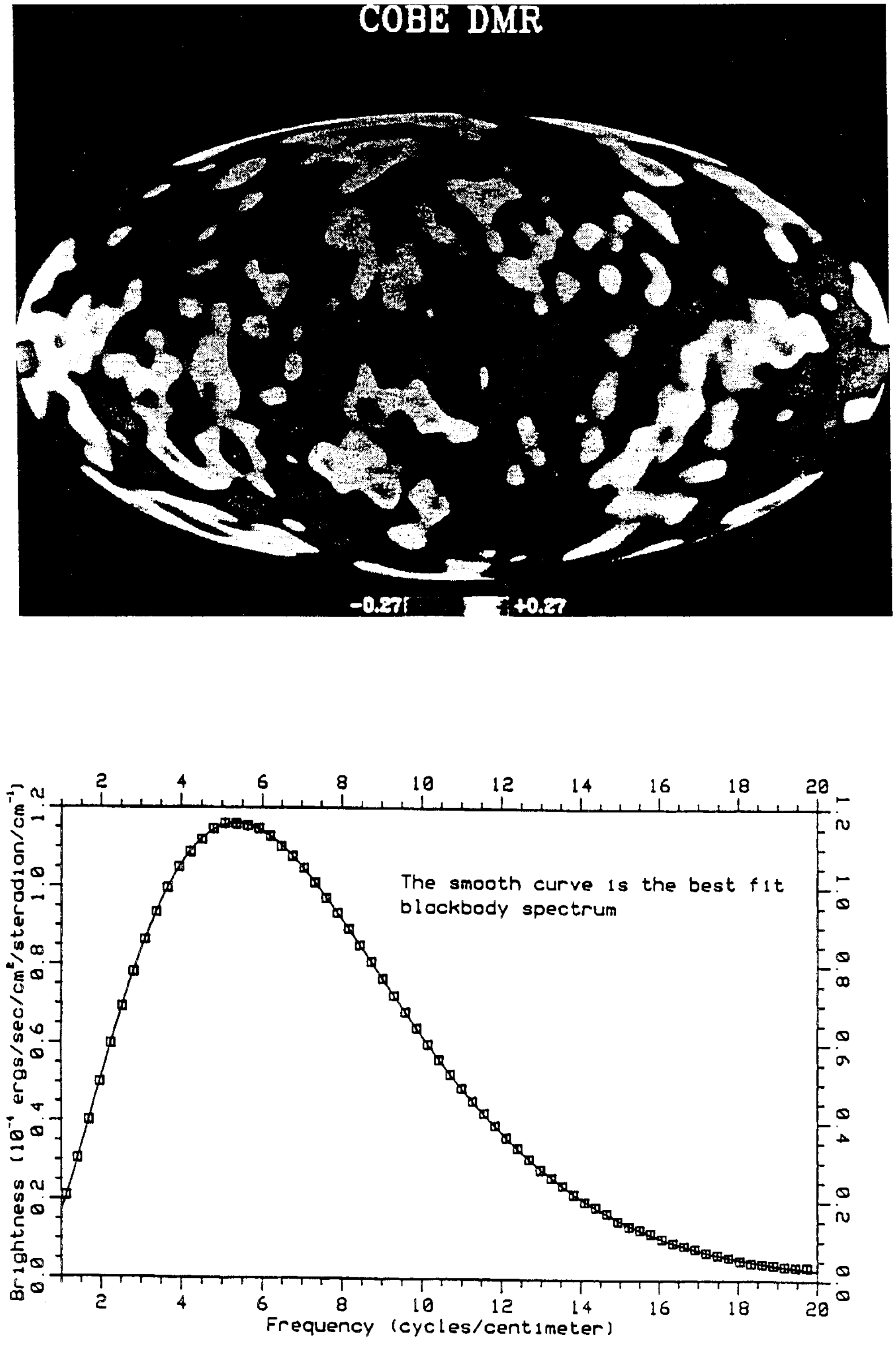


FIGURE 4 The top figure shows the distribution of temperature fluctuations in the cosmic background radiation. The bottom figure shows the exact match between the spectrum of the cosmic background radiation and a blackbody spectrum. COBE, Cosmic Background Explorer; DMR, differential microwave radiometer.
Source: Mather, J. C., and Boslough, J., "The Very First Light." Basic Books, New York, 1996.

section (II.A.2). In some cases most of the optical and ultraviolet light is absorbed and the total power of the galaxy can only be measured by observations in the infrared.

4. Galaxy Morphology

A very detailed and extensive classification process has been developed to characterize the shapes or morphologies of galaxies. This has become increasingly complex to describe the very large range of the optical appearance of galaxies. It is now being realized that this optical classification scheme may be very influenced by the distribution of dust and extremely luminous stars in a galaxy rather than by the actual distribution of stars. Near infrared images of galaxies generally show a much simpler and smoother structure which has led some morphologists to suggest that our classification process be revised and based on near infrared rather than optical images of galaxies. This was impractical in the past due to the limitations of infrared imaging mentioned at the beginning of this chapter. The presence of large format infrared cameras now makes such a change fairly easy in practice if not in preference.

III. TECHNIQUES FOR OBSERVING INFRARED RADIATION

Since the infrared region spans a factor of 1000 in wavelength, it is not surprising that there are several detector technologies needed to cover the entire infrared spectrum. Detector technology is changing fast, and a thorough discussion is impossible in this limited space. The following is a brief summary of some of the major technologies. Although in the last decade single-detector elements were employed for astronomical observations, that is rarely the case today. Most detectors are now arrays except for submillimeter receivers and bolometers, although there are now arrays containing a small number of individual bolometers.

A. Detector Arrays

Optical astronomy has long used large-format detector arrays with most of the arrays employing the charge-coupled device (CCD) technology. Unfortunately, the almost universal material for CCDs, silicon, is not a good infrared detector. There have been some devices that employed infrared detectors on top of CCDs, but these have generally been abandoned in favor of multiplexed direct readout devices. These arrays employ a large-format detector array that is indium bump bonded to a multiplexer (or mux) using silicon technology, as shown in Fig. 5.

Each detector pixel transfers charge to a multiplexer cell that contains a complete readout circuit. The signal from this circuit is sent to the output amplifiers via shift registers. An early version of the circuit for the detection and transfer of charge deposited by the radiation is shown Fig. 6.

The detector is represented by the diode symbol at the lower right of the circuit diagram just above the words DET SUB. The transistor circuit with the labels M1 through M4 are a set of switches that isolate the detector when observing, allow the voltage to be sensed when being read out, and reset the voltage to zero after the observation is finished. The voltage is due to the charge collected on the capacitance of the detector, which is the signal charge. The rest of the circuit contains the shift registers for addressing a particular pixel and the output amplifier that amplifies the detected voltage for measurement, but subsequent readout electronics are not shown.

Unlike CCDs the signal charge is not transferred from cell to cell. This has the general advantage of nondestructive readouts, which allows the signal to be monitored as it is built up. Unwanted events such as cosmic ray hits can then be detected and removed from the final signal. We will now take a quick look at some detector technologies in the various wavelength regime.

1. Near Infrared (1–5 μm)

There are two major detector materials in the near infrared, mercury cadmium telluride (Hg:Cd:Te), and indium antimony (In:Sb), although some lower efficiency detector arrays have been made with platinum silicide (PtSi). This wavelength region was the first to utilize arrays and is the most developed to date. Part of this has been due to the development in the early 1990s of Hg:Cd:Te arrays for the NICMOS on the HST. The NICMOS arrays contained 256×256 elements. Currently, 2048×2048 arrays of both Hg:Cd:Te and In:Sb are under development.

The advantage of Hg:Cd:Te good operation at liquid nitrogen temperatures (77 K), whereas In:Sb generally requires operation at liquid helium temperatures (4 K). In general, however, In:Sb has a lower dark current than Hg:Cd:Te arrays that can operate out to 5 μm. Hg:Cd:Te arrays can tune their long wavelength cut-offs by varying the amount of mercury in the detector. The shorter the cut-off wavelength, the lower the dark current.

2. Midinfrared (8–30 μm)

The wavelength region between 5 and 8 μm is blocked from ground-based observation due to atmospheric absorption. Therefore not much attention has been paid to that particular region, although it is available via space

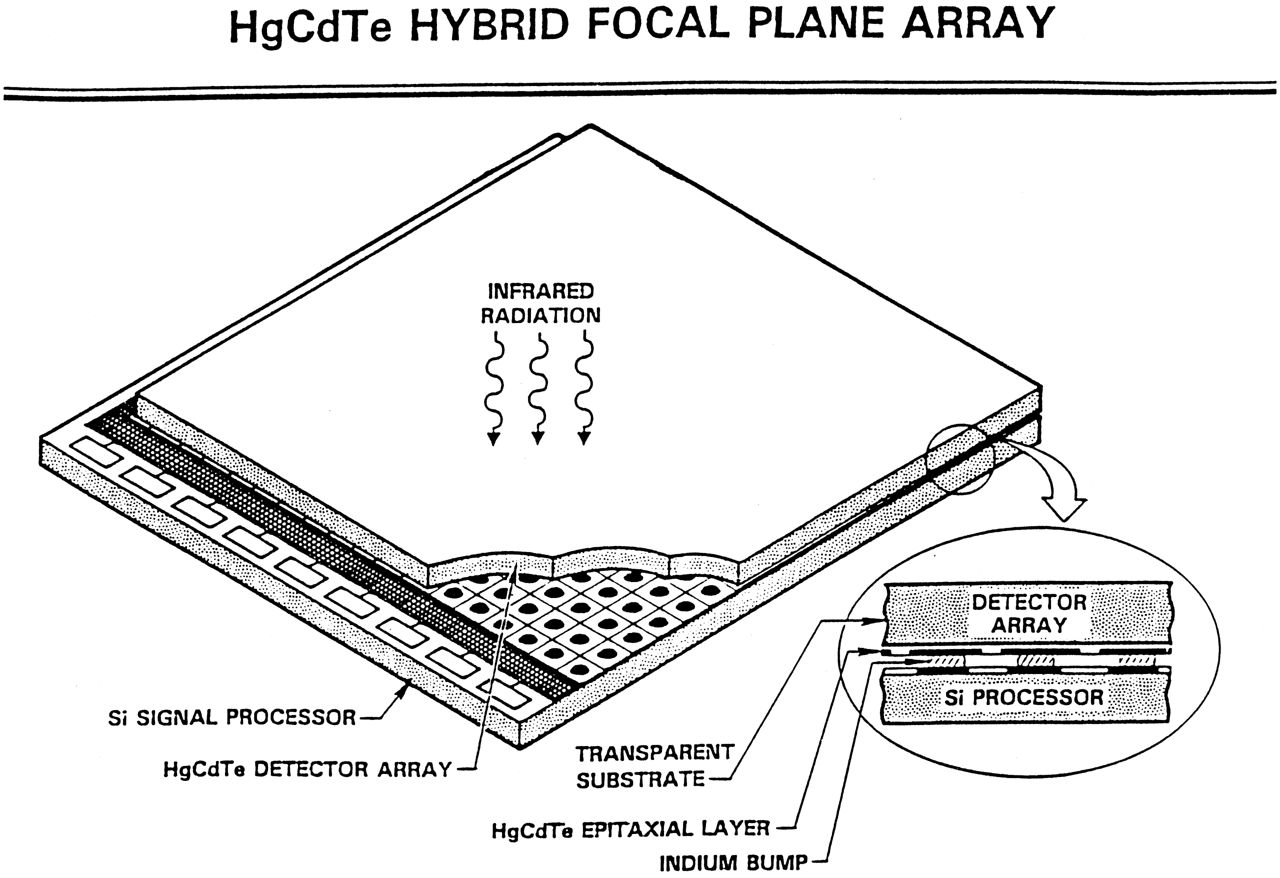


FIGURE 5 The structure of a hybrid detector array and multiplexer.

observation, and some space observatories (ISO and COBE) have made observations there. The major detector technologies in this region are the blocked impurity band (BIB) and impurity band conductor (IBC) detectors. These detectors have an additional layer of pure undoped material between the detection impurity or doped material and the electrical contacts. This layer increases the resistance and also the sensitivity of the detector. Examples of current detector types are silicon arsenic (Si:As) in both the BIB and IBC format and silicon antimony (Si:Sb) BIB detectors. All of these detectors now exist in array configurations. In addition, silicon gallium (Si:Ga) operating in a photoconductive mode is operating in a ground-based infrared camera. The multi band imaging photometer (MIPS) instrument on SIRTF will employ germanium gallium (Ge:Ga) array detectors out to a wavelength of 70 μm.

3. Far Infrared

The same detector material (Ge:Ga) used in midinfrared arrays can also be utilized for small arrays in the far infrared region. It was discovered that if a mechanical stress is placed on the detector material, it increased its wavelength response to significantly longer wavelengths. In general the requirement of applying mechanical stress has limited these units to small arrays only two pixels wide in one direction.

B. Bolometers

As its name implies, a bolometer is a detector designed for bolometric measurements or measurements that are sensitive to all wavelengths. In practice, however, the incoming radiation is limited to a given spectral region by a filter in the optical system. The thermometer used by Herschel to first detect near infrared radiation from the sun was a type of bolometer. The modern version is an individual Ge:Ga or doped Si detector suspended in a small enclosure. The detector is coated with material that is highly absorbing at the wavelengths of interest. The incident radiation absorbed by the detector raises its temperature. Since the electrical resistance of a bolometer is temperature sensitive, there is a change in the current carried by the thin suspending wires, which in turn produces a signal. The time duration of the signal is determined by the rate at which heat leaks from the bolometer. This is generally

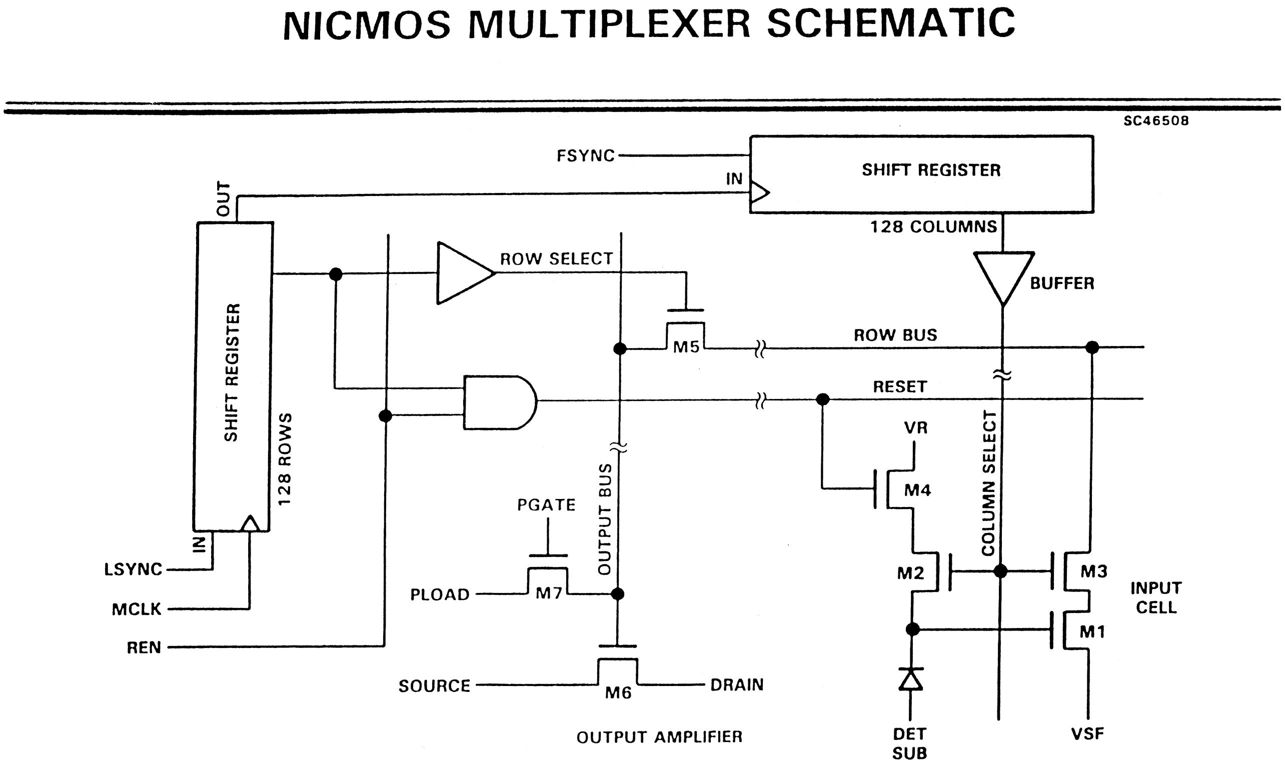


FIGURE 6 An electrical schematic of the unit cell for each pixel along with the shift registers and output amplifier.

controlled by the mounting system and is tuned to the expected observing conditions. Most far infrared detection systems are differential systems that compare the radiation from a source with the signal from a nearby piece of sky by moving the telescope between the two areas. The frequency of this movement or nodding will set the frequency requirements of the bolometer.

Bolometers have been produced in small arrays by simply mechanically placing them at the foci of an array of feed horns designed to channel far infrared radiation to them. Since bolometers are not as sensitive as other detectors and they are difficult to produce in large arrays, they are generally only used for far infrared astronomy where other detectors are not sensitive. They are operated at temperatures of 1 to 2 K or below in dewars containing liquid helium.

C. Submillimeter Astronomy

Submillimeter astronomy occupies the boundary between radio and infrared astronomy. As its name implies, the submillimeter wavelength region is at wavelengths less than 1 mm to about 0.3 mm or 300 μm. Most current submillimeter observations are carried out with special high-precision radio antennas that have smoother and more accurate shapes than antennas used at longer wavelengths. Detector systems for submillimeter observations can be either very high-frequency radio detection systems or bolometer-based systems such as described in the previous section. Observations in this wavelength region, particularly at 850 μm, have been very important for detecting star formation in galaxies that has been hidden by large amounts of dust. This work has been pioneered by the submillimeter common user bolometer array (SCUBA) on the James Clerk Maxwell Telescope. This telescope is on the summit of Mauna Kea, Hawaii, one of the highest and driest sites in the world. Even in the windows of transmission for submillimeter radiation the attenuation of signal by water vapor in the earth's atmosphere is quite high. For this reason there are only a few sites around the world that attempt ground-based submillimeter astronomy.

IV. INFRARED ASTRONOMICAL INSTRUMENTS

The existence of excellent infrared detectors solves only part of the problem of producing excellent infrared astronomical observations. The other part is producing high-quality infrared instruments to place the image or spectrum on the array. Although there are many types of instruments for measuring aspects of the radiation, such as polarization, we will concentrate in this chapter on two types of instruments, cameras and spectrometers.

A. Cameras

Infrared astronomical cameras relay the image formed by the telescope onto the detector array with the proper magnification for the desired observation. In this way they are no different from the cameras used in optical observations. The main difference in infrared cameras comes in the baffling and pupil stop requirements. Baffling is used to prevent stray light from contaminating the image, and pupil stops ensure that only light coming from the direction of the primary mirror falls on the detector. The problem with infrared cameras is that at the temperatures encountered in ground-based telescopes the baffles and pupil stops themselves become sources of contaminating infrared radiation. This problem is solved by placing the baffles and pupil stops inside the cold dewar along with the detector. Infrared cameras then have the requirement that a new pupil and image must be created inside the cold environment. A very simple schematic of this arrangement is shown in Fig. 7.

This arrangement refocuses the telescope focal plane onto the detector array to produce the image at the desired magnification. It also places an image of the telescope's primary mirror at the location of the cold pupil, where a cold stop masks out the warm areas around the primary mirror such as the mirror cell. The cold stop should also mask out the primary hole and perhaps the spiders that hold the secondary as they are also sources of stray infrared radiation. The transparent window is required to transmit the light into the dewar while maintaining the vacuum inside the dewar. The window itself is sometimes optically figured to be part of the reimaging system. The reimager is shown as a simple lens, but in practice infrared optics are generally mirrors rather lenses. This avoids any chromatic effects and utilizes the high reflectivities that can be achieved at infrared wavelengths. Since the background radiation is often far more intense than the observed source, careful baffling and placement of pupil stops is essential for good performance. Most cameras utilize a set of filters to take images in different wavelength bands. These filters must also be cold so that they do not radiate energy in the wavelengths they were designed to block. If mechanically possible, the best location for the filters is at the cold pupil spot. When located at the pupil, any spatial variations in the filters due to optical flaws or other reasons get spread out over the whole detector array rather than being imaged onto the detector, where they can be misinterpreted as spatial features in the image.

B. Spectrometers

Modern infrared spectrometers are almost universally dispersive spectrometers utilizing a dispersive element, such as a grating, grism, or prism. A *grism* is simply a grating and prism combined. Before the advent of good quality detector arrays, multiplexing instruments such as Fourier transform spectrometers (FTSs) were often used since they could observe spectra more efficiently with single detectors than scanning dispersive spectrometers that had to measure spectra one element at a time. Once arrays became available, dispersive spectrometers became the instrument of choice because the amount of light falling on a single pixel is much less than in an FTS. Since the

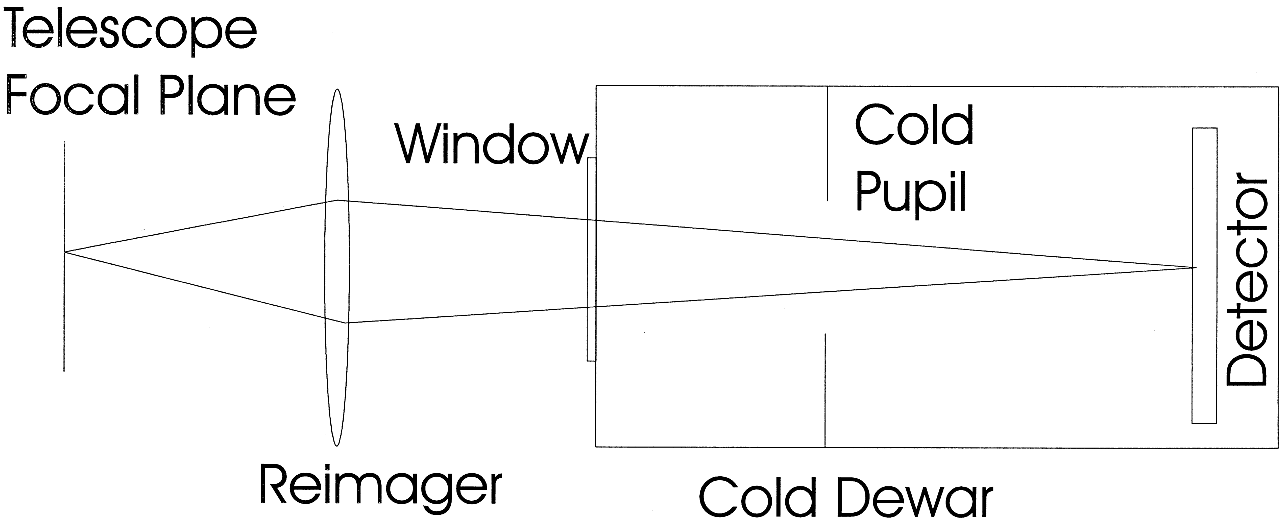


FIGURE 7 A simplified optical arrangement for an infrared camera.

photon noise is proportional to the number of photons detected, the dispersive spectrometer has a significantly better signal-to-noise ratio than a multiplexer such as an FTS. There are some applications where FTS systems do have some advantages. An example is very high-resolution solar astronomy where the signal levels are so high that signal-to-noise considerations are secondary. In this case the high-wavelength precision of an FTS is an advantage. In some specialized cases, Fabry–Perot spectrometers have been used. These spectrometers are useful for mapping out emission lines over a large area. A typical use might be finding the velocity structure of molecular hydrogen emission in a star-forming region.

As with infrared cameras, proper baffling and pupil stops are essential. Besides the dispersive element, the main difference in spectroscopic instruments is the requirement that the beam be well collimated when it encounters the disperser. Spectroscopic systems therefore generally have two main optical components. First there is a collimator that produces a collimated beam for the disperser and then a camera that focuses the dispersed light onto the detector array.

V. GROUND-BASED INFRARED ASTRONOMY

The vast majority of infrared astronomical observations have been made with ground-based telescopes. In general these are telescopes that have been primarily designed for observations at optical wavelength, although more recently there have been telescopes built that are optimized for infrared observations. Most of the achievements listed in section II came from ground-based observations. In this section we will discuss some of the unique aspects of ground-based infrared observations.

A. Limitations Due to the Earth's Atmosphere

Infrared observations from the ground encounter several obstacles that are either absent or far less of a problem at optical wavelengths. One of the major problems is blockage of a large portion of the infrared spectrum by the earth's atmosphere. As discussed in section I, the positions of the infrared bands are set by the transparent or semitransparent wavelength regions in the atmosphere. Figure 8 shows the transparency of the earth's atmosphere at both the altitude of the ground-based observatory on Mauna Kea mountain in Hawaii and from the altitude of a high-flying aircraft.

The atmospheric constituent that contributes the most absorption is water vapor. For this reason good infrared observatories are located in high dry sites. Other atmospheric gases such as carbon dioxide and methane also block some wavelengths. There is no way around these blockages except to move the telescope to much higher regions with airplanes, high-altitude balloons, or space craft.

At near infrared wavelengths, out to 2 μm, there is a severe problem from emission from the OH radical in our atmosphere. OH radicals in an excited state are produced by the chemical reaction $O_3 + H \rightarrow O_2 + \mathrm{OH}$ in the high earth's atmosphere. The majority of this emission comes from altitudes much higher than the highest mountains, so it cannot be overcome by site location. The emission is spectral line emission from the transitions of the OH molecule, but the lines are very closely spaced so only moderately high-resolution spectroscopy can see regions between the emission lines. The emission is also time and spatially variable, which complicates accurate subtraction of the emission from the observation. The sky brightness from OH emissions overwhelms all but the brightest sources at this wavelength.

At longer wavelengths, thermal emission from the atmosphere becomes a problem. Even in the infrared observational bands, the atmosphere is not completely transparent. This means that the atmosphere emits infrared radiation that adds noise to the observations. Most observations where this is a problem compare images or spectra of the desired source with images or spectra of a nearby region of sky off the source to remove the background sky contribution. This method, however, cannot remove the noise due to the photon statistics of the background emission. This noise is generally the largest contributor to the total noise of the measurement.

Another major problem is thermal emission from the telescope itself and the instrument used to measure the infrared signal. Objects at room temperature emit radiation throughout the infrared region, with a peak at 10 μm. The infrared detectors themselves are of course cooled to very low temperatures. The infrared instrument itself is often also cooled with the optical system contained inside a evacuated chamber called a dewar. The dewar is kept cold by adding a cryogenic fluid such as liquid nitrogen or liquid helium or by mechanical refrigerators. The telescope itself cannot be cooled. If it were cooled then atmospheric water vapor would condense on the mirrors and eventually ice over if the temperature was lowered far enough. Infrared instruments compensate for the thermal emission from the telescope by using optics that limit the detectors field of view to only the low emissivity telescope mirrors that reflect the observed radiation into the instrument.

B. New Large Telescopes

Infrared astronomy has received a substantial boost with the advent of several new 8-m diameter or larger ground-based telescopes at high dry sites. These telescopes have

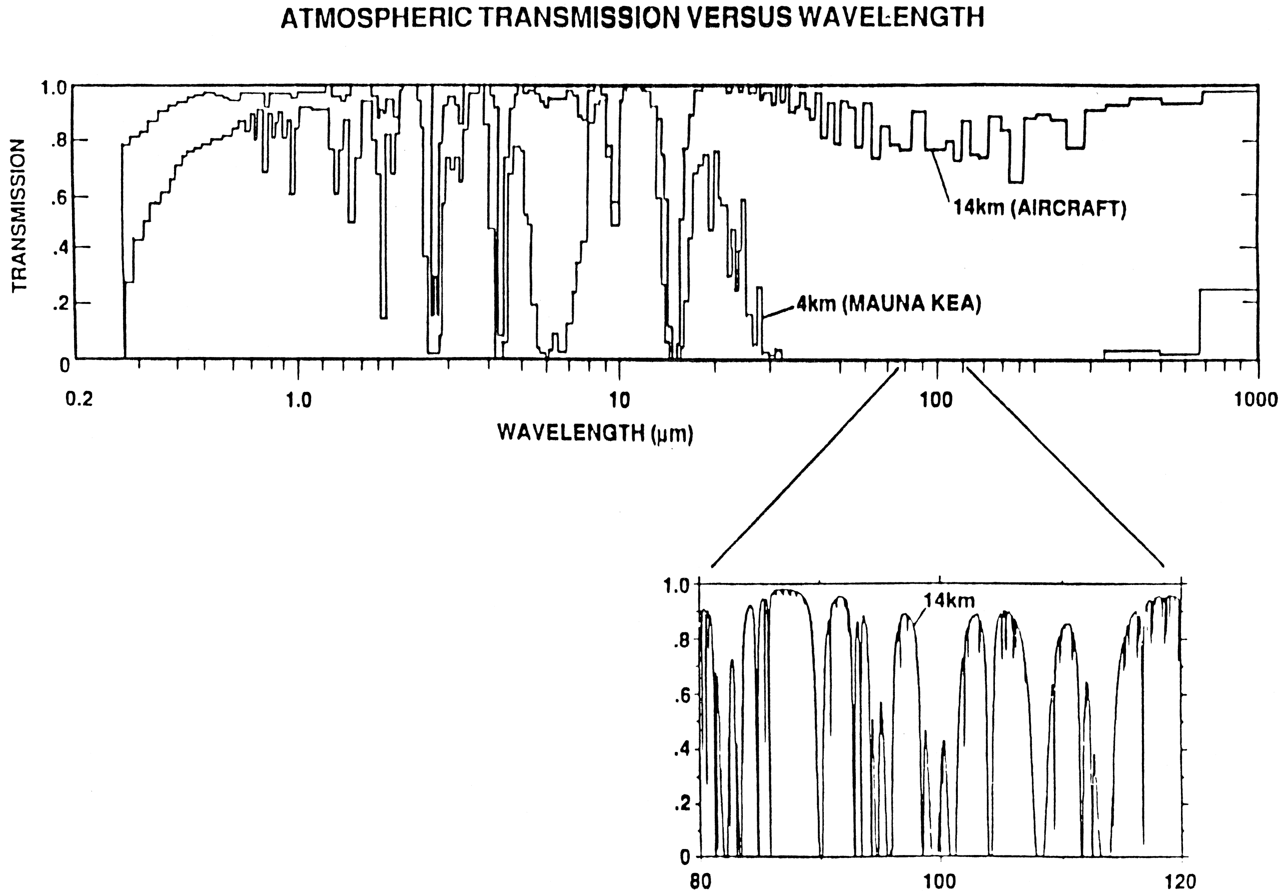


FIGURE 8 The transparency of the earth's atmosphere from a ground-based observatory and from a high-altitude aircraft by E. Erickson.
Source: Ericksen, E. F., SOFIA: Stratospheric observatory for infrared astronomy in next generation infrared space observatory, *in* "Next Generation Infrared Space Observatory" (B. Burnel, Ed.), pp. 62. Kluwer Academic Publishers, The Netherlands (1992) with kind permission from Kluwer Academic Publishers.

more than four times the light-gathering power than the previous 4-m class telescopes. When coupled with modern detector arrays they are advancing the sensitivity of infrared observations. One area of infrared astronomy where the new telescopes will be particularly effective is infrared spectroscopy. With the new low-noise and low-dark-current detectors, spectroscopic observations are generally limited by the statistical noise of the gathered photons even at moderate resolution. Due to Poisson statistics, photon noise is equal to the square root of the number of photons detected. In this case the signal-to-noise of the observation only increases as the square root of the observing time. This means that infrared spectroscopic observations on an 8-m telescope proceed 16 times faster than on a 4-m telescope. This is an extremely important advantage in executing follow-up spectroscopic observations on sources originally imaged by the NICMOS instrument and to be observed by the SIRTF instrument described in the section on space infrared astronomy.

C. Adaptive Optics

Recently there have been significant advances in correcting for the distortions caused by atmospheric seeing in ground-based telescopes. The basic technique is to measure the motion of the image or wave front with one optical system and then to adjust the optics in the imaging system to counteract the distortion. This can be done in various ways, but the technique in general is termed *adaptive optics*. The great significance to infrared astronomy of this technique is that the correction becomes progressively easier as the wavelength increases. There has been great success with this technique for images in the K band at 2.2 μm.

The adaptive optics correction depends on having a bright point source near the field of view to provide the information on the atmospheric distortion to the system. This source can be a bright astronomical source or an artificial point source created by a laser beam. The requirement for proximity is such that the measurement of the distortion occurs in the same region of the earth's atmosphere that the light from the field of view passes through. The characteristic distance in the atmosphere that has common distortion is referred to as *Fried's parameter*. It is this parameter that increases with wavelength, making correction easier at longer wavelengths and increases the field of view that can be corrected.

One drawback for infrared astronomy has been that the image motion or wave-front detection systems have generally operated on visible light. This has limited the technique's usefulness in regions highly obscured by dust, where bright visible light sources are hard to find. It also means that the image corrections are not done at the same wavelength as the image. This can cause loss of accuracy due to the differences in the indices of refraction between the visible and optical wavelengths. Use of infrared arrays in the image-motion systems can eliminate this problem.

The big gain of adaptive optics systems is that they can concentrate a significant amount of light into the diffraction-limited point-spread function (PSF) of a large telescope. This is very important for resolving objects such as close binaries or the nuclei of merging galaxies. It is also very important for infrared spectroscopy, where the size of the spectrometer is proportional to the size of the image.

The distribution of light in an image of a point object, such as a star, is called the PSF. The diffraction-limited PSF is a sharply peaked image that exploits the full diameter of the telescope. The seeing-limited PSF is the spread-out image caused by the distortions in the earth's atmosphere. The ratio of the light that is in the diffraction-limited peak of the PSF to the theoretical maximum peak is called the Strehl ratio. Excellent adaptive optics systems can achieve a Strehl ratio of about 60%, which leaves 40% of the light in the seeing-limited PSF. This spread-out light can cause problems in some types of observations that call for very high accuracy. That is one reason why space infrared telescope systems and adaptive optic systems complement each other.

D. South Pole Observatory

The South Pole is a unique location for astronomy and infrared astronomy in particular. Both pole regions have long periods of darkness and the opportunity to track celestial objects continuously. Only the South Pole, however, is on solid land, which allows a stable observatory to be established. Even more importantly the South Pole is quite high. The combination of the high altitude and extremely cold temperatures reduces the water vapor to very low levels. This means that the South Pole has the lowest attenuation of infrared and particularly of submillimeter radiation of any earth-based observatory. There are, of course, some environmental drawbacks for all but the heartiest of observers.

VI. SPACE INFRARED ASTRONOMY

There have been relatively few infrared space missions, partly because, for wavelengths longer than about 2 μm, the whole telescope must be cooled to very low temperatures to reduce infrared emission from the telescope itself. This is quite costly and requires the use of cryogens that limit the lifetime of the instrument or high-power consumption to provide refrigeration through mechanical or thermoelectric coolers. Techniques for passive radiative cooling are under development. They require large sunshades and orbits other than low-earth orbits, where blocking both the solar and terrestrial infrared emission is very difficult.

The advantages of doing infrared astronomy in space are quite large. The absence of the earth's atmosphere eliminates the image distortions, termed *seeing*, that plagues ground-based observations. It also eliminates the infrared emission from the atmosphere, which limits the sensitivity of ground-based telescopes. The opportunity of cooling the entire telescope, although costly as mentioned above, again greatly improves the sensitivity of space-based telescopes. Finally, since many regions of the infrared spectral range are blocked by absorption in the earth's atmosphere, space missions offer the only way to observe the entire infrared spectral range with no contamination by telluric absorption features.

A. Limitations of Space-Based Observations

Although vastly superior to ground-based observations, infrared observations from space are not completely background free. The primary source of infrared background radiation in space is reflection and emission by zodiacal dust. This is dust in our solar system that is distributed mainly in plane of the planets or zodiac. As shown in Fig. 9, there are two minima in the background radiation, one at 3 μm and another at 400 μm.

The minimum at 3 μm is at the crossover point between reflected and emitted zodiacal radiation. At shorter wavelengths the emission is dominated by reflected sunlight. At wavelengths longer than 3 μm, the main emission is thermal emission from the dust. The minimum at 400 μm is due to the fall off of the warmer emission from the zodiacal dust and extrasolar dust, termed *cirrus*, and the short wavelength falloff of the cosmic microwave background, which has a much colder temperature. These background minima are very useful regions for the observation of faint sources such as distant galaxies.

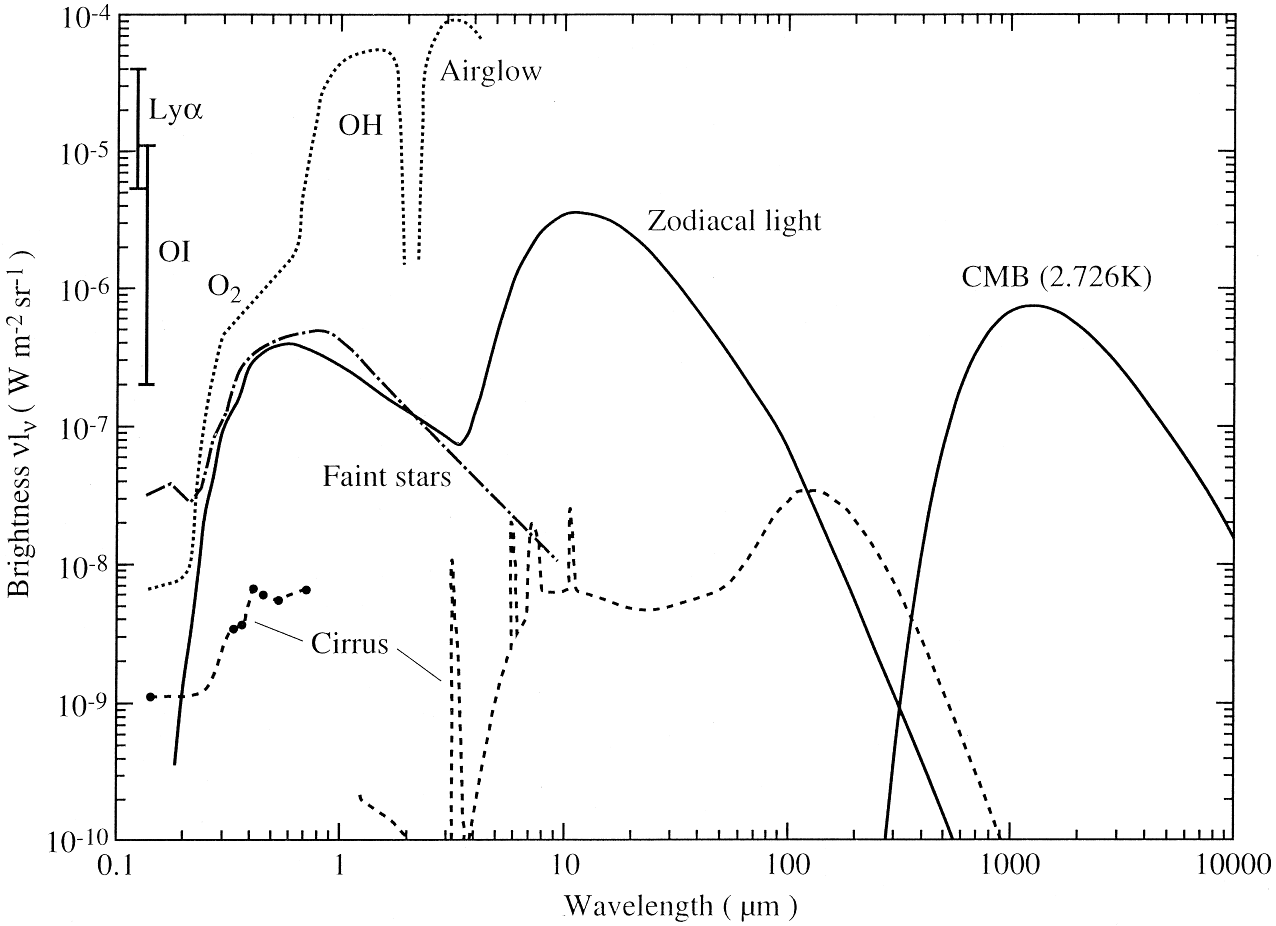


FIGURE 9 The emission spectrum as seen from space by Leinert et al. The O_2 and OH airglow are terrestrial emissions and are not seen from space.
Source: From Leinert, Ch. *et al. Astron. Astrophys. Suppl.* **127,** 1 (1998).

Although there have been few space infrared missions, the missions carried out to date have proved quite successful. Due to the advantages of space infrared observations, there are also several missions in the planning stage. We will review a few of the major missions here.

B. InfraRed Astronomical Satellite

The IRAS was the first orbital space mission devoted solely to observations at infrared wavelengths. The advantage of space observation are clearly demonstrated by noting that IRAS, with a diameter of only 60 cm, vastly outperformed ground-based observations with even the 5-m Mt. Palomar observatory, the largest major observatory at the time. IRAS, launched January 1983, surveyed the sky for a little less than a year. Observations of approximately 96% of the sky were made in broad photometric bands at 12, 25, 60 and 100 μm, with an array of discrete Ge detectors. Its nearly polar orbit meant that in the survey mode it covered the sky in strips that overlapped at the equator but had a significantly greater time of coverage near the poles. IRAS also operated part of the time in a pointed set of observations on objects or regions of particular interest.

The IRAS focal plane (Fig. 10) contained column detectors arranged so that the centers of the detectors were offset from each other by a little less than the length of the detectors. The detectors were rectangles with the long dimension parallel to the column line, which was perpendicular to the scan direction. Each column had two lines of detectors and a filter for one of the four wavelength bands. As the satellite scanned along the sky, sources would pass over the detectors producing a transient signal that repeated in each detector at a rate set by the rate of scan through the sky. Since the detector centers were offset, a source would pass through different parts of each detector as it ran down the column. Keeping track of the signal strengths and which detectors responded as different strips of sky were observed gave additional spatial information in the direction perpendicular to the scan. IRAS also contained a low spectral resolution spectrometer that

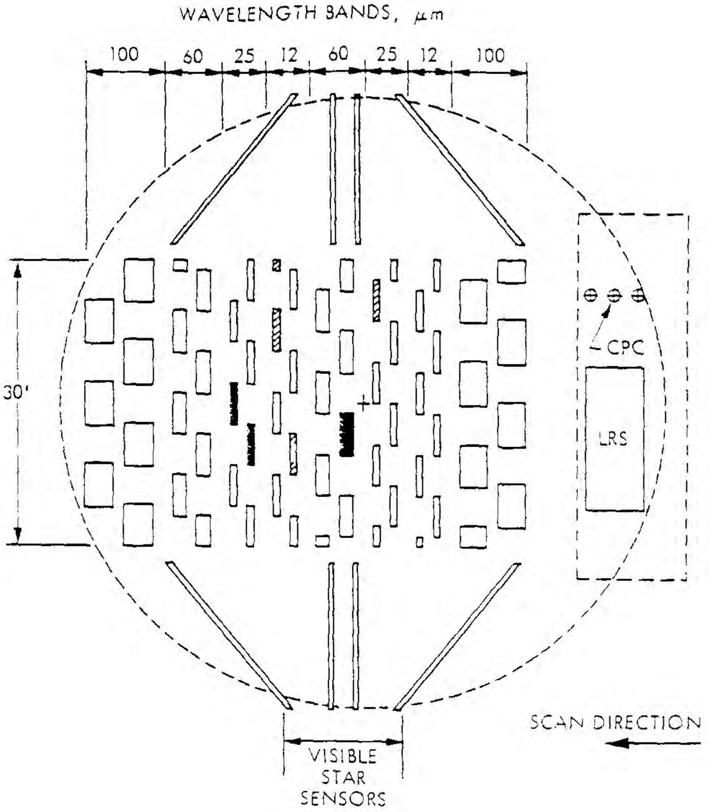


FIGURE 10 Above is a schematic of the InfraRed Astronomical Satellite focal plane. The different-sized rectangles represent the detectors.

operated in the 7.7- to 22.6-μm region. This spectrometer measured over 5000 sources with signal strengths greater than 10 Janskys.

IRAS established a then unprecedented database that is still being utilized at the present time. Catalogs of both extended and point sources were published from the survey. A very significant finding from these catalogs was the realization that many galaxies have the majority of their luminosity at mid- and far infrared wavelengths. IRAS also established that some stars that had been considered quite normal, such as Vega, were actually surrounded by large clouds of dust that might be the building blocks for planetary systems.

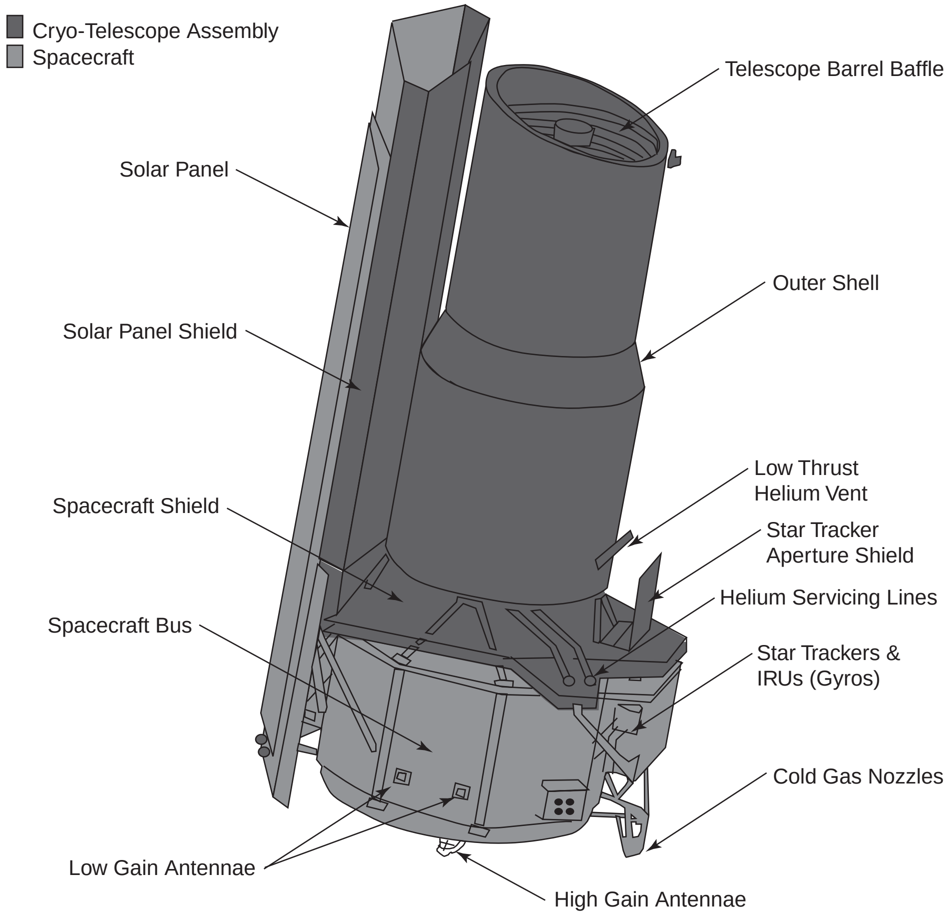
C. Cosmic Background Explorer

COBE has produced some of the most dramatic scientific achievements in infrared astronomy. COBE measured the spectral shape of the cosmic background and found it to be a perfect blackbody spectrum at a temperature of 2.37 K. COBE also made the first observations of structure in the distribution of cosmic background radiation that are the probable first step in changes that produced the galaxies, stars, and planets we know today from the primordial smooth distribution of matter produced by the Big Bang. (See the discussion in section II.A.5.) COBE had three instruments for observing the background radiation: the DMR, the FIRAS, and the diffuse infrared background experiment (DIRBE).

The DMR experiment utilized techniques more common to radio astronomy than infrared. The main part of the instrument consisted of sets of opposed antennas looking at different parts of the sky. These antennas looked for differential signals indicating different temperatures at different locations. FIRAS was a version of the FTS discussed earlier. It compared the spectral shape of the cosmic background radiation against the spectral shape of internal blackbody calibrators. The DIRBE instrument was more similar to standard infrared instruments than the other two COBE experiments. DIRBE used a suite of four different detector types to map out the sky at several infrared wavelengths. This instrument has contributed very important information on the contribution of various types of astronomical objects to the total background radiation in the universe.

D. Infrared Space Observatory

The ISO was launched in November 1995. Its various instruments operated over the wavelength range between 2.5



Observatory external configuration

FIGURE 11 Configuration of the Space InfraRed Telescope Facility spacecraft.

and 240 μm. With a primary mirror diameter of 60 cm, it was similar in size to IRAS but carried improved detectors and a more versatile instrument complement. ISO differed in its mission from IRAS. Instead of surveys as its main mission, ISO was designed primarily for pointed observations of objects of interest. Like COBE and IRAS, ISO was cooled by cryogens in order to operate in the mid- and far infrared bands. The liquid helium cryogens lasted until April 1998, providing an observing period of about $2\frac{1}{2}$ years. ISO was put into a highly elliptical orbit that provided significant observing time at large distances from the earth with data transmission to two ground stations. ISO was the first infrared space mission to offer observing opportunities to the entire community. The four instruments on the ISO mission were a combination of cameras and spectrometers described below. This mission also provided important information about the performance of some classes of infrared detectors in high radiation environments. In spite of the problems with some of the detectors, ISO was a highly successful mission whose database is an important tool in astrophysics. Its spectrometers demonstrated the richness of the mid- and far infrared spectral region. We will discuss below the ISO instruments.

FIGURE 12 Artist's Concept of the Stratospheric Observatory for Infrared Astronomy.

1. ISOPHOT

This instrument provided photometric, polarimetric, and spectrophotometry over the entire wavelength range of ISO. Small arrays of Si:Ga, Si:B, and Ge:Ga detectors also provided limited imaging capabilities. The main role of ISOPHOT was to provide accurate photometry of sources. It included an internal chopper and several calibration sources.

2. ISOCAM

ISOCAM provided the main imaging capability for the mission. It was split into two channels. The short wavelength channel was sensitive between 2.5 and 5.5 μm, and the long wavelength channel between 4 and 18 μm. These cameras provided imaging capability in several spectral regions that are inaccessible from the ground. The detector arrays were 32×32 pixels of In:Sb and Si:Ga. The short wavelength In:Sb array was operated in a charge-integrating mode, which has since been superseded by multiplexed readouts for much larger arrays. The long wavelength Si:Ga array was operated as a photoconductor. ISOCAM provided diffraction-limited optical performance with several filter options.

3. Short Wavelength Spectrometer

The short wavelength spectrometer (SWS) covered the spectral range between 2.38 and 42.5 μm, with a spectral resolution ranging from 1000 to 2000. It also carried a Fabry-Perot etalon to enhance the spectral resolution in the 11.4 to 44.5-μm region. Fabry-Perot etalons pass radiation in a narrow wavelength range that is altered by changing the spacing between the optical components. A combination of In:Sb, Si:Ga, Si:As, Si:Sb, and Ge:Be linear arrays provided the detectors for the large wavelength range covered by the instrument. Most of the arrays were 1×12 pixels, but the Si:Sb and Ge:Be arrays were 1×2. The SWS detector arrays were found to be very sensitive to the radiation environment encountered in space missions.

4. Long Wavelength Spectrometer

The long wavelength spectrometer (LWS) operates between 43.0 and 196.9 μm. Coupled with the SWS it provides continuous spectral coverage from 2.4 to 196.9 μm. This has been a great advantage in studying the emission of objects such as active galactic nuclei and starburst galaxies. The detector array is linear and consists of one Ge:Be, five unstressed Ge:Ga, and four stressed Ge:Ga photoconductive detectors. Like SWS, LWS also carried a Fabry-Perot etalon to increase the spectroscopic resolution of the instrument.

E. Near Infrared Camera Multiobject Spectrometer on the Hubble Space Telescope

The NICMOS is an instrument on board the HST. Although HST is the largest current telescope in space, with a 2.4-m primary mirror, it is not a cooled telescope. Its mirrors are maintained at a temperature of 18°C or about room temperature. For this reason, NICMOS limits its observations to wavelengths shorter than 2.5 μm. Observations at wavelengths longer than 2 μm are somewhat degraded by thermal emission from the telescope's mirrors. At wavelengths shorter than 2 μm, the NICMOS sensitivity is greater than the largest ground-based telescopes due to the absence of the bright OH background (discussed in section IV). The absence of the earth's atmosphere also means that NICMOS can achieve very high spatial resolution over its entire field of view. This advantage is also enjoyed by the optical instruments on HST. NICMOS has three separate cameras for observations at various spatial resolutions. It also has the capability of performing polarimetric and coronagraphic imaging and low-resolution spectroscopy.

The NICMOS detectors were cooled with a large block of solid nitrogen at 65 K, which sublimated away at the end of 1998. Revitalization of NICMOS with a mechanical cooling system is scheduled for late 2001. If this is successful, NICMOS can be operational for the remaining lifetime of HST. Sensitivity curves for all HST instruments are maintained by the Space Telescope Science Institute and are available on their web site: http://www.stsci.edu

F. Space InfraRed Telescope Facility

The SIRTF is a 85-cm diameter telescope (Fig. 11) that will be cooled to 5.5 K and specialize in mid- and far infrared observations. Scheduled to be launched on a Delta rocket in December of 2001, SIRTF will provide an excellent follow-up on the ISO observations and offer unique areas of science of its own. At this time SIRTF is expected to provide a significant increase in sensitivity and field of view over ISO, primarily due to an improved detector complement. The SIRTF orbit is a unique earth-trailing heliocentric orbit that will place the spacecraft at a considerable distance from Earth. This will facilitate the cooling of the telescope, which will be launched warm and then cooled once the proper orbital position is achieved. SIRTF contains three focal plane instruments described below.

1. Multiband Imaging Photometer

The MIPS for SIRTF will provide diffraction-limited imaging over the wavelengths between 20 and 180 μm. This range is covered by three detector arrays optimized for photometric bands centered at 24, 70, and 160 μm. The respective detector arrays are a 128×128 pixel Si:As BIB array, a 32×32 pixel gallium-doped Germanium (Ge:Ga) photoconductor, and a 2×20 pixel Ge:Ga array that has been mechanically stressed to extend its long wavelength coverage to 180 μm. MIPS also has low-resolution ($R = 14$–25) spectroscopic capabilities in the wavelength region between 52 and 99 μm. R is a measure of resolution and is given by $\lambda/\delta\lambda$, where λ is the wavelength and $\delta\lambda$ is the spectral resolution of two pixels in the Nyquist sampled case. A cryogenic scan mirror system will increase the efficiency of MIPS for mapping regions of sky larger than its field of view.

2. InfraRed Array Camera

The InfraRed Array Camera (IRAC) provides imaging at shorter wavelengths with bands centered at 3.6, 4.5, 5.8 and 8.0 μm. The 3.6 and 4.5 μm bands utilize two 256×256 pixel In:Sb detector arrays while the two longer bands utilize 256×256 pixel Si:As impurity band conduction detectors. Although there are four photometric bands IRAC has only two entrance apertures. Dichroic beam splitters allow the 3.6 and 5.8 band to share one aperture while the 4.5 and 8.0 bands share the other. This doubles the efficiency for multiband observations.

3. InfraRed Spectrograph

The InfraRed Spectrograph (IRS) provides the primary spectroscopic capability for SIRTF. It covers the wavelength range of 5.3 to 40.0 μm with low-resolution ($R = 60$–120) spectroscopy and the range between 10 and 37 μm at higher resolution ($R = 600$). IRS utilizes 128×128 Si:As and Si:Sb BIB arrays to cover its wavelength region.

VII. HIGH-ALTITUDE INFRARED ASTRONOMY

As we have seen above, infrared space missions are quite elaborate and expensive, both in time and in money. Throughout the history of infrared astronomy there has been a quest to gain some of the benefits of space missions but with less expense and with more frequent access than provided by space missions. As discussed earlier, the best observatory sites are on very high, very dry mountains, but this still leaves a large part of the infrared spectrum blocked from ground viewing. The obvious alternative has been observations from aircraft and high-altitude balloons. We will discuss some of these observations below.

A. Aircraft

The first aircraft to be used routinely for infrared observations was a Lear jet operated by NASA from the Ames Research Center. It was equipped with a 12-inch telescope that looked out from the side of the aircraft. The aircraft was modified to accommodate the telescope and for high-altitude performance, with flights typically reaching 50,000 feet. It could accommodate two observers and a pilot. With this platform observations of roughly 1 hr could be taken at high altitudes.

1. The Kuiper Airborne Observatory

The first large aircraft devoted solely to astronomical observations was the Kuiper Airborne Observatory (KAO). This was a C141 jet transport with a 36-inch telescope again looking out the side of the aircraft. The observatory could operate for roughly 8 hr at an altitude of 45,000 ft. This high altitude was needed to be above the tropopause, which is the boundary in earth's atmosphere between water-rich and water-poor air. Before IRAS and ISO, most of the available far infrared measurements were made with this aircraft. Unlike satellites, the KAO offered a platform for a large number of different types of instruments so that a wide range of studies could be made. The KAO was taken out of operation in 1995 to provide funding for the successor aircraft, the Stratospheric Observatory for Infrared Astronomy (SOFIA).

2. SOFIA

The SOFIA is a 747 SP modified to accommodate a side-looking 2.5-m infrared telescope (Fig. 12). The larger telescope will increase both the sensitivity and spatial resolution of the observations. SOFIA offers a combination of facility-maintained and observer-provided sets of instruments for observations. The location of the telescope opening section aft of the wing is unusual for aircraft observatories but was necessitated by the structure of the aircraft.

B. High-Altitude Balloons

As a less expensive, but less reliable, platform for infrared observations, high-altitude balloons have played an important role. Balloons offer much higher altitudes than aircraft and can provide observing times as long as several days under favorable conditions. The observations must be either preprogrammed or remotely controlled. Attitude control is somewhat harder with a balloon platform, and the descent phase can be hard on the equipment. Modern balloon astronomy flights are unmanned. An early NASA guide on the various methods of high-altitude observation listed one of the drawbacks of high-altitude manned balloon observations as "usually results in the death of the observer." Balloon flights are expected to play an important role in the mapping of the cosmic background radiation. The Maxima and Boomerang flights discussed in section II are just the beginning of these efforts.

VIII. SURVEYS

Up until the late 1990s, the only large-scale survey of the sky at near infrared wavelengths was at 2.2 μm and was complete over part of the sky to a limiting magnitude of $+3$. The IRAS survey went much deeper at longer wavelengths but was limited to a spatial resolution of about 1 min of arc. Recognizing the need for much deeper and more complete surveys, several groups have planned and are carrying out extensive surveys of the sky.

A. NASA Catalog of Infrared Observations

Although technically not a survey, the NASA Goddard Space Flight Center maintains a catalog of infrared observations that are not part of a massive survey such as the ones listed below. It contains a database of over 374,000 published infrared observations of more than 62,000 individual astronomical sources. It includes observations made between 1 and 1000 μm. Up until 1993, this heroic effort was published in a paper version with the last edition published as NASA Reference Publication RP-1294. Now the database is electronically available. The URL at the time of publication is http://ircatalog.gsfc.nasa.gov. This database is invaluable for anyone who would like to find the infrared characteristics of an object. A complete bibliography is part of the searchable database.

B. 2Mass

The 2Mass survey utilizes two 1.3-m telescopes in the Northern and Southern Hemispheres to survey the sky in the three near infrared bands of J, H, and K_s, a special K band filter centered at 2.17 μm. It will survey the entire sky with a spatial resolution of 2.0 arc sec. The limiting flux level in each band is about 1mJy, with a signal-to-noise ratio of 10. It uses the NICMOS detectors developed for the NICMOS instrument on HST (see section VI.E). The data from the 2-Mass Survey will be available electronically in three separate catalogs. The first is a digital atlas of the sky at a resolution of 4 arc sec. The second is point-source catalog of all unresolved sources, and the third is an extended source catalog of all resolved objects. The point-source catalog is expected to have about 300 million objects and the extended source catalog over 1 million objects. It has already contributed to the identification of a new stellar classification of stars fainter and redder than M dwarfs. The new classification has been designated L dwarfs.

C. Deep Near Infrared Southern Sky Survey

The Deep Near Infrared Southern Sky (DENIS) survey covers the southern sky in the I, J, and K bands. The I band is actually just shortward of 1 μm. The survey will cover the declinations between -88° and $+2^\circ$ with a 1-m telescope. A NICMOS detector array is used to image the J and K bands, and a Tektronix CCD detector is used for the I band. The resolution is 1 arc sec, and the limiting magnitudes at I, J and K are 18.0, 16.0, and 13.5.

D. Other Infrared Space and Ground-Based Surveys

Both the IRAS and COBE missions completed essentially all sky surveys at several wavelengths that were discussed in the sections devoted to the missions. That data combined with the near infrared ground-based surveys discussed above will provide a very valuable database.

There also exist limited-area surveys from both the ground and space. One of the best surveyed areas is the northern and southern Hubble Deep Fields discussed in section III.A.4. There are also surveys with ISOPHOT on ISO, termed FIRBACK, looking for light from highly luminous obscured galaxies. With the common usage of improved infrared detector arrays we expect these surveys to greatly increase in number.

IX. FUTURE

A. Large Ground-Based Telescopes and Adaptive Optics

In the next decade there will be several 8-m diameter or larger ground-based telescopes coming on-line. At the same time the availability of large-format detector arrays at even midinfrared wavelengths will greatly increase. These two factors alone will significantly improve the sensitivity and power of infrared observations, where power is defined as a product of sensitivity times field of view. The increase in sensitivity will be particularly large for spectroscopy, where the sky brightness is greatly reduced due to the dispersion of the light over many pixels. At the same time the push to observe the universe at increasingly larger redshifts will make infrared imaging and spectroscopy essential for conducting cosmological studies.

Also in the next decade the number and quality of adaptive optics systems will increase. Adaptive optic systems are planned or in use at both Keck telescopes, the VLT array of four telescopes, and the large binocular telescope (LBT). Infrared images at the diffraction limit of the large telescopes will be possible over almost all of the infrared spectral range. The implementation of laser-created point sources will also increase the area of sky available for adaptive optics imaging. In spectroscopic systems, adaptive optic systems will allow spectrometers designed for point sources to be much smaller due to the smaller input slit size provided by adaptive optics.

B. Next-Generation Space Telescope

The Next-Generation Space Telescope (NGST) will be an exceedingly powerful tool for infrared observation. Its planned mirror diameter is 8 m in contrast to HST's 2.4-m diameter. This alone will provide a large increase in sensitivity. Even more important than the increase in diameter is the expected very low temperature of the telescope. The telescope will be placed at the second Earth–Sun Lagrange point, which is quite distant from the earth compared to the low-earth orbit of HST. At that distance a large shade will block the radiation from both and the sun which allows the telescope to radiatively cool to a temperature of about 60 K. This immensely reduces the infrared emission from the telescope and thus greatly improves the sensitivity of the observations. At the same time the larger diameter of the telescope will make the images almost four times sharper than the already excellent HST images.

The main purpose of NGST will be the exploitation of the minimum of the sky background at a wavelength of 3.0 μm to make observations of very distant and very faint galaxies for cosmological purposes. The primary wavelength range of operation will be from 1.0 to 5.0 μm. Extensions to longer infrared wavelengths are under consideration, however. Instrumentation for the telescope is expected to include both cameras for imaging and spectrometers for spectral analysis.

Construction of NGST will be challenging. At present we do not have a launch vehicle that is capable of carrying a payload with an 8-m diameter. Unless such a launch vehicle becomes available, the primary mirror will have to be made in sections that are folded into place after the telescope is launched. This requires exquisite positioning accuracy to achieve the precision that maintains the quality of an diffraction-limited 8-m telescope.

ACKNOWLEDGMENTS

Rodger I. Thompson is supported in part by NASA grant NAG 5-3042 and the University of Arizona.

SEE ALSO THE FOLLOWING ARTICLES

COSMIC RADIATION • GALACTIC STRUCTURE AND EVOLUTION • INFRARED SPECTROSCOPY • INTERSTELLAR MATTER • PLANETARY SATELLITES, NATURAL • SOLAR SYSTEM, GENERAL • STELLAR SPECTROSCOPY • STELLAR STRUCTURE AND EVOLUTION • ULTRAVIOLET SPACE ASTRONOMY

BIBLIOGRAPHY

Allen, D. A. (1975). "Infrared, the New Astronomy," John Wiley & Sons, New York.

Bicay, M. D., Beichman, C. A., Cutri, R. M., and Madore, B. F. (eds.) (1999). "Astrophysics with Infrared Arrays: A Prelude to SIRTF," Astronomical Society of the Pacific, San Francisco.

Glass, I. S. (1999)."Handbook of Infrared Astronomy," Cambridge University Press, Cambridge.

Mather, J. C, and Boslough, J. (1996). "The Very First Light," Basic Books, New York.

McLean, I. S. (ed.) (1994). "Infrared Astronomy with Arrays: The Next Generation," Kluwer Academic Publishers, Dordrecht.

Minnitti, D., and Rix, H.-W. (eds.) (1996)."Spiral Galaxies in the Near-IR," Springer, Berlin.

Smoot, G., and Keay, D. (1993)."Wrinkles in Time," William Morrow and Company, Inc., New York.

Magnetic Fields in Astrophysics

Steven N. Shore
Indiana University South Bend

GLOSSARY

Ambipolar diffusion Generation of drift currents, due to differences in mass between ions and electrons, by charge separation.

Dynamo magnetic field generation The process whereby magnetic fields are amplified by fluid motions and restructured to change the symmetry of the poloidal field to toroidal.

Faraday effect The rotation of the plane of polarization by a bifringent medium. Produced astrophysically by the differential response of electrons in a magnetic field to right- and left-handed circular polarization.

Flux freezing High conductivity condition which transports magnetic energy as if the material were frozen to field lines.

Magnetic reconnection The process of magnetic to particle kinetic energy conversion, assumed to be due to annihilation of magnetic flux in thin current sheets.

Synchrotron emission Nonthermal radiation produced by free relativistic electrons spiraling in magnetic fields.

Units Solar masses ($M_\odot$) (2×10^{33} g); solar radius ($R_\odot$) (7×10^{10} cm); parsec (pc) (3.1×10^{18} cm $= 3.26$ light years).

Zeeman effect The splitting of atomic (or molecular) energy levels by an externally imposed magnetic field. Transitions between magnetic sublevels show differential displacement from the unperturbed wavelength depending on the field strength.

Zeeman polarimetry Photoelectric method for measuring the displacement in the hydrogen Balmer lines by measuring modulation in the line wing between two opposite circularly polarized states.

THIS ARTICLE reviews some of the theoretical and observational features of the role played by magnetic fields in cosmic bodies. Time-dependent fields are generated by dynamos driven by turbulence and differential rotation. In stars, this produces flaring, coronae, and starspots. Galactic-scale fields are important in regulating star formation and in structuring the interstellar medium. Extragalactic magnetic fields dominate emission

from active galaxies and may structure the gas in galaxy clusters.

I. INTRODUCTION

Magnetism, after gravity, is the most important force structuring cosmic environments. But astrophysically important magnetic fields are not forever. This truism dominates all research on their origin and evolution. The scales of time and length involved in stellar and galactic environments are so great that it is virtually impossible for any phenomenon to remain static over the entire lifetime of an object.

In this article, we will concentrate on extra-solar-system magnetic fields. The problem of remote sensing of such fields is highlighted in extraterrestrial environments, where *in situ* measurements are generally not possible. The sole exception to this is the sun, where we have sufficient spatial resolution in groundbased imaging and for which the magnetic field of the solar wind has been directly determined using spacecraft.

II. OBSERVATIONAL TECHNIQUES FOR MEASURING COSMIC MAGNETIC FIELDS

The measurement of magnetic fields in stars and in other cosmic sources has reached a high degree of refinement. Because of this ability to achieve high sensitivity in the determination of both magnetic-field strength and geometry, the importance of magnetic fields in astrophysics has become a central issue. New phenomena requiring magnetic activity are continually being discovered, and it appears that far from being a panacea invoked by astrophysicists for the explanation of otherwise mysterious effects, theoretical models can be increasingly constrained by direct observations. Before discussing the results of these measurements, we will outline some of the methods used for the study of cosmic magnetic fields.

A. Measurement of Magnetic Fields in Stars

The discovery of the Zeeman effect at the close of the nineteenth century proved to be the first essential step in the study of extraterrestrial magnetism. The classical derivation is based on Lorentz electron theory. An electron spirals about a magnetic-field line due to the Lorentz force,

$$\mathbf{F} = \frac{e}{c}\,\mathbf{v} \times \mathbf{B}, \tag{1}$$

where **B** is the magnetic field strength, **v** is the velocity of the electron, e is the electronic charge, and c is the speed of light. The spiral produces a specific helicity, due to the sign of the electron charge, so that circularly polarized radiation excites the orbits in one sense but not in the other. The effect is immediately applicable to the explanation of the Faraday effect (see the following) but also predicts that classical electrons orbiting the atomic nucleus should produce two oppositely circularly polarized emissions when seen longitudinally (that is, down the magnetic axis). The separation in wavelength of these two components is dependent on the force constraining the electron's motion and is consequently directly proportional to the strength of the local magnetic field, $\omega_L = eB/mc$. Thus, from the measurement of the separation of the circularly polarized, or $\sigma_\pm$ components, it is possible to determine the local longitudinal component of the magnetic field. In linearly polarized light, the line splits into three components (in the normal Zeeman effect), the central, undisplaced component being designated π.

The quantum mechanical derivation of the Zeeman effect is a bit more complicated, since it involves electron spin. The electron possesses an intrinsic magnetic moment, directly proportional to its spin. The magnetic moment of an atom in a specific state depends on the combined effect of all of the intrinsic moments of all of the configuration electrons and their coupling with the atomic nucleus. This combined moment gives rise to one component of the energy in the overall structuring of the atomic state but depends only on the intrinsic symmetries in the atom. An externally imposed magnetic field breaks the symmetry of the atomic potential, and the net magnetic moment either aligns or antialigns with this field. Thus, the interaction energy

$$E_B = \mu \cdot \mathbf{B} \tag{2}$$

depends on the sign of the direction cosine between the moment and the field. Spin quantization produces multiple components along the field direction, each of which has a different projected magnetic moment. The net effect of the external field is therefore to lift the *degeneracy* of the energy level, or in other words to split the unperturbed atomic state into several components that depend on the angular momentum of the state j. There are $2j + 1$ components to each state (regardless of the coupling scheme for the interacting bound electrons).

1. Atomic Parameters

Atomic lines result from the transitions between energy levels, and lines can form between substates that differ in the projected angular momentum m by $\Delta m = 0, \ \pm 1$. The maximum separation of the components thus depends on the combined separation of the individual states involved with the line formation:

$$\Delta\lambda_Z = \delta_L g B \lambda^2, \tag{3}$$

where δ_L is the classical Lorentz factor for the electron and g is the Lande factor, which for $L - S$ coupling is

$$g = 1 + \frac{J(J+1) - L(L+1) + S(S+1)}{2J(J+1)}, \qquad (4)$$

which depends on the orbital (L), spin (S), and total angular momentum of the state (J). The quantum mechanical result is that the line separation in frequency depends only on the intrinsic properties of the atomic state and is linearly dependent on the strength of the local magnetic field. For a strong enough field, the (possibly complex) pattern of subcomponents becomes classical again, looking like a Lorentz triplet. Neither of these effects changes the central wavelength of the transition. If the external field is so strong that the magnetic perturbation becomes comparable with the strength of the atomic potential, the mean energy of the atomic level shifts and the spectrum will be displaced in wavelength, depending on the quantum numbers for the states from which the lines arise. For hydrogen, the shift is proportional to n^2.

All of these effects are observed in astrophysical objects. For instance, G. E. Hale's application of the Zeeman effect to sunspot spectra soon after its original discovery demonstrated that solar activity is an essentially magnetic phenomenon. Since different atomic lines are differently affected by the external field, one can determine a mean field through the combination of measurements for a number of atomic transitions of different g value. These lines may also be formed in different regions of the stellar atmosphere, and at different depths depending on their excitation and ionization state, and thus can be used to map out the magnetic field as a function of position and height in the atmosphere. For solar-type stars, this is a more complicated task since we lack the ability to separate out the specific regions on the stellar photosphere. Experience with solar spectra proves here to be a valuable guide to the interpretation of the observations.

For the sun, and for other stars, both the splitting and the polarization of the Zeeman pattern have been extensively used to determine longitudinal field strengths. Since the σ-components are oppositely polarized, they will also produce a net polarization in the absorbed or scattered light of the photosphere. Viewed through oppositely circularly polarizing filters (quarter-wave plates), the centroid of the line shifts from $\lambda_{\sigma+}$ to $\lambda_{\sigma-}$. Thus the difference in the intensity of the line in the two states is a direct measure of the strength of the local field along the line of sight, provided the lines can be resolved.

Increased resolution of spectrographs equipped with CCD detectors now permit resolution of $\lambda/\Delta\lambda \sim 10^5$ to 10^6, with attendant reductions in the limits on global fields.

2. Zeeman Polarimetry

A very elegant technique, called Zeeman polarimetry, is now used to measure both solar and stellar magnetic fields. Consider a line that is only marginally resolved, either because of rotational broadening of the profile due to lack of spatial resolution of the stellar disk or because of intrinsic broadening mechanisms. The wings of the two σ-components may overlap, but because the fields are weak and the polarization states are orthogonal, there is no mixing of the light between the two profiles. Place a filter at one wavelength in the line wing and chop rapidly between the two polarization states using a Pockels cell (an electro-optical device that passes either right- or left-circularly polarized light along the extraordinary ray of the crystal depending on the sense of the applied electric field). At a fixed wavelength the two σ-components contribute different absorption to the profile, so the chopping produces a periodic intensity variation that can be detected photometrically with very high accuracy. Knowing the slope of the unresolved line wing at some wavelength λ_0, $(dI/d\lambda)_0$ the polarization is related to the variation in the intensity due to the switching between right- and left-circular polarizations:

$$p \equiv \frac{\Delta I}{I} = \frac{1}{I_0}\left(\frac{dI_\lambda}{d\lambda}\right)_0 \Delta\lambda \sim \Delta\lambda_Z \sim B_{\text{eff}}\lambda^2, \qquad (5)$$

where the effective magnetic field B_{eff} is the value measured for the longitudinal field. If the field reverses polarity, the sense of polarization switches sign, so it is possible to determine both the magnitude and the relative polarity of the field.

The power of the Zeeman polarimetric technique lies in its extreme sensitivity to weak fields. The technique can even be applied to stellar hydrogen lines despite their large Stark broadening. Instead of measuring the separation of the components by direct observation of the two profiles, as in the traditional Zeeman that passes either right- or left-circularly polarized light analyzer, which yields typical errors of hundreds of gauss even for sharp lines, this method achieves errors as low as 50 G for hydrogen lines in bright stars. The hydrogen line is also of singular value because it is the only low-ionization, strong line that displays the classical Zeeman effect. Continuing development of the polarimetric method has been of tremendous importance in the study of magnetic fields in a wide range of stars, from main sequence to white dwarfs.

3. Differential Zeeman Broadening

A final way of measuring stellar fields is by looking at the overall broadening of spectral lines of different Lande factor and Zeeman structure to discover broadening in excess of that expected from turbulence, rotation, and thermal broadening in the stellar atmosphere. If the field is complex, and there is no strong, global field, the polarimetric method will fail. A field may be present, but if its geometry is such as to produce no strongly directed field in the line of sight it will not produce a net polarimetric signal in the line wings. Instead, for such fields, one exploits the fact that, by chance, some atomic transitions produce Lande factors of zero. These lines do not split in the presence of a magnetic field even though their constituent energy levels do. The excess broadening of lines arising from the same ionization state over these *null lines* directly measures the mean surface field.

B. Measurement of Magnetic Fields in the Interstellar Medium

1. Direct Measurement of Zeeman Splitting for Radio Lines

For extremely weak fields, such as those observed in the interstellar medium, it is impossible to directly measure the Zeeman broadening using optical or infrared spectral lines. This is because of the λ^2 factor in the splitting. By going to very low frequency and using the strongest line observable in diffuse interstellar clouds, the 21-cm groundstate transition of neutral hydrogen, it is possible to measure fields as weak as tens of microgauss. Put differently, as the field becomes progressively weaker, the wavelength of the line must be matched to the expected field strength. Once the field is weaker than a few milligauss it is necessary to observe at millimeter or centimeter wavelengths, where few atomic lines are available. The HI interstellar Zeeman effect has been detected in a number of clouds, and provides an important handle on the strength of the field in the diffuse interstellar medium. However, the measurement suffers from the same limitations as the stellar polarimetric method—it is sensitive only to directed fields.

2. Faraday Effect

The free thermal of the phenomenon electrons in the interstellar medium behaves classically, and it is straightforward to apply the Lorentz theory to the transfer of polarized radiation through this medium. The index of refraction depends on the sense of polarization because the resonant frequency of the electron is moved to greater or lesser values depending on the sense of helicity of the particles around the local field direction. Thus, if the index of refraction of an electron gas is

$$n = 1 - \left(\frac{\omega_{\mathrm{p}}}{\omega}\right)^2, \qquad (6)$$

where

$$\omega_{\mathrm{p}} = \left(\frac{4\pi e^2 n_{\mathrm{e}}}{m}\right)^{1/2} \qquad (7)$$

is the *plasma frequency*, where n_{e} is the local electron number density, then the difference in the index of refraction between the two senses of circular polarization is

$$\Delta n = \frac{\omega_{\mathrm{p}}^2 \omega_{\mathrm{L}}}{\omega^3}, \qquad (8)$$

where ω_{L} is the Lorentz frequency eB/mc. As a signal propagates through this bifringent medium, its sense of polarization will change due to the projection of the B field along the line of sight, and the phase shift resulting from this field is the integrated action of the bifringence along the line of sight. The plane of polarization for linearly polarized light thus rotates through an angle:

$$\Delta\phi = \frac{\omega}{c}\int n\,ds \sim \lambda^2 \int n_{\mathrm{e}} B \cdot ds. \qquad (9)$$

Unfortunately, this Faraday rotation depends on both the projected field strength along the line of sight *and* on the local electron density of the magnetoactive medium. Variations in the orientation or strength of the magnetic field, or in the electron density, can produce changes in the rotation angle. Further, randomly oriented fields can produce no net change in the polarization, no matter how large the local field may be. Each polarization reversal reduces the overall rotation measure, and $\Delta\phi$ is reduced by $N^{1/2}$ for N random changes in the direction of the magnetic field along the line of sight. Independent estimates, using pulsar dispersion measurements, H_α diffuse line emission from the interstellar medium, and scintillation measurements of galactic and extragalactic centimeter wavelength point sources, help constrain the electron density along any observing direction, but only the Faraday effect can yield information about the magnetic field directly.

3. Dust and Interstellar Polarization

The discovery of interstellar polarization was accidental, made while Hall and Hiltner in 1947 were attempting to measure the intrinsic polarization resulting in hot stellar atmospheres from electron scattering. The correlation between extinction (also called color excess) and the degree of polarization identified the source as intervening dust between the observer and the star. Dust consists of material that may have either a permanent or induced magnetic moment and that is forced through collisions with charged and neutral particles to spin about the magnetic field direction. In doing so, it preferentially aligns with its largest moment of inertia orthogonal to the field direction. Scattering by nonspherical particles causes the direction of the polarization vector to lie parallel to the minimum cross section, hence the polarization is along the direction of the external field. The degree of alignment, produced by the magnetic torque $\mu \times \mathbf{B}$, is proportional to the magnetic field strength, but because of the collisional dealignment the degree of polarization is not a good measure of the local field strength. The position angle, on the other hand, is a direct measure of the local *direction* of the magnetic field.

Unlike the Faraday effect, dust scattering provides geometric information about the field structure. The orientation of the magnetic field in the diffuse interstellar medium has been determined largely by large-scale optical polarimetric mapping surveys. Advances in infrared instrumentation have allowed for the extension of these measurements into wavelength regimes at which interstellar clouds are less opaque, and it is even possible now to obtain information about the structures of magnetic fields within molecular clouds and in regions of active star formation, previously inaccessible to optical measurement. The technique of measuring dust polarization is insensitive to the spectral distribution of the background light source and is essentially a continuum measurement. Polarizations as small as 0.05% can be measured in this way using electronic cameras like CCDs over extended structures. Additional details are obtained through the observation of circular polarization of starlight, resulting from the variation in the position angle of the magnetic field across the line of sight and linear to circular conversion of the propagated photons.

4. Synchrotron Radiation

Synchrotron radiation is the only emission mechanism that is uniquely dependent on the presence of magnetic fields. A relativistic electron spiraling in an external magnetic field radiates photons at a characteristic frequency $\omega_{\mathrm{c}} = \gamma^2(eB/mc) = \gamma^2\omega_{\mathrm{L}}$, where $\gamma = E/mc^2$, which depends only on the local field strength. An ensemble of electrons with energy spectrum $N(E)$ radiates with an intensity $I(\omega) \sim B^{(p+1)/2}\omega^{-(p-1)/2}$ while even for an optically thick source at low frequencies, the spectrum is $I(\omega) \sim B^{-1/2}\omega^{5/2}$. Here, ω is the frequency of the radiation. The extra factor of $\omega^{1/2}$ is, in effect, because the electrons have an effective radiation temperature that depends on their energy.

In contrast, an optically thick thermal spectrum varies as ω^2 at low frequencies. The emission is polarized, since the electron radiates orthogonally to the local field direction; circular polarization results from viewing the source longitudinally. The synchrotron spectrum thus reflects the exponent of $N(E)$, so the identification of this radiation is generally straightforward, using the shape of its emitted spectrum. Another measure of the nonthermal nature of the emission is that the brightness temperature, the temperature required for a blackbody source to produce the same monochromatic flux, can be as high as 10^{11} K. Even for mildly relativistic electrons, it is possible to observe the synchrotron emission, which is strong in the radio spectrum at centimeter wavelengths at which most thermal emitters (except for hot optically thin plasmas) are not strong sources. Still less energetic electrons emit gyrocyclotron (or gyro-resonance) radiation when trapped in magnetic fields. Observed in solar flares and coronal loop structures where temperatures reach as high as 10^9 K, this emission can be due to an input *thermal* spectrum of electrons (a Maxwellian velocity distribution for a hot gas in a magnetic field) although the form of the emitted spectrum will not be thermal.

There are many strong cosmic sources of synchrotron radiation, both within the Galaxy and in extragalactic objects. Often, they are distinguished from thermal emission regions by their brightness temperatures and by their being linearly polarized. The degree of polarization depends on the shape of the electron spectrum and on the optical depth of the emitting volume, $P_{\mathrm{linear}} = 3(p+1)/(3p+7)$ if the source is optically thin, and $P_{\mathrm{linear}} = 3/(6p+13)$ if it is optically thick. Nonmagnetic, homogeneous, thermalized plasmas do not radiate polarized light.

III. OBSERVATIONAL RESULTS

A. Stellar Magnetic Fields

After planets, the most detailed information and the most extensive theoretical work are available for stellar magnetic fields. In the case of the sun, we have an enormous wealth of information about the spatial distribution and temporal evolution of magnetically related structures. One of the major discoveries of this past decade is the extent to which solar-type stars share many of the characteristics observed for solar active regions, and also the remarkable range of energy and length scale on which magnetic field activity can manifest itself in stellar envelopes. It should be borne in mind that stellar envelopes are probably the best places for the study of dynamo processes. They are observationally accessible, they vary on a timescale short compared with a human lifetime, and there are many plasma diagnostic spectral features with which to analyze the physical state of the star. Most of the work on dynamo field generation is constrained by the study of stellar and planetary fields, after which the models can be extrapolated to larger-scale phenomena, such as galactic field structures.

1. The Sun and Solar-Type Main Sequence Stars

The solar magnetic field is organized into complex structures, without a large-scale geometry other than an approximate polarity for the individual hemispheres. Structure exists on scales ranging from several tens of kilometers, about the size of individual convective elements, to large active regions that can be more than 10^4-km across. The strongest and most obvious fields are detected in sunspots, in which fields of several kilogauss are common. These regions are dominated by their local fields, which typically reach values of 1 to 3 kG in the umbra and are directed normally to the stellar surface. Peripheral, penumbral fields are weaker and are transverse to the surface. The internal fields of these spots are transient, decaying on timescales ranging from days to months depending on the area of the spot. Convective motion is suppressed within the spot, a primary cause of the relative darkness of the spot region in contrast with the convectively active photospheres. The field structures extend into the chromosphere and the corona above the active regions, often reaching heights of order 10^5 km, about 0.1 $R_\odot$. Such fields are entirely responsible for the thermal and dynamical structuring of the corona.

Sunspot activity, as is well known, behaves in a roughly cyclic manner with a mean timescale of about 11.5 years between successive maximum disk-covering fraction. The magnetic field of the sun, however, reverses on about twice this timescale, although the cycle closely resembles a strange attractor. The most prominent manifestation of this change is that the polarity of the leading spot in a given hemisphere is systematically of one polarity throughout one sunspot cycle and reverses during the next. The same is true in the more disorganized photospheric magnetic network, the umbral pores and uncoalesced magnetic flux that emerge over the majority of the solar disk.

The solar wind, and the entire outer solar atmosphere, are completely dominated by the magnetic field. Ultraviolet and X-ray images of the sun show that the corona is composed of numerous regions of heated, magnetically confined, plasma. The coronal geometry and rate of expansion are controlled by the variations of surface fields. Indeed, coronal activity appears to be a primarily magnetic phenomenon. For example, a solar flare involves the rapid acceleration of coronal electrons and protons to relativistic energies, along with radiative emission of a broad variety of spectral signatures ranging from gamma rays to radio, on timescales of several minutes. Coronal transients are also energetic and dramatic events in that they involve the dissipation and restructuring of magnetic fields in the corona. The most likely cause is magnetic field reconnection and the fast conversion of magnetic into radiative and particle kinetic energy.

Magnetic heating is also responsible for the high temperature of the corona, and, by extension, the heating of the outer atmospheres of solar-type stars to temperatures greater than 10^6 K. The dissipation of Alfven waves, magnetic disturbances that propagate along field lines with a velocity of $v_A = B/(4\pi\rho)^{1/2}$, where ρ is the mass density, and reconnection, taking place on many length scales in confined magnetic loops (like prominances), play important roles in the heating of the outer atmosphere. Reconnection and flaring activity also serve as an energy source for coronal activity, although the detailed mechanism is still a subject of considerable investigation. X-ray satellite observations, with Yokok and TRACE, reveal the complex structure of coronal magnetic fields and highlight the role of reconnection in the heating.

All stars with surface temperatures cooler than about 7500 K appear to have convective envelopes, the primary optical signature of which is the emission observed in the resonance lines of ionized calcium, Ca II, at $\lambda\lambda$3927 and 3933 Å. The strength of this emission appears to decline with age, as does the rotational velocity of the star, thus supporting the contention that magnetic activity is the primary agent for the production of chromospheres and coronae in low-mass stars. We shall return to this point later.

Ultraviolet spectra also show the presence of chromospheric and coronal regions in late-type stellar atmospheres. All late-type main sequence and more evolved stars for which ultraviolet spectra have been obtained below λ3300 Å show the emission-line signatures of hot envelopes. The Mg 11 λ2800 Å doublet is especially good as a chromospheric indicator, since it is an analog to the Ca II doublet. Additional support for hot exospheres comes from the C IV λ1550 Å doublet, C II λ1334 Å, O I λ1304 Å, and the Si III and C III intercombination lines near 1900 Å. These lines also vary in response to flare activity. They are also far more sensitive than the optical lines to the local heated structure in the outer atmosphere, so are tied closely to the active regions. As these heated regions are carried by rotation across the line of sight, the UV emission lines vary in intensity and also in velocity structure. While this is better studied in binary systems (see Section III.A.2), the variation in the UV and optical lines in single stars form much of the basis for the interpretation of magnetic activity in normal stars.

Many of the late-type stars close enough to be detected by X-ray satellites have been observed. The X-ray luminosities L_{XR} are typically 10^{-4} L_{bol} and the coronal temperatures are consistent with the solar coronal value. Support for the link between magnetic activity and coronal emission comes from the correlation between XR luminosity and stellar rotation frequency $L_{\mathrm{XR}} \sim \Omega^n$, where n appears to be between unity and 2.

Very low-mass main sequence (hydrogen core-burning) stars, less than about 0.5 $M_\odot$, have been observed to display flaring activity on a scale about 10^5 or higher than that of the sun. These so-called flare stars, or dMe stars, show many of the same emission processes observed in solar flares, although the timescales and energies involved are greater than seen on the sun. To date, however, no fields have been directly measured. Instead, activity cycles have been observed in the periodic variations of Ca II emission intensity, and individual regions have been mapped on many main sequence stars by observing the modulation of the line strength on a timescale of weeks to months, comparable to the rotation period of the stars. One indirect measure of the field strengths is provided by the excess broadening of lines with large Lande factors over that observed for small g-factor lines, and from the periodic variation in the optical brightness of some of the dMe stars due to star-spot activity.

2. Magnetic Chemically Peculiar Main Sequence Stars

The magnetic chemically peculiar stars, also called the Ap or Bp stars, are confined to the effective temperatures, on the main sequence, between roughly 7500 and 25,000 K and have masses ranging from about 1.5 to about 10 $M_\odot$. These stars are notable for their strong absorption lines of anomalously abundant species, like the rare earths and silicon, and also for abnormally weak lines of helium. The Ap/Bp stars typically rotate more slowly than field stars of the same mass. They show strong magnetic fields ranging from the limits of detectability of about 100 G to more than 30 kG. The strongest field yet observed in a main sequence star, HD 215441, measured by the transverse Zeeman effect, is 32 kG. About 50 bright stars have been directly measured to possess these strong fields. Typically the field has a strong dipolar component that is oblique to the rotational axis by a random angle between 0° and 90°. The field of HD 37776 is especially notable in having a predominantly quadrupolar field that is also oblique. The field bears a remarkable resemblance to that observed by Voyager 2 at Neptune. The magnetic fields often must be fit using decentered dipoles, probably indicating that most of the magnetic stars have strong nondipolar components.

An important difference between these fields and those observed in late-type stars is that the chemically peculiar star fields show no time dependence. The rotational phase of the dipole remains constant, in some cases over more than 3000 rotation periods (as in α^2 CVn), and there is no evidence for changes in the surface distribution of overabundance patches. Thus these fields seem to be a relic from an earlier stage of evolution and not to be continually generated by any dynamo activity. While their frequency in the main sequence stellar population is not well known, it appears that as large a fraction as 10% of all main sequence stars in this mass range have some strong surface magnetism. One possibility is that the stars were initially rapid rotators in the pre-main-sequence stage of their evolution, and generated strong dynamo fields while in possession of deep convective envelopes. Subsequent to this contraction stage, when the envelopes turn radiative and the stellar surface temperature increases on its way toward the main sequence, these fields might be preserved and additionally amplified by stellar contraction. The field would also torque the star down, constraining mass loss to large radii and increasing the angular momentum loss of the star over that which would be seen for a nonmagnetic star.

3. Magnetic Activity in Close Binary Stars

Two classes of close binary stars display evidence for magnetic activity, both connected with solar-type stars. Main sequence contact binaries of the W UMa type are generally low-mass systems of large mass ratio. The lower mass star in the system is generally ≤ 1.5 $M_\odot$ and therefore has a convective envelope. These stars, of the W-type in the taxonomy of contact systems, show a light curve instability and emission spectrum in the UV that has been interpreted as arising from dynamo activity in the lower mass star. Dark waves move systematically through the light curve, indicative of regions of lower (about 300 K or more) effective temperature that may occupy a few percent of the photosphere of the pair. Chromospheric emission is seen for these systems. The orbital periods are typically about 0.5 day. These stars have rotational rates about a factor of 20 to 100 slower than this, hence the expectation that these close binaries should show strong magnetic fields. Nonthermal radio emission and flaring activity have been reported for several of these systems, but these reports are still controversial. The best available evidence that these stars maintain dynamo fields comes from the photometric signature of dark waves.

Considerably more certain identification of large-scale magnetic field generation is available for the RS CVn and Algol-type binaries. These are also short-period systems, ranging from about 0.5 to 40 days, which consist of evolved primaries and main-sequence secondaries. In the classical Algol systems, the more evolved star is actually the lower mass member of the binary, having lost considerable mass both to the unevolved star and from the system as a whole; these evolved stars are currently in contact with their Roche surface and losing mass onto the main sequence member. Both radio and X-ray flaring have been observed in these systems, the most notable being Algol(β Per). Ultraviolet observations taken during eclipse show strong chromospheric emission spectra for the cool star in the system. Because the evolved star is optically the fainter member of the system, starspot activity is not directly observable in the light curve.

Better evidence for dynamo activity comes from the RS CVn stars, which like Algols have evolved members. In most of the RS CVn systems, however, the more massive star is the photometric primary and also the more evolved member. These stars display several characteristics of enhanced magnetic activity when compared with normal stars: (1) Ca 11 emission (by as much as a factor of 10); (2) dark waves in their light curves that show cyclic variations on timescales of years to decades; (3) strong UV chromospheric signatures (C II, O I, Si III], and C III] are especially strong for their spectral class); (4) hot coronas with temperatures up to 10^7 K; and (5) most important, strong flaring activity, reaching luminosities greater than 10^{27} erg s^{-1} over timescales up to weeks (the most extensively studied being the 1978 radio flare of HR 1099). Ultraviolet flares have been observed in a number of RS CVn binaries with energies higher than 10^{35} ergs over timescales of nearly one day. In Algol systems, X-ray flare energies have detected up to 10^{35} erg, comparable with the UV output. In contrast, the strongest flares observed for the sun emit about 10^{22} erg s^{-1} in the radio and have a total bolometric energy of about 10^{25} erg. More typically, the total radio luminosities in RS CVn flares are of order 10^{30} erg, comparable with those observed in the sun. These radio flares may last for several hours, while solar radio flares are rapid, of order 10^3 seconds.

The RS CVn systems are especially important because the optical and UV spectra of the primary can be directly studied, and individual active regions can be followed as they rotate across the line of sight. The technique, called *Doppler imaging*, has been applied to a number of systems, notably AR Lac, HR 1099, and λ And. The RS CVn stars are also rotating much faster than single stars of the same mass and effective temperature, by up to a factor of 10. For a few systems, FF Aql and V471 Tau, ultraviolet observations have revealed the presence of prominencelike material located above the limb of the evolved star. The temperatures of this material (above 10^4 K) and densities (above 10^7 cm^{-3}) support the contention that it is trapped in magnetic loops located above active regions in the photosphere. The similarity of the Algol and RS CVn stars is taken as a strong argument for the ubiquity of magnetic activity in any star with deep envelope convection, provided the rotational frequency is large, and connects the enhanced activity in these stars to their more normal, single, main sequence counterparts.

4. Pre–Main Sequence Stars

The photometric and high-energy signatures of magnetic-field generation have been observed in several T Tau stars, which are pre-main-sequence stars still in the process of contraction toward the main sequence before the onset of hydrogen core burning. The rotation periods obtained for these stars on the basis of their light curves are shorter than 7 days. For instance, V410 Tau has a period of about 4 days. The starspot characteristics are quite similar to those encountered in the W Uma and RS CVn systems, although they have not displayed the strong radio flares seen in the latter; they do, however, show strong X-ray flaring activity (especially several of the sources observed in the ρ Oph molecular cloud).

A few stars belonging to the Herbig Ae/Be subclass of pre-main-sequence stars display time variable chromospheric features in both the optical and ultraviolet. These are fairly hot (>7000 K) clearly unevolved stars that are found in regions of recent star formation but are usually not embedded in their parent cloud. The profiles of several resonance lines, notably Ca II, Mg II, and C IV, are modulated on timescales of about 40 h. This seems to indicate the presence of active regions, or at least surface inhomogeneities. This result is surprising since the Ae/Be stars have high effective temperatures, about 10,000 K, and so do not have extensive convective envelopes, which normally appears to preclude such magnetic activity. These stars do not show flares, unlike the cooler T Tau stars, and they are not strong radio or IR sources. Nonetheless, they show classic symptoms of magnetic activity evidenced by their spectral features.

Indirect evidence of the role of magnetic fields in the pre-main-sequence stage of evolution comes from the discovery of bipolar mass outflows from these systems. Many T Tau stars are associated with jetlike CO and Hα flows, which are collimated and supersonic relative to the cloud medium (velocities as high as 50 km s^{-1}, or upwards of Mach 50, have been observed). The theoretical explanation of these flows involves magnetic torquing of accretion disks surrounding the nascent protostar, which generates Alfvén waves that accelerate disk material off the disk and preferentially out the poles. Such fields are assumed to originate within the molecular clouds, by amplification of the field of the galaxy, and to be transported as frozen-influx onto the protostar by the accreting matter.

5. White Dwarf Magnetic Fields

White dwarfs are the remnant of the evolution of low and intermediate ($\leq 5\ M_\odot$) mass stars, formed by core contraction during the last stages of nuclear burning. Since they reach densities in excess of 10^8 g cm^{-3} and typically have radii of order 0.01 $R_\odot$ they were expected, by simple theory, to be able to possess very strong surface fields. Any weak field remnant in the core at the time of contraction, should the field be able to survive this stage of evolution, would be amplified by flux-freezing to very large values, of order 10^6 G. Such fields have been discovered in a number of isolated white dwarf systems, by measurement of polarization properties both of the continuum (due to scattering by electrons in such strong fields) and by the Zeeman effect they produce in photospheric lines (through the use of Zeeman polarimetry).

There is also an important class of white dwarf for which the magnetic field plays an important role in structuring the accretion of matter onto the star. These are the AM Herculis stars, extremely close binary systems with periods of order hours in which the primary, a low-mass star, has expanded sufficiently to transfer mass onto the surface of its companion white dwarf. If the accreter is magnetic, the infalling mass funnels preferentially to the poles where it accretes in a column of hot, shocked, plasma onto the white dwarf. Like the Ap and Bp stars, these dwarfs have obliquely oriented fields and by the rotation of the star, the accretion columns vary in their angle to the line of sight. There are two observable effects of this rotation. One is that the strength and velocity of emission lines formed in the polar columns vary with phase of rotation. A second is that the lines thus formed are polarized by the strong fields and can be studied using the same Zeeman polarimetric techniques used for more ordinary and weaker fields. Observation shows that the field strengths in these systems are in the range of 1–10 MG, and that the rotation of the white dwarf is synchronous with the orbital period.

The presence of magnetic fields in white dwarf stars is important as a physical process for many binary star systems, notably cataclysmics like novas and dwarf novas. Accretion of matter from a circulating disk accumulated in the environment of the white dwarf transfers angular momentum to the star and spins it up. The tidal coupling between the dwarf and its companion produces spin-orbit coupling, which then alters the angular velocity of the binary, causing it to shrink to compensate for the loss of orbital angular momentum. Thus the presence of a strong field on the accreter can drive the evolution of the star and maintain the mass transfer in excess of that which would be expected solely from nuclear evolution of the companion star. A number of classical novas, especially DQ Her (N 1936) and V1500 Cyg (N 1975), have magnetic white dwarf components.

6. Neutron Stars and Pulsars

The strongest fields encountered in astrophysical environments are found for neutron stars, the end product of stellar core collapse at the start of the supernova event for massive stars. These objects, having about 1 $M_\odot$ but radii of order 10 km, have been observed to have fields upwards of 10^{12} G. Two indirect means have been used to infer the strengths of these fields, which defy direct detection because of the lack of spectral diagnostics on the neutron star surface. Pulsars were first observed as radio sources rotating with periods between tens of milliseconds and seconds, which display both pulsed radio emission and slow secular increase in the rotation rate. It is this spindown that serves as a measure of the surface field strength. A rotating magnetic dipole loses energy at a rate:

$$\frac{dE}{dt} \sim \Omega^4 B^2, \tag{10}$$

since the rate of radiation depends on $\ddot{\mu}^2$, the square of the second derivative of the magnetic dipole moment. Since the rotational energy is $\frac{1}{2}/\Omega^2$, where I is the stellar moment of inertia, the rate of spindown $\dot{\Omega}$ is proportional to B^2. This is measured by the spindown time $\Omega/\dot{\Omega}$ and provides fields of such large magnitude. Such large (TG) fields are also required for the production of the emission observed at all wavelengths from pulsars. In addition, for neutron stars still embedded in supernova remnants, the emission of low-frequency (at a frequency Ω) electromagnetic radiation can power synchrotron emission from the surrounding plasma (as in the Crab Nebula). Such supernova remnants are called *pleirons*, after the Greek word for crab. The observed emission from these remnants is consistent with the model.

The weakest fields observed to date for neutron stars have come from the discovery of the millisecond pulsars, neutron stars with rotation periods of order 1.5 to 10 ms. While such rapid rotation would have been expected naively only for extremely young pulsars, all indications are that these stars are rather old. They show breaking ages of more than 10^9 years; in fact, to date, spindown has not been observed for any of these systems. The theoretical explanation for their rapid rotation and slow spindown is that they have far weaker magnetic fields than previously detected for neutron stars, of order 10^9 to 10^{10} G. While such fields may be the product of decay of a stronger initial field, considerable controversy currently surrounds the theory of neutron star magnetic field decay. Rather, these stars appear to have formed with initially weak fields. Their rapid rotation is presumed to have originated from accreted matter from a one-time close binary companion. The lack of a strong field ensures that the high angular frequency is preserved for a long time, making the binary look “forever rotationally young.”

B. Interstellar Magnetic Fields

The source for stellar magnetism, at least in the protostellar phase, must lie in the environment. Measurements of dust-induced polarization in nearby molecular clouds, especially Taurus-Auriga and ρ Ophiuchus, show that the magnetic field is complex and pervasive throughout the parent cloud in which star formation is occurring.

Additional evidence for large-scale magnetism in the interstellar medium comes from the radio observation of supernova remnants, strong synchrotron sources that show the effects of the expanding blast wave in the redirection and shock-amplification of the ambient galactic field.

There are two methods for measuring the magnetic field of the interstellar medium. The most direct method is the detection of the Zeeman splitting of the 21-cm line of neutral hydrogen, or the splitting of the ground state transitions of the OH molecule. Both of these are present in low-density clouds that pervade the interstellar medium and that have low enough internal temperatures that the lines are not too broad for the measurement.

Individual HI clouds having temperatures of order 100 K display internal fields of order 10 μG, larger than that inferred for the low-density gas and consistent with flux-freezing in the clouds. OH transitions, which are collisionally excited and arise from masers that are pumped in the presence of strong infrared sources, provide similar field measurements.

The magnetic field in the diffuse interstellar medium is significantly lower than that observed in the clouds. One indication of the presence of magnetic fields in molecular clouds comes from the observation that line widths in these clouds are far larger than would be expected from the cloud temperatures. The sound speed in a typical molecular cloud that has a temperature of 30 or 40 K (measured using infrared emission from the embedded dust) is about 0.5 km s^{-1}. Yet line widths are often observed in excess of 2 km s^{-1}, and it is possible that this indicated the presence of turbulent Alfvén waves within the cloud. Such turbulence is also apparently needed to support the clouds against gravitational collapse. Fields of order milligauss appear to be required to supply this turbulence. Theoretical models of collapse of magnetic clouds show that flux conservation during cloud formation is capable of amplifying the microgauss fields observed in the diffuse medium to these large values.

Observations of the galactic center show filamentary structures extending from the core nonthermal source, Sgr A, that extend radially to about 50 pc and have field strengths of order 1 mG. These structures, which show both clustered arclike emission regions and isolated filaments extending up to 50 pc from the plane, may be similar to the structures observed in the active nuclei of external galaxies, and their connection with the nuclear molecular clouds argues for the ubiquity of magnetic activity in all portions of the interstellar medium and its central role in the structuring of galactic scale activity.

The Orion Molecular Cloud (OMC-1) is perhaps the best studied galactic region of recent star formation. The magnetic field has been measured in the cloud by a variety of methods. All of these methods show that the weakest fields are of order of a few microgauss, while the strongest fields exceed about 100 μG. The condition of flux-freezing provides for such large amplifications, assuming that they grow like $\rho^{1/2}$.

The strongest interstellar fields are observed in OH masers. These sources are especially difficult to measure with Zeeman effect because beam smearing usually limits the accuracy of the polarization determination. Since the sources are very compact, it is necessary to use very long baseline interferometry to separate out the various (circularly) polarized components. The fields observed in a number of maser knots range from 2 to 10 mG. The conversion between magnetic field strength and velocity width for the normal Zeeman pattern for the OH lines, which have a wavelength of about 18 cm (1.6 GHz) is approximately 0.3 km s^{-1} mG^{-1} so that for a 10 mG field, the lines are completely resolved *even in the total intensity*. This means that it is possible to measure the total field—much as in the case of Babcock's star HD 215441—not just the projected field along the line of sight. In fact, the field is responsible for the desaturation of the maser. OH observations are the most certain test to date of the scaling of the magnetic field of any interstellar region with ambient density, and show that the field follows the $\rho^{1/2}$ scaling law.

The role played by magnetic fields in the support and structuring of molecular clouds is highlighted by molecular observations. These indicate that the line widths observed in CO, CS, and other strong molecular tracers of the density and dynamics of the cloud cores are far in excess of the thermal widths. The primary mechanism for support of the clouds against gravitational collapse is turbulence supported by Alfvén waves trapped within the clouds, which provides the equivalent of an added pressure $(\delta B^2 > /4\pi)$. It is also possible that the fields help regulate star formation by serving as the primary mechanism for transferring angular momentum away from the collapsing cores of the clouds and assisting accretion of matter. As previously mentioned, the observation of bipolar molecular outflows associated with newly formed stars supports the contention that magnetic fields, amplified in the process of collapse, are responsible for the growth of the protostellar object.

C. Galactic Scale Magnetic Fields

In the 1980s, with the advent of aperture synthesis radio telescopes, the study of the large-scale structure of magnetic fields in external galaxies has undergone explosive growth.

The synchrotron emission from the diffuse interstellar medium is the most direct measure of the overall magnetic energy of a galaxy. Perhaps the most significant result in the 1980s is the study of the large-scale structure of magnetic fields in galaxies and on scales of clusters of galaxies. This cannot be accomplished optically, and must be performed using aperture synthesis techniques. Cosmic ray electrons are the primary agents for producing the diffuse synchrotron radiation in galaxies. Because of their distribution throughout the galactic interstellar medium, and the fact that their motion is tied to the magnetic field structure, they serve as excellent tracers of the field configurations.

There is little evidence for magnetic fields in the diffuse interstellar medium of elliptical galaxies. Many of these galaxies, however, display large radio structures on scales up to megaparsecs. These appear the result of mass and flux transport from active nuclei deep within the parent galaxy, likely arising in the vicinity of a massive central black hole. Imaging reveals that many galaxies, especially the central cD-type galaxies in large clusters, have relativistic jets directed orthogonally to a central accretion disk. Evidence for transported magnetic fields comes from the synchrotron luminosity of the jets and their attendant terminating radio lobes, and the magnetic field geometry is suggested by the polarization of the emission. The fields observed along the jets appear to be initially radial and directed along the jet axis, while at large distances (kpc) from the core, the fields become more helical. Synchrotron losses appear to be concentrated toward the boundaries of the jets, especially in M87, the central cD in the Virgo cluster and one of the most extensively imaged nearby (20 Mpc) galaxies. The mechanism for the generation and transport of these fields is an open question. It is possible, however, to estimate the magnetic field strength using a *minimum energy* argument. This analysis assumes that the total energy of the relativistic fluid is minimized with respect to the local field. When applied to the large-scale structures observed in radio galaxies, the derived fields are usually in the range of 100 μG to a few milligauss.

In spiral galaxies, there is abundant evidence for large-scale magnetic fields. Relativistic particles supplied by stellar activity, supernovas, and local acceleration within the interstellar medium, serve as synchrotron emitting markers of the local field intensity and direction. This diffuse emission has been mapped for the nearby spirals, and appears to be well aligned with the spiral structure. Several nearby spirals, especially M83, have been shown to have large-scale fields ordered by the galactic differential rotation. In external galaxies, as in our own, the field shows complex structure, consisting of holes, loops reaching several kiloparsecs above the plane, and ordering on scales ranging from a few parsecs to the size of the galaxy.

The study of magnetic fields in external galaxies is complicated by the resolution of currently available radio telescopes. We see structure on all scales in the interstellar medium of our galaxy from subparsec to kiloparsec lengths, and the magnetic field is structured on similar scales. Across an external galaxy, the average synthesized radio beam is between 0.1 and a few arcsec, depending on the radio telescope and its configuration of antennas. This corresponds to scales ranging from a few parsecs to several hundred parsecs per beam even for relatively nearby galaxies within about 100 Mpc. At present, it is only possible to place limits on the extent of the small-scale structure, and the total measured field is uncertain at best. For the large-scale field, on the kiloparsec scale, the general statements appear more secure. Several galaxies have been mapped, and generally the geometry of magnetic fields determined from the synchrotron emission corresponds well with that obtained from optical (dust) polarization. This is especially well represented by M 31 = NGC 224, M 51 = NGC 5195, and M 81. However, recent results for M 83 = NGC 5236 do not show correspondence with the spiral arm tracers. It appears that in actively star-forming galaxies. the presence of H 11 regions and supernovas distorts the magnetic structures from that ordered by the *grand design* spiral pattern. As larger radio telescopes become operational, especially the Very Long Baseline Array (VLBA) now under construction in the United States, the detailed mapping of the nonthermal emission in nearby galaxies should be able to reach the sensitivity and spatial/resolution to study the complex small-scale structures in the ISM of external galaxies.

IV. SOME THEORETICAL DEVELOPMENTS ON MAGNETIC FIELD GENERATION AND DECAY

A. Dynamo Generation of Magnetic Fields

The discovery of the activity cycle of the sun, and of stars, supports the contention that magnetic fields in cosmic objects must be continually regenerated. The decay of the field, due to the finite electrical conductivity of matter, is sufficiently rapid to require that if a field is observed at all, it must not be primordial.

The electric field in a moving medium is due to two contributing physical processes, one from the induced field due to the motion if there is already a magnetic field present, and the other due to the finite rate of dissipation of the field from the potential difference and finite electrical conductivity. That is, if a current is set up, the strength of the current is assumed due to an electric field.

Therefore

$$\mathbf{E}' = \mathbf{E} - \frac{1}{c}\mathbf{v} \times \mathbf{B}, \tag{11}$$

which by Ohm's law relates the electric field to the current:

$$\mathbf{E} = \frac{1}{\sigma}\mathbf{J}. \tag{12}$$

The Maxwell equations govern the temporal evolution of the fields. For the electric field, the rate of change of the field with time is

$$\frac{\partial \mathbf{E}}{\partial t} = c\nabla \times \mathbf{B} - 4\pi \mathbf{J}, \tag{13}$$

while for the magnetic field we have:

$$\frac{\partial \mathbf{B}}{\partial t} = -c\nabla \times \mathbf{E}. \tag{14}$$

In other words, the two effects that induce the formation of a magnetic field are the displacement current and the free charges that form **J**, while the decay of the magnetic field is due to the generation of a spatially variable electric field. If we assume that the material is in motion, then the form of the equations is invariant, but the motion of a charged medium through a magnetic field produces the appearance of an electric field, thus adding to the decay term for the field because the generation of the electric field is at the expense of the magnetic field already present in the medium.

Assuming that the electric field in the moving medium is vanishingly small—that the conductivity is very high (although not infinite)—we obtain

$$\begin{aligned}\frac{\partial \mathbf{B}}{\partial t} &= \nabla \times \mathbf{v} \times \mathbf{B} - \frac{c}{\sigma}\nabla \times \mathbf{J} \\ &= \nabla \times (\mathbf{v} \times \mathbf{B}) + \eta\nabla^2 \mathbf{B} + \nabla \times \mathscr{E}. \end{aligned} \tag{15}$$

Here the magnetic *diffusion coefficient* is $\eta = 4\pi c^2/\sigma$, and $\mathscr{E}$ is the electromotive force. This last equation, the so-called *dynamo equation,* is of the most interest to us. The first term is that of a generator, the fluid motions being used to produce a magnetic field from the preexisting field. The latter is a diffusive term, resulting from the transport of magnetic field through the fluid by random motions.

B. Flux-Freezing

Any attempt to maintain an internal electric field in a perfectly conducting medium will be thwarted by the mobility of the charges, which immediately move to cancel any potential difference. The timescale for this cancellation is very short in comparison with the timescales for the field to begin building in the medium; that is, they take place on times short in comparison with the actual fluid motions, so there will be no net electric field. The meaning of the diffusive term to the evolution of a magnetic field can thus be explained more clearly by thinking about the effect of mass motions in a magnetic field. For a net current to result from the fluid motion in some medium, there must be uncanceled potential differences that manage to survive within the fluid. In a highly conducting medium, there is a simpler amplification mechanism for the field that acts even without recourse to a dynamo. The magnetic field appears to move as if "frozen" into the medium, the spatial energy density of the magnetic field precisely following the fluid density. So if the density increases locally, so does the magnetic field.

To see this, assume that σ is the value usually quoted for the conductivity of stellar plasma, $>10^{15}$ s^{-1}. This conductivity is large enough to ensure that the magnetic diffusion term effectively vanishes. Then,

$$\frac{d\mathbf{B}}{dt} = -\mathbf{B}\nabla \cdot \mathbf{v}. \tag{16}$$

From the continuity equation, we have

$$\frac{d\rho}{dt} = -\rho\nabla \cdot \mathbf{v}. \tag{17}$$

Here, we have used the convective, or co-moving, derivative: $d/dt = \partial/\partial t + \mathbf{v} \cdot \nabla$. The magnetic flux is therefore simply a scalar multiple *f* of the mass density, or $B \approx fp$. Thus, if the density is locally increased, so is the magnetic field strength. The magnetic field seems to move with the fluid, hence the appellation "frozen."

Now imagine that there is a small deviation from perfect conductivity. As the fluid moves, there will be some slippage of the mass through the field. This appears to change the magnetic field strength in the co-moving fluid. At the same time, the fact that the field is changing induces the formation of a potential difference which, in a finite conductivity environment, induces the generation of a current. All of this is at the expense of the magnetic fluid. The field consequently decreases in local strength and will do so everywhere throughout the fluid in time. In fact, the form of the magnetic field decay is the same as that of the heat equation, so the field can be said to be diffusively lost. The energy simply goes over into heat, since the field generates dissipative currents that lose their energy through collisions throughout the fluid, and the result is the gradual fading away of the field with time.

The characteristic timescale for the decay of the field depends on the scale length for the field generation and the dissipation scale for the currents. To see this, look at the dimensionless form of the dynamo equation, but now ignore the effects of the fluid motions. The equation is linear in the field strength, which means that we can scale it by any arbitrary value of the field at any time. The change with time is related to the second derivative in space, so that we can relate the timescale for the change in the field strength τ to the length scale L by

$$\tau = \frac{4\pi L^2}{\eta} \to 4\pi L^2 \frac{\sigma}{c^2}. \tag{18}$$

Again, notice that in the case of an infinitely conducting medium the time for the decay is infinitely long. We have used only one possible representation for the magnetic diffusion coefficient, however, and we shall shortly see that this is one of the longer estimates.

The field throughout a magnetized body will therefore decay with time, unless it is regenerated or the medium undergoes steady collapse at a rate that is rapid compared with the decay rate. Eventually, regardless of the artifice, the body will be unable to prevent the dissipation of its internal field without constant mechanical, that is dynamo, input.

C. Building Dynamos: Differential Rotation and Turbulence

The simplest way to imagine that a cosmic body will generate a magnetic field is if it is rotating. Since most plasmas have high conductivities, the amplification of an external field by collapse or compression (like shocks), coupled with rotation should generate a strong Lorentz force. This field is not, however, supported for indefinite lengths of time. Instead, Cowling's theorem—also known as the *antidynamo* theorem—states that no stationary (time-independent) axisymmetric dynamo is possible. But shear is essential to breaking the spherical symmetry.

The symmetry is further broken, and the effect of the rotation translated into a poloidal field, through the combined action of circulation and turbulence. An initially axisymmetric field is sheared by differential rotation, and if it is initially cylindrical (B_z) or poloidal (B_r, B_θ), then an azimuthal field (B_ϕ) results. Here r and θ are the radius and latitude, respectively. A poloidal field results from a toroidal potential field, $\mathbf{B}_\mathrm{p} = \Delta \times \mathbf{A}_\phi$, so that the toroidal magnetic field results from a distortion of the poloidal field. Finally, to convert the toroidal field back into a toroidal potential, some additional symmetry breaking is required. Turbulence in a rotating medium has vorticity, or handedness, which is parallel to the local angular-velocity vector and neither radial nor even hemispherically symmetric.

In an electrically conducting fluid, buoyant turbulent cells produce a helical twist to the toroidal field and induce a poloidal conversion. This is the basis of the $\alpha-\omega$ dynamo model. The electromotive force is $\mathscr{E} = \alpha\mathbf{B}$, where α is related to the velocity correlation function and essentially measures the amplitude of velocity fluctuations in the fluid. In a fluctuating medium, the velocity breaks into a mean component $\mathbf{V}$ plus a fluctuating part $\mathbf{u}$ (which has a vanishing mean value, but for which $\langle u^2 \rangle$ does not vanish). Here the brackets $\langle \cdots \rangle$ represent ensemble averages over the turbulent spectrum of the eddies in the fluid. The magnetic-field evolution depends on both the mean field $\mathbf{B}$ and the fluctuating part $\mathbf{b}$, and the dynamo equation becomes

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{V} \times \mathbf{B}) + \nabla \times \langle \mathbf{u} \times \mathbf{b} \rangle + \eta \nabla^2 \mathbf{B}, \tag{19}$$

where $\mathscr{E} = \langle\, \mathbf{u} \times \mathbf{b} \rangle = \alpha\mathbf{B}$. Thus α represents the fluctuations in the fluid, and describes schematically the way that this feeds back into the magnetic field strength. The evolution of the fluctuating part of the field is given by

$$\frac{\partial \mathbf{b}}{\partial t} = \nabla \times (\mathbf{u} \times \mathbf{B} \times \mathbf{V} \times \mathbf{b}) + \nabla \times (\mathbf{u} \times \mathbf{b}\, \langle \mathbf{u} \times \mathbf{b} \rangle) + \eta \nabla^2 \mathbf{b}. \tag{20}$$

Then $\alpha \sim l_0 \, \langle u_2 \rangle / \eta$ and therefore depends on the velocity fluctuation spectrum. The turbulence therefore controls the small-scale structure, and the differential rotation (shear) controls the large-scale, ordered field and provides the symmetry breaking necessary to generating the dipole.

D. The Dynamo Number and Scaling Relations

A schematic estimate of the strength of the dynamo components, and an approximate scaling law, results from the quantitative side of this picture. Differential field stretching causes poloidal to toroidal conversion, which takes place at a rate $\delta v_\phi / L$. Vorticial motion of rising convective eddies transforms toroidal to locally poloidal field at a rate Γ, which is a pseudo-scalar quantity whose sign depends on the hemisphere. The dynamo equations simplify by dimensional analysis. For the poloidal field, which is given by a vector potential field,

$$\mathbf{B}_\mathrm{p} = \nabla \times \mathbf{A}_\phi \to \frac{A_\phi}{L}. \tag{21}$$

The rate at which the poloidal field dissipates by turbulence and diffusion is, in the steady state, balanced by the conversion of toroidal to poloidal flux:

$$-\eta \nabla^2 \mathbf{A}_\phi \approx \Gamma \mathbf{B}_\phi \to -\eta \frac{A_\phi}{L^2} \sim -\Gamma B_\phi. \tag{22}$$

Finally, the poloidal field is wrapped to form the azimuthal field at a rate that depends on the shear $\delta\Omega$

$$-\eta \nabla^2 \mathbf{B}_\phi \approx \nabla \times (\nu_\phi \times \mathbf{B}_\mathrm{p}) \to -\eta \frac{B_\mathrm{p}}{L^2} \sim \frac{\delta \nu_\phi}{L} B_\phi \sim \frac{\Delta\Omega}{B_\mathrm{p}}. \tag{23}$$

The dimensionless number

$$N_{\mathrm{D}} = \frac{\Gamma \Delta \Omega L^2}{\eta^2}, \tag{24}$$

called the *Dynamo Number*, serves as the scaling parameter for the generation by the α—Ω process. Notice that N_{D} is independent of the magnetic field, and that in the steady-state case it is of order unity. For large values, the field will not be steady (what is usually meant by an active dynamo).

Observationally, it appears that X-ray emission from late-type stars correlates with a slightly different measure of the dynamo activity. This is the Rossby number, which actually measures the convective (buoyant) timescale compared with the rotation rate: $\mathrm{Ro} = \Omega\tau_{\mathrm{c}}$, where τ_{c} is the convective turnover time for some large depth in the stellar envelope. When Ro is large, either the rotation is very rapid or the turnover line is large (the gravity is low and the stellar envelope is very distended). Empirically, UV emission line strengths and L_{XR} correlate with Ro, although the precise law is still debated. It seems reasonable that the Rossby number measures at least something of the α–ω dynamo activity. The Γ factor in N_{D} is a measure of the vorticity in the buoyant cells, so depends on both Ω and τc. The turbulent diffusion coefficient η scales as $L^2\tau_{\mathrm{c}}^{-1}$ for some characteristic length L. However, the determination of an empirical scaling relation, drawn as it is from very indirect measures of the magnetic field strength and its rate of generation and structure, does not provide a serious constraint on the dynamo theories. In general, the most that can be said from the measurement of activity in late-type stars is that a dynamo must be acting, that the field must be constantly emerging anew at the photosphere, and that the mechanism must depend on surface gravity, effective temperature, and rotational frequency.

E. Planetary Magnetic Fields as Tests for Dynamo Mechanisms

Planetary magnetic fields are the only ones for which it is possible to obtain direct *in situ* measurements of the detailed field configuration and time variability. The terrestrial field is probably the best studied. It displays complex time-variable structure; most important are the long-term global polarity reversals on timescales between 10^4 and 10^6 years. These appear to be random and have been modeled as chaotic. The earth's dipole field maintains its same orientation, inclined to the rotational axis by about 11°, but its higher order moments vary with different timescales, and the geometry of the field changes slowly as a result. For scaling purposes, the earth's magnetic dipole moment is 7.9×10^{30} G cm^3 and polar field strength is about 0.6 G.

With the exception of Mercury, Mars, and Venus, all of the planets studied by spacecraft have strong magnetic fields. Neptune and Uranus have highly inclined fields, similar to those seen in the chemically peculiar stars, and also have strong nondipole components. The Jovian field is about 4 G, it has a dipole moment about 500 times that of Earth, it is inclined similarly (about 9.5°). These planets also display polar auroral rings that strikingly resemble those on Earth, which can be imaged in the UV and optical from Earth orbit with HST. These clinch the link between trapped magnetospheric high energy particles and ionization of the upper atmosphere and support other measurements of the magnetic field geometry and structure. Saturn is the anomaly. Although its field is about the same strength as Earth's (0.2 G), and its magnetic dipole moment is 4.6×10^{28} G cm^3, it is almost perfectly aligned (about a one-degree inclination to the rotation axis). It is also interesting to note that no changes in the field configuration or strength were detected for either Jupiter or Saturn between the Pioneer and Voyager flybys of these two planets. The groundbased radio measurements of Jupiter, while not very sensitive to the precise field strength, are also consistent with little change in the dominantly dipole geometry through the 1980s. It is now possible to compare the previous *in situ* measurements of the Jovian field with those from the Galileo probe, since these sample a timescale of nearly 20 years. Cassini will provide similar information for Saturn beginning in 2004. Uranus and Neptune have similar magnetic fields, both being highly inclined ($>40°$) and with polar field strengths of order 0.2 G. The extreme offset required for the Neptune field, discovered by Voyager 2 in the 1989 flyby, makes this planet an especially interesting test case for dynamo models and seems to indicate the presence of a complicated field geometry, more complex than observed for any of the other planets.

The absence of a strong field on Venus, despite its otherwise terrestrial bulk properties, is probably consistent with the dynamo mechanism. The planet rotates about a factor of 250 times more slowly than Earth. Mercury rotates slowly and is too small to support a strong convective core, but it does have a very detectable dipole moment of 2.4×10^{22} G cm^3. Its field is very small, about 0.002 G, and is nearly aligned and probably a relic from the earlier stages of planetary evolution. Mars rotates with nearly the same period as Earth, but it is smaller and may only support a very small convective core. Mars has displayed vulcanic activity in the past, evidence for core or mantle convection, but the planet does not possess even a very weak intrinsic magnetic field.

Polarity reversals were observed in the sun in the first decades of the twentieth century, long before they were recognized in the earth. The explanation for the reversal of the terrestrial field, therefore, lends support to the dynamo explanation for the magnetic fields in late-type stars and serves to link the study of dynamo processes across a large scale of mass and size of cosmic objects.

F. Galactic-Scale Dynamos

The formation of a large-scale field by dynamo action requires the conversion of toroidal magnetic fields to poloidal configurations. The presence of turbulence and differential rotation on the galactic scale, and the requirement that these act in concert in stars to produce the observed fields, thus invite a comparison between the two environments. While the length scales are much larger for galactic-size field structures, the timescales over which they are generated are also longer. Thus the scaling of the dynamo process to the interstellar medium as a whole does not seem an outlandish idea and has been attempted by a number of groups. (A good critical review of this was done in 1999 by R. Kulsrud.) Most of the details of the problem are still to be understood, but in broad outline the galactic dynamo is largely indistinguishable from a stellar one; the notably differences are the planarity of the differentially rotating disk and the nonthermal nature of the turbulence and buoyancy of the medium. These models have the advantage that α can be determined from the turbulence of the diffuse interstellar medium and that Ω is specified by differential galactic rotation.

G. Ambipolar Diffusion

The relation between the strength of a magnetic field and the density of the ambient medium strictly holds only if the medium is highly ionized. In stellar plasmas, this is almost certainly a good assumption, as it also appears to be for large-scale extragalactic radio jets and lobes. But in interstellar clouds, where the densities may become large enough and the opacity high enough, ultraviolet radiation is effectively screened out of the cloud cores. Thus, in the densest part of a molecular cloud, the ionization is expected to be due primarily to cosmic ray penetration of the cloud, leaving ionization fractions $<10^{-5}$ in those regions where the density is $>10^5$ cm^{-3}. When the temperature in this region falls below about 100 K, it becomes locally unstable to gravitational collapse. A trapped magnetic field will be transported inward with the collapsing gas, but because of the low conductivity will not be amplified as rapidly as one would expect from flux freezing. Instead of increasing in strength as $\rho^{2/3}$ or ρ, the matter separates from the field, and models show that $B_{\text{collapse}} \sim \rho^{1/2}$. The separation of the field and the material is due to ambipolar diffusion, the tendency of ions to separate in an external magnetic field, and also the field-line slippage relative to the gravitationally contracting medium. Direct measurements support this approximate scaling relation, although models constrain the exponent n in the relation $B_{\text{collapse}} \sim \rho^n$ to lie only between about $\frac{1}{3}$ and $\frac{2}{3}$.

H. Magnetic Reconnection

Magnetic reconnection is presumed to occur when regions of oppositely directed polarity are brought into contact by turbulent motions. This mixing results in the annihilation of the field within a small volume and the transformation of magnetic energy into kinetic energy through the generation of extremely strong local electric fields within very small-scale (of order 100 to 1000 km) lengths.

The rate at which the field destruction takes place is a matter of debate, and significantly affects estimates of the heating of the confining plasma. The fastest timescale, the Alfvén wave crossing time m appears to be too short (μs) to explain the acceleration observed in solar flares. These require an extended period of electric-field amplification during the annihilation phase. Theoretical particle simulations suggest that strong MHD turbulence is probably present during the reconnection phase of magnetic field line merging, and that the electrons are energized through being scattered off of these waves.

Reconnection processes have been implicated in the acceleration of particles in the terrestrial magnetotail and in the generation of disconnection events in comets, when the plasma tail of a comet separates from the coma as it crosses regions in the solar wind of oppositely directed magnetic polarity (sector boundaries). There is strong evidence, from fast observations of stellar flares, that the frequency dependence and energetics of large stellar active regions mimic those observed on the sun, where reconnection is more secure. Finally, it may be possible to see field lines merging in radio lobes and jets of external galaxies. Large-scale filamentation may be associated with this same instability.

V. FINAL WORDS

The theoretician H. C. van de Hulst has remarked that “magnetic fields are to astrophysics as sex is to psychology,” suggesting that magnetic fields are seen as the universal explicans for all cosmic phenomena. As long as measurements remained inexact and theory was only qualitative, this could certainly have been true in fact as well as in jest. The late twentieth century has seen a vast improvement in the available techniques for the measurement of magnetic-field-related phenomena, especially the introduction of high-resolution imaging polarimeters, improvements in spectrograph design and detector sensitivity, and the availability of polarimetric instruments over a broad wavelength regime going from the ultraviolet into the radio. Supercomputers have grown increasingly important in magnetohydrodynamic modeling of stellar and galactic-scale objects, and predictions are well enough refined to be testable. Plasma codes for particle simulations of reconnection and small-scale processes are reaching a mature stage and can confront many of the radio observations of solar and stellar flare processes. Radio imaging allows now for the study of structures in extragalactic sources with a dynamic range of as much as 10^5. And in the past decade, high spatial and spectral resolution space observations have finally become routine with the launch of the Hubble Space Telescope, ASTRO, FUSE, and ISO. We finally see magnetic fields as the necessary product of astrophysical processes.

SEE ALSO THE FOLLOWING ARTICLES

Binary Stars • Galactic Structure and Evolution • Geomagnetism • Interstellar Matter • Neutron Stars • Pulsars • Solar System, Magnetic and Electric Fields • Stellar Structure and Evolution

BIBLIOGRAPHY

Beck, R., Brandenburg, A., Moss, D., Shukurov, A., and Sokoloff, D. (1996). Galactic magnetism: Recent developments and perspectives. *Annu. Rev. Astron. Astrophys*. **34,** 155.

Belton, M., West, R. A., and Rahe, J. (eds.) (1989). "Time-Variable Phenomena in the Jovian System," NASA SP494.

Borra, E. F., Landstreet, J. D., and Mestel, L. (1982). *Annu. Rev. Astron. Astrophys.* **20,** 191.

Chanmugam, G. (1992). Magnetic fields in degenerate stars. *Annu. Rev. Astron. Astrophys*. **30,** 143.

Cravens, T. E. (1997). "Physics of Solar System Plasmas," Cambridge Univ. Press, Cambridge, UK.

Crutcher, R. (1999). Magnetic fields in molecular clouds: Observations confront theory. *Astrophys. J*. **520,** 706.

Hartmann, L., and Noyes, R. (1987). Rotation and magnetic activity in main sequence stars. *Annu. Rev. Astron. Astrophys*. **25,** 271.

Krause, F., and Raedler, K. H. (1980). "Mean Field Magnetohydrodynamics and Dynamo Theory," Pergamon Press, Oxford.

Kulsrud, R. (1999). A critical review of galactic dynamos. *Annu. Rev. Astron. Astrophys*. **37,** 37.

La Rosa, T. N., Kussim, N. E., Lazio, T. J. W., and Hyman, S. D. (2000). Wide-field 90 cm VLA images of the galactic center region. *Astron. J*. **119,** 207.

Moffatt, H. K. (1978). "Magnetic Field Generation in Electrically Conducting Fluids," Cambridge Univ. Press, Cambridge.

Morris, M., and Serabyn, E. (1996). The galactic center environment. *Annu. Rev. Astron. Astrophys*. **34,** 645.

Myers, P. C., Goodman, A. A., Gusten, R., and Heiles, C. (1995). Observations of magnetic fields in diffuse clouds. *Astrophys. J*. **442,** 177.

Parker, E. N. (1979). "Cosmical Magnetic Fields," Oxford Univ. Press. London.

Priest, E. R. (1984). "Solar Magnetohydrodynamics," Reidel, Dordrecht, The Netherlands.

Priest, E. R., and Forbes, T. (2000). "Magnetic Reconnection: MHD Theory and Applications," Cambridge Univ. Press, Cambridge, UK.

Verschuur, G., and Kellerman, K., ed. (1988). "Galactic and Extragalactic Radio Astronomy," 2nd ed. Springer-Verlag, New York.

Zel'dovich, Y. B., Ruzmaikin, A. A., and Sokoloff, D. D. (1983). "Magnetic Fields in Astrophysics," Gordon & Breach, New York.

Zwaan, C. (1987). Elements and pattern in the solar magnetic field. *Annu. Rev. Astron. Astrophys*. **25,** 83.

Millimeter Astronomy

Jeffrey G. Mangum
National Radio Astronomy Observatory

GLOSSARY

Correlator An electronic device that performs the multiplication and averaging of the astronomical signals from a radio telescope. Correlators are used both to process spectral measurements from single antenna telescopes and to process the signals from a collection of electronically connected telescopes.

Interferometer A collection of electronically connected telescopes used to make high spatial resolution astronomical measurements.

Isomerism The existence of more than one substance having a given molecular composition and mass but differing in structure. For example, HNC is an isomer of HCN.

Jansky Unit of flux density used in radio astronomy. 1 jansky (Jy) is equal to 10^{-26} W m^{-2} Hz^{-1}.

MILLIMETER ASTRONOMY involves the study of astrophysical phenomena through observations at wavelengths from about 0.3 to 4.5 mm (frequencies from about 65 to 900 GHz). Millimeter astronomical investigations are conducted within a wide range of astronomical categories, including planetary astrophysics, studies of the interstellar medium and its inhabitants, and investigations of the properties of external galaxies. Millimeter astronomical observations use the measurements of the emission from dust grains and the spectroscopic signature from interstellar molecules to study the physical, chemical, and dynamical state of the universe.

I. MILLIMETER ASTRONOMICAL OBSERVATORIES

There are two basic types of millimeter astronomical observatory: single antenna facilities and interferometers. Both can be operated either on the surface of the earth or as space-based observatories, although to date only single antenna measurements have been conducted from satellites. Both types of observatory operate in fundamentally the same way, by collecting millimeter wavelength radiation from a celestial object and processing that signal into radiant flux information about the object under study. When a group of two or more single antennas are electronically linked and commanded to observe the same celestial object, the group of antennas has a spatial resolution equivalent to a single antenna with a diameter equal to the separation of the individual elements. This allows millimeter interferometers to make astronomical measurements with very high spatial resolution not easily obtainable with single antenna millimeter telescopes.

TABLE I Millimeter Astronomical Observatories

Name	Location	Latitude (degrees)	Elevation (m)	Telescopes	Diameter (m)	λ (mm)
CSO	Mauna Kea, HI	+19	4000	1	10.4	0.3–1.0
IRAM	Pico Veleta, Spain	+37	2920	1	30	0.8–3.5
JCMT	Mauna Kea, HI	+19	4000	1	15	0.3–0.8
NRAO	Kitt Peak, AZ	+32	1914	1	12	1.0–4.5
NRO	Nobeyama, Japan	+36	1350	1	45	1.3–15.0
OSO	Onsala, Sweden	+57	24	1	20	3.0, 13.0
SEST	La Silla, Chile	−29	2347	1	15	0.9–4.3
SMT	Mt. Graham, AZ	+33	3186	1	10	0.3–1.0
UMASS	Amherst, MA	+42	314	1	14	3.0
BIMA	Hat Creek, CA	+41	1043	10	6	1, 3
NMA	Nobeyama, Japan	+36	1350	6	10	1, 2, 3
OVRO	Owens Valley, CA	+37	1216	6	10.4	1, 3
PdBI	Plateau de Bure, France	+45	2560	5	15	1.2–3.7
SMA	Mauna Kea, HI	+19	4000	8	6	0.3–0.8

Table I lists the characteristics of the currently operating millimeter single antenna and interferometer facilities in the world. Note that most of these facilities are located at high altitude sites. This is owing to the fact that millimeter astronomical observations made from the surface of the earth must contend with the earth's atmosphere. Many of the molecular constituents of the earth's atmosphere are very good absorbers of millimeter wavelength emission. Figure 1 shows the millimeter absorption spectrum of the earth's atmosphere, displayed as signal transmission as a function of frequency. The frequencies in Fig. 1 that show low transmission coincide with energy transitions from O_2 and H_2O. The atmospheric scale heights of O_2 and H_2O are approximately 8000 and 2000 m, respectively. Therefore, by locating millimeter astronomical observatories at high altitude sites, measurements made with these facilities are less adversely affected by the absorptive properties of the earth's atmosphere.

II. MILLIMETER ASTRONOMICAL OBSERVING TECHNIQUES

A. Single Antenna Techniques

The basic millimeter single antenna receiving system contains the components shown in Fig. 2. Since the astronomical signals which millimeter telescopes are trying to detect are often thousands of times weaker than the atmospheric and instrumental emission that accompanies the astronomical signal, all measurements made with single antennas involve some kind of switching.

1. Position Switching

In position switching, the telescope acquires total power measurements at the source position and at a nearby reference position by pointing the telescope and integrating on each position in succession. Since the entire telescope structure must be moved to each position, these measurements are usually made over 30–60 sec time scales. To produce the final astronomical signal, the individual source and reference measurements are differenced. Figure 3 shows a pictorial representation of a position switched observation. As long as the received signal from the instrumentation or the atmosphere has not changed significantly during the time between a source and its respective reference measurement, position switching will produce reliable astronomical measurements.

2. Beam Switching

In beam switching, the telescope acquires total power data at signal and reference positions by repositioning one of the optical components of the telescope system, usually the secondary mirror (see Fig. 3). Since the optical components are usually relatively small, they can be switched and positioned accurately at a rapid rate, often several Hz. This allows for rapid switching between source and reference positions, which often allows for better subtraction of the fluctuating emission from the earth's atmosphere, but is limited by the relatively small reference position offsets attainable with most telescope optics.

3. Frequency Switching

In this observing mode, a reference spectrum is obtained by shifting the center frequency of the source spectral

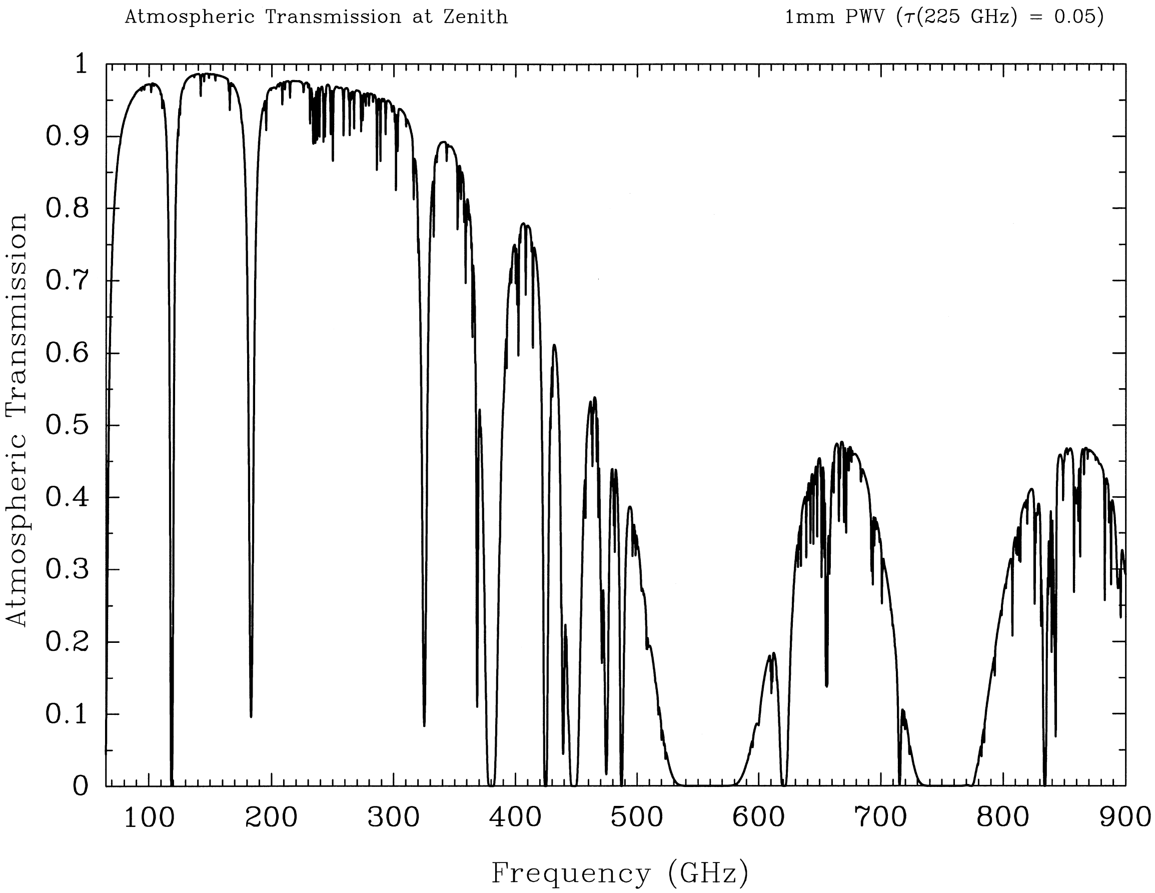


FIGURE 1 Millimeter emission spectrum from the earth's atmosphere from an elevation of 4000 m. The scaling from precipitable water vapor column to atmospheric opacity at a frequency of 225 GHz is shown in the upper right.

measurement. Since it is not necessary to move the telescope or any of the telescope optics, on-source integration time is increased through "in-band" switching. Frequency switching also alleviates the need to find an emission-free reference position when observing in a (spatially) complex emission region. Frequency switching also entails less system overhead than most other observing modes. In principle, frequency switching can be done by switching the frequency of the local oscillator (LO) or an intermediate frequency (IF) oscillator. A pictorial description of frequency switching is shown in Fig. 3. If the frequency shift is small enough, usually less than 50 MHz, the spectral line will appear in both the source and reference spectra. When the resultant spectrum is formed, the line will appear twice, once in emission and once in absorption. The spectrum can be "folded" to obtain a $\sqrt{2}$ improvement in signal-to-noise. The primary drawback of frequency switching is that the spectral baselines are generally not as good (flat) as with position or beam switching. This is because the two frequency positions each have their own spectral bandpass shapes that do not cancel in the computation of the final spectrum, thereby leaving a residual standing wave in the overlapped spectrum. A number of techniques, including focus modulation and beam peak scattering, can be used to dampen this residual standing wave.

B. Interferometric Techniques

A basic (two antenna element) interferometer is shown in Fig. 4. Measurements with millimeter interferometers are usually done by having all telescopes in the interferometer measure the millimeter signal from an object. The independent but phase-coherent signals from each antenna are then multiplied and integrated (together referred to as *correlation*), which results in interference fringes containing the amplitude and phase information of the astronomical signal. As the earth rotates, the interferometer receives varying responses to the structure in the source being measured, and in the process builds up a two (spatial) or three (spatial and spectral) dimensional picture of the astronomical source.

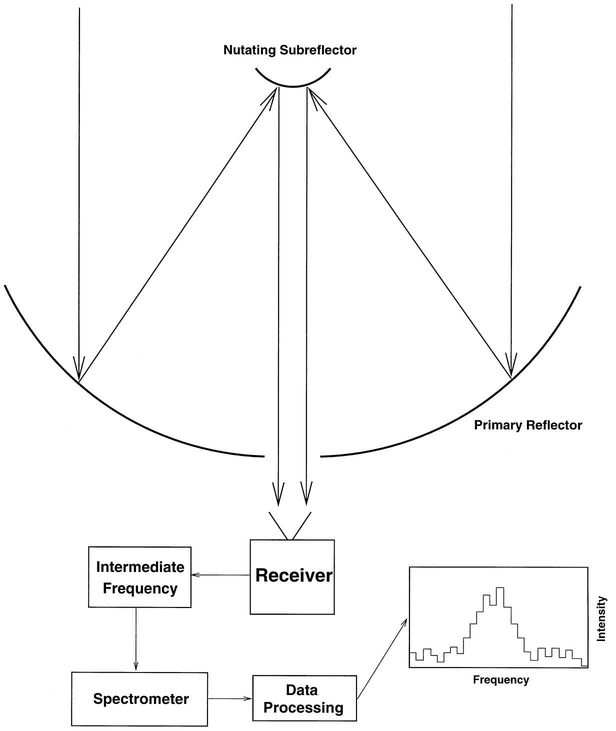


FIGURE 2 Basic components of the single antenna millimeter observatory. Following signal collection by the *primary reflector* and *nutating subreflector*, the *receiver* amplifies the astronomical signal. The *intermediate frequency* (*IF*) then downconverts the received signal to a frequency that can be processed by the *spectrometer*. The output from the spectrometer, following data processing, is astronomical data.

In principle, millimeter wavelength interferometers operate in the same way as centimeter wavelength interferometers, but with the complication that millimeter interferometers are sensitive to atmospheric effects. The varying atmosphere above a millimeter interferometer influences the received signals through absorption and by introducing fluctuations in the received phase of a measurement. Therefore, as with millimeter single antenna observations, it is advantageous to locate millimeter interferometers at high mountain sites. Atmospheric phase fluctuations produce a blurring of the image produced by a millimeter interferometer. Since these atmospheric phase fluctuations seem to depend upon the distance between the antennas, increasing the separation between the antennas in a millimeter interferometer leads to increasingly blurry images for a given level of atmospheric stability.

To correct for the detrimental affects of atmospheric phase fluctuations, a number of correction techniques have been developed. One technique uses a bright point source in the image as a fringe reference point. This technique is often referred to as *self-calibration*. Since bright point sources are not always available within a given image, techniques which simultaneously measure the atmospheric emission at wavelengths which provide information regarding the amount of water in the Earth's atmosphere have been developed in recent years. These atmospheric phase correction schemes have proven to be quite successful, allowing for phase corrections which improve the phase noise in a measurement by factors of 2–3. Future improvements in these phase correction systems will likely yield higher levels of improvement to the received astronomical signals.

III. SCIENCE AT MILLIMETER WAVELENGTHS

A. Star Formation

1. Molecular Clouds

As early as the mid-1920s, optical astronomers identified regions in our galaxy that possessed moderate to high levels of visual extinction. Simple interstellar molecules, such as CH, CH^+, and CN, were discovered in these regions by optical astronomers as early as 1937–1941, indicating the presence of an interstellar medium containing regions with densities large enough to support the formation and existence of molecules. The discovery of complex molecules had to await the advent of radio astronomy. In 1968, water (H_2O) and ammonia (NH_3) were discovered toward the regions identified by optical astronomers as having high visual extinction. The existence of these molecules indicated the presence of dense, cool, opaque regions that were eventually recognized as the sites of star formation.

Within the general evolution of spiral galaxies, the star formation process is catalyzed by the continued condensation of gas and dust from the interstellar medium. In this *molecular cloud* phase, vast regions of the interstellar medium are characterized by high densities, low temperatures, and kinematic motions indicating the existence of infall and outflow. Our galaxy contains about $10^9\ M_\odot$ of molecular gas. The majority of this gas is located in giant molecular clouds containing approximately $10^{4-6}\ M_\odot$ of material. A cloud with support only through kinematic motions of the gas will collapse upon itself if the mass exceeds the Jeans mass,

$$M_J = \left(\frac{\pi k T_\mathrm{K}}{\mu_{m_\mathrm{H}} G}\right)^{1.5} \rho^{-0.5},$$

while a molecular cloud without pressure support will free-fall collapse on a timescale given by

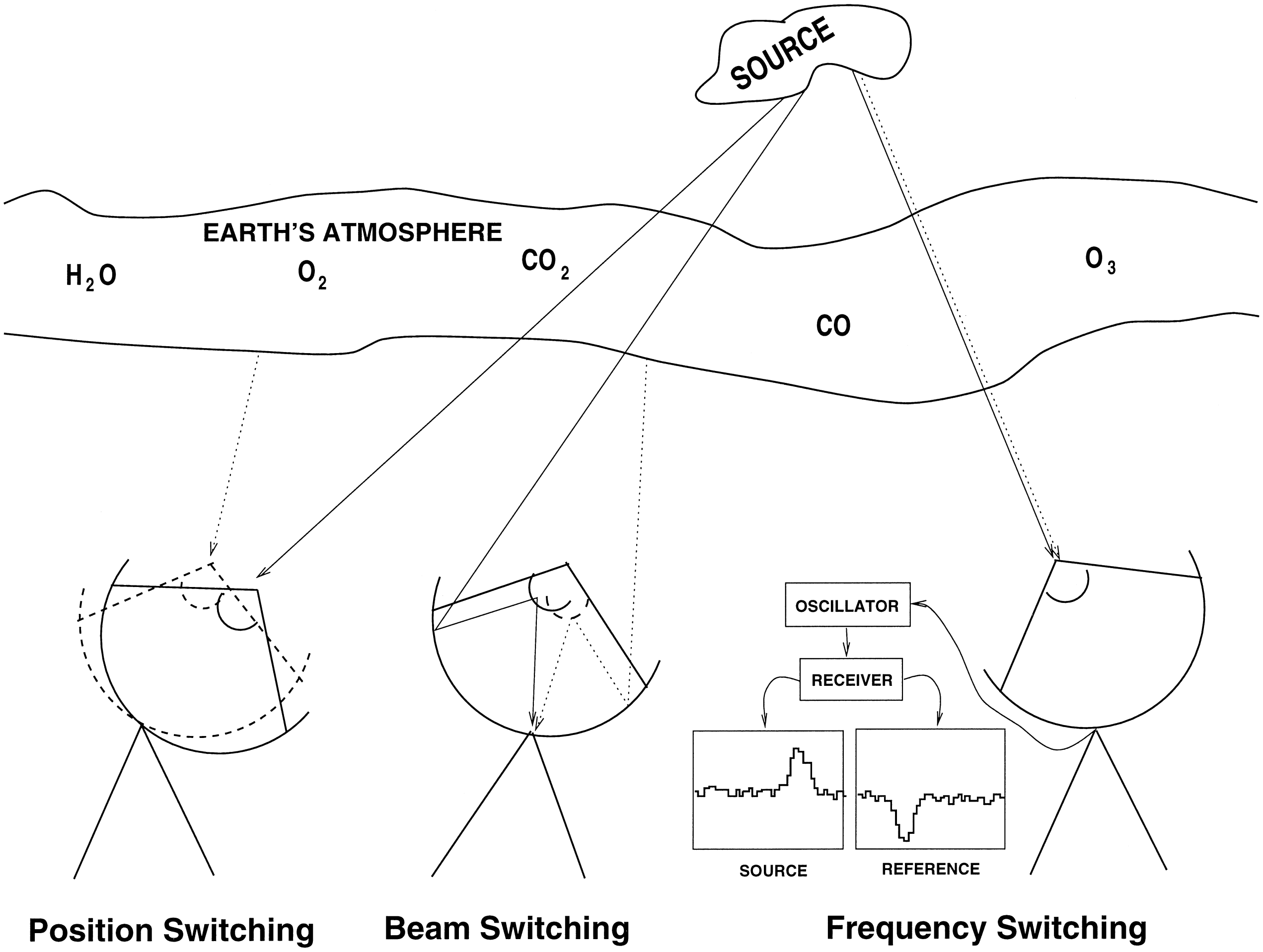


FIGURE 3 Graphical description of the position, beam, and frequency switching observing techniques.

$$t_{\mathrm{ff}} = \left(\frac{3\pi}{32G\rho}\right)^{0.5},$$

where k is Boltzmann's constant (1.380658×10^{-16} erg K^{-1}), T_{K} is the kinetic temperature of the gas (K), μ is the mean mass per particle (which is equal to 2.29 in a cloud which is 100% molecular composed of 25% He by mass), m_{H} is the mass of a hydrogen atom (g), G is the gravitational constant (6.67259×10^{-8} cm^3 g^{-1} sec^{-2}), and ρ is the mass density (g cm^{-3}). Since the moderate to high densities and low temperatures found in these giant molecular clouds indicate that most of them should be collapsing, indicating a star formation efficiency that does not agree with the observed star formation rate, there must be some form of additional support in these regions.

Two mechanisms have been considered as candidates for the additional structural support for molecular clouds: turbulence and magnetic fields. The theoretical and observational investigations into the importance of each of these mechanisms in the physical evolution of molecular clouds currently reach inconclusive results regarding the importance of each.

2. Measurements of Physical Conditions

The study of molecular cloud stability and evolution leads naturally to studies of the physical and chemical evolution of the star formation process. Fundamental to this study of the star formation process is the characterization of the physical conditions in the gas and dust comprising these regions. For the gas, volume density n (cm^{-3}), kinetic temperature T_{K} (K), chemical composition X, turbulent motion Δv (km sec^{-1}), and magnetic field strength B (Gauss) are fundamental physical quantities. For the dust, the dust temperature T_{d} (K), dust volume density n_{d} (cm^{-3}), and dust opacity κ describe the physical conditions representative of the dust component of a molecular cloud. Note that all of these quantities are dependent upon time and position.

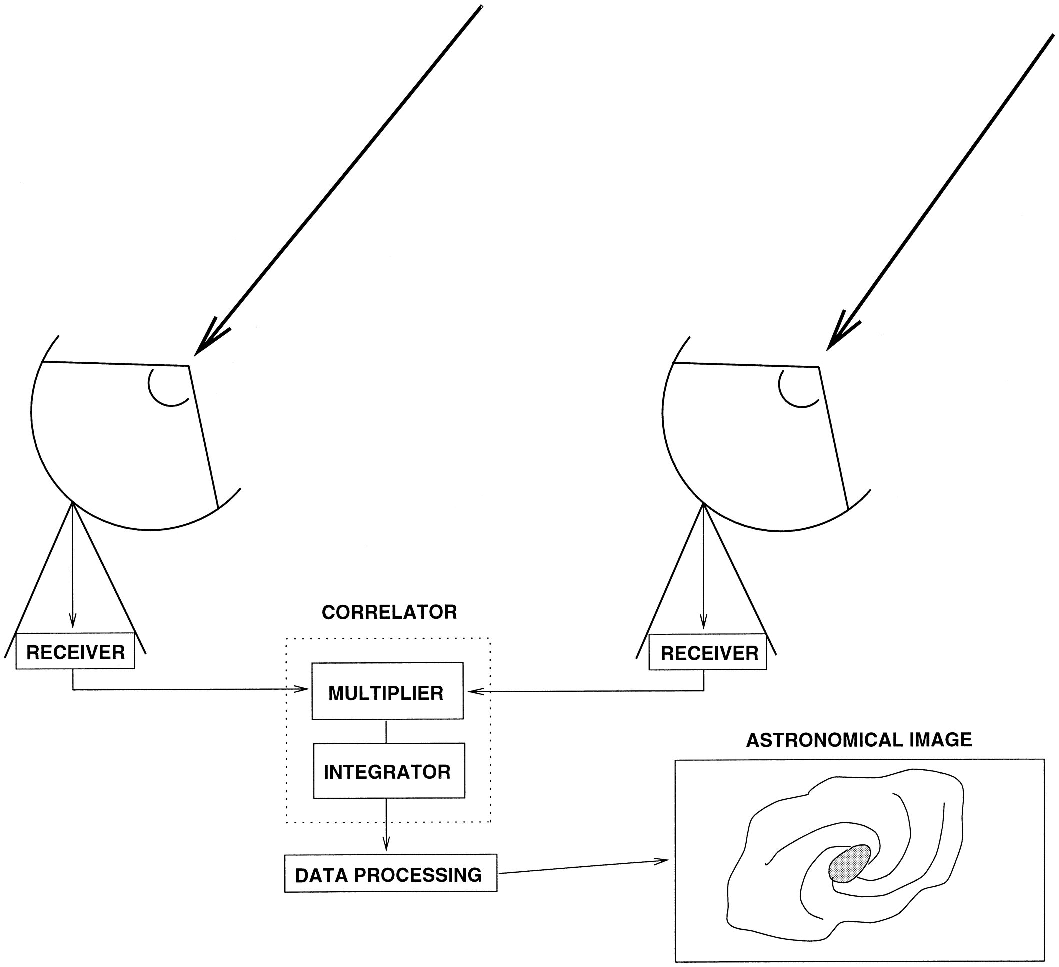


FIGURE 4 A basic interferometer. After the signals from an astronomical object are received and amplified by the two antennas, they are combined in *a correlator*. The *correlator* multiplies and integrates the two signals, yielding interference fringes that contain the correlated amplitude and phase information from the astronomical source. This information is then processed in a computer to produce an astronomical image.

Most of the material in molecular clouds is in the form of H_2, which owing to its lack of a permanent dipole moment has no easily observable rotational transitions. It can be observed through rovibrational and fluorescent transitions, but only within environments which are very specific, such as shocks and regions containing high levels of ultraviolet emission. Therefore, the principal component of molecular clouds is effectively unmeasurable. This fact forces astronomers to use trace constituents, other molecules and dust, to measure the physical conditions in molecular clouds.

a. Molecular emission as a tracer of physical conditions. The primary constituent of molecular clouds, H_2, is also the main collision partner with other molecular inhabitants of these regions. These collisions lead to the excitation of rotational transitions in a variety of molecules, many of which emit at observable millimeter wavelengths. The most abundant molecule after H_2 is carbon monoxide ($^{12}C^{16}O$, usually simply written CO). It was the first molecule discovered at millimeter wavelengths by Wilson, Penzias, and Jefferts in 1970 using the National Radio Astronomy Observatory 36 ft (now 12 m) millimeter telescope located on Kitt Peak, Arizona. It has been used extensively as a probe of the volume density and kinetic temperature in molecular clouds through measurements of its lowest two rotational transitions at 115.271 and 230.538 GHz. CO has proven to be a very good tracer of the global physical conditions in molecular clouds, but for more compact regions with a larger number of particles along the line of sight [referred to as the *column density* (N) of a particular molecule], it loses its sensitivity to the bulk of the gas as the opacity in the measured transitions rises. Fortunately, there are other less abundant molecular tracers, including isotopomers (isotopic variants) of CO, such as ^{13}CO, C^{18}O, and ^{13}C^{18}O, which prove to be better probes of these high column density environments.

There are a wide variety of molecules that can be used as tracers of the volume density and kinetic temperature in molecular clouds. The choice of molecular probe depends upon what environment one wishes to study. For example, to measure the physical conditions in the dense cores of molecular clouds, it is best to choose a molecular tracer that is particularly sensitive to the prevalent conditions in this environment. A useful guide used to calculate the sensitivity of a transition to volume density is the *critical density* n_{crit}, which is the volume density required to collisionally excite a transition assuming optically thin conditions,

$$\begin{aligned} n_{\text{crit}} &= \frac{A_{ij}}{C_{ij}} \\ &= \frac{64\pi^4 \nu_{ij}^3}{3hc^3 g_i C_{ij}} |\overrightarrow{\mu_{ji}}|^2 \\ &= \frac{64\pi^4 \nu_{ij}^3}{3hc^3 g_i C_{ij}} S\mu^2, \end{aligned}$$

where A_{ij} is the spontaneous emission (Einstein A) coefficient for level i, C_{ij} is the collisional deexcitation rate per molecule in level i, g_i is the upper state degeneracy, $|\overrightarrow{\mu_{ji}}|$ is the dipole moment matrix element for the transition, S is the line strength for the transition, μ is the dipole moment for the molecule, and the other terms have their usual meanings. Critical densities for common molecules such as CS, HCN, and H$_2$CO are in the range 10^{4-8} cm^{-3} for a kinetic temperature of 10 K.

Therefore, a simple detection of a transition from one of these molecules implies the existence of dense gas. A second consideration is to choose molecules that allow one to derive accurate measures of the volume density and kinetic temperature in a molecular cloud. Since the collisional excitation of molecular transitions is dependent upon the coupled effects of volume density and kinetic temperature, it is often necessary to use molecules that allow one to decouple these effects. The ability to decouple these physical effects depends upon the properties of the molecular structure. There are three basic types of molecules in this regard; linear, symmetric rotor, and asymmetric rotor. Figures 5–7 show the energy level structure for these three types of molecules. As can be seen from Fig. 5, linear molecules have one ladder of energy levels, the transitions between which are excited by the coupled effects of volume density and kinetic temperature. In general, linear molecules are used to derive the volume density in a molecular cloud by assuming a kinetic temperature or by using a calculation of the kinetic temperature based on measurements of the transitions from another molecule. The energy level structures for the symmetric and asymmetric rotor molecules shown in Figs. 6 and 7 indicate a more complex structure. Like linear molecules, the strengths of transitions within a given ladder (designated by the "K" rotational quantum numbers) are dependent upon the coupled effects of volume density and kinetic temperature. A comparison of the strengths of transitions from the same J levels but from different K ladders, though, is dependent only on the kinetic temperature, thus making it possible to derive a decoupled measurement of the kinetic temperature in a molecular cloud. In general, then, symmetric and asymmetric molecules have molecular level properties that, when the appropriate transitions are compared, allow *decoupled* measurements of the volume density and kinetic temperature in a molecular cloud. Linear molecules do not possess these decoupling properties, thus requiring an independent measurement of either the volume density or kinetic temperature.

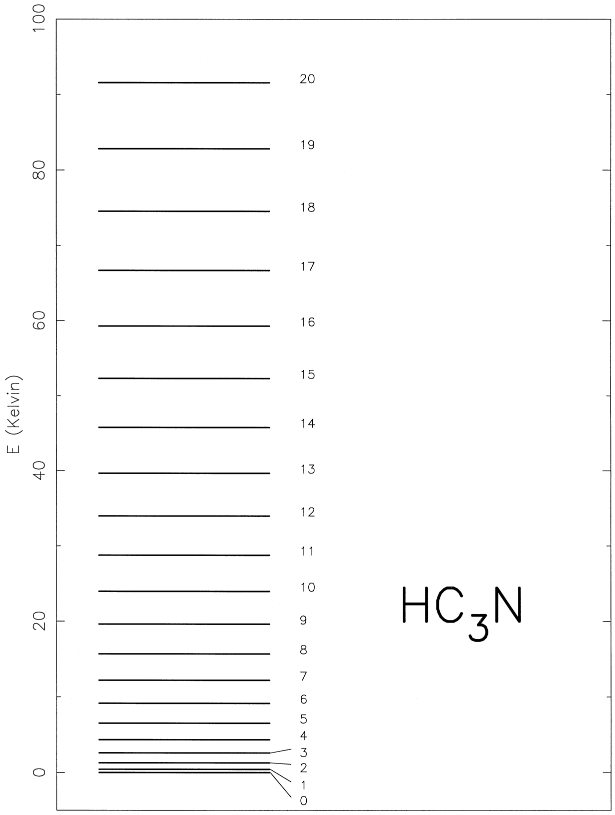


FIGURE 5 Rotational energy level diagram for HC$_3$N, a typical linear molecule. The rotational quantum number "J" and associated energy above the ground state are indicated for each level.

By comparing the measured intensities of a variety of transitions from a given molecule with molecular line intensity predictions from a molecular cloud model,

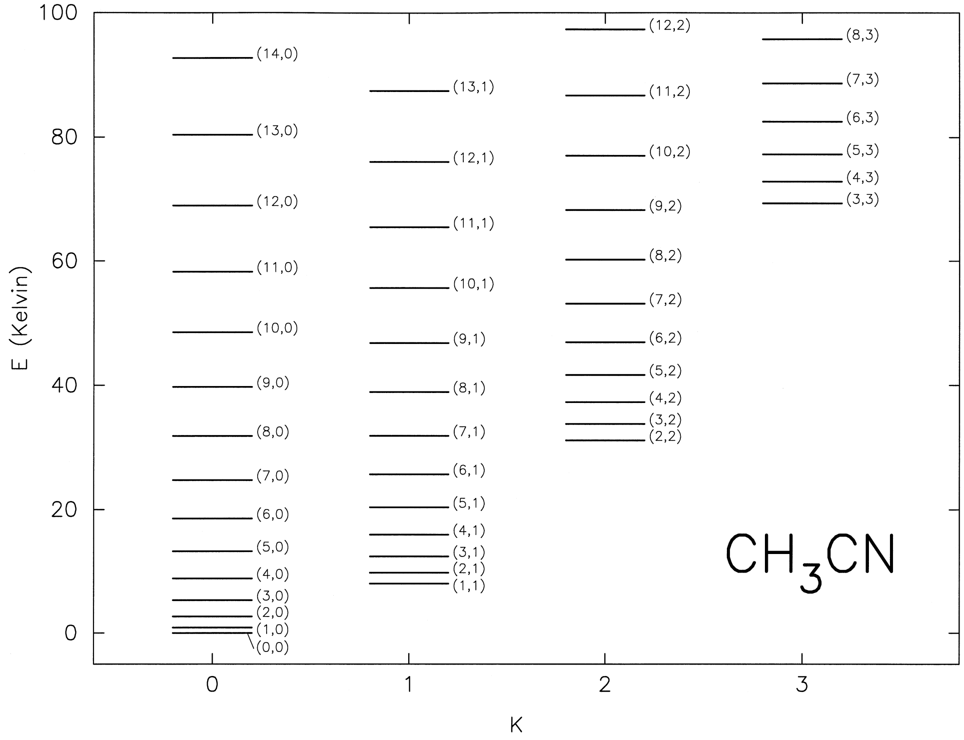


FIGURE 6 Rotational energy level diagram for CH_3CN, a typical symmetric rotor molecule. The rotational quantum numbers J and K, along with the associated energy above the ground state, are indicated for each level.

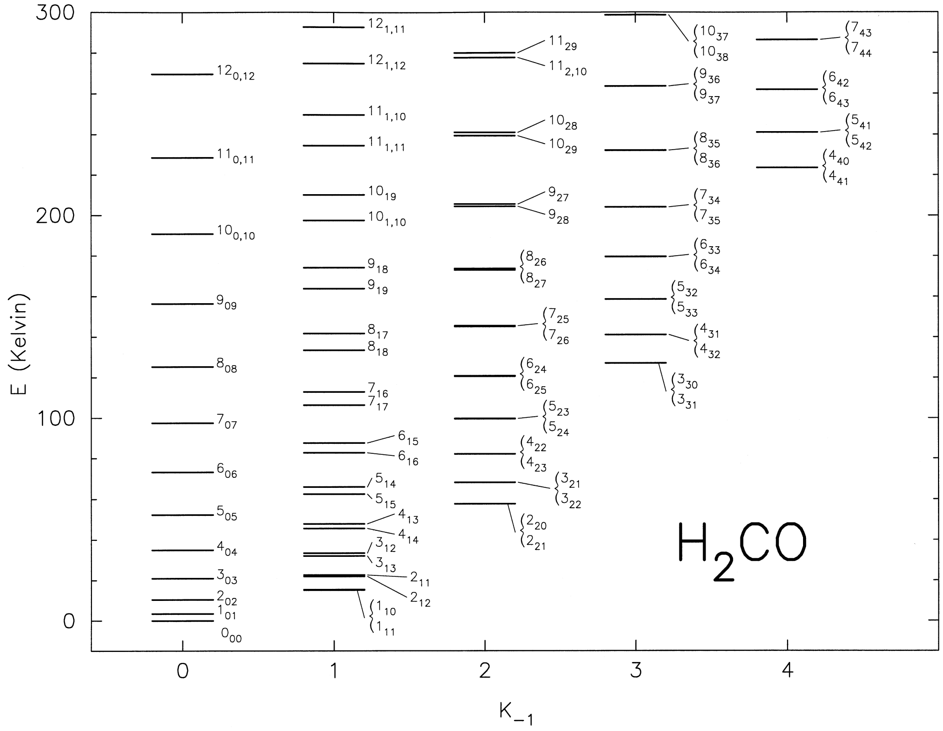


FIGURE 7 Rotational energy level diagram for H_2CO, a typical asymmetric rotor molecule. The rotational quantum numbers J, K_{-1}, and K_{+1}, along with the associated energy above the ground state, are indicated for each level.

estimates of the volume density and kinetic temperature within a molecular cloud can be made. In general, these estimates reveal that the volume densities in the regions where stars form within molecular clouds exceed 10^4 cm^{-3}, while the kinetic temperatures range from 10 to 300 K. These models also indicate that many molecular cores possess density gradients, suggestive of a structure that could evolve into a collapsing protostar.

b. Kinematics. The shape of a spectral line is determined by the radial velocity structure along the line of sight through a molecular cloud. The measured widths of spectral lines are generally larger than the thermal width, indicating that the velocity fields within molecular clouds are dominated by Doppler broadening owing to turbulence:

$$\begin{aligned}\Delta v &= v_{\text{therm}} + v_{\text{turb}}\\ &= 2\sqrt{\frac{2\ln 2kT_{\text{K}}}{m}} + v_{\text{turb}}\end{aligned}$$

where Δv is the full width at half maximum of the spectral line, T_{K} is the kinetic temperature of the gas, and m is the mass of the particles which make up the molecular cloud (principally molecular hydrogen). Unfortunately, a detailed derivation of the spectral line shape has proven elusive, owing to the effects of kinetic temperature and volume density gradients, spatial structure, and radiative transfer effects within the molecular cloud. Measurements of the line center velocity as a function of position over a molecular cloud do indicate that, in general and on large scales, they are neither collapsing nor rotating. This is not the case on small scales. Evidence for rotation and collapse of molecular cloud cores on 0.1 pc scales have yielded interesting clues to the details of the star formation process. Measurements of cloud core rotation indicate that in magnitude it is only 2% of the gravitational potential energy before collapse, making it relatively unimportant to the overall dynamics of a molecular cloud core.

The physical nature of the turbulent component of a spectral line in a molecular cloud is currently a source of considerable debate. Physical processes that have been suggested as sources of the turbulence in molecular clouds are expanding HII regions, supernova remnants, cloud–cloud collisions, galactic differential rotation, and stellar winds. Unfortunately, for all of these processes there are theoretical problems with coupling the energy produced into turbulence.

c. Magnetic fields. An understanding of the magnetic field properties of molecular clouds is an important aspect of the overall physical understanding of molecular clouds given their apparent role in providing dynamical support in these environments. There are three methods that have been used to measure the magnetic field strength and direction within molecular clouds: atomic and molecular Zeeman effect splitting, which tell us about the line-of-sight magnetic field component (B_{Z}); polarization of the emission from dust grains, which gives us information about the component of the magnetic fields perpendicular to the line-of-sight ($B_{\perp}$); and measurements of spectral line emission polarization, which also tells us about $B_{\perp}$. Measurements of the Zeeman splitting in atoms and molecules have concentrated on studies of HI, OH, and CN, yielding typical values for B_Z of 10–20 μG within regions with volume densities of approximately 10^3 cm^{-3}. At higher volume densities, magnetic fields as large as 700 μG have been measured. Millimeter dust continuum emission polarization levels of at most a few percent have shown that the magnetic field within the high volume density ($n \geq 10^6$ cm^{-3}) cores of molecular clouds is perpendicular to the major axis of the high density structures (such as disks) and parallel to the outflows associated with these objects. Although the possibility of there being measurable levels of polarization of thermal millimeter spectral line emission has been known for years, it has only recently been detected. The percentage of measured polarized emission is equivalent to that detected through millimeter continuum polarimetry.

Turbulence and magnetic fields cannot simply be applied as independent solutions to the problem of cloud support since turbulence should tangle magnetic fields, thus reducing their effectiveness as a source of support. The theory of magnetohydrodynamic turbulence within molecular clouds has shown that magnetic fields should slow the decay of turbulent motions if these motions are less than the propagation speed along the magnetic field lines, referred to as the *Alfvén velocity*. However, the stability of magnetic support in the presence of turbulence has been called into question, and the interplay between cloud stability and dynamics drives our understanding of the importance of magnetic fields as a means of molecular cloud support. Future measurements of the magnetic field and direction in molecular clouds should clarify their influence on the overall dynamics and evolution of these regions.

3. Dust Emission as a Tracer of Physical Conditions

Dust is also a viable probe of physical conditions in molecular clouds through attenuation measurements at ultraviolet through near-infrared wavelengths and by using measurements of its emission at far-infrared through radio wavelengths. As noted earlier, extinction of background starlight can be used as a probe of the dust column density along the line of sight. Studies of the extinction of optical starlight probe molecular clouds at low extinctions ($A_{\mathrm{v}} \leq 5$), thus giving information about the outer skins of molecular clouds. Recently, surveys of the near-infrared extinction have probed much deeper into molecular clouds, as much as $A_{\mathrm{v}} \sim 30$ magnitudes. These near-infrared surveys have revealed much regarding the inner structure of molecular clouds.

Dust also reveals its presence through emission at millimeter wavelengths. That the opacity of dust continuum emission possesses a power-law dependence on wavelength,

$$\kappa_\nu \propto \nu^\beta,$$

where $\beta \sim 1{-}2$, allows that dust emission at long wavelengths can trace large column densities in molecular clouds. Millimeter emission from dust can therefore probe larger column densities in molecular clouds, as deeply as $A_{\mathrm{v}} \sim 10^4$ magnitudes. Since the dust emission depends upon the physics of the dust as follows:

$$S_\nu = \frac{2h\nu^3\Omega_{\mathrm{s}}}{c^2}\left\{\frac{1-\exp\left[-\tau_0\left(\frac{\nu}{\nu_0}\right)^\beta\right]}{\exp\left(\frac{h\nu}{kT_{\mathrm{D}}}\right)-1}\right\},$$

where S_ν is the emission flux at frequency ν, Ω_{s} is the source solid angle, τ_0 is the optical depth at frequency ν_0, β is the dust emissivity power law, and T_{D} is the dust temperature, an independent estimate of the kinetic temperature within a molecular cloud can be obtained assuming that the gas and dust are coupled ($T_{\mathrm{D}} \simeq T_{\mathrm{K}}$). The column density and mass of a molecular cloud can also be inferred from dust emission measurements as follows:

$$N = N_\lambda\left(\frac{\lambda}{\lambda_0}\right)^\beta \tau_\lambda,$$

$$M = \frac{C_\nu D^2 S_\nu c^2}{2h\nu^3}\left[\exp\left(\frac{h\nu}{kT_{\mathrm{D}}}\right)-1\right],$$

where C_ν is a coefficient that describes the physical size and millimeter emission properties of the dust grains at frequency ν, D is the distance to the molecular cloud, c is the speed of light, and the other terms have been previously defined. Therefore, measurements of the dust continuum emission from molecular clouds adds complementary physical information to that obtainable through molecular emission measurements.

4. Classification of Sources and Evolutionary Scenarios

In order to understand the evolutionary characteristics of the star formation process, it is convenient to have a source classification system. A classification scheme that used the infrared spectral index for $\lambda > 2\ \mu$m was developed in the 1980s to characterize infrared sources. This classification scheme contained three categories, Class I–III, which are distinguished by a decreasing amount of emission at longer wavelengths. Once detailed models of the evolution of the star formation process for an isolated low-mass core were developed, these categories became associated with stages in the star formation process. Later, a fourth younger class had to be added to this system, Class 0, when it became apparent that there were a significant number of sources with even larger amounts of emission at longer wavelengths. Figure 8 shows a pictorial description of these classes. Figure 9 shows an integrated CO 2–1 emission image from a portion of the Ophiuchus star formation region, which contains a rich variety of protostellar and young stellar objects.

5. Isolated and Clustered Star Formation Properties

The star formation process is thought to proceed in two different physical environments: isolated regions containing a single, usually lower mass ($M < 10\ M_\odot$), object evolving toward the formation of a star, and a group or cluster of dense cores, usually of higher mass ($M \geq 10\ M_\odot$), which appear to be forming stars. The detailed evolutionary sequence described in Section 4 has been developed mainly to explain the isolated star formation process given its simplicity. Observations of the molecular spectral line and continuum emission from these isolated star formation regions are currently being used to test these detailed evolutionary sequences. Specifically, studies of the core collapse signature from spectral line measurements, multifrequency dust emission measurements of the emission signature from dense cores, and imaging studies of the molecular spectral line and continuum emission from these regions are being used to provide stringent tests of the theories of isolated star formation.

Our understanding of the clustered star formation process is quite a bit different than that for isolated star formation. High mass stars represent a much smaller fraction of the stellar population than low mass stars, so the nearest examples of high mass star formation are far more distant than their low mass cousins. Fortunately, though, stellar luminosity is a strong function of mass, which makes it relatively easy to detect high mass stars at large distances.

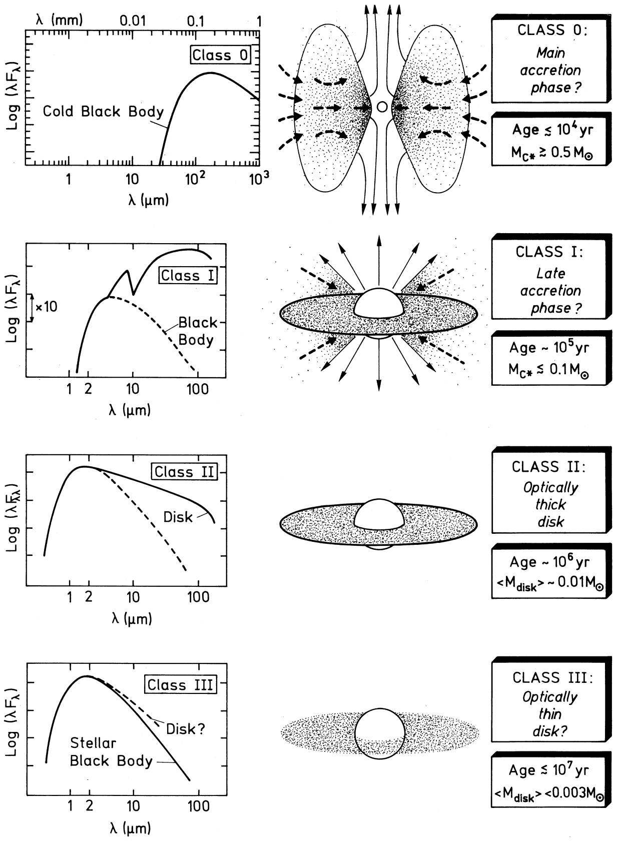


FIGURE 8 Pictorial description of the star formation evolutionary process. The column on the left of the figure depicts the radiant flux of each class as a function of frequency. The emission from physically distinct components which compose each class are indicated. The column in the center of the figure presents a pictorial description of each class. Arrows indicate mass inflow and outflow. The column on the right of the figure lists the class name, its primary physical component, its age, and the mass of the central accreting protostar (for Class 0 and I) or the mass of the stellar disk (for Class II and III). [From André, P. (1994). *In* "Proceedings of the XIIIth Moriond Astrophysics Meeting, The Cold Universe" (T. Montmerle, C. J. Lada, I. F. Mirabel, and J. Tran Thanh Van, eds.), Editions Frontiéres, Gif-sur-Yvette, France.]

The quantity of observational information on clustered high mass star formation is equivalent or greater than that from isolated low mass star formation regions. This information is often difficult to interpret, though, owing to the fast evolution of massive stars in a clustered environment, the strong influence that high mass stars exert on their surroundings, and the poor spatial separation of multiple star formation events in these regions afforded by current millimeter astronomical instrumentation. The general picture of low mass protostellar evolution described in Section 4 is likely to be generally valid for high mass star formation in clusters, but the higher densities and more kinematically dynamic environments in which high mass star formation proceeds forces the process to evolve more quickly than in a low mass environment. In general, our overall understanding of clustered star formation is quite a bit more incomplete than it is for isolated star formation, but it is believed that the two evolutionary sequences have many similarities.

B. Evolved Stars

During the evolution of many stars, mass loss during the giant branch phase modifies a star's physical properties. Studies of the stellar envelopes that result from this mass loss process indicate that evolved stars are a significant contributor of processed material to the interstellar medium. Since this processed material is eventually incorporated into the next generation of stars, an understanding of evolved stars and their mass loss mechanisms is key to understanding the entire star formation process.

Toward the end of their lifetimes, low and intermediate mass stars ($1M_\odot \quad M \quad 10\ M_\odot$) go through a high mass loss phase. As a result, a circumstellar envelope is formed that can often be detected with spectral line and continuum observations at millimeter wavelengths. These objects include (1) classical Mira and semiregular variable stars, (2) OH/infrared (OH/IR) and carbon stars, (3) early spectral type (warm) stars of type B–F, and (4) planetary nebulae (PN). Stellar evolution theory places variables, OH/IR, and carbon stars on the asymptotic giant branch (AGB). These stars lose mass at very high rates of up to $10^{-5}\ M_\odot$ per year through continuous flows and episodic outbursts. The total galactic rate of mass loss from evolved stars could be as high as $0.35\ M_\odot$ per year, making them a significant contributor to the cycle of star formation and the return of processed material to the interstellar medium. The mass loss mechanism in these objects is not well understood, but a prime suspect is radiation pressure on grains that are condensing in the cool extended atmospheres of these objects. Planetary nebulae have evolved beyond the AGB stage, characterized by the ejection and ionization of the star's envelope.

The morphology and extent of circumstellar envelopes is determined by the dynamics of the mass loss process and by the stellar and interstellar radiation field. Since the envelope's density structure is approximately known, the time exposure of the interstellar radiation field can be estimated, and the distribution of molecular photoproducts can be measured. Stellar envelopes represent an unparalleled astrochemistry and radiative transfer laboratory. For this reason, evolved stars have been prime targets for the detection and study of rare molecules in the overall investigation of interstellar chemistry.

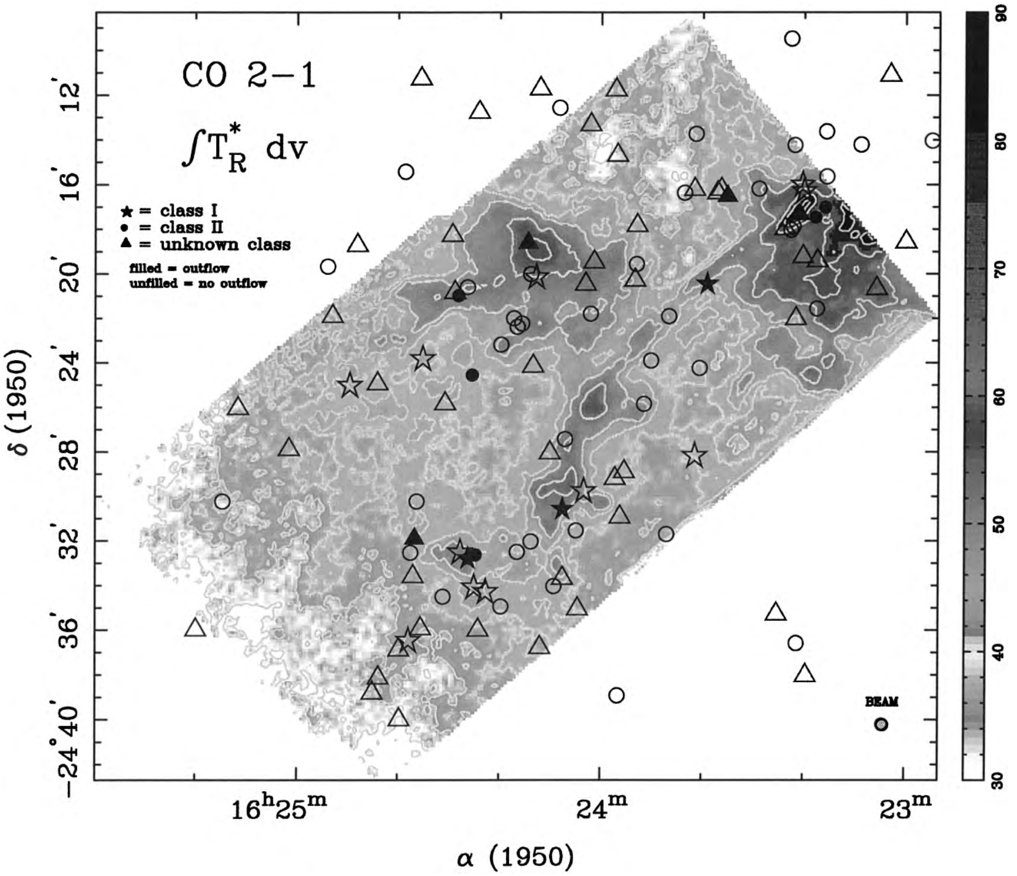


FIGURE 9 Spectrally integrated CO 2–1 emission image of a portion of the Ophiuchus star formation region. The intensity scale is in K ∗ km sec^{-1}. Symbols identify the positions of the young stellar objects in the region, some of which possess molecular outflows.

C. Chemistry and the Role of Interstellar Molecules

There are currently 120 known interstellar molecules, 82 of which were discovered at millimeter wavelengths, and the majority of which are organic (see Table II). Cosmic atomic abundances are generally reflected in the relative abundances of interstellar molecules, with H, C, O, and N being highly important in molecules. The atomic abundances of S, Si, P, Na, Al, Cl, F, K, and Mg are seen in fewer species, partly owing to reduced abundance and partly owing to inefficient chemistry. Carbon chemistry is the dominant form in the interstellar medium, with long carbon chains playing a dominant role among the organic compounds.

Because of the very low densities and temperatures prevalent in the interstellar medium, molecules cannot be formed by the same processes that produce them on earth. Astrochemistry appears to be dominated by three chemical reaction regimes: ion-molecule, grain surface, and shock chemistry.

1. Ion-Molecule Chemistry

Gas-phase reactions involving molecular ions are the best understood and are thought to be the most important of the three chemical regimes in the interstellar medium. Ion-molecule chemistry appears to explain fairly well the formation of the simpler interstellar molecules (those containing less than 5 atoms), and recently the formation of a few of the more complex molecules have been explained by ion-molecule reactions.

2. Grain Surface Chemistry

Catalytic reactions on the surfaces of interstellar dust grains, followed by release into the gas phase, can explain the formation of the most abundant molecule, H_2, and appears to be successful at explaining the formation of many of the more complex interstellar molecules. Unfortunately, grain chemistry processes are poorly understood and difficult to simulate. Grain surfaces are also recognized as repositories for many interstellar molecules that

TABLE II Known Interstellar Molecules (January 2001) (Courtesy Barry Turner)

Inorganic molecules (stable)

Diatomic		Triatomic	4 Atom	5 Atom
H_2	HCl	H_2O	NH_3	SiH_4
CO	PN	H_2S		
CS	NaCl	SO_2		
NO	AlCl	OCS		
NS	KCl	N_2O		
SiO	AlF			
SiS	HF			

Organic Molecules (stable)

Alcohols	Aldehydes and ketones	Acids	Hydrocarbons
CH_3OH	H_2CO	HCN	C_2H_2
CH_3CH_2OH	CH_3CHO	HCOOH	C_2H_4
	H_2CCO	HNCO	CH_4
	$(CH_3)_2CO$		
	HCCCHO		
	$HOCH_2CHO$		

Amides	Esters and ethers	Organosulfur
NH_2CHO	CH_3OCHO	H_2CS
NH_2CN	$(CH_3)_2O$	HNCS
NH_2CH_3		CH_3SH

Paraffin derivatives	Acetylene derivatives	Other
CH_3CN	HC_3N	CH_2NH
CH_3CH_2CN	CH_3C_2H	CH_2CHCN
		C_6H_6

Unstable molecules

Radicals		Ions	Rings	Carbon chains	Isomers
CH	C_3H	CH^+	SiC_2	HC_5N	HNC
CN	C_3N	HCO^+	C_3H_2	HC_7N	CH_3NC
OH	C_3O	N_2H^+	C_3H	HC_9N	HCCNC
SO	C_4H	$HOCO^+$	C_2H_4O	$HC_{11}N$	HNCCC
HCO	C_5H	HCS^+	SiC_3	CH_3C_3N	MgNC
C_2	C_6H	H_3O^+		CH_3C_4H	MgCN
C_2H	C_2S	$HCNH^+$		CH_3C_5N	NaCN
C_3	C_3S	H_2D^+		SiC_4	
HNO	CH_2CN	HOC^+		H_2CCC	
CP	SiC	SO^+		H_2CCCC	
SiH_2	CH_2	CO^+		C_7H	
NH_2	C_2O	$H_3{}^+$		C_8H	
HCCN	C_5N	H_2COH^+		H_2C_6	
	SH	HC_3NH^+		SiC_3	
	CH_3	CH_2D^+			
	SiCN				

may have been produced during a cooler and denser phase of the parent molecular cloud. These grain repositories are thought to be composed of grain surface ices, which when warmed by a nearby star formation event release the stored molecules into the interstellar medium.

3. Shock Chemistry

Strong shocks produced by the expanding ionized envelopes of massive stars and supernova remnants heat and compress the interstellar medium, leading to conditions ripe for many high-temperature chemical reactions. Like grain surface chemistry, reactions within the shock chemistry environment are difficult to simulate, but are progressing toward a physical framework that can be compared to observations. While OH and H_2O are prominent products of ion-molecule chemistry as well as shock chemistry, SiO, and SiS are predominently produced in shocks.

In nonquiescent regions, the chemistry is generally time dependent. This means that the chemical abundance of many molecules are a function of time and evolutionary state of the parent molecular cloud. With a more thorough understanding of the chemical processes at work in the interstellar medium, studies of the abundances of interstellar molecules will in future be coupled to the physical evolution of the interstellar medium to allow a better determination of the ages of various types of processes which occur in the interstellar medium.

A fundamental role played by molecules in the interstellar medium is as one of the cooling catalysts for the star formation process. Molecular clouds that exceed the maximum stable mass for a cloud with only thermal support, known as the Jeans mass (see Section A1), are predicted to gravitationally collapse. In order for this collapse to proceed, molecules and dust are required to radiate away energy released by the gravitational collapse. Many of the stars formed will eventually produce novae and supernovae, which further enrich the interstellar medium with molecules that can be used as catalysts to future star formation events.

D. Solar System

1. The Sun

Studies of the millimeter emission from the sun have concentrated on the identification and characterization of the nonthermal millimeter emission produced by solar flares. Only the most energetic electrons (energies >1 MeV) accelerated in solar flares produce millimeter emission. Millimeter interferometric observations made in recent years have shown that the electron population that produces the millimeter emission is different than the electron population that produces the keV-energy bremsstrahlung at x-ray frequencies. These observations suggest that two distinct electron acceleration mechanisms may be at work in solar flares.

2. Comets

Molecular spectral line and millimeter continuum measurements of comets are used to study the physical conditions within the comae and nuclei of these objects using the same basic techniques that are applied to these kinds of measurements made toward molecular clouds. In general, millimeter observations of comets are much more difficult to make given the rather small apparent angular size (generally less than 10″) of most comets. In any event, inspired by the recent visitation by comet Hale–Bopp, which was the most massive comet ever observed, quite a few millimeter continuum and spectral line observations of comets have been made in the past decade.

For example, millimeter continuum observations of comets P/Halley and Hale–Bopp have been used to measure the dust particle properties of cometary nuclear matter. These measurements have shown that the dust particle sizes in these comets are greater than 1 mm and that the mass of dust contained in the radiating grains is 10^{10}–10^{12} kg. Molecular spectral line observations of approximately 28 molecular species have been made toward a number of comets, with the recent visitation by the extremely large comet Hale–Bopp accounting for over half of these. Many of the molecules observed in comets are thought to be "parent" species, or species deposited onto icy grain mantles in the core during the formation of the comet. These species are released into the coma of a comet when it receives enough heating from the sun to sublimate them into the gas phase, thus allowing them to be observable through their rotational transitions. Thus, measurements of parent molecular species in comets are a direct measure of the chemical properties of the material from which comets were made, which is thought to be the same material from which the solar system formed. It is theorized that parent species chemically react with other parent species and atoms in the coma gas to produce other molecules, called "daughter" species. Since the intensity and distribution of a molecular species in a comet can be used to derive the physical conditions within the cometary nucleus and coma, studies of molecules in comets are a key element to an understanding of the physical structure of comets. An example of the variety in spatial distribution found in the molecular emission from comets is shown in Fig. 10.

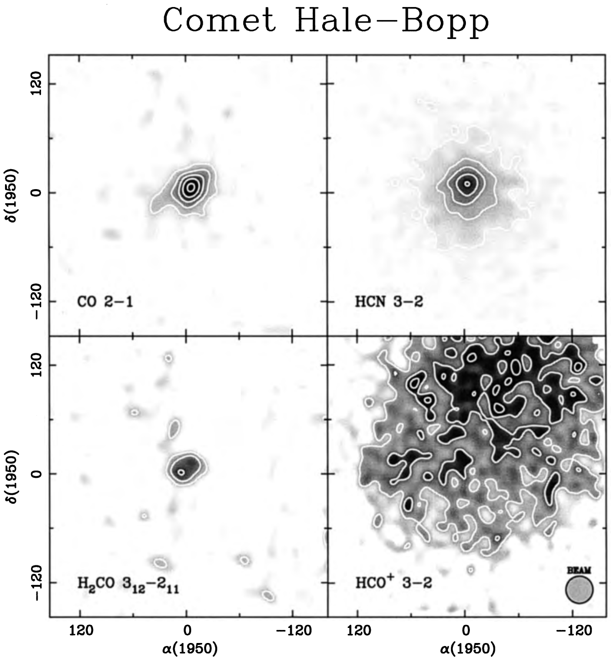


FIGURE 10 Spectrally integrated intensity images of the CO 2–1, HCN 3–2, $H_2CO\ 3_{12}-2_{11}$, and HCO^+ 3–2 emission from comet Hale–Bopp. Taken near perihelion in the spring of 1997, these four images reveal the chemical diversity of molecular emission from comets.

3. Planets

The millimeter wavelength emission from planets is made up of a combination of the black body continuum emission from the surface of the planet and atmospheric molecular spectral line emission. Spectral line profiles owing to millimeter wavelength molecular transitions from planetary atmospheres are used to derive its atmospheric temperature and abundance structure. Examples include measurements of CO, HDO, HCN, HC_3N, and CH_3CN from the atmospheres of Mars, Venus, Jupiter, Titan, Saturn, Uranus, and Neptune. These measurements are then used to derive the atmospheric chemistry and transport mechanisms. For example, measurements of the deuterated water (HDO) distribution in the atmosphere of Mars have been used to trace variations in the abundance of water as a function of martian latitude and season. Recent measurements point to the presence of significantly more water vapor near the north polar cap during northern summer on Mars, consistent with the sublimation of water deposits in the north polar cap. Coupled with measurements of the martian atmospheric temperature profile, the vertical atmospheric water profile has been derived. Comparison with the water vapor measurements made by the Viking lander in the 1970s indicates that the global mean water vapor content above 5 km is significantly lower today. Either the water vapor resides at very low altitudes in the Martian atmosphere or the entire atmosphere is 10 times drier than it was 20 years ago.

4. Asteroids

Observations of asteroids at millimeter wavelengths are limited to measurements of their continuum emission. The dust emission properties from an asteroid can be used to derive the dielectric properties of its surface material, which is a direct indicator of the physical properties of the asteroid surface. This information is complementary to that which one can obtain through optical and infrared observations, but the accuracy of many of the physical measurements made using millimeter observations is often better owing to the millimeter emission process in these objects. For example, at millimeter wavelengths the absorption and emission of radiation happen in the surface layers of the asteroid at a depth of just a few wavelengths. This causes the thermal inertia to be much larger at millimeter wavelengths than it is at optical and infrared wavelengths, making the time scales for changes in the millimeter emission properties much longer (tens of hours) than the measured rotation periods of asteroids (a few hours). Another simplifying factor that makes millimeter observations more sensitive to the asteroid surface physical conditions is the fact that the emissivity at millimeter wavelengths is very close to unity. This implies that the observed brightness temperature of an asteroid is equal to its equilibrium physical temperature.

The rather simple millimeter emission properties of asteroids have made them a very attractive option for accurate flux calibration at millimeter wavelengths. Current flux calibration at millimeter wavelengths is based on models of the millimeter emission from Mars extrapolated to measurements of other planets such as Jupiter, Saturn, Uranus, and Neptune. Unfortunately, the millimeter emission from Mars is complicated by the poorly understood effects of the polar ice caps, the longitudinal dependence of the disk temperature, and the atmospheric dust storms, making the uncertainty in the Mars-based flux calibration scale 5–10%. Asteroids can provide a flux calibration standard that is accurate to better than 5%. Their small apparent angular size of less than 2″ has limited their utility in this capacity as this angular size generally represents a small fraction of the spatial resolution of a telescope that operates at millimeter wavelengths. Advances in millimeter observing instrumentation in the next decade will overcome this limitation and further develop the use of asteroids as millimeter wavelength flux calibrators.

E. Extragalactic Astrophysics

Studies of the dust and molecular components of galaxies parallel the analyses of the physical properties in star formation regions in our own galaxy. Most studies of the molecular and dust component of galaxies concentrate on the morphological and evolutionary impact of these components. For example, stars form within the dense molecular clouds that inhabit galactic disks, and these stars contribute significantly to the total luminosity of their parent galaxy. Measurements of the CO and dust continuum emission from a wide variety of galaxies have allowed scientists in recent years to study the global content of H_2 as a function of morphological type, luminosity, and environment in these objects. Following are some of the areas of research into extragalactic astrophysics done at millimeter wavelengths.

1. Mass Determination in External Galaxies

For the reasons discussed in Section III.A, molecular emission is a good tracer of H_2. Since CO is the most abundant molecule after H_2, it has become the de facto standard for deriving the molecular mass in galaxies. The molecular mass is derived from measurements of the CO emission by first measuring the CO luminosity from a molecular cloud in a galaxy:

$$L_{\rm CO} = D^2 \int I_{\rm CO}\, d\Omega,$$

where D is the distance to the galaxy, $I_{\rm CO}$ is the CO brightness temperature integrated over the line profile

$$I_{\rm CO} = \int T_{\rm CO}\, dv,$$

and Ω is the angular size of the molecular cloud. Assuming for simplicity a spherical cloud that is in virial equilibrium, such that for a cloud of mass $M_{\rm c}$ the linewidth is given by

$$\Delta v = \sqrt{\frac{GM_{\rm c}}{R}},$$

the cloud mass $M_{\rm c}$ is derived from the CO luminosity with the following:

$$M_{\rm c} = \frac{L_{\rm CO}}{T_{\rm CO}}\sqrt{\frac{4\rho}{3\pi G}},$$

where ρ is the mass density of the gas. One obtains the total CO luminosity, and therefore the total molecular mass, for a galaxy by integrating over all of the individual molecular clouds in the galaxy. A key assumption used when relating the total CO luminosity to the molecular mass is the constancy from galaxy to galaxy of the ratio $\sqrt{\rho}/T_{\rm CO}$. Studies of this ratio in molecular clouds in our own galaxy, in addition to studies of the CO emission from external galaxies, indicate that this factor varies by about at least a factor of 2, indicating a reasonable accuracy to the proportionality between CO luminosity and molecular mass.

2. Morphological Studies

The spatial distribution of molecular gas in a variety of galaxies has been used to address a variety of morphological issues. Studies of spiral structure have considered the relationships between the HI, H_2, and HII gas in spiral arms, the nature of spiral arm streaming motions, and the possible existence of higher order resonance patterns within spiral structure. One result from these studies is that in spiral galaxies the radial distributions of CO emission generally peak in the center and decrease monotonically with radius, while the atomic distributions (as traced by HI emission) in these objects show a central depression followed by a relatively constant distribution with increasing radius. An excellent example of the information available through millimeter spectral line studies of spiral arm structure in external galaxies is the image of CO 1–0 in M51 shown in Fig. 11. This image shows a clear definition of the molecular gas associated with the major spiral arms in M51. A comparison with an optical image obtained with

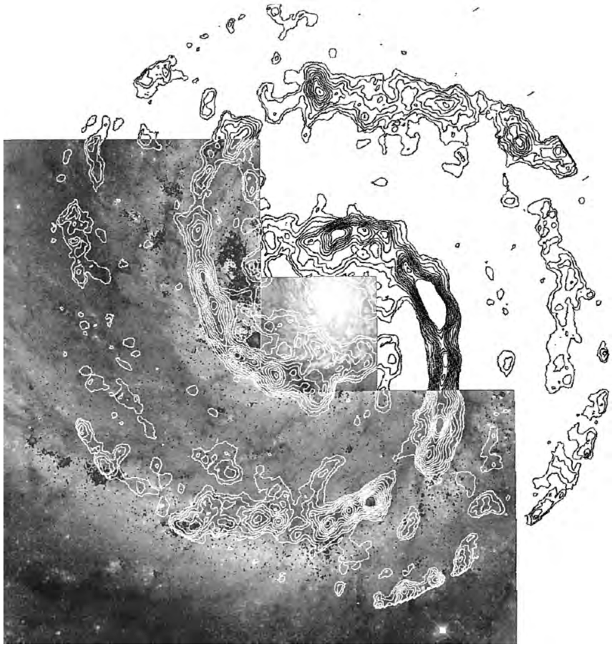


FIGURE 11 CO emission contours shown superimposed on a Hubble Space Telescope image of M51. The CO image, which was constructed by combining 19 individual pointings of the Owens Valley Radio Observatory millimeter array, has a resolution of 2.3″. [From Aalto *et al.* (1999). *Astrophysical Journal* **522,** 165.]

the Hubble Space Telescope shows the positional coincidence between the molecular gas and the visible dust lanes in this galaxy. The CO 1–0 imaging of M51 has revealed spiral arm molecular cloud associations with masses of 10^7–10^8 $M_\odot$, spatial extents of approximately 150 pc, and streaming motions in the range 20–50 km sec^{-1}. Imaging studies of the millimeter dust continuum emission from M51 show a good correlation with the distribution of CO emission in this galaxy.

Barred spiral galaxies have been found to have an enhancement of CO emission along the optical bar. Additionally, many galaxies that do not show barred morphology in the optical have been found to exhibit bar-like structures in their central regions. For example, the nearby spiral galaxy IC342 contains a molecular bar, but no apparent optical bar (see Fig. 12). Spiral arms that trace the locations of young stars are also apparent in high resolution imaging studies of the CO emission in these objects.

During the past five years morphological studies of spiral galaxies have been significantly advanced through two major millimeter interferometric surveys of the CO 1–0 emission in nearby galaxies. Using the OVRO and BIMA millimeter arrays, complete images of the CO emission from samples of 20 and 44 galaxies, respectively, were made at resolutions of 2–4″ (OVRO) and ∼7″ (BIMA). Both of these surveys incorporated single antenna measurements (NRO 45 m measurements for OVRO, NRAO 12 Meter Telescope measurements for BIMA) into their interferometer data to provide a complete sample of the CO emission in each galaxy. These surveys have made it possible to compare the detailed morphologies within each sample and derive physical models that describe the evolution of these morphologies. For example, the OVRO/NRO survey has found that barred spirals exhibit a higher degree of central gas concentration, but a lower overall molecular gas surface density than normal spirals. This property of barred spiral galaxies confirms theoretical predictions of rapid radial transport of gas toward the center of this type of spiral galaxy. The lower overall molecular surface density in barred spiral galaxies is thought to be caused by a more rapid expenditure of the material which feeds star formation in these galaxies.

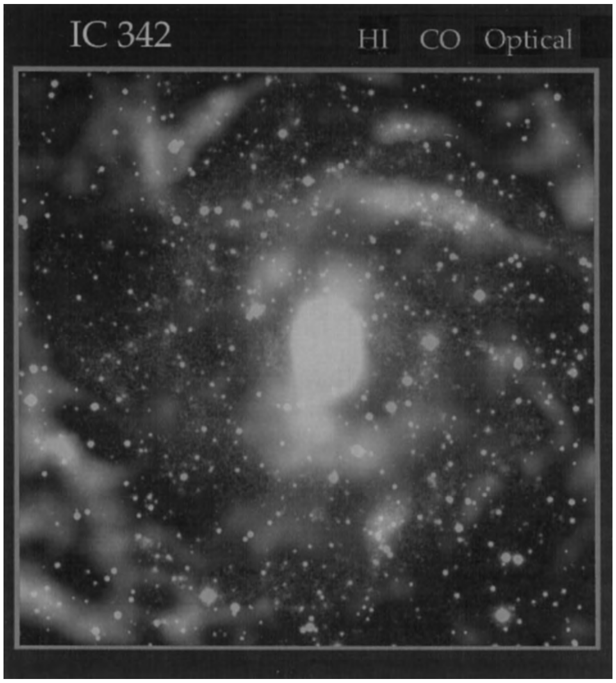


FIGURE 12 Overlay of the CO 1–0 (green) integrated intensity obtained using the NRAO 12 Meter Telescope, a Very Large Array HI (red) emission image, and the Palomar Observatory Sky Survey red optical image (blue) from the nearby Scd galaxy IC342. Note that the CO and optical emission coexist in the central part of the galaxy, where the HI emission is very weak. Note also that a number of spiral arms apparent in the HI emission from the outer parts of the galaxy can be traced smoothly into spiral arms apparent in the CO emission in the central parts of the galaxy. Photo courtesy of Jean Turner.

3. Galactic Evolution

Studies of galactic evolution have focused on the comparison between the atomic (HI) and molecular (H_2) gas properties and star formation rates as a function of environment, luminosity, and galaxy type. The general conclusions from these studies are as follows:

1. The sum of the galactic atomic and molecular masses range from 10^6 to 5×10^{10} $M_\odot$.
2. The ratio of molecular to atomic gas mass $M_{\mathrm{H}_2}/M_{\mathrm{HI}}$ decreases as a function of morphological type for spiral galaxy types Sa–Sd.
3. The ratio of total neutral gas mass to dynamical mass $M_{\mathrm{gas}}/M_{\mathrm{dynamic}}$ increases from 4% for early type (Sa) spiral galaxies to 25% for late type (Sd) galaxies.
4. The global star formation rates and efficiencies for spiral galaxies do not show a strong dependence on morphology.

Of special consideration are galaxies that are interacting with other galaxies. Active galactic nuclei (AGN) and ultraluminous infrared galaxies (ULIRGs) represent two categories of galactic systems whose evolution has been modified by an interaction event. Galaxy–galaxy interactions are known to disrupt the stellar component of galactic disks and enhance the star formation rate. The molecular component of these systems appears also to be disrupted, leading in many cases to an order-of-magnitude increase in the star formation efficiency, with a more efficient production of massive stars.

4. Galactic Nuclei

The most dominant molecular component in galaxies is their nuclei. In recent years the level of sensitivity available for studies of galactic nuclei has allowed for not only the detection of CO in these nuclei, but also studies of the spatial distribution of CO and other molecules. Studies utilizing measurements of the spatial distribution of HCN and CS from a number of nearby active galactic nuclei and starburst galaxies have resulted in a better understanding of the physical properties, kinematics, and mass distribution of the dense gas in these regions. The information gathered from these studies has also been used to investigate the relationships among various types of "active" galaxies, such as AGN, ULIRGs, and starburst galaxies.

Studies of nearby galaxies have provided information on the physical properties of the dense gas in galactic nuclei, which in turn have provided a useful comparison to studies of the dense gas properties in our own galaxy. For example, the nearby starburst galaxy NGC 253 has been studied with high spatial resolution multitransition measurements of CO and HCN. The spatial distribution of the molecular clouds imaged in NGC 253 is similar to that found in the central regions of our own galaxy, but the kinetic temperature and volume density is 100 K and 10^4–10^5 cm^{-3}, warmer and denser than average galactic molecular clouds. Similar studies of the dense gas distribution and properties in other nearby galaxies such as M82, NGC 1068, and M51 have yielded comparative information useful in defining the detailed physical characteristics of these galaxies.

5. High Redshift Galaxies

Important constraints on the evolution of galaxies have been provided by the search for and characterization of distant galaxies. These searches, conducted using CO as a tracer of molecular gas, have pointed to the existence of massive quantities of dust and molecular gas out to redshifts as high as 4.69. This high redshift corresponds to ages within a few billion years of the Big Bang, thus indicating enrichment of molecular gas through chemical processing at a very early epoch. Currently 12 galaxies have been measured with redshifts greater than 2. Imaging studies of these objects have provided information on their gas and dynamical mass, information that cannot be gleaned from observations at other wavelengths. The gas mass fraction in these systems is used as a measure of their evolutionary state, providing a constraint on the epoch of initial star and galaxy formation. A characterization of the physical properties of the high redshift systems is also being used to establish whether galaxies formed hierarchically or as systems that were given their present day masses during formation.

F. Cosmology

In addition to the cosmological implications of studies of the millimeter emission properties of galaxies, the structure of the cosmic microwave background radiation (CMBR) and direct measurements of the Hubble constant H_0 have been made using millimeter astronomical techniques. The recent measurements of the Sunyaev Zel'dovich effect (SZE) toward a sample of 27 clusters using 26–36 GHz receivers on the BIMA and OVRO millimeter arrays have provided some of the most direct measurements of several cosmological parameters. The SZE is the spectral distortion of the CMBR owing to inverse Compton scattering of photons by electrons in hot gas confined to the deep gravitational potential produced by galaxy clusters. A fundamental property of SZE emission is that the observed brightness is a function of the cluster properties and is independent of its distance. Therefore, SZE measurements can yield a very accurate measure of the Hubble constant H_0 and the matter density of the universe Ω_M. Current results from these SZE measurements indicate that $H_0 = 67$ km sec^{-1} Mpc^{-1} and $\Omega_M h_{100} \sim 0.25 \pm 0.06$ (h_{100} is the Hubble constant normalized to a value of 100 km sec^{-1} Mpc^{-1}). Note that the distance-independent nature of these SZE measurements makes them the most accurate way in which to measure H_0 and Ω_M astronomically.

SEE ALSO THE FOLLOWING ARTICLES

CLOUD PHYSICS • COMETARY PHYSICS • COSMOLOGY • GALACTIC STRUCTURE AND EVOLUTION • INTERSTELLAR MATTER • MAGNETIC FIELDS IN ASTROPHYSICS • RADIO-ASTRONOMY INTERFEROMETRY • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS • STELLAR STRUCTURE AND EVOLUTION

BIBLIOGRAPHY

André, P., Ward-Thompson, D., and Barsony, M. (2000). From Pre-Stellar Cores to Protostars: The Initial Conditions of Star Formation. *In* "Protostars and Planets IV" (V. Mannings, A. P. Boss, and S. S. Russell, eds.), in press. University of Arizona Press, Tucson, AZ.

Bachiller, R. (1996). "Bipolar molecular outflows from young stars and protostars," *Ann. Rev. Astron. Astrophys.* **34,** 111–154.

Butler, B. J., and Gurwell, M. A. (2001). Solar System Science with ALMA. *In* "Science with the Atacama Large Millimeter Array" (A. Wootten, ed.), in press, Astronomical Society of the Pacific, San Francisco, CA.

de Pater, I. (1990). "Radio images of the planets," *Ann. Rev. Astron. Astrophys.* **28,** 347–399.

Evans, N. J., II (1999). "Physical conditions in regions of star formation," *Ann. Rev. Astron. Astrophys.* **37,** 311–362.

Neininger, N. (1999). Interferometric Observations of Nearby Galaxies. *In* "The Physics and Chemistry of the Interstellar Medium" (V. Ossenkopf, J. Stützki, and G. Winnewisser, eds.), pp. 42–49. GCA–Verlag, Herdecke, Germany.

Sargent, A. I., and Welch, W. J. (1993). "Millimeter and submillimeter interferometry of astronomical sources," *Annual Reviews of Astronomy and Astrophysics*. **31,** 297–343.

Scoville, N. Z., and Sargent, A. I. (2000). Imaging Spectroscopy at mm-Wavelengths. *In* "Imaging the Universe in Three Dimensions: Astrophysics with Advanced Multi-Wavelength Imaging Devices" (W. van Breugel and J. Bland-Hawthorn, eds.), pp. 236–247. ASP Conference Series Vol. 195, Astronomical Society of the Pacific, San Francisco.

Turner, B. E. (1992). Interstellar Medium, Molecules. *In* "The Astronomy & Astrophysics Encyclopedia" (S. Maran and C. Sagan, eds.), p. 378, The Astronomy and Astrophysics Encyclopedia. van Nostrand, New York.

van Dishoeck, E. F., and Blake, G. A. (1998). "Chemical evolution of star-forming regions," *Ann. Rev. Astron. Astrophys.* **36,** 317–368.

Neutrino Astronomy

Raymond J. Davis, Jr.
Alfred K. Mann
University of Pennsylvania (Emeritus)

GLOSSARY

Carbon–nitrogen cycle Sequentially ordered set of fusion reactions of hydrogen with carbon and nitrogen isotopes, serving to combine four hydrogen atoms into helium.

Čerenkov light Light radiated by a charged particle traversing a medium. Čerenkov radiation occurs when the particle has a velocity greater than the velocity of light in that medium.

Leptons General term for the light elementary particles—electrons, muons, tauons, and their associated neutrinos—that interact with nuclei and with each other by the weak force.

Main sequence Narrow region on a plot of the luminosity versus surface temperature of stars (Hertzsprung–Russell diagram) where the hydrogen-burning stars are located.

Muon Unstable, weak-interacting, charged particle (+ or −) with a mass 207 times that of an electron. It decays to an electron and two neutrinos, $\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$, with a mean lifetime of 2.2×10^{-6} sec.

Neutrino oscillation Hypothetical process that would allow different neutrino types, or flavors, to change into one another.

Proton–proton chain Set of nuclear fusion reactions that converts four hydrogen atoms into helium. This chain is the dominant energy producer in the sun.

Solar model Theoretical calculation of the internal structure of the sun that follows the changes in structure, temperature, and composition as the sun ages.

Weak interaction or weak force One of the fundamental forces in nature that governs the interaction of leptons.

THE NEUTRINO is an electrically neutral elementary particle that has the unique capability of penetrating matter on the scale of stellar dimensions and densities. Neutrinos are produced abundantly in nuclear processes (including beta-decay) in the interior of stars and planets and by a variety of processes in the cosmos. The study of neutrino radiation from stars is a direct means of observing the energy production mechanism occurring in their interiors. In addition, intermediate mass stars at the end of their nuclear life, i.e., after exhausting their nuclear fuel, suffer a catastrophic collapse (supernovae) and emit an intense burst of neutrinos. Furthermore, neutrinos are a component of cosmic radiation and of cosmic black body radiation, and, if they have nonzero mass, they may conceivably contribute significantly to the total mass of the universe.

For these reasons, the detection of neutrinos from such sources is a subject of considerable interest in astronomy and cosmology. Because of their weak interaction with matter, neutrinos are difficult to observe directly. Available techniques require massive detectors on the order of 100–10,000 tons to record a neutrino event rate between 0.1 and 1.0 event per day arising from not-too-distant sources. Because of this constraint, studies have been limited primarily to observations of neutrino radiation from the sun which were carried out to verify and understand the hydrogen fusion reactions that are the source of the sun's energy. Also, an intense pulse of antineutrinos was observed for the first time on February 23, 1987, emanating from a supernova (SN1987A) in the Large Magellanic Cloud (LMC), a nearby satellite galaxy 165,000 light years distant. In this article, we discuss principally the observations of solar neutrinos and the supernova event in 1987 and briefly mention other neutrino sources in the cosmos.

I. DETECTING NEUTRINOS

A. Classification of Neutrinos as Elementary Particles

Neutrinos belong to the family of elementary particles called leptons (originally, "smaller mass" particles). Leptons, of which there are three known subfamilies, are distinguished from other elementary particles by a unique property: They are coupled to all other particles by the weak force, which results in extremely feeble interactions of leptons with the nuclei of atoms, with other elementary particles, and with themselves. Charged leptons, of course, may interact with other charged particles and fields through the much stronger electromagnetic force. As elementary particles, leptons carry an angular momentum of $\frac{1}{2}$ unit of spin and are therefore regarded as fermions (particles with $\frac{1}{2}$ unit of spin). With zero or very small mass, all neutrinos spin in a left-handed direction, and all antineutrinos spin in a right-handed direction. The three lepton families are designated by their charged members: the electron, e^- (and e^+); the muon, μ^- (and μ^+); and tauon, τ^- (and τ^+). Each charged set has a negative and a positive member, referred to as particle and antiparticle, respectively. Associated with each set of charged members is a corresponding set of neutral particles. These are the neutrinos and antineutrinos: ν_e, $\bar{\nu}_e$, ν_μ, $\bar{\nu}_\mu$, ν_τ, $\bar{\nu}_\tau$. The lepton families are listed in Table I. Recently, it has been demonstrated experimentally that only these three relatively low mass, stable lepton families exist in nature.

TABLE I The Lepton Families

	Electron family	Muon family	Tauon family
Particle	$e^-, l_e = +1$	$\mu^-, l_\mu = +1$	$\tau^-, l_\tau = +1$
Antiparticle	$e^+, l_e = -1$	$\mu^+, l_\mu = -1$	$\tau^+, l_\tau = -1$
Particle	$\nu_e, l_e = +1$	$\nu_\mu, l_\mu = +1$	$\nu_\tau, l_\tau = +1$
Antiparticle	$\nu_e, l_e = -1$	$\nu_\mu, l_\mu = -1$	$\nu_\tau, l_\tau = -1$

The masses of neutrinos may be zero and, in any event, are much smaller than the masses of their charged partners. The electron, muon, and tauon have masses of 0.511, 105.7, and 1784 MeV, respectively, and the muon and tauon have mean lives of 2.2×10^{-6} sec and 3.0×10^{-13} sec, respectively. Many experiments have been performed in an effort to measure the masses of the three neutrinos. The most sensitive experimental measurement published to date reports an upper limit on the electron neutrino mass of a few electron volts, or less than $1/50{,}000$th the mass of the electron. New experiments to improve this limit continue to be performed. Only crude upper limits have been set directly on the masses of the muon and tauon neutrinos from the kinematics of their decay modes: 0.2 and 35 MeV, respectively. It is generally presumed that the masses of all neutrinos are very much smaller than those upper limits, probably too small to be measured directly. A cosmological argument suggests that the sum of the three known stable neutrino masses should be less than a constant times the square of the Hubble constant, or roughly 100 eV. Later, we discuss means of determining mass differences between neutrino types, or flavors, from observations of neutrinos from the sun. These studies using the sun as a primary source of electron neutrinos have led to tests of neutrino flavor conservation, neutrino masses, and other fundamental neutrino properties.

Neutrinos are created in nuclear processes and in various elementary particle interactions. The most familiar process is nuclear beta-decay, in which an unstable nucleus simultaneously emits an electron (beta-ray) and a neutrino. This process may be visualized as an unstable nucleus radiating its energy by creating a pair of leptons: a neutrino and an electron. It is referred to as beta-minus decay when an electron (e^-) is emitted with an antineutrino ($\bar{\nu}_e$) or beta-plus decay when a positron (e^+) is emitted with a neutrino (ν_e). In another beta-decay process, called electron capture, one of the orbital electrons in an atom is absorbed by the nucleus and a neutrino is emitted. Examples of these processes are

$$^{14}\mathrm{C} \rightarrow {}^{14}\mathrm{N} + e^- + \bar{\nu}_e, \text{ beta-minus decay} \quad (1)$$

$$^{11}\mathrm{C} \rightarrow {}^{11}\mathrm{B} + e^+ + \nu_e, \text{ beta-plus decay} \quad (2)$$

$$e^- + {}^{37}\mathrm{Ar} \rightarrow {}^{37}\mathrm{Cl} + \nu_e, \text{ electron capture} \quad (3)$$

In these examples, the guiding principle is the principle of lepton number conservation. In any process, the total number of leptons and antileptons does not change; the number before and after is conserved. Table I lists the assigned lepton numbers for each lepton family. The lepton number is positive for a lepton and negative for an antilepton. By applying the lepton numbers for the electron family, the principle of lepton conservation is exhibited in the three examples given.

The principle of lepton conservation applies to each of the three families, and in addition, the leptons of each family appear to be separately conserved. It has been shown, for example, that when a neutrino of the muon type (ν_μ) is absorbed in a nucleus, a muon is emitted, never an electron, for example,

$$\nu_\mu + \text{Fe} \rightarrow \text{Co} + \mu^-.$$

It was by this process that the muon neutrino was discovered and shown to be a different particle from the electron neutrino. The principle of lepton conservation for each lepton type (or flavor) has been tested experimentally in many different ways and appears to be valid, but it is possible that separate lepton conservation is not rigorously true and additional tests are still needed. A superior way of testing these lepton properties is by studying the neutrino spectrum from the sun and cosmic rays. This topic will be discussed further in Section II.

B. Detection of Neutrinos

The weak force has both a charged and a neutral component, arising from the exchange of massive intermediating particles. The existence of these particles, the $W^\pm$ for charged currents, and the Z^0 for neutral currents, was directly demonstrated in 1983. The theory of weak-interaction processes is now well understood and allows one to calculate the interactions of neutrinos with nuclei and electrons. Using this theoretical foundation, it is possible to determine the sensitivity of detectors to fluxes of neutrinos as a function of neutrino energy. The interaction probability is expressed as the equivalent target area of a nucleus, or electron, or other particle that is presented to a neutrino. It is referred to as the cross section (σ) and is usually given in units of square centimeters per atom or particle.

C. Interactions with Nuclei

Neutrinos may be absorbed in nuclei with the emission of an electron, a muon, or a tauon, depending on the incident neutrino type. These are called inverse-beta-decay processes because they are, in the case of the electron neutrino, the inverse of normal radioactive beta-decay. Neutrino (ν_e) absorption by an inverse beta-decay reaction is illustrated by two examples, one for antineutrinos and one for neutrinos:

$$\bar{\nu}_e + \text{p} \rightarrow \text{n} + e^+ \tag{4}$$

$$\nu_e + {}^{37}\text{Cl} \rightarrow {}^{37}\text{Ar} + e^- \tag{5}$$

Reaction (4) was used by Reines and Cowan in the first experiment to detect antineutrinos. The antineutrinos originated from the beta-minus decay of fission products in a nuclear reactor. The same reaction was also used in detectors to observe antineutrinos from a collapsing star. Reaction (4) is the inverse of the beta-decay of the neutron,

$$\text{n} \rightarrow \text{p} + e^- + \bar{\nu}_e,$$

and requires that the antineutrino have sufficient energy to provide the mass difference between the neutron and the proton and to create the positron, a minimum total energy of 1.804 MeV. The minimum energy needed to carry out the reaction is called the threshold energy, E(thresh). Given the half-life of the neutron (~15 min), the threshold energy, the spin change (none in this case), and the weak-interaction coupling constant, the theory of weak interactions may be used to calculate the cross section for reaction (4) as a function of the antineutrino energy. For example, a 5-MeV antineutrino would have a cross section of 2.4×10^{-43} cm^2/H atom. With this cross-section value, we can calculate the distance that an antineutrino or neutron will penetrate through ordinary water. A beam of 5-MeV neutrinos would be reduced in intensity by 50% in traveling through 45 light years of water!

Reaction (5) is used for detecting neutrinos from the sun. It is the inverse of the electron capture of ^{37}Ar [see reaction (3)], a decay process with a half-life of 35 days. Again, the ν_e capture cross section can be calculated, knowing the half-life, the threshold energy (0.814 MeV), the spins of the ^{37}Cl and ^{37}Ar nuclei, the weak interaction coupling constant, and the overlap of the wave functions of orbital electrons in the nucleus. In complex nuclei such as ^{37}Cl, the neutrino absorption also produces ^{37}Ar in various excited states. The excited states decay rapidly to the ground state by emitting gamma-rays. In the ^{37}Cl–^{37}Ar case, the transition probabilities to all relevant excited states can be evaluated, so that the cross section for capture can be calculated accurately over the range of neutrino energies expected from the sun. One excited state in ^{37}Ar, the so-called analog state with an excitation energy of 4.98 MeV, is of particular importance. Neutrinos with energy greater than 5.79 MeV will feed this state and have a higher ν_e capture cross section.

These two examples of neutrino absorption reactions illustrate the procedures used in calculating the neutrino capture cross sections. The neutrino capture cross section for reaction (4) is the only one that has been measured experimentally with good accuracy (±5%). An approximate measurement of the cross section of the reaction

$$\nu_e + \mathrm{d} \rightarrow \mathrm{p} + \mathrm{p} + e^- \tag{6}$$

has also been performed (≈30%); the cross section of this reaction can also be calculated with confidence. This reaction is of potential importance in observing neutrinos from the sun. For all other neutrino capture reactions considered for radiochemical neutrino detectors, one must rely on a theoretical calculation. There is a particular difficulty in calculating the capture cross section for complex nuclei to produce the product nucleus in an excited state. In these cases, one must resort to nuclear reaction studies and theoretical nuclear models to estimate the cross section.

D. Interactions with Electrons

Neutrinos of all types, ν_e, ν_μ, and ν_τ, can scatter elastically from an electron. This process is a result of the weak-interaction coupling between the neutrino and the charged lepton, a coupling that may have a charged and a neutral current component. The cross section for scattering from electrons depends on the neutrino type, ν_e, ν_μ, ν_τ, and whether it is a neutrino or an antineutrino. Only the electron neutrino is coupled to the electron by both the charged and the neutral currents, whereas other neutrino types are coupled by neutral currents. These reactions may be represented schematically as follows:

$$\nu_e + e^- \rightarrow \nu_e + e^- \text{ (recoil), charged and neutral current}$$

$$\sigma(\nu_e) = 0.933 \times 10^{-43}\ (\mathrm{E}_\nu/10\ \mathrm{MeV})\ \mathrm{cm}^2 \tag{7}$$

$$\nu_\mu + e^- \rightarrow \nu_\mu + e^- \text{ (recoil), neutral current}$$

$$\sigma(\nu_\mu \text{ or } \nu_\tau) = 0.159 \times 10^{-43}\ (\mathrm{E}_\nu/10\ \mathrm{MeV})\ \mathrm{cm}^2 \tag{8}$$

$$\bar{\nu}_e + e^- \rightarrow \bar{\nu}_e + e^- \text{(recoil), charged and neutral current}$$

$$\sigma(\nu_e) = 0.388 \times 10^{-43}\ (\mathrm{E}_\nu/10\ \mathrm{MeV})\ \mathrm{cm}^2 \tag{9}$$

$$\bar{\nu}_\mu + e^- \rightarrow \bar{\nu}_\mu + e^- \text{(recoil), neutral current}$$

$$\sigma(\bar{\nu}_\mu \text{ or } \bar{\nu}_\tau) = 0.130 \times 10^{-43}\ (\mathrm{E}_\nu/10\ \mathrm{MeV})\ \mathrm{cm}^2 \tag{10}$$

Monoenergetic electron neutrinos, when scattered from electrons, produce a flat recoil energy spectrum. On the other hand, monoenergetic electron antineutrinos produce a spectrum of electron recoils that decreases inversely with the energy. The maximum energy of the recoil electron corresponds approximately to the energy of the neutrino, disregarding the initial binding energy of the struck electron. An important characteristic of the ν-electron scattering process is that the recoil electron moves in approximately the same direction as the incoming neutrino. Therefore, the neutrino-electron scattering process can be used to determine the direction as well as the energy of the neutrino. However, measuring neutrino energies and directions is difficult because of background processes. One must reduce the flux of cosmic rays, external gamma-radiation, and energetic neutrons to low levels. This can be accomplished by going deep underground, by using massive self-shielding, and by particle detectors to observe and veto charged particles that enter the active region of the detector from outside. In addition, radioactive elements that produce energetic electrons by beta-decay must be removed from the target and construction materials.

The method to distinguish the direction of the electron is to observe the Čerenkov light emitted by the recoil electron. When an electron (or any other charged particle) passes through a medium traveling faster than light travels in that medium, a cone of light is emitted around the direction of the electron. The cone of Čerenkov light is observed by a large number of light-sensitive detectors (photomultiplier tubes) located on the walls of the detector. This technique was employed in very large detectors filled with several thousand tons of water to search for the decay of the proton. Such detectors have recently turned principally to the detection of neutrinos.

II. NEUTRINOS FROM THE SUN

A. Theoretical Solar Models

Our sun is classified as a dwarf G-type main sequence star which is generating energy primarily by the fusion of hydrogen into helium. The overall hydrogen fusion process can be represented:

$$4\mathrm{H} \rightarrow {}^4\mathrm{He} + 2e^+ + 2\nu_e + \text{ gamma-radiation} \\ + \text{kinetic energy.} \tag{11}$$

The fusion of hydrogen into helium can be accomplished by two separate reaction sequences. known respectively as the proton–proton chain of reactions and the carbon–nitrogen cycle (Table II). It can be observed from Table II that the proton–proton chain has three competing branches, designated for reference as PP-I, PP-II, and PP-III. All stars generate energy by the proton–proton chain in the early stage of their evolution. The sun, with an age of 4.7 billion years, is still generating energy chiefly by this mechanism and will continue to do so for a few billion years more. As a star ages, its internal temperature increases, and the carbon–nitrogen cycle begins to play a more prominent role in energy production.

TABLE II The Proton–Proton Chain and Carbon–Nitrogen Cycle

Reaction		Neutrino energy (MeV)	Neutrino flux[a] (cm^{-2} sec^{-1})
	The proton–proton chain		
PP-I	$H + H \rightarrow d + e^{+} + \nu_e$ (99.75%)	0–0.420 spectrum	6.0×10^{10}
	or		
	$H + H + e^{-} \rightarrow D + \nu_e$ (0.75%)	1.44 line	1.4×10^{8}
	$D + H \rightarrow {}^{3}He + \gamma$		
	${}^{3}He + {}^{3}He \rightarrow {}^{2}H + {}^{4}He$(87%)		
PP-II	${}^{3}He + {}^{4}He \rightarrow {}^{7}Be + \gamma$(13%)		
	${}^{7}Be + e^{-} \rightarrow {}^{7}Li + \nu_e$	0.861 (90%) line	4.7×10^{9}
	${}^{7}Li + H \rightarrow \gamma + {}^{8}Be$ $\hookrightarrow 2\,{}^{4}He$	0.383 (10%) line	
PP-III	${}^{7}Be + H \rightarrow {}^{8}B + \gamma$(0.017%)		
	${}^{8}B \rightarrow {}^{8}Be + e^{\tau} + \nu_e$ $\hookrightarrow 2\,{}^{4}He$	0–14.1 spectrum	5.8×10^{6}
	The carbon–nitrogen cycle		
$H - {}^{12}C \rightarrow {}^{13}N + \gamma$			
${}^{13}N \rightarrow {}^{13}C + e^{+} + \nu_e$		0–1.20 spectrum	6.1×10^{8}
$H + {}^{13}C \rightarrow {}^{14}N + \gamma$			
$H + {}^{14}N \rightarrow {}^{15}O + \gamma$			
${}^{15}O \rightarrow {}^{15}N + e^{+} + \nu_e$			
$H + {}^{15}N \rightarrow {}^{12}C + {}^{4}He$		0–1.73 spectrum	5.2×10^{8}

[a] From Bahcall and Ulrich (1988). *Rev. Mod. Phys.* **60,** 297.

All of the reactions shown in Table II are used in solar model calculations and are the only reactions considered.

The nuclear reactions in Table II have been studied extensively in the laboratory. It has been established that these are the only reactions that are important in the hydrogen fusion processes. However, the primary reaction that initiates the proton–proton chain,

$$p + p \rightarrow d + e^{+} + \nu_e, \qquad (12)$$

has too low a cross section to be measured directly in the laboratory, but it is possible to calculate accurately the cross section at thermal energies. All the reactions are exothermic, producing the energies listed in Table II. Six of the reactions produce neutrinos; two emit monoenergetic neutrinos; and the other four emit a spectrum of neutrinos from near-zero energy to the maximum energy noted.

To determine the rates of these reactions in the sun, detailed calculations must be made of the temperatures, particle densities, and chemical composition of the different regions of the solar interior. Knowing the rates of the neutrino-producing reactions, one can calculate the energy spectrum and fluxes of the neutrinos emitted by the sun.

The internal conditions in the sun are derived from a complex solar model calculation, which follows the evolution of the sun from its initial formation to its ultimate state as a cooling white dwarf star, devoid of fuel. The sun is the only star for which mass, radius, luminosity, surface temperature, and age are well known. These parameters are used in the model and provide boundary conditions. It is assumed, with good physical justification, that the primitive sun was well mixed and heated by the release of gravitational energy. For this reason, it is thought that the initial elemental composition throughout its mass was identical with that now observed in the photosphere. The elemental composition is introduced in the solar model as the ratio of the abundance of each element heavier than helium to the abundance of hydrogen. This ratio is a directly measured quantity for each element. The principal elements heavier than helium, in order of their weight percent, are O, C, Fe, Ne, N, Si, Mg, and S. The total percentage of these elements is 1.8% by mass. The helium composition is not directly entered in the model but is calculated from the model. The initial helium composition calculated for the standard model is around 25% by mass. As the sun evolves, the helium content in the core gradually increases as a result of the fusion reactions.

The chemical composition of the sun is an important factor in determining the internal temperatures, because an increase in the composition of the heavier elements affects the rate of energy transport. The transmission of

TABLE III Calculated Solar Neutrino Fluxes and Cross Sections for the Reaction $\nu_e + {}^{37}\mathrm{Cl} \rightarrow {}^{37}\mathrm{Ar} + e$

Neutrino source	Flux on earth[a] ($\mathrm{cm}^{-2}\ \mathrm{sec}^{-1}$)	Capture cross[b] section (cm^2)	Capture rate[c] (SNU)
$\mathrm{H} + \mathrm{H} \rightarrow \mathrm{d} + e^{+} + \nu_e$	6.0×10^{10}	0	0 (0)
$\mathrm{H} + \mathrm{H} + e^{-} \rightarrow \mathrm{d} + \nu_e$	1.4×10^{8}	1.56×10^{-45}	0.2 (0.2)
^{7}Be decay	4.7×10^{9}	2.38×10^{-46}	1.1 (1.0)
^{8}B decay	5.8×10^{6}	1.08×10^{-42}	6.1 (4.1)
^{13}N decay	6.1×10^{8}	1.66×10^{-46}	0.1 —
^{15}O decay	5.2×10^{8}	6.61×10^{-46}	0.3 —
		Total	7.9 ± 0.9 (5.8 ± 1.3)

[a] From Bahcall, J. N., and Ulrich, R. K. (1988). *Rev. Mod. Phys.* **60,** 297.
[b] From Bahcall, J. N. (1978). *Rev. Mod. Phys.* **50,** 881.
[c] From reference *a* above and (in parentheses) Turck-Chieze *et al.* (1988). *Astrophys. J.* **335,** 415.

energy depends on a number of processes in which the thermal radiation interacts with unbound electrons and with atoms in various states of ionization and excitation. These processes must be calculated from simplified atomic models and photon-electron scattering theory. The complex calculations of opacity are provided by scientists from Los Alamos and Livermore National Laboratories. Their results are extensively employed in astronomical calculations.

The structure of the sun is expressed by equations of hydrodynamic equilibrium. That is, at every radius shell, the downward gravitational forces are balanced by the outward kinetic and radiation pressures. The calculations follow the initial collapse of the gravitating mass of the sun to the main sequence. The only additional energy is that introduced by the nuclear fusion reactions. When a star, such as the sun, is on the main sequence, where it remains for more than 6 billion years, all of the energy is provided by the hydrogen fusion reactions. The nuclear reaction rates are derived from the kinetic energy of the reactants, assuming a Maxwellian velocity distribution and a nuclear barrier penetration factor. Experimental nuclear reaction cross sections are used. They are usually measured in the laboratory at energies above 100 keV and extrapolated to the energies corresponding to the temperatures in the sun's interior (less than 1 keV). The theoretically calculated value of the primary p–p reaction is used.

The theoretical forecasts of the neutrino fluxes used to compare with experimental observations are those from the so-called standard solar model (SSM). This model relies on the values of the solar mass, radius, luminosity, and age (4.7 billion years) and on the best-considered values of the nuclear reaction cross sections. In addition, there are some special assumptions. It is presumed that the sun is not rotating, or differentially rotating, rapidly enough in its interior to affect its internal structure or dynamics. Processes that could mix the solar interior, such as diffusion or periodic hydrodynamic oscillation, are not taken into account. Magnetic fields are not regarded as sufficiently intense to affect the sun in any important way. A simplified convective theory is used for determining the transport of energy in the convective zone, the outer dynamic region of the sun.

Table III shows the neutrino flux at the earth for each neutrino source in the proton–proton cycle and the carbon–nitrogen cycle, obtained by standard solar model calculations. It may be noticed that the total neutrino flux at the earth is 6.6×10^{10} neutrinos/cm^{-2} sec^{-1}, and 92% of the flux can be attributed to the low-energy neutrinos from the proton–proton reaction.

B. The Chlorine Solar Neutrino Experiment

In the early 1960s, it was clear that the neutrinos from the sun could be observed. This conclusion was reached after it was realized that the PP-II and PP-III branches were an important part of the energy generation of the sun and sufficiently energetic to drive the neutrino capture reaction

$$\nu_e + {}^{37}\mathrm{Cl} \rightleftarrows {}^{37}\mathrm{Ar} + e^{-}. \qquad (13)$$

This reaction has too large a threshold energy (0.816 MeV) for observing the abundant neutrinos from the p–p reaction but is suitable for measuring the neutrino flux from the PP-II and PP-III branches and from the carbon–nitrogen cycle. A radiochemical method is used which employs a large volume of a suitable chemical compound of chlorine. The radioactive product is removed, purified, and placed in a small detector for measuring the decay of ^{37}Ar back to ^{37}Cl (half-life, 35 days). There are several reasons for choosing the Cl–Ar method for solar neutrino detection: chlorine compounds are inexpensive; the ^{37}Ar can be recovered by a simple efficient chemical procedure; ^{37}Ar has a convenient half-life of 35 days; and the decay of ^{37}Ar is easily measured.

A radiochemical detector does not observe or characterize a neutrino capture event. It is only capable of measuring a radioactive product that has accumulated over a period of time. In designing a detector, it is therefore important to consider all possible nuclear processes that could produce ^{37}Ar in the detector and to be sure that they are small compared to the production expected from solar neutrinos. In general, these background processes are small for radiochemical detectors. It is necessary, however, to carry out the measurement deep underground to reduce the ^{37}Ar production by cosmic ray muons to a level that is low compared to that expected from solar neutrino capture.

During the period 1965–1967, Brookhaven National Laboratory built a chlorine solar neutrino detector in the Homestake Gold Mine at Lead, South Dakota, to make a quantitative test of the theory of solar energy generation. The detector uses 380,000 liters (615 tons) of tetra-chloroethylene (C_2Cl_4) as the target material. This liquid is an inexpensive, commerical, drycleaning fluid, known as perchloroethylene. The experiment was installed 1480 m (4850 ft) underground in a specially designed chamber. Figure 1 shows the experimental arrangement.

It was learned, in the first few years of operation, that the ^{37}Ar production rate in the detector was lower than expected from the SSM and, in fact, was comparable to the expected background rate in the detector. It was necessary to improve the sensitivity of detection, which was accomplished by improving the counting technique so that ^{37}Ar decay events could be distinguished clearly from background events. After these improvements, measurements from 1970 to 1990 produced a positive result. A radiochemical detector observes the sum of the production rate from all solar neutrino sources having an energy above the threshold energy. Table III lists the theoretical neutrino flux and neutrino capture cross section of ^{37}Cl for each neutrino reaction that occurs in the sun. The chlorine experiment would respond to the sum of flux × cross section for each of these sources. The expected neutrino capture rate in the chlorine experiment, derived from the most recent theoretical calculations, is between 5.8 and 7.9×10^{-36} captures per second per ^{37}Cl atom. This rate is usually expressed in solar neutrino units (SNU); an SNU is defined as 10^{-36} captures per second per atom. It is difficult to express quantitatively the uncertainties in the solar model prediction, since the model may not correctly represent in detail the structure and dynamics of the sun. If however, one uses the combined errors in the astronomical and nuclear reaction data employed in the calculation, an uncertainty of about 1 SNU is obtained.

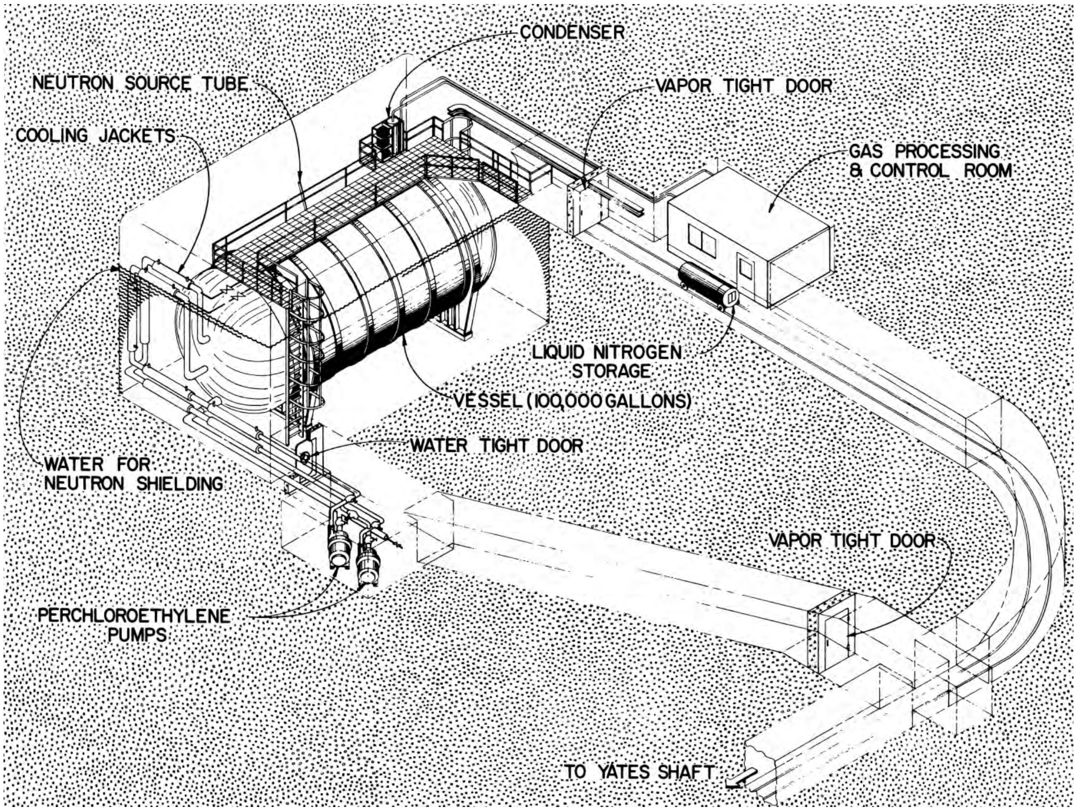


FIGURE 1 The chlorine (^{37}Cl) solar neutrino detector in the Homestake Gold Mine, Lead, South Dakota.

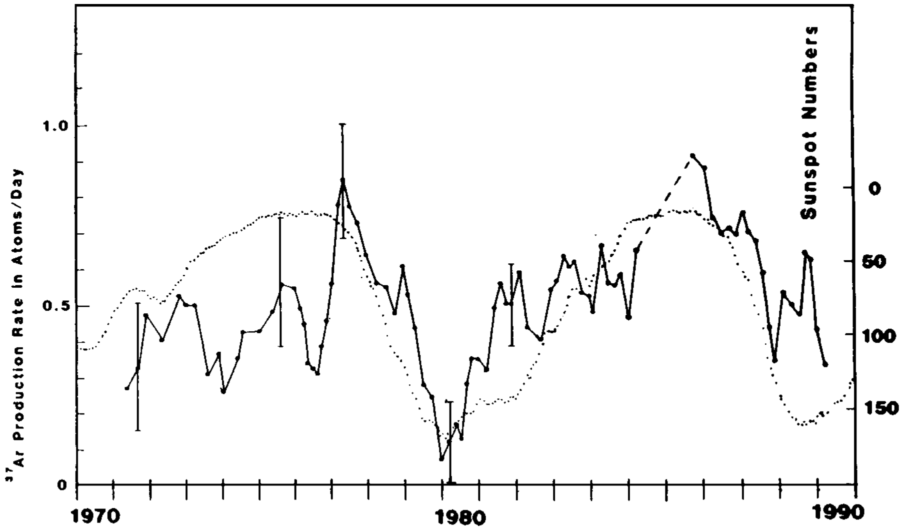


FIGURE 2 Plots showing a five point running average of ^{37}Ar production, solid circles; smoothed sunspot numbers, dotted curve; both against time in years.

The average rate observed by the chlorine detector from 1970 to 1990 was 2.3 ± 0.3 SNU. The average rate is a factor of 3–4 below the theoretical predictions in Table III.

It is of interest to see if there are any systematic changes in the signal rate of the chlorine detector with time. Figure 2 shows the data smoothed by taking a five point running average. In this figure, the solar activity cycle, as measured by the monthly average sunspot numbers, is plotted with inverted scale. The data suggest that from 1977 to 1988 the neutrino capture rate might have been inversely correlated with the quantity of sunspot numbers. The greatest change in rate occurs at about the time the sunspot cycle turns on. The statistical significance of this correlation is reasonably good; the correlation coefficient is 0.8. This effect has not been observed in other, later experiments, however. Moreover, it is generally believed that the solar energy processes do not change in short periods of time ($\approx 10^5$ years) and the solar core is regarded as a constant (in time) source of neutrinos.

C. The Kamiokande-II Solar Neutrino Detector

The Kamiokande-II detector, shown schematically in Fig. 3, was an imaging water Čerenkov detector of useful mass 2140 metric tons, of which the central 680 tons make up the fiducial mass for observation of the ^{8}B solar neutrinos. Kamiokande-II (K-II) was the successor to the nucleon decay experiment (nde) located in the Kamioka mine in Gifu prefecture, Japan. K-II was a collaboration among the Japanese universities Tokyo, Niigata, Tokai, Osaka, and Kobe, together with the National Laboratory for High Energy Physics of Japan (KEK), and with the University of Pennsylvania in the United States. Solar neutrinos were detected through the reaction $\nu_e + e^- \rightarrow \nu_e + e^-$ and by measuring the initial position and vector momentum of the recoiling electron. It is difficult to detect electrons with energy less than about 6 MeV in such large water Čerenkov detectors, and consequently, the energy threshold for solar neutrino detection was approximately that value. On the other hand, the kinematics of the elastic scattering reaction imposes an angular constraint on the recoiling electron, $\theta_e^2 \leq 2m_e/E_\nu$, which implies that the recoiling electron direction is closely aligned with the incident neutrino direction. This, combined with the imaging property of the detector, made it possible to project the incident neutrinos back to the sun and to establish the sun as their origin.

The properties of the K-II detector relevant to the detection of solar neutrinos are given in Table IV. The resolutions in energy, position, and angle in Table IV

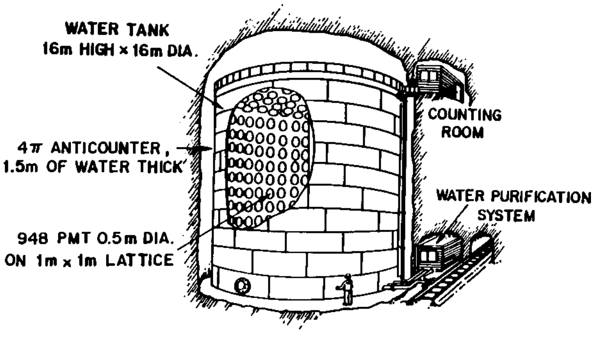


FIGURE 3 Schematic drawing of the Kamiokande-II detector in the Kamioka Mine in Japan, showing the location of the water purification system, anticounter, and electronics hut.

TABLE IV Properties of the Kamiokande-II Detector at Low Energies

Trigger threshold (at present)[a]	6.1 MeV
Energy resolution (10 MeV)	±20%
Vertex position resolution (10 MeV)	±1.0 m
Angular resolution (10 MeV)	±28
Energy calibration uncertainty	≤3%

[a] Detection efficiency was 50% at 6.1 MeV and 90% at 7.9 MeV over the fiducial volume of 680 tons. Trigger required ≥20 photomultiplier tubes to give signal within 100 nsec.

were limited by the low intensity of Čerenkov radiation, by scattering of the Čerenkov light in the detector water, and principally by multiple scattering of the low energy recoiling electrons in the water. The energy calibration was achieved through measurement of photons from a radioactive source inserted in the detector, through measurement of the energy spectra of electrons from the decays of cosmic ray muon-induced spallation products in the detector water, and through measurement of the spectrum of electrons from the decays of cosmic ray muons that stop in the detector.

For the low energy electrons produced by solar neutrino interactions, there are backgrounds arising from natural radioactivity, primarily radon dissolved in the detector water, from gamma-rays emitted by radioactive elements in the rock of the cavity housing the detector, and from the beta-decays of the muon-induced spallation nuclei in the water.

After major reductions in the backgrounds by means of fiducial volume and event selection criteria, the ratio of the observed integrated (mostly remaining background) event rate to the corresponding event rate predicted by the standard solar model was roughly 20:1. The final criterion to extract the ^{8}B solar neutrino signal from the background was the angular correlation of the neutrino signal with the direction of the sun.

Figure 4 shows the distribution in cos θ_{sun} for $E_e \geq 9.3$ MeV for a data sample of 1040 live detector days, where θ_{sun} is the measured angle of the trajectory of each observed electron with respect to the radius vector from the sun (cos $\theta_{\text{sun}} = 1$ corresponds to the direction from the sun to the earth). The solid histogram in the figure gives the shape of the solar neutrino signal expected from a Monte Carlo simulation based on the energy and angular resolution of the detector. For simplicity, the area under the histogram is made equal to the numerical prediction of the solar model calculation of Bahcall and Ulrich. One sees that the cos θ_{sun} distribution is enhanced in the direction from the sun above the isotropic background, but the magnitude of the enhancement is less than that expected from the solar model calculation.

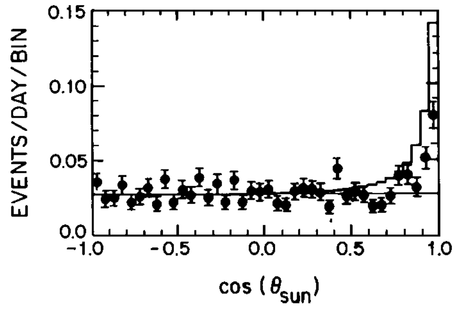


FIGURE 4 Plot of the cosine of the angle between the electron direction and a radius vector from the sun, showing the signal from the sun plus an isotropic background. This plot is for $E_e \geq 9.3$ MeV and the time period is January 1987 through April 1990, a total of 1040 live detector days.

The intensity of the observed signal relative to the standard solar model value, denoted by Data/SSM, was obtained directly from the data in Fig. 4 and was Data/SSM = 0.46 ± 0.05 (stat.) ±0.06 (syst.), where again the SSM referred to the calculation of Bahcall and Ulrich. If the SSM calculation of Turck Chieze *et al.* was used as the basis of comparison, the result was Data/SSM = 0.70 ± 0.08 (stat.) ±0.09 (syst).

The energy spectrum of the final state (recoiling) electrons induced by the solar ν_e after background subtraction was also determined and is shown in Fig. 5 along with the best fit of a calculated spectrum based on the cross section for the reaction $\nu_e e \rightarrow \nu_e e$, the known shape of the energy distribution of the ν_e from ^{8}B decay, and on the measured energy resolution of the K-II detector. The shapes of the data and the histogram in Fig. 5 are seen to be in good

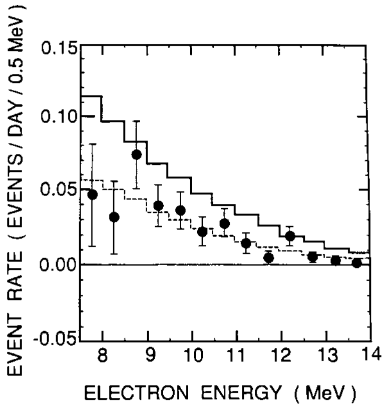


FIGURE 5 Differential electron total energy distribution of the events produced by ^{8}B solar neutrinos. The dashed histogram is the best fit to the data of a Monte Carlo calculation based on $\sigma(\nu_e e \rightarrow \nu_e e)$, the known shape of the neutrino flux from ^{8}B decay, and the energy resolution of the detector. The solid histogram has the shape dictated by the considerations above but the area predicted by the SSM in Bahcall and Ulrich (see Table III).

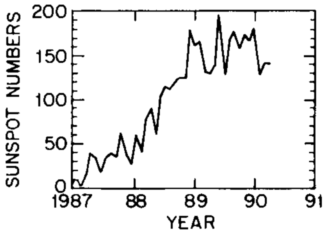


FIGURE 6 Plot of sunspot numbers vs time in the period 1987–1990.

agreement and are independent of the magnitude of the total ^{8}B neutrino flux prediction of the SSM.

The K-II data were also of interest in connection with a possible time variation of the ^{8}B solar neutrino flux because the data-taking time extended over a period in which sunspot activity, reflecting the solar magnetic cycle, rose steeply from a minimum value at the end of solar magnetic cycle 21 to a maximum value approximately 15 times larger at the peak of solar cycle 22, as shown in Fig. 6. The time dependence of the K-II data is shown in Fig. 7, in which the data are separated into five time intervals, each approximately 200 live detector days. The reduced chi-squared value calculated under the assumption of a constant flux with respect to time is 0.40, which corresponds to a confidence level of 81% in the validity of the assumption.

Other possible short-time variations of the neutrino flux, such as day/night, seasonal, and semiannual variations, are also of interest. Results of the searches for these variations in the K-II data were, within statistical errors, also negative.

The totality of the ^{8}B solar neutrino data from the K-II detector provided clear two-part evidence for a neutrino signal from ^{8}B production and decay in the sun; namely, the directional correlation of the neutrino signal with the sun, and the consistency of the differential electron energy spectrum of the signal in shape and energy scale with that expected from ^{8}B decay. Accordingly, the mechanism of energy generation in the sun based on the fusion reactions that give rise to ^{8}B as a by-product, and which were suggested by the 20 years of data from the ^{37}Cl detector, would appear to be unequivocally confirmed by the K-II detection of neutrinos that could only have originated in the core of the sun.

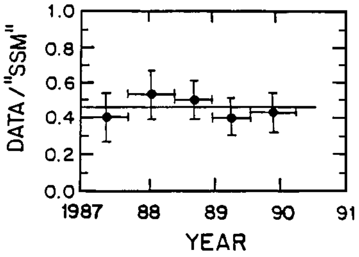


FIGURE 7 Plot from the K-II detector of the solar neutrino flux in five time intervals, each approximately 200 live detector days, from January 1987 through April 1990. The earliest two points are $E_e \geq 9.3$ MeV, and the latest three points are with $E_e \geq 7.5$ MeV.

D. Later Experiments

During the last decade of the 20th century, there were two additional radiochemical solar neutrino experiments, both using gallium, and two additional real-time, directional (electronic) experiments using massive water Čerenkov detectors, all following the methods established by the pioneer experiments described in detail above.

E. The Gallium Experiments

The gallium experiments stemmed from the intense interest in the scientific community to devise an experiment to measure the low energy neutrinos from the reaction that initiates the fusion cycle, the p–p reaction. A gallium detector ($E_{thresh} = 0.233$ MeV) utilizes the neutrino capture reaction:

$$\nu_e + {}^{71}\mathrm{Ga} \rightarrow {}^{71}\mathrm{Ge} + e^-.$$

The produced ^{71}Ge is extracted from many tons of gallium; the isolated germanium is purified; chemically converted to germane gas, Ge_4; and placed in a miniature proportional counter to observe the radioactive decay of ^{71}Ge (half-life of 11.4 days).

Two independent gallium experiments were performed, one by a collaboration of Soviet scientists from the Institute of Nuclear Research of Moscow with the American institutions of the Brookhaven and Los Alamos National Laboratories and the University of Pennsylvania. The experiment known by the acronym SAGE was located in the institute's underground laboratory in the Baksan Valley in the north Caucasus mountains. In this Soviet-American experiment, 60 tons of gallium metal (melting point 30°C) were used. The germanium was extracted by an oxidizing hydrochloric acid solution.

A second gallium experiment was prepared by the GALLEX collaboration of several institutions in Western Europe and the United States. It used 30 tons of gallium chloride-hydrochloric acid solution, from which germanium is easily removed as germanium tetrachloride by purging with nitrogen gas, an extraction technique developed earlier at Brookhaven National Laboratory. This experiment was in the Gran Sasso underground laboratory in Italy.

A gallium experiment responds to the entire spectrum of solar neutrinos; it is important to determine the neutrino capture cross section in ^{71}Ga to produce ^{71}Ge in various excited states. The ground state is easily calculated from the known half-life of ^{71}Ge, but the contribution to excited states is more difficult to determine. The proton–proton reaction alone would contribute 70 SNU. The corresponding total neutrino capture rate in an experiment using 30 tons of gallium is expected to be only 1.2 atoms of ^{71}Ge per day, which, after correction for detection efficiency, would correspond to about four observed ^{71}Ge decays per month.

F. The Electronic Detector Experiments

The electronic detector solar neutrino experiments that followed the K-II experiment were larger, more elaborate versions of the K-II detector. There were two of them, one of which has published its results and one on the verge of doing so. In addition, two other experimental measurements utilizing nuclear reactor antineutrinos, with energy spectra overlapping the solar neutrino energy spectrum, have reported results that bear importantly on the interpretation of the solar neutrino data.

G. SuperKamiokande

A Japanese-American collaboration that succeeded the original one responsible for K-II constructed the detector known as Superkamiokande (SuperK) between 1990, when K-II had largely outlived its usefulness, and 1995. The properties of K-II and SuperK—which employed the same detection method—differ in their size and in the fractional area covered by photomultiplier tubes (PMTs); SuperK is 50 kilotons in volume compared with 3 kilotons for K-II, and PMT cover 40% of the internal area of SuperK compared with 20% for K-II. The solar neutrino results from both experiments, shown in Table V with all other solar neutrino results, are in good agreement within their reported errors, although the statistical errors quoted by SuperK are significantly lower than those from K-II and the precision accordingly is significantly greater.

H. Results from Solar Neutrino Experiments

The available results from the solar neutrino experiments discussed above are given in Table V, which summarizes both the measured and the calculated values of the solar neutrino fluxes in several energy regions of the solar neutrino spectrum. The discrepancies between the measured and calculated solar neutrino fluxes in Table V are not likely to be explained by an error in the current mathematical simulation of the sun. They are thought to be the result of long-suspected properties of neutrinos which allow them to oscillate spontaneously between energy states after traversing a given distance in free space or in matter with appropriate electron density. The enhancement of neutrino oscillations in the presence of matter, the MSW effect (see suggested readings), is expected to be particularly important in the sun, where the variation of electron density with the solar radius is likely to satisfy the conditions for enhancement and to induce a level crossing between two neutrino states. The question of neutrino oscillations is further addressed below.

III. ATMOSPHERIC (OR COSMIC RAY) NEUTRINOS

Atmospheric (cosmic ray) neutrinos are extraterrestrial neutrinos which originate in the collisions of the primary proton component of cosmic rays with the earth's atmosphere, giving rise to mesons that decay to a muon and a neutrino. The muons are the penetrating charged leptons of the cosmic ray spectrum that reach the earth; the neutral, even more penetrating neutrinos are their leptonic birth partners and also the muons' decay products. The energy spectrum of the atmospheric neutrinos is shown in Fig. 8, which extends to a much higher region than the solar neutrinos.

TABLE V Summary of Recent Measured and Calculated Solar Neutrino Fluxes

Experiment or calculation	^{37}Cl → ^{37}Ar (SNU)[a]	^{71}Ga → ^{71}Ge (SNU)[a]	^{8}Bv flux (10^6 cm^{-2} s^{-1})	Reference
Homestake (DAVIS)	2.33 ± 0.25			*Annu. Rev. Nucl. Part. Sci.* **39,** 467 (1989)
GALLEX (HAMPEL)		$69.7 \pm 6.7^{+3.9}_{-4.5}$		*PL* **B388,** 384 (1996)
SAGE (ABDURASHITOV)		73^{+18+5}_{-16-7}		*PL* **B328,** 234 (1994)
Kamiokande-II (FUKUDA)			$2.80 \pm 0.19 \pm 0.33$	*PRL* **77,** 1683 (1996)
Superkamiokande (FUKUDA)			$2.42 \pm 0.05^{+0.09}_{-0.07}$	*PRL* **81,** 1158 (1998)
Bahcall *et al.*	$7.7^{+1.2}_{-1.0}$	129^{+8}_{-6}	$5.15^{+0.19}_{-0.14}$	*PL* **B433,** 1 (1998)
Brun	7.18	127.2	4.82	*Astrophys. J.* **506,** 913 (1998)
Dar	4.1 ± 1.2	115 ± 6	2.49	*Astrophys. J.* **468,** 933 (1996)
Bahcall	$9.3^{+1.2}_{-1.4}$	137^{+8}_{-7}	$6.6(100^{+0.14}_{-0.17})$	*Rev. Mod. Phys.* **67,** 781 (1995)
Turck-Chiese	6.4 ± 1.4	123 ± 7	4.4 ± 1.1	*Astrophys. J.* **408,** 347 (1993)
Bahcall	$8.0 \pm 3.0^{+}$	$132^{+21\dagger}_{-17}$	$5.69(1.00 \pm 0.43)^{b}$	*Rev. Mod. Phys.* **64,** 885 (1992)

[a] 1 SNU (solar neutrino unit) = 10^{-36} captures per atom per second.
[b] 3σ error.

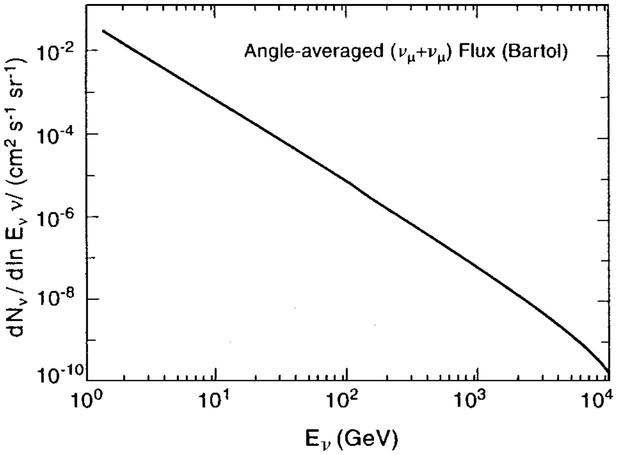


FIGURE 8 Calculated atmospheric neutrino flux, $\phi(\nu_\mu + \bar{\nu}_\mu)$, as a function of E_ν.

The three-body decay of muons includes a charged electron and a neutrino and antineutrino, each of the type or flavor to conserve a separate lepton number, i.e., one muon type and one electron type, while the meson (pion or kaon) decay produces almost exclusively muon type neutrinos or antineutrinos (see Table I). As a consequence, atmospheric neutrinos interacting in a detector such as K-II or SuperK would be expected to produce, on average, twice as many muons as electrons. The double ratio: R_{meas} (no. of muons/no. of electrons) divided by R_{calc} (no. of muons/ no. of electrons) was explored in K-II and subsequently in Super K, and the results are shown in Table VI, where it is seen that apart from one experiment (Frejus; less than three standard deviations from the average of the other experiments), the double ratio $R_{\text{meas}}/R_{\text{calc}}$ is significantly different from the expected value of unity.

TABLE VI Summary of the Ratio of Ratios $R(\mu/e) \equiv$ Measured Ratio (μ/e)/Expected Ratio (μ/e) from atmospheric neutrinos

Author	$R(\mu/e)$	Reference
SuperK		*PRL* **81,** 1562 (1998)
Fukuda	$0.63 \pm 0.03 \pm 0.05$ (sub-GeV)	
	$0.65 \pm 0.05 \pm 0.08$ (multi-GeV)	
Soudan		*PL* **B391,** 491 (1997)
Allison	$0.72 \pm 0.19^{+0.05}_{-0.07}$	
Frejus		*Z Ph* **C66,** 417 (1995)
Daum	$1.00 \pm 0.15 \pm 0.08$	
K-II		*PL* **B335,** 237 (1994)
Fukuda	$0.06^{+0.06}_{-0.05} \pm 0.05$ (sub-GeV)	
	$0.57^{+0.08}_{-0.07}$ (multi-GeV)	
IMB		*PR* **D46,** 3720 (1992)
Becker-Szendy	Close to Kamiokande value for same energy limits[b]	

[a] Systematic effects, such as flux uncertainties, tend to cancel in $R(\mu/e)$, which should equal unity in the absence of new physics.
[b] See Beier (1992). *PL* **B283,** 446.

IV. NEUTRINO OSCILLATIONS

A. Interpretation of Data

The discrepancies in both Tables V and VI probably stem from neutrino oscillations. Briefly, the idea originated in the speculation that the energy eigenstates of the three known neutrinos might differ from the weak eigenstates of the neutrinos, ν_e, ν_μ, and ν_τ, that are found in the laboratory, e.g., in nuclear beta-decay, and the decays of mesons and the weak intermediate vector bosons. Assuming each neutrino of a definite flavor (or weak eigenstate) is a coherent superposition of mass (or energy) eigenstates, and taking into account the time evolution of the energy eigenstates, one finds for the probability of a two-flavor oscillation, say, $\nu_e \leftrightarrow \nu_\mu$,

$$P(\nu_e \leftrightarrow \nu_\mu; L) = \sin^2 2\theta_{e\mu} \sin^2 \times \left[1.27\Delta m_{12}^2(\text{eV}^2)L(\text{meters})/E(\text{MeV})\right]. \quad (14)$$

Correspondingly, the probability that a beam initially of ν_e remains ν_e after traversing a distance L is

$$P(\nu_e \leftrightarrow \nu_e; L) = 1 - P(\nu_e \leftrightarrow \nu_\mu; L) - P(\nu_e \leftrightarrow \nu_\tau; L). \quad (15)$$

In Eqs. (14) and (15), L is the neutrino source to detector distance; E is the neutrino energy, $\Delta m_{12}^2 \approx |m_1^2 - m_2^2|$; and m_1, m_2 represent the mass eigenstates ν_1 and ν_2 which make up ν_e and ν_μ. The amplitude, $\sin 2\theta_{e\mu}$, specifies the strength of neutrino mixing. Most experiments searching for oscillations have been analyzed, assuming that only two flavors mix appreciably, and the numerical result is presented in terms of the regions allowed and forbidden in the parameter space of Δm^2 and $\sin^2 2\theta$. Eq. (15) describes all of the experiments in which a fraction of the initial beam has disappeared and which have therefore found allowed regions in that space.

The results from the astroneutrino experiments in Tables V and VI are consistent with different numerical descriptions of neutrino oscillations. One finds solutions in which Δm^2 is of the order of 10^{-5} eV^2 from two MSW solutions and 10^{-10} eV^2 from the so-called solar neutrino vacuum oscillation solution, nominally representing a $\nu_e \leftrightarrow \nu_\mu$ oscillation. A fourth solution yields 10^{-3} eV^2

from the atmospheric neutrino oscillation data nominally from $\nu_\mu \leftrightarrow \nu_\tau$. In three of the allowed solutions, $\sin^2 2\theta$ is large ($\sin^2 2\theta \gtrapprox 0.8$), while in one of the MSW solutions, $\sin^2 2\theta$ lies between 10^{-3} and 10^{-2}.

There is reason to believe that the above properties reflect the major neutrino oscillation channels, although only one of the three solar neutrino solutions will ultimately turn out to be valid. Neutrino searches for oscillations at nuclear reactors in France, in southern California in the United States, and in Japan, all past or near the data-taking stage, will ultimately help clarify the choice of solutions. More importantly, at least one of those experiments or the experiments discussed below may provide a conclusive demonstration that the astroneutrino data is correctly interpreted by the neutrino oscillation explanation.

B. Sudbury Solar Neutrino Experiment

During the 10-year period in which the data in Tables V and VI were acquired, the electronic solar neutrino experiment in the Sudbury Neutrino Observatory (SNO) in Canada was completed and began data taking. The heart of SNO is a transparent acryclic sphere holding 1000 tons of heavy (deuterated) water with which to observe the reaction initiated by ν_e from the ^{8}B in the sun.

$$\nu_e + \mathrm{d} \rightarrow e^- + \mathrm{p} + \mathrm{p}, \tag{16}$$

$$\nu_e + \mathrm{d} \rightarrow \nu_e + \mathrm{n} + \mathrm{p}, \tag{17}$$

and

$$\nu_x + e^- \rightarrow \nu_x \rightarrow e^-. \tag{18}$$

Note that reaction (3) will be induced by ν_e, ν_μ, and ν_τ, although with different rates. Measurement of the rates of reactions (1)–(3) should go far toward a definitive solution of the neutrino oscillation problem and, ultimately, toward absolute numerical values of neutrino masses.

Another solar neutrino electronic experiment directed primarily at a measurement of the flux of 0.861-MeV neutrinos from the reaction $^7\mathrm{Be} + e^- \rightarrow {}^7\mathrm{Li} + \nu_e$ is known as Borexino; it should help to unravel the integrated energy spectrum measurements from the radiochemical detectors. There are also less advanced plans for further detailed study of the solar neutrino spectrum in proposed electronic experiments which seek to explore the neutrinos from the primary fusion reaction $\mathrm{p} + \mathrm{p} \rightarrow \mathrm{d} + e^+ = \nu_e$.

C. Long Baseline Terrestrial Experiments

The data in Table VI are currently taken as the strongest evidence for the existence of neutrino oscillations. Despite having been observed in two independent atmospheric neutrino experiments—K-II and SuperK—a need is felt to confirm them by a long baseline experiment using terrestrial neutrinos from a particle accelerator. There are three very long baseline experiments, utilizing accelerator-produced ν_μ beams, that are aimed at a definitive test of neutrino oscillations. The neutrino source in one of them is from a newly built, external, 10-GeV proton beam at KEK, the detector is SuperK, 250 km away; the experiment is referred to as K2K. This experiment is already taking data. It will most likely duplicate the disappearance of ν_μ that was observed in the atmospheric neutrino data and probably do so as a function of muon energy. That information should be sufficient to demonstrate the validity of the neutrino oscillation interpretation of the disappearance experiments. Two other experiments—using high energy neutrinos, one (MINOS) from Fermilab and the other (ICARUS) from CERN—also plan for detectors very far away, about 730 km in each case. The Fermilab beam will be directed at the Soudan Mine in northern Minnesota; and the CERN beam will be directed at the Apennine Mountains joining east and west Italy. Both seek to observe a $\nu_\mu \leftrightarrow \nu_\tau$ oscillation directly through the appearance of ν_τ interactions at the end of the beam lines. Success depends on their ability to devise and construct ν_{tau} detectors adequate for the purpose.

A linear array of the original Kamiokande-style detectors, arranged on a long baseline (50–100 km), might also track the sinusoidal dependence on distance of the flavor content of a ν_μ beam. This disappearance experiment would bypass the difficult problem of devising a ν_τ detector for the Fermilab and CERN experiments and would still serve the equivalent purpose of demonstrating neutrino oscillations conclusively.

V. CONCLUSIONS

The observations of extraterrestrial neutrinos from the sun by different experimental methods have confirmed the details of the nuclear processes in the core of the sun and have verified its internal structure and method of energy generation. Moreover, the neutrinos from the earth's atmosphere and from the sun appear to exhibit the phenomenon of neutrino oscillations which, if fully verified, will conclusively demonstrate that at least one of the neutrinos in the current model of elementary leptons has nonzero mass and that at least two of them can spontaneously exchange flavor.

Neutrinos observed from the supernova SN1987A, approximately 55 kiloparsecs from earth, have quantitatively confirmed the current description of type II supernovae and the particle physics and astrophysics principles on which the description is based.

Together, these detailed observations have established experimental neutrino astronomy, or, perhaps more appropriately, neutrino astrophysics, as a source of fundamental information relating to the internal structure and behavior of stars. As more powerful neutrino detectors become available and permit observation of other neutrino radiation from the sky, we expect that neutrino astronomy will be an increasingly valuable adjunct of the older astronomy employing electromagnetic radiation as its means of observation.

SEE ALSO THE FOLLOWING ARTICLES

COSMIC RADIATION • DARK MATTER IN THE UNIVERSE • GAMMA-RAY ASTRONOMY • GRAVITATIONAL WAVE ASTRONOMY • NEUTRON STARS • NUCLEAR CHEMISTRY • PARTICLE PHYSICS, ELEMENTARY • SOLAR PHYSICS • STELLAR STRUCTURE AND EVOLUTION • SUPERNOVAE

BIBLIOGRAPHY

Bahcall, J. N. (1994). “Solar Neutrinos: First Thirty Years,” Addison–Wesley, Reading, MA.

Faessler, A. (1998). “Progress in Particle and Nuclear Physics: Neutrinos in Astro, Particle and Nuclear Physics,” Vol. 40, Pergamon, Elmsford, NY.

Livio, M. (2000). “Unsolved Problems in Stellar Evolution,” Cambridge Univ. Press, Cambridge, UK.

Prialnik, D. (2000). “An Introduction to the Theory of Stellar Structure and Evolution,” Cambridge Univ. Press, Cambridge, UK.

Suzuki, Y., and Totsuka, Y., eds. (1999). “Neutrino Physics and Astrophysics,” North-Holland, Amsterdam.

Winter, K. (2000). “Neutrino Physics,” 2nd ed., Cambridge Univ. Press, Cambridge, UK.

Planetary Radar Astronomy

Steven J. Ostro
California Institute of Technology

GLOSSARY

Aliasing Overlapping of echo at different frequencies or at different time delays.

Antenna gain Ratio of an antenna's sensitivity in the direction toward which it is pointed to its average sensitivity in all directions.

Circular polarization ratio Ratio of echo power received in the same sense of circular polarization as transmitted (the SC sense) to that received in the opposite (OC) sense.

Doppler shift Difference between the frequencies of the radar echo and the transmission, caused by the relative velocity of the target with respect to the radar.

Echo bandwidth Dispersion in Doppler frequency of an echo, that is, the width of the echo power spectrum.

Ephemeris Table of planetary positions as a function of time (plural: ephemerides).

Klystron Vacuum-tube amplifier used in planetary radar transmitters.

Radar albedo Ratio of a target's radar cross section in a specified polarization to its projected area; hence, a measure of the target's radar reflectivity.

Radar cross section Most common measure of a target's scattering efficiency, equal to the projected area of that perfect metal sphere that would give the same echo power as the target if observed at the target's location.

Scattering law Function giving the dependence of a surface element's radar cross section on viewing angle.

Synodic rotation period Apparent rotation period of a target that is moving relative to the observer, to be distinguished from the "sidereal" rotation period measured with respect to the fixed stars.

Time delay Time between transmission of a radar signal and reception of the echo.

PLANETARY RADAR ASTRONOMY is the study of solar system entities (the Moon, asteroids, and comets as well as the major planets and their satellites and ring systems) by transmitting a radio signal toward the target and then receiving and analyzing the echo. This field of research has primarily involved observations with Earth-based radar telescopes, but also includes certain experiments with the transmitter and/or the receiver on board a spacecraft orbiting or passing near a planetary object. However, radar studies of Earth's surface, atmosphere, or ionosphere from spacecraft, aircraft, or the ground are not considered part of planetary radar astronomy. Radar studies of the Sun involve such distinctly individual methodologies and physical considerations that solar radar astronomy is considered a field separate from planetary radar astronomy.

I. INTRODUCTION

A. Scientific Context

Planetary radar astronomy is a field of science at the intersection of planetology, radio astronomy, and radar engineering. A radar telescope is essentially a radio telescope equipped with a high-power radio transmitter and specialized electronic instrumentation designed to link transmitter, receiver, data acquisition, and telescope-pointing components together in an integrated radar system. The principles underlying operation of this system are not fundamentally very different from those involved in radars used, for example, in marine and aircraft navigation, measurement of automobile speeds, and satellite surveillance. However, planetary radars must detect echoes from targets at interplanetary distances ($\sim 10^5$–10^9 km) and therefore are the largest and most powerful radar systems in existence.

The advantages of radar observations in astronomy stem from the high degree of control exercised by the observer on the transmitted signal used to illuminate the target. Whereas virtually every other astronomical technique relies on passive measurement of reflected sunlight or naturally emitted radiation, the radar astronomer controls all the properties of the illumination, including its intensity, direction, polarization, and time/frequency structure. The properties of the transmitted waveform are selected to achieve particular scientific objectives. By comparing the properties of the echo to the very well known properties of the transmission, some of the target's properties can be deduced. Hence, the observer is intimately involved in an active astronomical observation and, in a very real sense, performs a controlled laboratory experiment on the planetary target.

Radar delay-Doppler and interferometric techniques can spatially resolve a target whose angular extent is dwarfed by the antenna beamwidth, thereby bestowing a considerable advantage on radar over optical techniques in the study of asteroids, which appear like "point sources" through ground-based optical telescopes. Furthermore, by virtue of the centimeter-to-meter wavelengths employed, radar is sensitive to scales of surface structure many orders of magnitude larger than those probed in visible or infrared regions of the spectrum. Radar is also unique in its ability to "see through" the dense clouds that enshroud Venus and the glowing gaseous coma that conceals the nucleus of a comet. Because of its unique capabilities, radar astronomy has made essential contributions to planetary exploration for a third of a century.

B. History

Radar technology was developed rapidly to meet military needs during World War II. In 1946, soon after the war's conclusion, groups in the United States and Hungary obtained echoes from the Moon, giving birth to planetary radar astronomy. These early postwar efforts were motivated primarily by interest in electromagnetic propagation through the ionosphere and the possibility for using the Moon as a "relay" for radio communication.

During the next two decades, the development of nuclear weaponry and the need for ballistic missile warning systems prompted enormous improvements in radar capabilities. This period also saw rapid growth in radio astronomy and the construction of huge radio telescopes. In 1957, the Russia launched *Sputnik* and with it the space age, and in 1958, with the formation by the U.S. Congress of the National Aeronautics and Space Administration (NASA), a great deal of scientific attention turned to the Moon and to planetary exploration in general. During the ensuing years, exhaustive radar investigations of the Moon were conducted at wavelengths from 0.9 cm to 20 m, and the results generated theories of radar scattering from natural surfaces that still see wide application.

By 1963, improvements in the sensitivity of planetary radars in both the United States and Russia had permitted the initial detections of echoes from the terrestrial planets (Venus, Mercury, and Mars). During this period, radar investigations provided the first accurate determinations of the rotations of Venus and Mercury and the earliest indications for the extreme geologic diversity of Mars. Radar images of Venus have revealed small portions of that planet's surface at increasingly fine resolution since the late 1960s, and in 1979 the Pioneer Venus Spacecraft Radar Experiment gave us our first look at Venus's global distributions of topography, radar reflectivity, and surface slopes. During the 1980s, maps having sparse coverage but resolution down to ~1 km were obtained from the Soviet *Venera 15* and *16* orbiters and from ground-based observations with improved systems. Much more recently, the *Magellan* spacecraft radar revealed most of the planet's surface with unprecedented clarity, revealing a rich assortment of volcanic, tectonic, and impact features.

The first echoes from a near-Earth asteroid (1566 Icarus) were detected in 1968; it would be nearly another decade before the first radar detection of a main-belt asteroid (1 Ceres in 1977), to be followed in 1980 by the first detection of echoes from a comet (Encke). During 1972 and 1973, detection of 13-cm-wavelength radar echoes from Saturn's rings shattered prevailing notions that typical ring particles were 0.1 to 1.0 mm in size—the fact that decimeter-scale radio waves are backscattered efficiently requires that a large fraction of the particles be larger than a centimeter. Observations by the Voyager spacecraft confirmed this fact and further suggested that particle sizes extend to at least 10 m.

In the mid-1970s, echoes from Jupiter's Galilean satellites Europa, Ganymede, and Callisto revealed the manner in which these icy moons backscatter circularly polarized waves to be extraordinarily strange and totally outside the realm of previous radar experience. We now understand that those echoes were due to high-order multiple scattering from within the top few decameters of the satellites' regoliths, but there remain important questions about the geologic structures involved and the nature of electromagnetic interactions with those structures.

The late 1980s saw the initial detections of Phobos and Titan; the accurate measurement of Io's radar properties; the discovery of large-particle clouds accompanying comets; the dual-polarization mapping of Mars and the icy Galilean satellites; and radar imaging of asteroids that revealed an extraordinary assortment of radar signatures and several highly irregular shapes, including a "contact-binary" near-Earth asteroid.

During the 1990s, the novel use of instrumentation and waveforms yielded the first full-disk radar images of the terrestrial planets, revealing the global diversity of small-scale morphology on these objects and the surprising presence of radar-bright polar anomalies on Mercury as well as Mars. Similarities between the polarization and albedo signatures of these features and those of the icy Galilean satellites argue persuasively that Mercury's polar anomalies are deposits of water ice in the floors of craters that are perpetually shaded from sunlight by Mercury's low obliquity. The first time-delay-resolved ("ranging") measurements to Ganymede and Callisto were carried out in 1992. That same year, delay-Doppler images of the closely approaching asteroid 4179 Toutatis revealed this strange object to be in a very slow, non-principal-axis spin state and provided the first geologically detailed pictures of an Earth-orbit-crossing asteroid. This decade also saw the first intercontinental radar observations and the beginning of planetary radar in Germany, Japan, and Spain. By 2000, the list of small planetary objects detected by radar included 7 comets, 37 main-belt asteroids, and 58 near-Earth asteroids (Table I).

Perhaps the most far-reaching recent development is the upgrading of the Arecibo telescope's sensitivity by over an order of magnitude. At this writing, Arecibo has just returned to operation, and radar astronomy is beginning a new era of major contributions to planetary science.

II. TECHNIQUES AND INSTRUMENTATION

A. Echo Detectability

How close must a planetary target be for its radar echo to be detectable? For a given transmitted power P_T and antenna gain G, the power flux a distance R from the radar will be $P_\mathrm{T}G/4\pi R^2$. We define the target's radar cross section, σ, as 4π times the backscattered power per unit of solid angle per unit of flux incident at the target. Then, letting λ be the radar wavelength and defining the antenna's effective aperture as $A = G\lambda^2/4\pi$, we have the received power defined as follows:

$$P_\mathrm{R} = P_\mathrm{T}GA\sigma\big/(4\pi)^2 R^4. \qquad (1)$$

This power might be much less than the receiver noise power, $P_\mathrm{N} = kT_\mathrm{S}\,\Delta f$, where k is Boltzmann's constant, T_S is the receiver system temperature, and Δf is the frequency resolution of the data. However, the mean level of P_N constitutes a background that can be determined and removed, so P_R will be detectable as long as it is at least several times larger than the standard deviation of the random fluctuations in P_N. These fluctuations can be shown to have a distribution that, for usual values of Δf and the integration time Δt, is nearly Gaussian with standard deviation $\Delta P_\mathrm{N} = P_\mathrm{N}/(\Delta f\ \Delta t)^{1/2}$. The highest signal-to-noise ratio, or $\mathrm{SNR} = P_\mathrm{R}/\Delta P_\mathrm{N}$, will be achieved for a frequency resolution equal to the effective bandwidth of the echo. As discussed in the following, that bandwidth is proportional to $D/\lambda P$, where D is the target's diameter and P is the target's rotation period, so let us assume that $\Delta f \sim D/\lambda P$. By writing $\sigma = \hat{\sigma}\pi D^2/4$, where the radar albedo $\hat{\sigma}$ is a measure of the target's radar reflectivity, we arrive at the following expression for the echo's signal-to-noise ratio:

$$\mathrm{SNR} \sim (\text{system factor})(\text{target factor}(\Delta t)^{1/2} \qquad (2)$$

where

$$\begin{aligned} \text{system factor} &\sim P_\mathrm{T}A^2\big/\lambda^{3/2}T_\mathrm{S} \\ &\sim P_\mathrm{T}G^2\lambda^{5/2}\big/T_\mathrm{S} \end{aligned} \qquad (3)$$

and

$$\text{target factor} \sim \hat{\sigma}D^{3/2}P^{1/2}/R^4. \qquad (4)$$

The inverse-fourth-power dependence of SNR on target distance is a severe limitation in ground-based observations, but it can be overcome by constructing very powerful radar systems.

B. Radar Systems

The world has two active planetary radar facilities: the Arecibo Observatory (part of the National Astronomy and Ionosphere Center) in Puerto Rico and the Goldstone

TABLE I Radar-Detected Planetary Targets

Year of first detection	Planets, satellites, rings	Main-belt asteroids	Near-Earth asteroids	Comets
1946	Moon			
1961	Venus			
1962	Mercury			
1963	Mars			
1968			1566 Icarus	
1972			1685 Toro	
1973	Saturn's rings			
1974	Ganymede			
1975	Callisto Europa		433 Eros	
1976	Io		1580 Betulia	
1977		1 Ceres		
1979		4 Vesta		
1980		7 Iris		Encke
		16 Psyche	1862 Apollo	
1981		97 Klotho	1915 Quetzalcoatl	
		8 Flora	2100 Ra-Shalom	
1982		2 Pallas		Grigg-Skjellerup
		12 Victoria		
		19 Fortuna		
		46 Hestia		
1983		5 Astraea	1620 Geographos	IRAS-Araki-Alcock
		139 Juewa	2201 Oljato	Sugano-Saigusa-Fujikawa
		356 Liguria		
		80 Sappho		
		694 Ekard		
1984		9 Metis	2101 Adonis	
		554 Peraga		
		144 Vibilia		
1985		6 Hebe	1627 Ivar	Halley
		41 Daphne	1036 Ganymed	
		21 Lutetia	1866 Sisyphus	
		33 Polyhymnia		
		84 Klio		
		192 Nausikaa		
		230 Athamantis		
		216 Kleopatra		
		18 Melpolmene		
1986		393 Lampetia	6178 1986DA	
		27 Euterpe	1986JK	
			3103 Eger (1982BB)	
			3199 Nefertiti	
1987		532 Herculina	1981 Midas	
		20 Massalia	3757 1982XB	
1988	Phobos	654 Zelinda	3908 1980PA	
		105 Artemis		

continues

TABLE I (*Continued*)

Year of first detection	Planets, satellites, rings	Main-belt asteroids	Near-Earth asteroids	Comets
1989	Titan		4034 1986PA	
			1989JA	
			4769 Castalia (1989PB)	
			1917 Cuyo	
1990		78 Diana	1990MF	
		194 Prokne	1990OS	
			4544 Xanthus (1989FB)	
1991		324 Bamberga	1991AQ	
		796 Sarita	6489 Golevka (1991JX)	
			1991EE	
1992			5189 1990UQ	
			4179 Toutatis	
1994			4953 1990MU	
1995			2062 Aten (1976AA)	
1996			1992QN	Hyakutake (C/1996 B2)
			1993QA	
			2063 Bacchus	
			1996JG	
			1991CS	
			4197 1982TA	
1997			7341 1991VK	
			7482 1994PC1	
			1997BR	
1998			1998BY7	
			6037 1988EG	
			4183 Cuno	
			1998KY26	
1998			1998 KY26	C/1998 K5
1998			8201 1994 AH2	
1998			1998 ML14	
1999			10115 1992 SK	
1999			1999 FN19	
1999			1999 GU3	
1999			1999 FN53	
1999			1999 JM8	
1999			1999 NW2	
1999			1999 RQ36	
1999			1999 TY2	
1999			1999 TN13	

Solar System Radar in California. Radar wavelengths are 13 and 70 cm for Arecibo and 3.5 and 13 cm for Goldstone; with each instrument, enormously more sensitivity is achievable with the shorter wavelength. The upgraded Arecibo telescope has twice the range and will see three times the volume of Goldstone, whereas Goldstone sees twice as much sky as Arecibo and can track targets at least three times longer. Figure 1 shows the relative sensitivities of planetary radar systems as a function of target declination.

The Arecibo telescope (Fig. 2) consists of a 305-m-diameter, fixed reflector whose surface is a 51-m-deep section of a 265-m-radius sphere. Moveable feeds designed to correct for spherical aberration are suspended from a triangular platform 137 m above the reflector and can be aimed toward various positions on the reflector, enabling

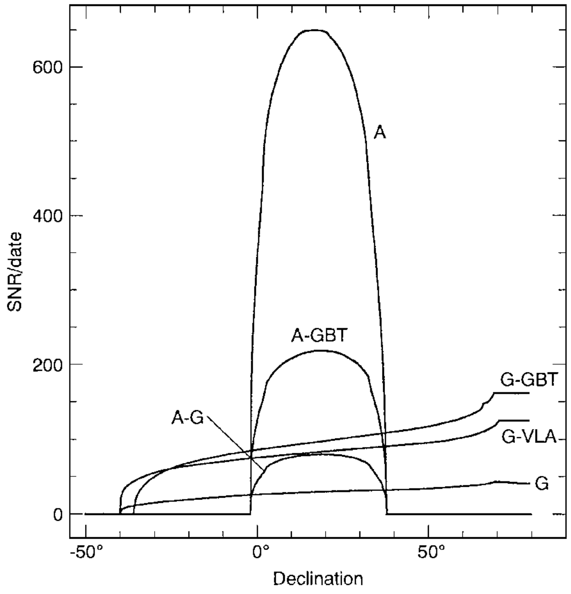


FIGURE 1 Sensitivities of planetary radar systems. Curves plot the single-date, signal-to-noise ratio of echoes from a typical 1-km asteroid ($\hat{\sigma} = 0.1$, $P = 5$ hr) at a distance of 0.1 AU for the upgraded Arecibo telescope (A), Goldstone (G), and bistatic configurations using those instruments and the Very Large Array (VLA) or the Greenbank Telescope (GBT).

the telescope to point within about 20° of the overhead direction (declination 18.3°N). Components of the recent upgrade included a megawatt transmitter, a ground screen to reduce noise generated by radiation from the ground, and replacement of most of the old single-frequency line feeds with a Gregorian reflector system (named after the 17th-century mathematician James Gregory) that employs 22-m secondary and 8-m tertiary subreflectors enclosed inside a 26-m dome.

The Goldstone main antenna, DSS-14 (DSS stands for Deep Space Station), is part of the NASA Deep Space Network, which is run by the Jet Propulsion. Laboratory (JPL). It is a fully steerable, 70-m, parabolic reflector (Fig. 3). Bistatic (two-station) experiments employing transmission from DSS-14 and reception of echoes at the 27-antenna Very Large Array (VLA) in New Mexico have synthesized a beamwidth as small as 0.24 sec of arc versus 2 min of arc for single-dish observations with Arecibo or Goldstone. Bistatic experiments using DSS-14 transmissions and reception of echoes at DSS-13, a 34-m antenna 22 km away, have been conducted on several very close targets. In coming years, bistatic observations between Arecibo and Goldstone, or using transmission from Arecibo or Goldstone and reception at the 100-m

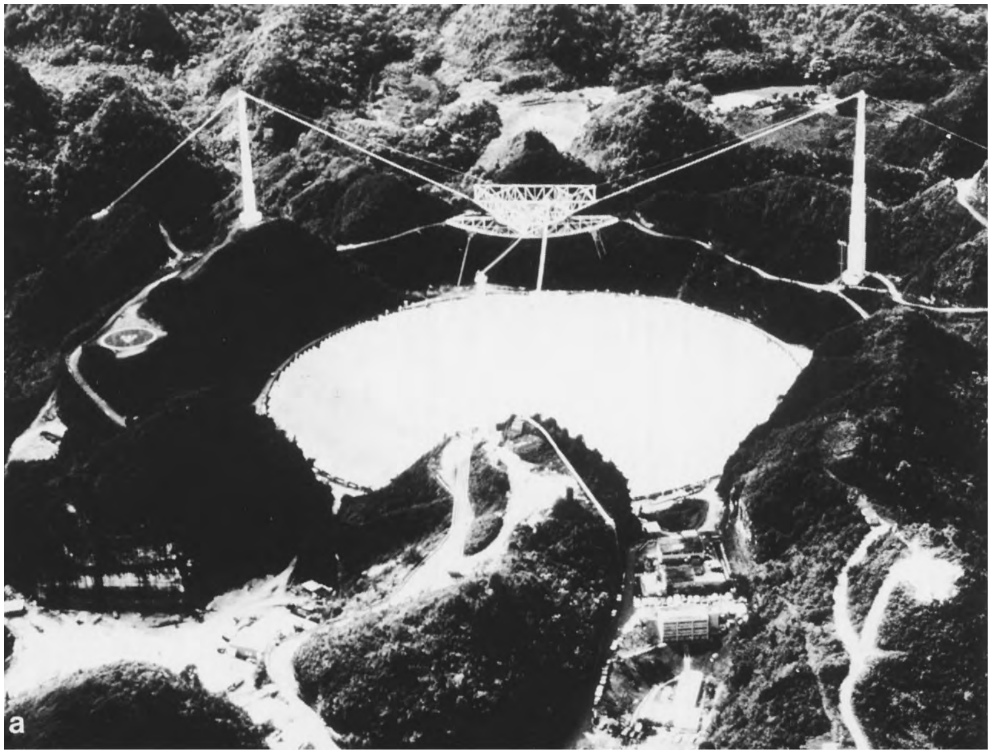


FIGURE 2 The Arecibo Observatory in Puerto Rico. (a) Aerial view prior to the recent upgrade. The diameter of the spherical reflector is 305 m. (b) A close-up view of the structure suspended above the reflector, showing the old 430-MHz line feed and the radome that encloses the new Gregorian secondary and tertiary subreflectors.

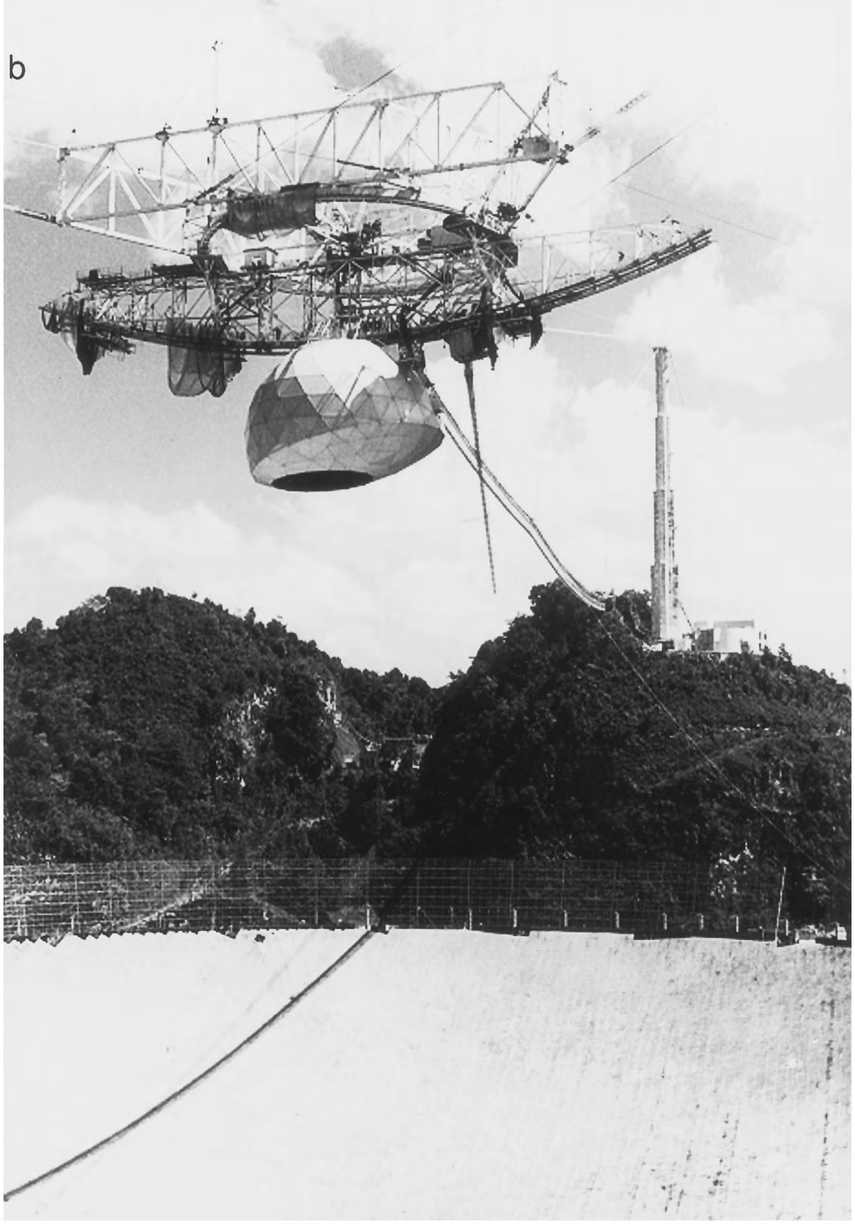


FIGURE 2 (*Continued*)

Greenbank Telescope, now under construction in West Virginia, should prove advantageous for outer planet satellites and nearby asteroids and comets.

Figure 4 is a simplified block diagram of a planetary radar system. A waveguide switch, a moveable subreflector, or a moveable mirror system is used to place the antenna in a transmitting or receiving configuration. The heart of the transmitter is one or two klystron vacuum-tube amplifiers. In these tubes, electrons accelerated by a potential drop of some 60 kV are magnetically focused as they enter the first of five or six cavities. In this first cavity, an oscillating electric field at a certain radio frequency (RF, e.g., 2380 MHz for Arecibo) modulates the electrons' velocities and hence their density and energy flux. Subsequent resonant cavities enhance this velocity bunching (they constitute what is called a "cascade amplifier") and about half of the input DC power is converted to RF power and sent out through a waveguide to the antenna feed system and radiated toward the target. The other half of the input power is waste heat and must be transported away from the klystron by cooling water. The impact of the electrons on the collector anode generates dangerous X-rays that must be contained by heavy metal shielding surrounding the tube, a requirement that further boosts

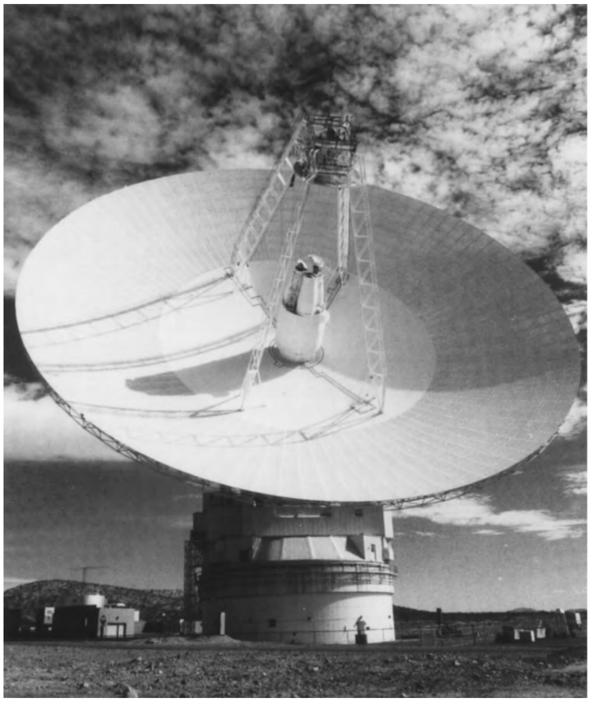


FIGURE 3 The 70-m Goldstone Solar System Radar antenna, DSS-14, in California.

the weight, complexity, and hence cost of a high-power transmitter.

In most single-antenna observations, one transmits for a duration near the roundtrip propagation time to the target (i.e., until the echo from the beginning of the transmission is about to arrive) and then receives for a similar duration. In the "front end" of the receiving system, the echo signal is amplified by a cooled, low-noise amplifier and converted from RF frequencies (e.g., 2380 MHz for Arecibo at 13 cm) down to intermediate frequencies (IF, e.g., 30 MHz), for which transmission line losses are small, and passed from the proximity of the antenna feed to a remote control room containing additional stages of signal-processing equipment, computers, and digital recorders. The signal is filtered, amplified, and converted to frequencies low enough for analog voltage samples to be digitized and recorded. The frequency down-conversion can be done in several stages using analog devices called superheterodyne mixers, but in recent years it has become possible to do this digitally, at increasingly higher frequencies. The nature of the final processing prior to recording of data on a hard disk or magnetic tape depends on the nature of the radar experiment and particularly on the time/frequency structure of the transmitted waveform. Each year, systems for reducing and displaying echoes in "real time" and techniques for processing recorded data are becoming more ambitious as computers get faster.

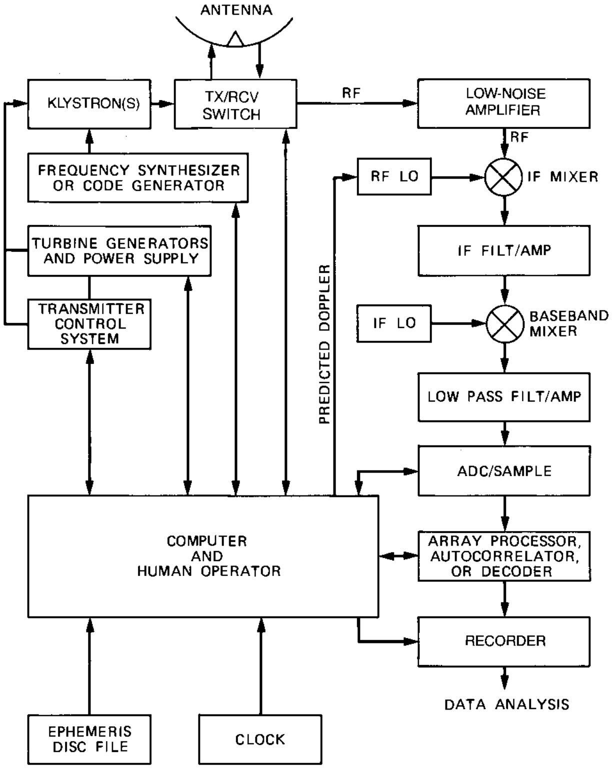


FIGURE 4 Block diagram of a planetary radar system. RF LO and IFLO denote radio frequency and intermediate frequency local oscillators, and ADC denotes analog-to-digital converter.

C. Echo Time Delay and Doppler Frequency

The time between transmission of a radar signal and reception of the echo is called the echo's roundtrip time delay, τ, and is of order $2R/c$, where c is the speed of light, 299,792,458 m $\sec^{-1}$. Since planetary targets are not points, even an infinitesimally short transmitted pulse would be dispersed in time delay, and the total extent $\Delta\tau_{\mathrm{TARGET}}$ of the distribution $\sigma(\tau)$ of echo power (in units of radar cross section) would be D/c for a sphere of diameter D and in general depends on the target's size and shape.

The translational motion of the target with respect to the radar introduces a Doppler shift, ν, in the frequency of the transmission. Both the time delay and the Doppler shift of the echo can be predicted in advance from the target's ephemeris, which is calculated using the geodetic position of the radar and the orbital elements of Earth and the target. The predicted Doppler shift can be removed electronically by continuously tuning the local oscillator used for RF-to-IF frequency conversion (see Fig. 4). Sometimes it is convenient to "remove the Doppler on the uplink" by modulating the transmission so that echoes return at a fixed frequency, the predicted Doppler must be accurate enough to avoid smearing out the echo in delay, and this requirement places stringent demands on the quality of the observing ephemeris. Time and frequency measurements are critical because the delay/Doppler distribution of echo power is the source of the finest spatial resolution and also because delay and Doppler are fundamental dynamical observables. Reliable, precise time/frequency measurements are made possible by high-speed data acquisition systems and stable, accurate clocks and frequency standards.

Because different parts of the rotating target will have different velocities relative to the radar, the echo will be dispersed in Doppler frequency as well as in time delay. The basic strategy of any radar experiment always involves measurement of some characteristic(s) of the function $\sigma(\tau, \nu)$, perhaps as a function of time and perhaps using more than one combination of transmitted and received polarizations. Ideally, one would like to obtain $\sigma(\tau, \nu)$ with very fine resolution, sampling that function within intervals whose dimensions, $\Delta\tau \times \Delta\nu$, are minute compared to the echo dispersions $\Delta\tau_{\mathrm{TARGET}}$ and $\Delta\nu_{\mathrm{TARGET}}$. However, one's ability to resolve $\sigma(\tau, \nu)$ is necessarily limited by the available echo strength. Furthermore, as described in the next section, an intrinsic upper bound on the product $\Delta\tau\ \Delta\nu$ forces a trade-off between delay resolution and Doppler resolution for the most commonly used waveforms. Under these constraints, many planetary radar experiments employ waveforms aimed at providing estimates of one of the marginal distributions, $\sigma(\tau)$ or $\sigma(\nu)$. Figure 5 shows the geometry of delay-resolution cells and Doppler-resolution cells for a spherical target and sketches the relation between these cells and $\sigma(\tau)$ and $\sigma(\nu)$. Delay-Doppler measurements are explored further in the following.

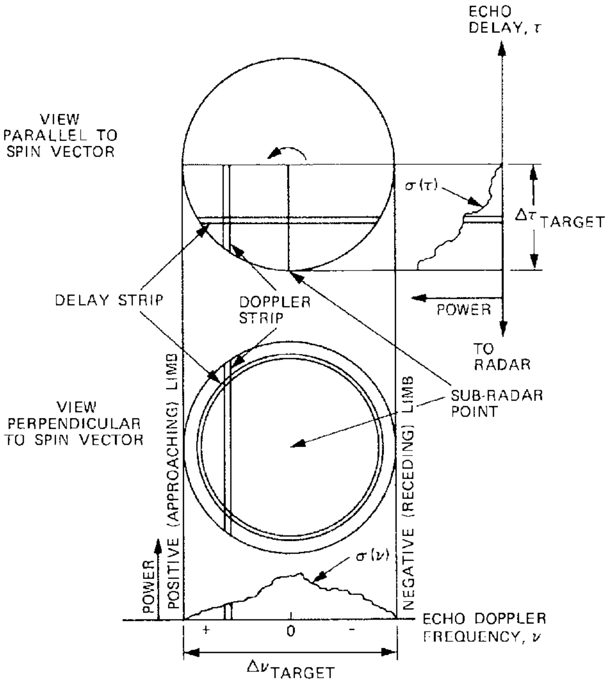


FIGURE 5 Time-delay and Doppler-frequency resolution of the radar echo from a rotating spherical target.

D. Radar Waveforms

In the simplest radar experiment, the transmitted signal is a highly monochromatic, unmodulated continuous-wave (cw) signal. Analysis of the received signal comprises Fourier transformation of a series of time samples and yields an estimate of the echo power spectrum $\sigma(\nu)$, but contains no information about the distance to the target or $\sigma(\tau)$. To avoid aliasing, the sampling rate must be at least as large as the bandwidth of the low-pass filter (see Fig. 4) and usually is comparable to or larger than the echo's intrinsic dispersion $\Delta\nu_{\mathrm{TARGET}}$ from Doppler broadening. Fast-Fourier transform (FFT) algorithms, implemented via software or hardwired (e.g., in an array processor), greatly speed the calculation of discrete spectra from time series and are ubiquitous in radar astronomy. In a single FFT operation, a string of N time samples taken at intervals of Δt seconds is transformed into a string of N spectral elements with frequency resolution $\Delta\nu = 1/N\ \Delta t$. Most planetary radar targets are sufficiently narrowband for power spectra to be computed, accumulated, and recorded directly on disk or magnetic tape at convenient intervals. In some situations, it is desirable to record time samples on tape and Fourier-analyze the data later, perhaps using FFTs of different lengths to obtain spectra at a variety of frequency resolutions. In others, it is convenient to pass the signal through an autocorrelator to record autocorrelation functions, and then apply FFTs to extract spectra.

To obtain delay resolution, one must apply some sort of time modulation to the transmitted waveform. For example, a short-duration pulse of cw signal lasting 1 μsec would provide delay resolution of 150 m. However, the echo would have to compete with the noise power in a bandwidth of order 1 MHz so the echo power from many consecutive pulses would probably have to be summed to yield a detection. One would not want these pulses to be too close together, however, or there would be more than one pulse incident on the target at once and interpretation of echoes would be insufferably ambiguous. Thus, one arranges the pulse repetition period t_{PRP} to exceed the target's intrinsic delay dispersion $\Delta\tau_{\mathrm{TARGET}}$, ensuring that the echo will consist of successive, nonoverlapping "replicas" of $\sigma(\tau)$ separated from each other by t_{PRP}. To generate this "pulsed cw" waveform, the transmitter is switched on and off while the frequency synthesizer (see Fig. 4) maintains phase coherence from pulse to pulse. Then Fourier transformation of time samples taken at the same position within each of N successive replicas of $\sigma(\tau)$ yields the power spectrum of echo from a certain delay resolution cell on the target. This spectrum has an unaliased bandwidth of $1/t_{\text{PRP}}$ and a frequency resolution of $1/Nt_{\text{PRP}}$. Repeating this process for a different position within each replica of $\sigma(\tau)$ yields the power spectrum for echo from a different delay resolution cell, and in this manner one obtains the delay-Doppler image $\sigma(\tau, \nu)$.

In practice, instead of pulsing the transmitter, one usually codes a cw signal with a sequence of 180° phase reversals and cross-correlates the echo with a representation of the code (e.g., using the decoder in Fig. 4), thereby synthesizing a pulse train with the desired values of Δt and t_{PRP}. With this approach, one optimizes SNR because it is much cheaper to transmit the same average power continuously than by pulsing the transmitter. Most modern ground-based radar astronomy observations employ cw or repetitive, phase-coded cw waveforms.

A limitation of coherent-pulsed or repetitive, binary-phase-coded cw waveforms follows from combining the requirement that there never be more than one echo received from the target at any instant (i.e., that $t_{\text{PRP}} > \Delta\tau_{\text{TARGET}}$) with the antialiasing frequency requirement that the rate $(1/t_{\text{PRP}})$ at which echo from a given delay resolution cell is sampled be no less than the target bandwidth $\Delta\nu_{\text{TARGET}}$. Therefore, a target must satisfy $\Delta\tau_{\text{TARGET}}\,\Delta\nu_{\text{TARGET}} \leq 1$ or it is "overspread" (Table II) and cannot be investigated completely and simultaneously in delay and Doppler without aliasing, at least with the waveforms discussed so far. Various degrees of aliasing may be "acceptable" for overspread factors less than about 10, depending on the precise experimental objectives and the exact properties of the echo.

How can the full delay-Doppler distribution be obtained for overspread targets? Frequency-swept and frequency-stepped waveforms have seen limited use in planetary radar; the latter approach has been used to image Saturn's rings. A new technique uses a nonrepeating, binary-phase-coded cw waveform. The received signal for any given delay cell is decoded by multiplying it by a suitably lagged replica of the code. Developed for observations of the highly overspread ionosphere, this "coded-long-pulse" or "random-code" waveform redistributes delay-aliased echo power into an additive white-noise background. The SNR is reduced accordingly, but this penalty is acceptable for strong targets.

III. RADAR MEASUREMENTS AND TARGET PROPERTIES

A. Albedo and Polarization Ratio

A primary goal of the initial radar investigation of any planetary target is estimation of the target's radar cross section, σ, and its normalized radar cross section or "radar albedo," $\hat{\sigma} = \sigma/A_{\text{p}}$, where A_{p} is the target's geometric projected area. Since the radar astronomer selects the transmitted and received polarizations, any estimate of σ or $\hat{\sigma}$ must be identified accordingly. The most common approach is to transmit a circularly polarized wave and to use separate receiving systems for simultaneous reception of the same sense of circular polarization as transmitted (i.e., the SC sense) and the opposite (OC) sense. The handedness of a circularly polarized wave is reversed on normal reflection from a smooth dielectric interface, so the OC sense dominates echoes from targets that look smooth at the radar wavelength. In this context, a surface with minimum radius of curvature very much larger than λ would "look smooth." SC echo power can arise from single scattering from rough surfaces, multiple scattering from smooth surfaces or subsurface heterogeneities (e.g., particles or voids), or certain subsurface refraction effects. The circular polarization ratio, $\mu_{\text{C}} = \sigma_{\text{SC}}/\sigma_{\text{OC}}$, is thus a useful measure of near-surface structural complexity or "roughness." When linear polarizations are used, it is convenient to define the ratio $\mu_{\text{L}} = \sigma_{\text{OL}}/\sigma_{\text{SL}}$, which would be close to zero for normal reflection from a smooth dielectric interface. For all radar-detected planetary targets, $\mu_{\text{L}} < 1$ and $\mu_{\text{L}} < \mu_{\text{C}}$. Although the OC radar albedo, $\hat{\sigma}_{\text{OC}}$, is the most widely used gauge of radar reflectivity, some radar measurements are reported in terms of the total power $(\text{OC} + \text{SC} = \text{OL} + \text{SL})$ radar albedo $\hat{\sigma}_{\text{T}}$, which is four times the geometric albedo used in optical planetary astronomy. A smooth metallic sphere would have $\hat{\sigma}_{\text{OC}} = \hat{\sigma}_{\text{SL}} = 1$, a geometric albedo of 0.25, and $\mu_{\text{C}} = \mu_{\text{L}} = 0$.

If μ_{C} is close to zero (see Table II), its physical interpretation is unique, as the surface must be smooth at all scales within about an order of magnitude of λ and there can be no subsurface structure at those scales within several $1/e$ power absorption lengths, L, of the surface proper. In this special situation, we may interpret the radar albedo as the product $g\rho$, where ρ is the Fresnel power-reflection coefficient at normal incidence and the backscatter gain g depends on target shape, the distribution of surface slopes with respect to that shape, and target orientation. For most applications to date, g is <10% larger than unity, so the radar albedo provides a reasonable first approximation to ρ. Both ρ and L depend on very interesting characteristics of the surface material, including bulk density, porosity, particle size distribution, and metal abundance.

If μ_{C} is ~0.3 (e.g., Mars and some near-Earth asteroids), then much of the echo arises from some backscattering mechanism other than single, coherent reflections from large, smooth surface elements. Possibilities include multiple scattering from buried rocks or from the interiors of concave surface features such as craters or reflections from very jagged surfaces with radii of curvature much less than a wavelength. Most planetary targets have values

TABLE II Characteristics of Selected Planetary Radar Targets[a]

Target	Minimum echo delay[b] (min)	Radar cross section (km^2)	Radar albedo $\hat{\sigma}_0$	Circular polarization ratio, μ_C	Maximum dispersions[c]		
					Delay (msec)	Doppler (Hz)	Product
Moon	0.04	6.6×10^5	0.07	0.1	12	60	0.7
Mercury	9.1	1.1×10^6	0.06	0.1	16	110	2
Venus	4.5	1.3×10^7	0.11	0.1	40	110	4
Mars	6.2	2.9×10^6	0.08	0.3	23	7600	170
Phobos	6.2	22	0.06	0.1	0.1	100	10^{-2}
1 Ceres	26	2.7×10^4	0.05	0.0	3	3100	9
2 Pallas	25	1.7×10^4	0.08	0.0	2	2000	4
12 Victoria	15	2.3×10^3	0.22	0.1	0.5	590	3
16 Psyche	28	1.4×10^4	0.31	0.1	0.8	2200	2
216 Kleopatra	20	7.1×10^3	0.44	0.0	?	750	?
324 Bamberga	13	2.9×10^3	0.06	0.1	0.8	230	0.2
1685 Toro	2.3	1.7	0.1	0.2	0.02	14	10^{-4}
1862 Apollo	0.9	0.2	0.1	0.4	0.01	16	10^{-4}
2100 Ra-Shalom	3.0	1.0	0.1	0.2	0.01	5	10^{-4}
2101 Adonis	1.5	0.02	<0.3	1.0	?	2	?
4179 Toutatis	0.4	1.3	0.24	0.3	0.01	1	10^{-5}
4769 Castalia	0.6	0.2	0.15	0.3	0.01	10	10^{-4}
6178 1986DA	3.4	2.4	0.6	0.1	12	15	0.2
1998 KY26	0.09	2.5×10^{-5}	0.01 to 0.1	0.5	0.0001	15	0.0015
IAA[d] nucleus	0.5	2.4	0.04?	0.1	?	4	?
IAA coma	0.5	0.8	?	0.01	?	600	?
HYA[e] nucleus	1.7	0.11	?	0.5	?	12	?
HYA coma	1.7	1.3	?	<1	?	3000	?
Io	66	2×10^6	0.2	0.5	12	2400	29
Europa	66	8×10^6	1.0	1.5	10	1000	11
Ganymede	66	1×10^7	0.6	1.4	18	850	15
Callisto	66	5×10^6	0.3	1.2	16	330	5
Saturn's rings	134	10^8–10^9	0.7	0.5	1600	6×10^5	10^6

[a] Typical 3.5- to 13-cm values. Question marks denote absence of radar data or of prior information about target dimensions.

[b] For asteroids and comets, this is the minimum echo time delay for radar observations to date.

[c] Doppler dispersion for transmitter frequency of 2380 MHz (λ13 cm). The product of the dispersions in delay and Doppler is the overspread factor at 2380 MHz.

[d] IAA denotes comet IRAS-Araki-Alcock.

[e] HYA denotes comet Hyakutake (C/1996 B2).

of $\mu_C < 0.3$ at decimeter wavelengths, so their surfaces are dominated by a component that is smooth at centimeter to meter scales.

The observables $\hat{\sigma}_{OC}$ and μ_C are disk-integrated quantities, derived from integrals of $\sigma(\nu)$ or $\sigma(\tau)$ in specific polarizations. How their physical interpretation profits from knowledge of the functional forms of $\sigma(\nu)$ and $\sigma(\tau)$ is shown below.

B. Dynamical Properties from Delay/Doppler Measurements

Consider radar observation of a point target a distance R from the radar. As noted above, the "roundtrip time delay" between transmission of a pulse toward the target and reception of the echo would be $\tau = 2R/c$. It is possible to measure time delays to within 10^{-7} sec. Actual delays encountered range from $2\frac{1}{2}$ sec for the Moon to $2\frac{1}{2}$ hr for Saturn's rings. For a typical target distance ~1 astronomical unit (AU), the time delay is ~1000 sec and can be measured with a fractional timing uncertainty of 10^{-9}, that is, with the same fractional precision as the definition of the speed of light.

If the target is in motion and has a line-of-sight component of velocity toward the radar of v_{LOS}, the target will "see" a frequency that, to first order in v_{LOS}/c, equals $f_{TX} + (v_{LOS}/c)f_{TX}$, where f_{TX} is the transmitter frequency. The target reradiates the Doppler-shifted signal, and the radar receives echo whose frequency is, again to first order, given by the following:

$$f_{\mathrm{TX}} + 2(v_{\mathrm{LOS}}/c) f_{\mathrm{TX}}.$$

That is, the total Doppler shift in the received echo is

$$2v_{\mathrm{LOS}} f_{\mathrm{TX}}/c = v_{\mathrm{LOS}}/(\lambda/2)$$

so a 1-Hz Doppler shift corresponds to a velocity of half a wavelength per second (e.g., 6.3 cm $\sec^{-1}$ for λ12.6 cm). It is not difficult to measure echo frequencies to within 0.01 Hz, so v_{LOS} can be estimated with a precision finer than 1 mm $\sec^{-1}$. Actual values of v_{LOS} for planetary radar targets can be as large as several tens of kilometers per second, so radar velocity measurements have fractional errors as low as 10^{-8}. At this level, the second-order (special relativistic) contribution to the Doppler shift becomes measurable; in fact, planetary radar observations have provided the initial experimental verification of the second-order term.

By virtue of their high precision, radar measurements of time delay and Doppler frequency are very useful in refining our knowledge of various dynamical quantities. The first delay-resolved radar observations of Venus, during 1961–1962, yielded an estimate of the light-second equivalent of the astronomical unit that was accurate to one part in 10^6, constituting a thousandfold improvement in the best results achieved with optical observations alone. Subsequent radar observations provided additional refinements of nearly two more orders of magnitude. In addition to determining the scale of the solar system precisely, these observations greatly improved our knowledge of the orbits of Earth, Venus, Mercury, and Mars and were essential for the success of the first interplanetary missions. Radar observations still contribute to maintaining the accuracy of planetary ephemerides for objects in the inner solar system and have played an important role in dynamical studies of Jupiter's Galilean satellites. For newly discovered near-Earth asteroids, whose orbits must be estimated from optical astrometry that spans short arcs, a few radar observations can mean the difference between successfully recovering the object during its next close approach and losing it entirely. Even for near-Earth asteroids with secure orbits, delay-Doppler measurements can shrink the positional error ellipsoid significantly for decades or even centuries.

Precise interplanetary time-delay measurements have allowed increasingly decisive tests of physical theories for light, gravitational fields, and their interactions with matter and each other. For example, radar observations verify general relativity theory's prediction that for radar waves passing nearby the Sun, echo time delays are increased because of the distortion of space by the Sun's gravity. The extra delay would be $\sim$100 μsec if the angular separation of the target from the Sun were several degrees. (The Sun's angular diameter is about half a degree.) Since planets are not point targets, their echoes are dispersed in delay and Doppler, and the refinement of dynamical quantities and the testing of physical theories are tightly coupled to estimation of the mean radii, the topographic relief, and the radar scattering behavior of the targets. The key to this entire process is resolution of the distributions of echo power in delay and Doppler. In the next section, we consider inferences about a target's dimensions and spin vector from measurements of the dispersions ($\Delta\tau_{\mathrm{TARGET}}$, $\Delta\nu_{\mathrm{TARGET}}$) of the echo in delay and Doppler. Then we examine the physical information contained in the functional forms of the distributions $\sigma(\tau)$, $\sigma(\nu)$, and $\sigma(\tau, \nu)$.

C. Dispersion of Echo Power in Delay and Doppler

Each backscattering element on a target's surface returns echo with a certain time delay and Doppler frequency (see Fig. 5). Since parallax effects and the curvature of the incident wave front are negligible for most ground-based observations (but not necessarily for observations with spacecraft), contours of constant delay are intersections of the surface with planes perpendicular to the line of sight. The point on the surface with the shortest echo time delay is called the subradar point; the longest delays generally correspond to echoes from the planetary limbs. As noted already, the difference between these extreme delays is called the dispersion, $\Delta\tau_{\mathrm{TARGET}}$, in $\sigma(\tau)$, or simply the "delay depth" of the target.

If the target appears to be rotating, the echo will be dispersed in Doppler frequency. For example, if the radar has an equatorial view of a spherical target with diameter D and rotation period P, then the difference between the line-of-sight velocities of points on the equator at the approaching and receding limbs would be $2\pi D/P$. Thus the dispersion of $\sigma(\nu)$ would be $\Delta\nu_{\mathrm{TARGET}} = 4\pi D/\lambda P$. This quantity is called the bandwidth, B, of the echo power spectrum. If the view is not equatorial, the bandwidth is simply $(4\pi D \sin\alpha)/\lambda P$, where the "aspect angle" α is the acute angle between the instantaneous spin vector and the line of sight. Thus, a radar bandwidth measurement furnishes a joint constraint on the target's size, rotation period, and pole direction.

In principle, echo bandwidth measurements obtained for a sufficiently wide variety of directions can yield all three scalar coordinates of the target's intrinsic (i.e., sidereal) spin vector $\mathbf{W}$. This capability follows from the fact that the apparent spin vector $\mathbf{W}_{\mathrm{app}}$ is the vector sum of $\mathbf{W}$ and the contribution ($\mathbf{W}_{\mathrm{sky}} = \dot{\mathbf{e}} \times \mathbf{e}$, where the unit vector $\mathbf{e}$ points from the target to the radar) from the target's plane-of-sky motion. Variations in $\mathbf{e}$, $\dot{\mathbf{e}}$, and hence $\mathbf{W}_{\mathrm{sky}}$, all of which are known, lead to measurement of different values of $\mathbf{W}_{\text{app}} = \mathbf{W} + \mathbf{W}_{\text{sky}}$, permitting unique determination of all three scalar components of $\mathbf{W}$.

These principles were applied in the early 1960s to yield the first accurate determination of the rotations of Venus and Mercury (Fig. 6). Venus's rotation is retrograde with a 243-day sidereal period that is close to the value (243.16 days) characterizing a resonance with the relative orbits of Earth and Venus, wherein Venus would appear from Earth to rotate exactly four times between successive inferior conjuctions with the Sun. However, two decades of ground-based observations and ultimately images obtained by the *Magellan* spacecraft while in orbit around Venus have conclusively demonstrated nonresonance rotation: the period is 243.0185 ± 0.0001 days. To date, a satisfactory explanation for Venus's curious spin state is lacking.

For Mercury, long imagined on the basis of optical observations to rotate once per 88-day revolution around the Sun, radar bandwidth measurements (see Fig. 6) demonstrated direct rotation with a period (59 days) equal to two-thirds of the orbital period. This spin–orbit coupling is such that during 2 Mercury years, the planet rotates three times with respect to the stars but only once with respect to the Sun, so a Mercury-bound observer would experience alternating years of daylight and darkness.

What if the target is not a sphere but instead is irregular and nonconvex? In this situation, which is most applicable to small asteroids and cometary nuclei, the relationship between the echo power spectrum and the target's shape is shown in Fig. 7. We must interpret D as the sum of the distances r_+ and r_- from the plane ψ_0 containing the line of sight and the spin vector to the surface elements with the greatest positive (approaching) and negative (receding) line-of-sight velocities. In different words, if the planes ψ_+ and ψ_- are defined as being parallel to ψ_0 and tangent to the target's approaching and receding limbs, then ψ_+ and ψ_- are at distances r_+ and r_- from ψ_0. Letting f_0, f_+, and f_- be the frequencies of echoes from portions of the target intersecting ψ_0, ψ_+, and ψ_-, we have $B = f_+ - f_-$. Note that f_0 is the Doppler frequency of hypothetical echoes from the target's center of mass and that any constant-Doppler contour lies in a plane parallel to ψ_0.

It is useful to imagine looking along the target's pole at the target's projected shape, that is, its poleon silhouette S. D is simply the width, or "breadth," of this silhouette (or, equivalently, of the silhouette's convex envelope or "hull," H) measured normal to the line of sight (see Fig. 7). In general, r_+ and r_- are periodic functions of rotation phase ϕ and depend on the shape of H as well as on the projected location of the target's center of mass, about which H rotates. If the radar data thoroughly sample at least 180° of rotational phase, then in principle one can

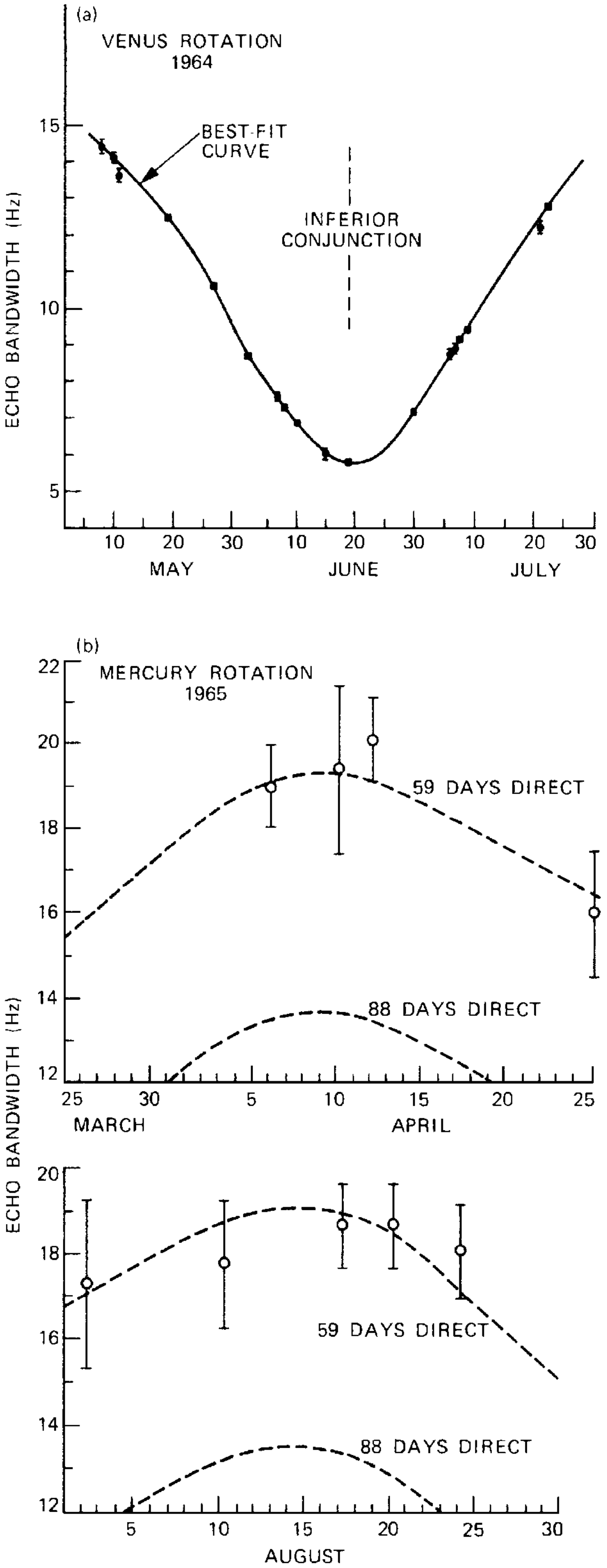


FIGURE 6 Measurements of echo bandwidth (i.e., the dispersion of echo power in Doppler frequency) used to determine the rotations of (a) Venus and (b) Mercury. [From Dyce, R. B., Pettengill, G. H., and Shapiro, I. I. (1967). *Astron. J.* **72,** 351–359.]

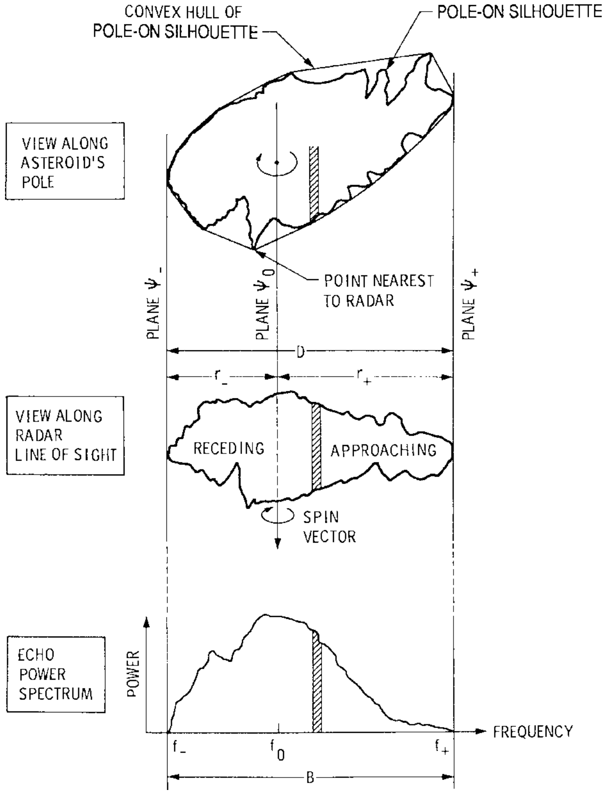


FIGURE 7 Geometric relations between an irregular, nonconvex rotating asteroid and its echo power spectrum. The plane ψ_0 contains the asteroid's spin vector and the asteroid–radar line. The cross-hatched strip of power in the spectrum corresponds to echoes from the cross-hatched strip on the asteroid.

determine $f_+(\phi)$ and $f_-(\phi)$ completely and can recover H as well as the astrometrically useful quantity f_0. For many small, near-Earth asteroids, pronounced variations in $B(\phi)$ reveal highly noncircular pole-on silhouettes (see Fig. 8 and Section III.J).

D. Topography on the Moon and Inner Planets

For the Moon, Mercury, Mars, and Venus, topography along the subradar track superimposes a modulation on the echo delay above or below that predicted by ephemerides, which generally are calculated for a sphere with the object's a priori mean radius. Prior to spacecraft exploration of these objects, there were radar-detectable errors in the radii estimates as well as in the target's predicted orbit. These circumstances required that an extended series of measurements of the time delay of the echo's leading edge be folded into a computer program designed to estimate simultaneously parameters describing the target's orbit, mean radius, and topography. These programs also contain parameters from models of wave propagation through the interplanetary medium or the solar corona, as well as parameters used to test general relativity, as noted above.

Radar has been used to obtain topographic profiles across the Moon and the inner planets. For example, Fig. 9 shows a three-dimensional reconstruction of topography derived from altimetric profiles obtained for Mars in the vicinity of the giant shield volcano Arsia Mons. The altimetric resolution of the profiles is about 150 m (1 μsec in delay), but the surface resolution, or footprint, is very coarse ($\sim$75 km). Figure 10 shows altitude profiles across impact basins on Mercury. The *Magellan* radar altimeter, with a footprint typically 20 km across and vertical resolution on the order of tens of meters, has produced detailed topographic maps of most of Venus.

E. Angular Scattering Law

The functional forms of the distributions $\sigma(\tau)$ and $\sigma(\nu)$ contain information about the radar scattering process and about the structural characteristics of the target's surface. Suppose the target is a large, smooth, spherical planet. Then echoes from the subradar region (near the center of the visible disk; see Fig. 5), where the surface elements are nearly perpendicular to the line of sight, would be much stronger than those from the limb regions (near the disk's periphery). This effect is seen visually when one shines a flashlight on a smooth, shiny ball—a bright glint appears where the geometry is right for backscattering. If the ball is roughened, the glint is spread out over a wider area and, in the case of extreme roughness, the scattering would be described as "diffuse" rather than "specular."

For a specular target, $\sigma(\tau)$ would have a steep leading edge followed by a rapid drop. The power spectrum $\sigma(\nu)$ would be sharply peaked at central frequencies, falling off rapidly toward the spectral edges. If, instead, the spectrum were very broad, severe roughness at some scale(s) comparable to or larger than λ would be indicated. In this case, knowledge of the echo's polarization properties would help to ascertain the particular roughness scale(s) responsible for the absence of the sharply peaked spectral signature of specular scattering.

By inverting the delay or Doppler distribution of echo power, one can estimate the target's average angular scattering law, $\sigma_0(\theta) = d\sigma/dA$, where dA is an element of surface area and θ is the "incidence angle" between the line of sight and the normal to dA. For the portion of the echo's "polarized" (i.e., OC or SL) component that is specularly scattered, $\sigma_0(\theta)$ can be related to statistics describing the probability distribution for the slopes of surface elements. Examples of scattering laws applied in planetary radar astronomy are the Hagfors law:

$$\sigma_0(\theta) \sim C(\cos^4\theta + C\sin^2\theta)^{-3/2}, \qquad (5)$$

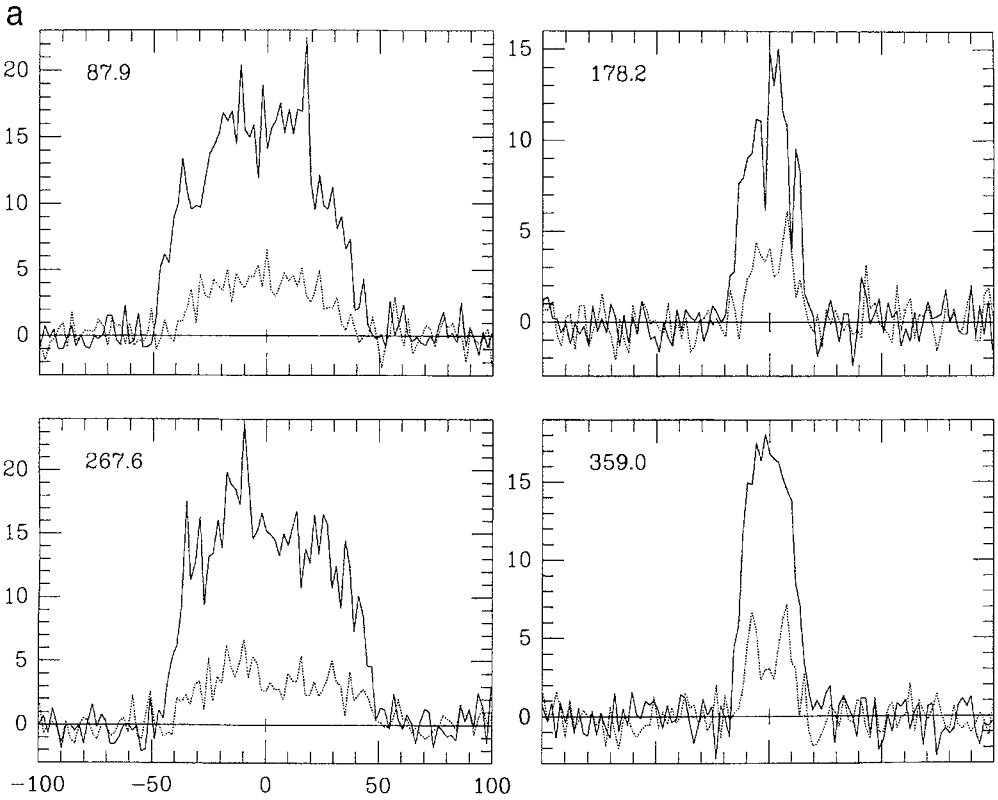


FIGURE 8 Constraints on the shape of near-Earth asteroid 1620 Geographos from Goldstone radar echoes. (a) Spectra obtained at phases of bandwidth extrema. OC (solid curve) and SC (dotted curve) echo power is plotted versus Doppler frequency. (b) Comparison of an estimate of the hull (*H*) on the asteroid's pole-on silhouette (*S*) with an estimate of *S* itself. The white curve is the cw estimate of *H* and the X marks the projected position of the asteroid's center of mass (COM) with respect to *H*. That curve and the X are superposed on an estimate of *S* from delay-Doppler images; the bright pixel is the projection of the COM determined from analysis of those images. The absolute scales and relative rotational orientations of the two figures are known: border ticks are 1 km apart. The offset between the X and the bright pixel is a measure of the uncertainty in our knowledge of the COM's delay-Doppler trajectory during the experiment. In the diagram at right, the arrows point to the observer at phases of lightcurve maxima (*M*1 and *M*2) and minima (*m*1 and *m*2). [From Ostro, S. J., *et al.* (1996). *Icarus* **121,** 46–66.]

the Gaussian law:

$$\sigma_0(\theta) \sim [C \exp(-C \tan^2 \theta)]/\cos^4 \theta, \qquad (6)$$

and the Cosine law:

$$\sigma_0(\theta) \sim (C+1)\cos^{2C} \theta, \qquad (7)$$

where $c^{-1/2} = S_0 = \langle \tan^2 \theta \rangle^{1/2}$ is the adirectional rms slope.

Echoes from the Moon, Mercury, Venus, and Mars are characterized by sharply peaked OC echo spectra (Fig. 11). Although these objects are collectively referred to as "quasispecular" radar targets, their echoes also contain a diffusely scattered component and have full-disk circular polarization ratios averaging about 0.07 for the Moon, Mercury, and Venus, but ranging from 0.1 to 0.4 for Mars, as discussed below.

Typical rms slopes obtained at decimeter wavelengths for these four quasispecular targets are around 7° and consequently these objects' surfaces have been described as "gently undulating." As might be expected, values estimated for S_0 increase as the observing wavelength decreases. For instance, for the Moon, S_0 increases from ~4° at 20 m to ~8° at 10 cm to ~33° at 1 cm. At optical wavelengths, the Moon shows no trace of a central glint, that is, the scattering is entirely diffuse. Thisphenomenon arises because the lunar surface (Fig. 12) consists of a regolith (an unconsolidated layer of fine-grained particles) with much intricate structure at the scale of visible wavelengths. At decimeter wavelengths, the ratio of diffusely scattered power to quasispecularly scattered power is about one-third for the Moon, Mercury, and Venus, but two to three times higher for Mars. This ratio can be determined by assuming that all the SC echo is diffuse and then calculating the diffusely scattered fraction (x) of OC echo by fitting to the OC spectrum a model based on a "composite" scattering law, for example,

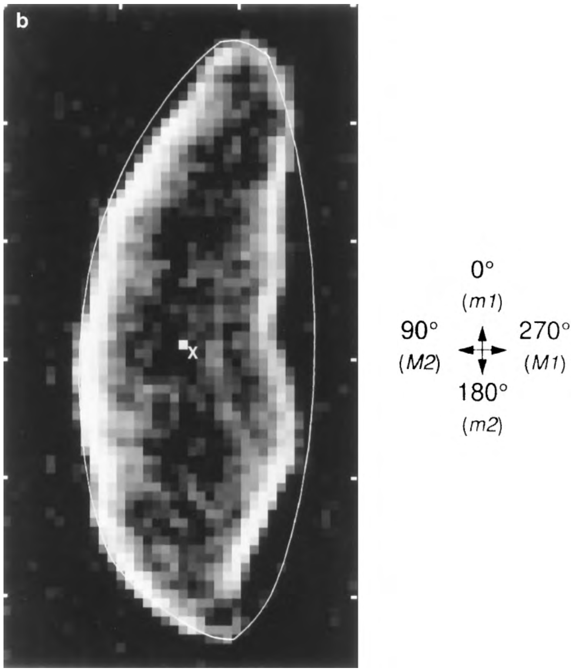


FIGURE 8 (*Continued*)

$S_0(\theta) = x\sigma_{\mathrm{DIF}}(\theta) + (1 - x)\sigma_{\mathrm{QS}}(\theta)$. Here $\sigma_{\mathrm{QS}}(\theta)$ might be the Hagfors law and usually $\sigma_{\mathrm{DIF}}(\theta) \sim \cos^m \theta$; when this is done, estimated values of m usually fall between unity (geometric scattering, which describes the optical appearance of the full Moon) and 2 (Lambert scattering).

For the large, nearly spherical asteroids 1 Ceres and 2 Pallas (see Section III.J), the closeness of μ_C to zero indicates quasispecular scattering, but the OC spectra, rather than being sharply peaked, are fit quite well using a Cosine law with C between 2 and 3 or a Gaussian law with C between 3 and 5. Here we can safely interpret the diffuse echo as due to the distribution of surface slopes, with S_0 between 20° and 50°. OC echo spectra obtained from asteriod 4 Vesta and Jupiter's satellite Io have similar shapes, but these objects' substantial polarization ratios ($\mu_C \sim 0.3$ and ~ 0.5, respectively) suggest that small-scale roughness is at least partially responsible for the diffuse echoes. Circular polarization ratios between 0.5 and 1.0 have been measured for several asteroids (see Table II) and parts of Mars and Venus, implying extreme decimeter-scale roughness, perhaps analogous to terrestrial lava flows (Fig. 13). Physical interpretations of the diffusely scattered echo employ information about albedo, scattering law, and polarization to constrain the size distributions, spatial densities, and electrical properties of wavelength-scale rocks near the surface, occasionally using the same theory of multiple light scattering applied to radiative transfer problems in other astrophysical contexts.

F. Radar Mapping of Spherical Targets

The term "radar image" usually refers to a measured distribution of echo power in delay, Doppler, and/or up to two angular coordinates. The term "radar map" usually refers to a display in suitable target-centered coordinates of the residuals with respect to a model that parameterizes the target's size, shape, rotation, average scattering properties, and possibly its motion with respect to the delay-Doppler ephemerides. Knowledge of the dimensions of the Moon and inner planets has long permitted conversion of radar images to maps of these targets. For small asteroids, the primary use of images is to constrain the target's shape (see Section III.J).

As illustrated in Fig. 5, intersections between constant-delay contours and constant-Doppler contours on a sphere constitute a "two-to-one" transformation from the target's surface to delay-Doppler space. For any point in the northern hemisphere, there is a conjugate point in the southern hemisphere at the same delay and Doppler. Therefore, the source of echo in any delay-Doppler resolution cell can be located only to within a twofold ambiguity. This north–south ambiguity can be avoided completely if the radar beamwidth (~2 arcmin for Arecibo at 13 cm or Goldstone at 3.5 cm) is comparable to or smaller than the target's apparent angular radius, as in the case of observations of the Moon (angular radius ~15 arcmin). Similarly, no such ambiguity arises in the case of side-looking radar observations from spacecraft (e.g., the *Magellan* radar), for which the geometry of delay-Doppler surface contours differs somewhat from that in Fig. 5. For ground-based observations of Venus and Mercury, whose angular radii never exceed a few tens of arcseconds, the separation of conjugate points is achievable by either (1) offsetting the pointing to place a null of the illumination pattern on the undesired hemisphere or (2) interferometrically, using two receiving antennas, as follows.

The echo waveform received at either antenna from one conjugate point will be highly correlated with the echo waveform received at the other antenna from the same conjugate point. However, echo waveforms from the two conjugate points will be largely uncorrelated with each other, no matter where they are received. Thus, echoes from two conjugate points can, in principle, be distinguished by cross-correlating echoes received at the two antennas with themselves and with each other, and performing algebraic manipulations on long time averages of the cross product and the two self products.

The echo waveform from a single conjugate point will experience slightly different delays in reaching the two

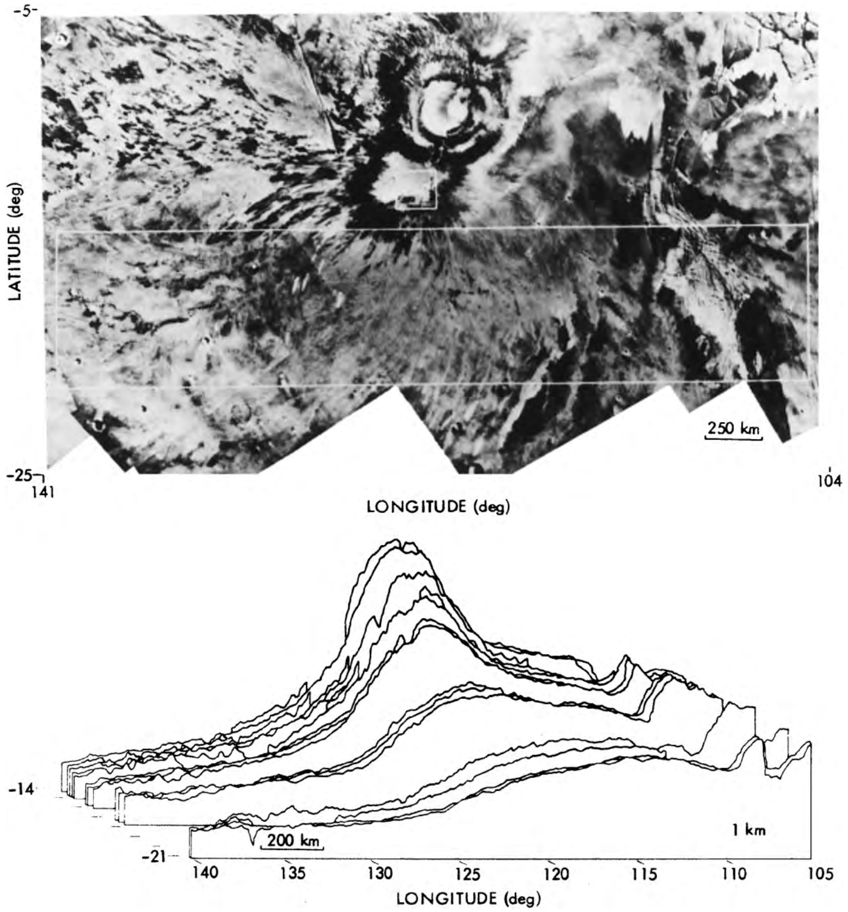


FIGURE 9 Topographic contours for the southern flank (large rectangle) of the Martian shield volcano Arsia Mons, obtained from radar altimetry. [From Roth, L., Downs, G. S., Saunders, R. S., and Schubert, G. (1980). *Icarus* **42,** 287–316.]

antennas, so there will be a phase difference between the two received signals, and this phase difference will depend only on the geometrical positions of the antennas and the target. This geometry will change as the Earth rotates, but very slowly and in a predictable manner. The antennas are best positioned so contours of constant phase difference on the target disk are as orthogonal as possible to the constant-Doppler contours, which connect conjugate points. Phase difference hence becomes a measure of north–south position, and echoes from conjugate points can be distinguished on the basis of their phase relation.

The total number of "fringes," or cycles of phase shift, spanned by the disk of a planet with diameter D and a distance R from the radar is approximately $(D/R)(b_{\mathrm{PROJ}}/\lambda)$, where b_{PROJ} is the projection of the interferometer baseline normal to the mean line of sight. For example, Arecibo interferometry linked the main antenna to a 30.5-m antenna about 11 km farther north. It placed about seven fringes on Venus, quite adequate for separation of the north–south ambiguity. The Goldstone main antenna (see Fig. 3) has been linked to smaller antennas to perform three-element as well as two-element interferometry. Tristatic observations permit one to solve so precisely for the north–south location of a given conjugate region that one can obtain the region's elevation relative to the mean planetary radius. Goldstone tristatic interferometry of the Moon's polar regions has produced topographic maps with 150-m spatial and 50-m height resolution as well as correspondingly detailed backscatter maps (Fig. 14). Altimetric information can be extracted also from bistatic observations using the time history of the phase information, but only if the variations in the projected baseline vector are large enough.

In constructing a radar map, the unambiguous delay-Doppler distribution of echo power is transformed to planetocentric coordinates, and a model is fit to the data, using a maximum-likelihood or weighted-least-squares estimator. The model contains parameters for quasispecular and diffuse scattering as well as prior information about the target's dimensions and spin vector. For Venus, effects of the dense atmosphere on radar wave propagation must also

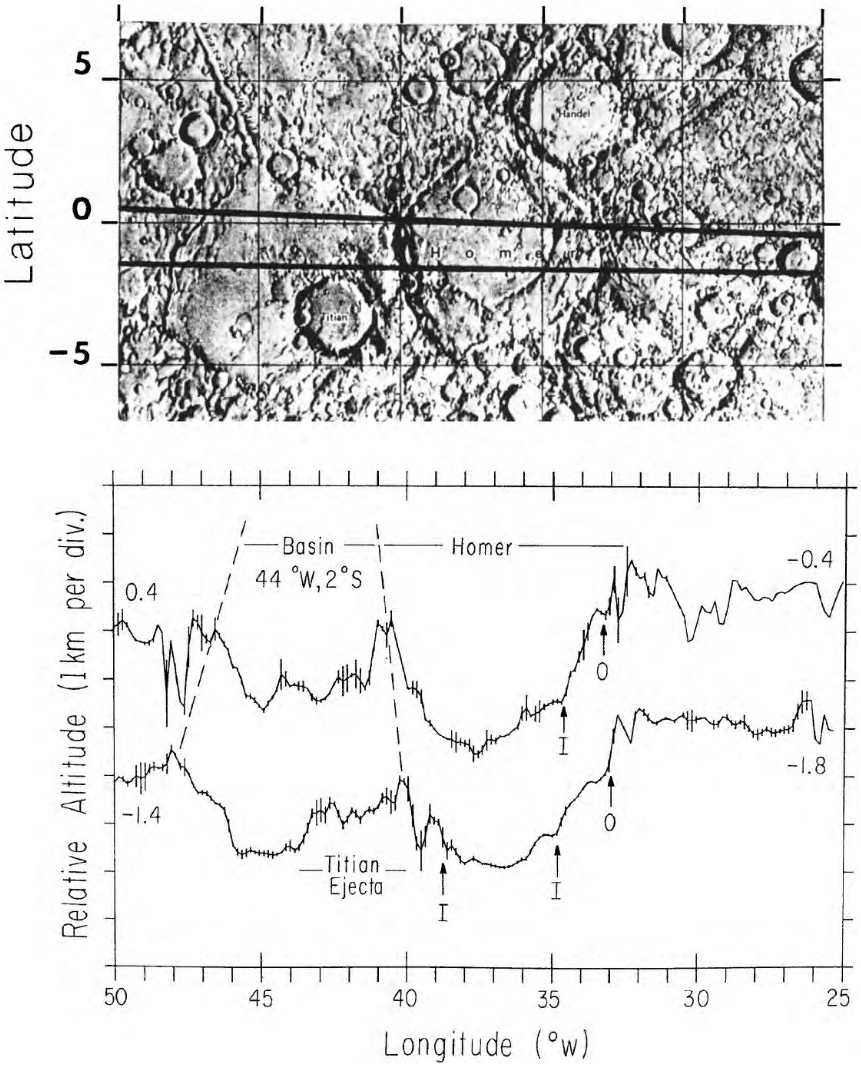


FIGURE 10 Mercury altitude profiles (bottom) showing topography across Homer Basin and a large, unnamed basin to the west, estimated from observations whose subradar tracks are shown on the USGS shaded-relief map (top). Broken lines indicate approximate locations of the basin rims as seen in *Mariner 10* images. Arrows locate Homer's inner/outer (I/O) basin rings. [From Harmon, J. K., Campbell, D. B., Bindschadler, D. L., Head, J. W., and Shaprio, I. I. (1986). *J. Geophys. Res.* **91,** 385–401.]

be modeled. Residuals between the data and the best-fit model constitute a radar reflectivity map of the planet. Variations in radar reflectivities evident in radar maps can be caused by many different physical phenomena, and their proper interpretation demands due attention to the radar wavelength, echo polarization, viewing geometry, prior knowledge about surface properties, and the nature of the target's mean scattering behavior. Similar considerations apply to inferences based on disk-integrated radar albedos.

Delay-Doppler interferometry is not currently feasible for targets like the Galilean satellites and the largest asteroids, which are low-SNR and overspread (see Section II.D). A different, "Doppler mapping" technique, developed for these spherical targets, reconstructs the global albedo distribution from cw echo spectra acquired as a function of rotation phase and at an arbitrary number of subradar latitudes. To visualize how Doppler mapping works, note that a target's reflectivity distribution can be expanded as a truncated spherical harmonic series and that the distribution of echo power in rotation phase and Doppler frequency can be obtained as a linear, analytic function of the series coefficients. Estimation of those coefficients from an observed phase-Doppler distribution can be cast as a least-squares problem to form a linear imaging system. Doppler mapping works best when the

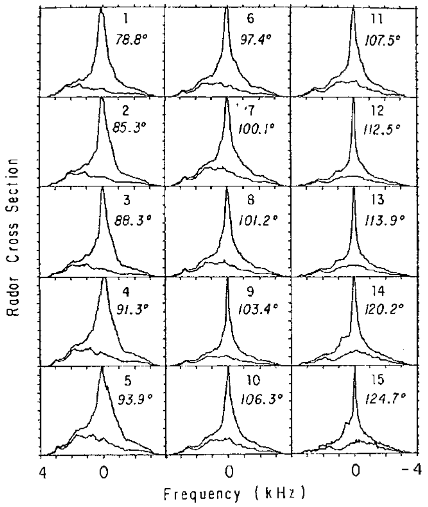


FIGURE 11 Mars 13-cm radar echo spectra for subradar points along 22° north latitude at the indicated west longitudes obtained in the OC (upper curves) and SC (lower curves) polarizations. In each box, spectra are normalized to the peak OC cross section. The echo bandwidth is 7.1 kHz. Very rough regions on the planet are revealed as bumps in the SC spectra, which move from positive to negative Doppler frequencies as Mars rotates. [From Harmon, J. K., Campbell, D. B., and Ostro, S. J. (1982). *Icarus* **52,** 171–187.]

limbdarkening is minimal and is ideally suited to overspread targets whose echoes are too weak for the random-code method. Removal of the north–south ambiguity in Doppler images (or in single-antenna delay-Doppler images) is possible if the data sample nonequatorial subradar latitudes because then the Doppler (or delay-Doppler) versus rotation-phase trajectories of conjugate points are different.

A novel kind of radar observation using Goldstone as a 3.5-cm cw transmitter and the Very Large Array in New Mexico as a synthetic aperture receiver has been used for direct measurement of the angular distribution of echo power. The 27-antenna VLA constitutes a 351-baseline interferometer whose synthesized beamwidth is as narrow as 0.24 arcsec at the Goldstone transmitter frequency, providing resolutions of roughly 80, 40, and 70 km for Mercury, Venus, and Mars at closest approach. (The finest resolutions in published ground-based delay-Doppler maps of those planets are approximately 15, 1, and 40 km, but the maps cover small fractions of the surface at such fine resolutions.) Radar aperture synthesis has produced full-disk images of those planets and Saturn's rings that are free from north–south ambiguities, avoid problems related to overspreading, and permit direct measurement of local albedo, polarization ratio, and scattering law. The Goldstone–VLA system has achieved marginal angular resolution of main-belt asteroid echoes and could readily image any large-particle comet clouds as radar detectable as the one around comet IRAS-Araki-Alcock. Goldstone–VLA resolution of asteriods and comet nuclei is impaired by the VLA's coarse spectral resolution ($>$380 Hz) and inability to accommodate time-modulated waveforms. Figures 14–18 show examples of radar maps constructed using various techniques.

G. Radar Evidence for Ice Deposits at Mercury's Poles

The first full-disk (Goldstone–VLA) radar portraits of Mercury surprisingly revealed anomalously bright polar features with $\mu_C > 1$, and subsequent delay-Doppler imagery from Arecibo established that the anomalous radar echoes originate from interiors of craters that are perpetually shaded from sunlight because of Mercury's near-zero obliquity (see Fig. 15). The angle between the orbital planes of Mercury and Earth is 7°, so portions of the permanently shadowed regions are visible to Earth-based radars. Most of the south pole anomalies are confined to the floor of the 155-km crater Chao Meng-Fu. At each pole, bright radar features in regions imaged by *Mariner 10* correlate exactly with craters; numerous radar features lie in the hemisphere not imaged by that spacecraft. [No convincing radar evidence has been found for ice in perpetually shadowed lunar craters (Fig. 14). If ice exists on the Moon, it is likely to have extremely low concentration in the soil.]

Similarities between the radar-scattering properties of the Mars and Mercury polar anomalies and those of the icy Galilean satellites (see Section III.K) support the inference that the radar anomalies are deposits of water ice. Temperatures below 120 K in the permanent shadows are expected and are low enough for ice to be stable against sublimation for billions of years. Temperatures several tens of kelvins lower may exist inside high-latitude craters and perhaps also beneath at least 10 cm of optically bright regolith. Plausible sources of water on Mercury include comet impacts and out-gassing from the interior. It has been noted that most water vapor near the surface is photodissociated, but that some molecules will random-walk to polar cold traps. Ices of other volatiles, including CO_2, NH_3, HCN, and SO_2, might also be present.

H. Venus Revealed by Magellan

The *Magellan* spacecraft entered Venus's orbit in August 1990 and during the next 2 years explored the planet with a single scientific instrument operating as a radar imager, an altimeter, and a thermal radiometer. *Magellan's* imaging resolution (~100 m) and altimetric resolution (5 to 100 m) improves upon the best previous spacecraft and

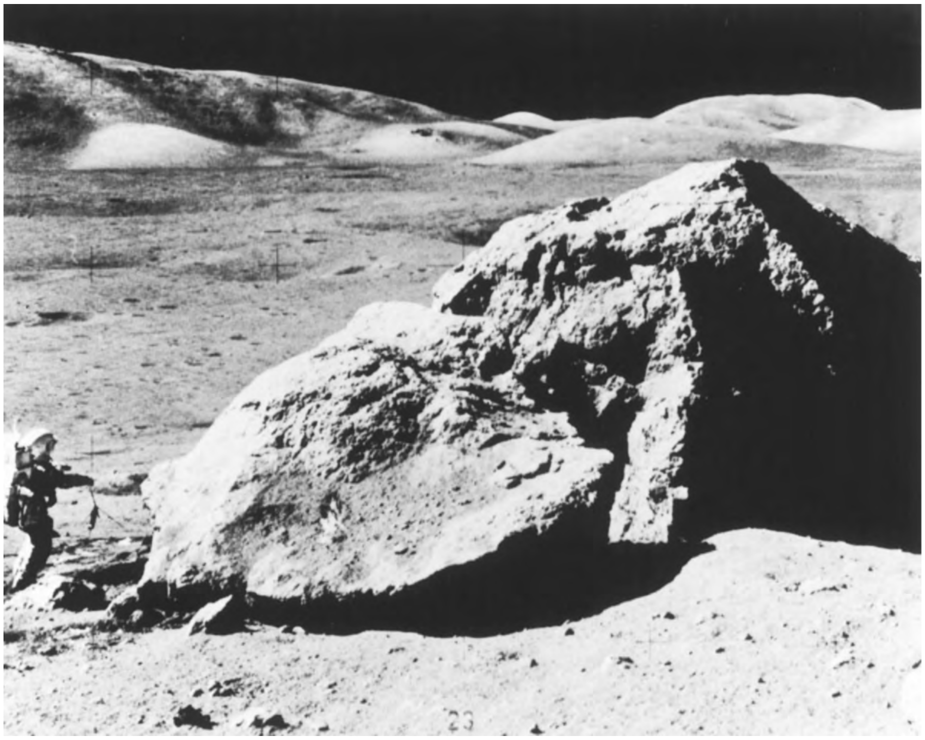


FIGURE 12 Structure on the lunar surface near the *Apollo 17* landing site. Most of the surface is smooth and gently undulating at scales much larger than a centimeter. This smooth component of the surface is responsible for the predominantly quasispecular character of the Moon's radar echo at $\lambda \gg 1$ cm. Wavelength-scale structure produces a diffuse contribution to the echo. Wavelength-sized rocks are much more abundant at $\lambda \sim 4$ cm than at $\lambda \sim 10$ m (the scale of the boulder being inspected by astronaut H. Schmitt), and hence diffuse echo is more substantial at shorter wavelengths.

ground-based measurements by an order of magnitude, and does so with nearly global coverage. Analysis of *Magellan's* detailed, comprehensive radar reconnaissance of Venus's surface topography, morphology, and electrical properties ultimately will revolutionize our understanding not only of that planet, but of planetary geology itself.

FIGURE 13 This lava flow near Sunset Crater in Arizona is an example of an extremely rough surface at decimeter scales and is similar to terrestrial flows yielding large circular polarization ratios at decimeter wavelengths.

Venus's surface contains a plethora of diverse tectonic and impact features, but its formation and evolution have clearly been dominated by widespread volacnism, whose legacy includes pervasive volcanic planes, thousands of tiny shield volcanoes, monoumental edifices, sinuous lava flow channels, pytoclastic deposits, and pancakelike domes. The superposition of volcanic signatures and elaborate, complex tectonic forms records a history of episodic crustal deformation. The paucity of impact craters smaller than 25 km and the lack of any as small as a few kilometers attests to the protective effect of the dense atmosphere. The multilobed, asymmetrical appearance of many large craters presumably results from atmospheric breakup of projectiles before impact. Atmospheric entrainment and transport of ejecta are evident in very elongated ejecta blankets. Numerous craters are surrounded by radar-dark zones, perhaps the outcome of atmospheric pressure-wave pulverization and elevation of surface material that upon resettling deposited a tenuous and hence unreflective "impact regolith." Figure 17 shows examples of *Magellan* radar images.

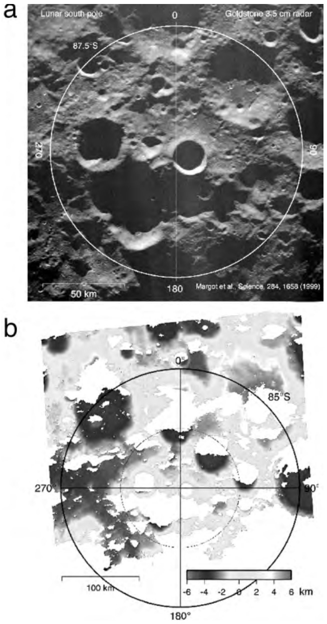


FIGURE 14 Digital elevation model (a) and radar backscatter map (b) of the lunar south polar region from Goldstone tristatic interferometry. Absolute elevations with respect to a 1738-km-radius sphere are shown. The elevation map establishes that the interiors of many of the craters visible in (b) are in permanent shadow from solar illumination. (a) Reprinted with permission from J. M. Margot, D. B. Campbell, R. F. Jurgens, and M. A. Slade, *Science* **284,** 1658–1660, copyright 1999 American Association for the Advancement of Science. (b) Courtesy J.-L. Margot.

I. The Radar Heterogeneity of Mars

Ground-based investigations of Mars have achieved more global coverage than those of the other terrestrial targets because the motion in longitude of the subradar point on Mars (whose rotation period is only 24.6 hr) is rapid compared to that on the Moon, Venus, or Mercury and because the geometry of Mars's orbit and spin vector permits subradar tracks throughout the Martian tropics. The existing body of Mars radar data reveals extraordinary diversity in the degree of small-scale roughness as well as in the rms slope of smooth surface elements. For example, Fig. 19 shows the variation in OC echo spectral shape as a function of longitude for a subradar track along ~16°S latitude. Slopes on Mars have rms values from less than 0.5° to more than 10°. Chryse Planitia, site of the first *Viking Lander*, has fairly shallow slopes (4°–5°) and, in fact, radar rms slope estimates were utilized in selection of the *Viking Lander* (and *Mars Pathfinder*) sites, and their accuracy was confirmed by the spacecraft results.

Diffuse scattering from Mars is much more substantial than for the other quasispecular targets and often accounts for most of the echo power, so the average near-surface abundance of centimeter-to-meter-scale rocks presumably is much greater on Mars than on the Moon, Mercury, or Venus. Features in Mars SC spectra first revealed the existence of regions of extremely small-scale roughness (see

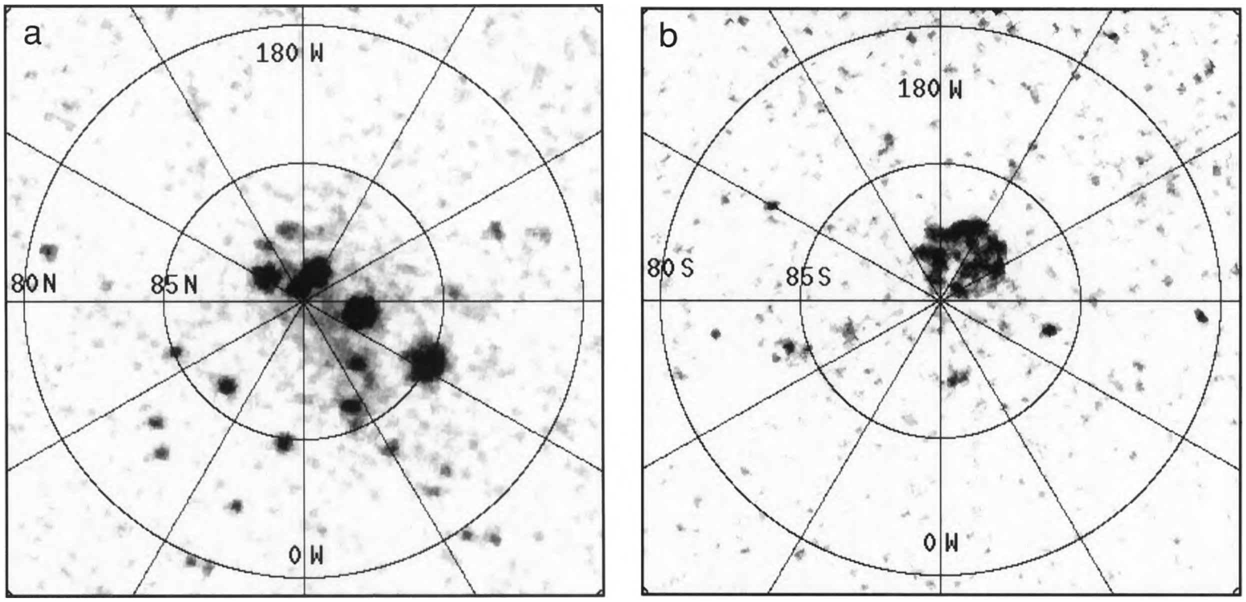


FIGURE 15 Arecibo 13-cm delay-Doppler images of the (a) north and (b) south poles of Mercury, taken in the SC polarization. The resolution is 15 km. The radar-brightness regions are shown here as dark. [From Harmon, J. K., Slade, M. A., Velez, R. A., Crespo, A., Dryer, M. J., and Johnson, J. M. (1994). *Nature* **369,** 213–215. Copyright 1994 Macmillan Magazines Limited.]

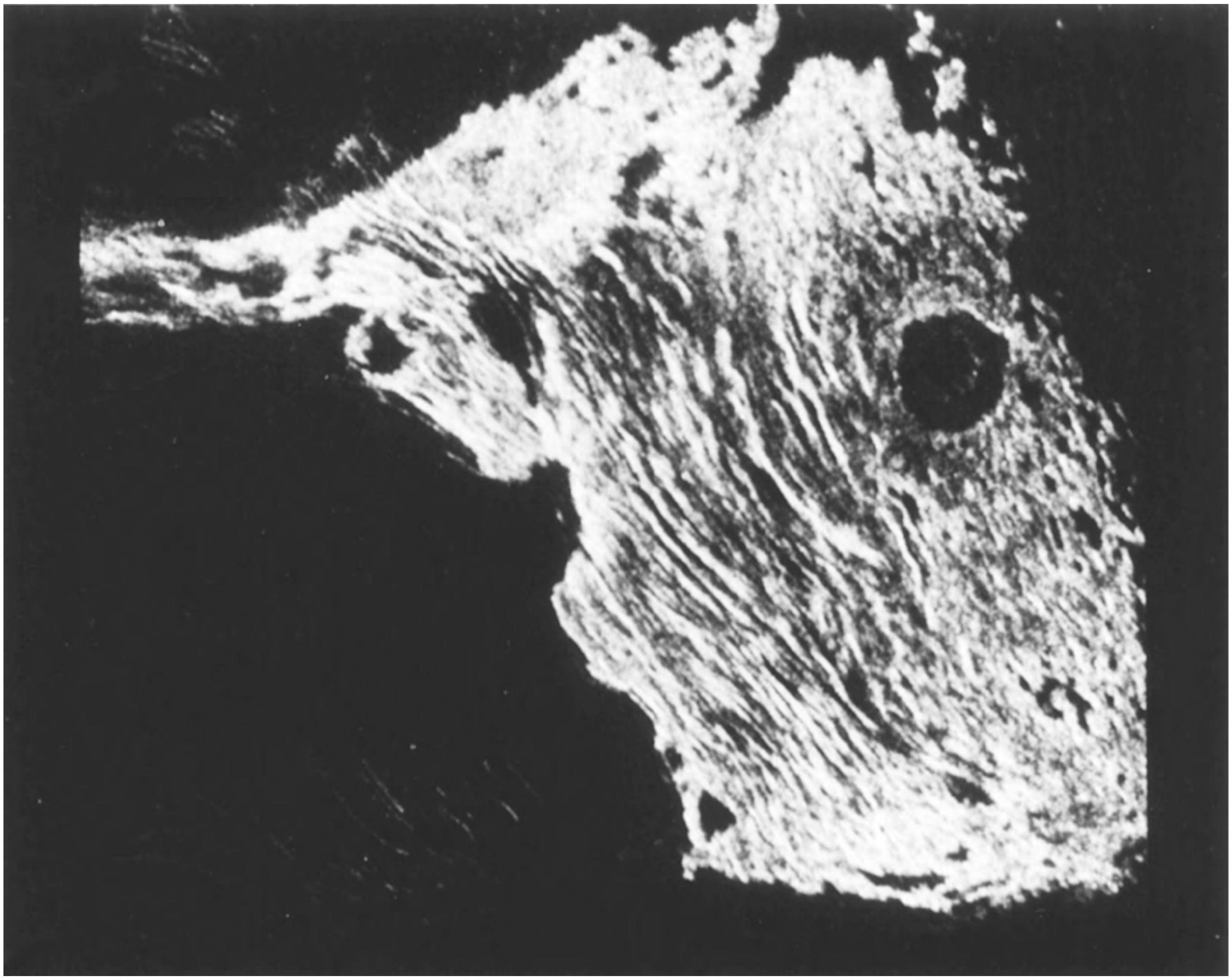


FIGURE 16 Arecibo delay-Doppler OC radar map of Maxwell Montes on Venus. (Courtesy of D. B. Campbell.)

Fig. 11), and the trajectory of these features' Doppler positions versus rotation phase suggested that their primary sources are the Tharsis and Elysium volcanic regions. The best terrestrial analog for this extremely rough terrain might be young lava flows (see Fig. 13). Goldstone–VLA images of Mars at longitudes that cover the Tharsis volcanic region (see Fig. 18) confirm that this area is the predominant source of strong SC echoes and that localized features are associated with individual volcanoes. A 2000-km-long band with an extremely low albedo cuts across Tharsis; the radar darkness of this "Stealth" feature probably arises from an underdense, unconsolidated blanket of pyroclastic deposits ~1 m deep. The strongest SC feature in the Goldstone–VLA images is the residual south polar ice cap, whose scattering behavior is similar to that of the icy Galilean satellites (Section III.K). Arecibo observations of Mars, including Doppler-only and coded-long-pulse delay-Doppler mapping, have charted the detailed locations and fine structure of Mars SC features (see Fig. 18b).

J. Asteroids

Echoes from 37 main-belt asteroids (MBAs) and 58 near-Earth asteroids (NEAs) have provided a wealth of new information about these objects' sizes, shapes, spin vectors, and surface characteristics such as decimeter-scale morphology, topographic relief, regolith porosity, and metal concentration. During the past decade, radar has been established as the most powerful Earth-based technique for determining the physical properties of asteroids that come close enough to yield strong echoes.

The polarization signatures of some of the largest MBAs (e.g., 1 Ceres and 2 Pallas) reveal surfaces that are smoother than that of the Moon at decimeter scales but much rougher at some much larger scale. For example, for

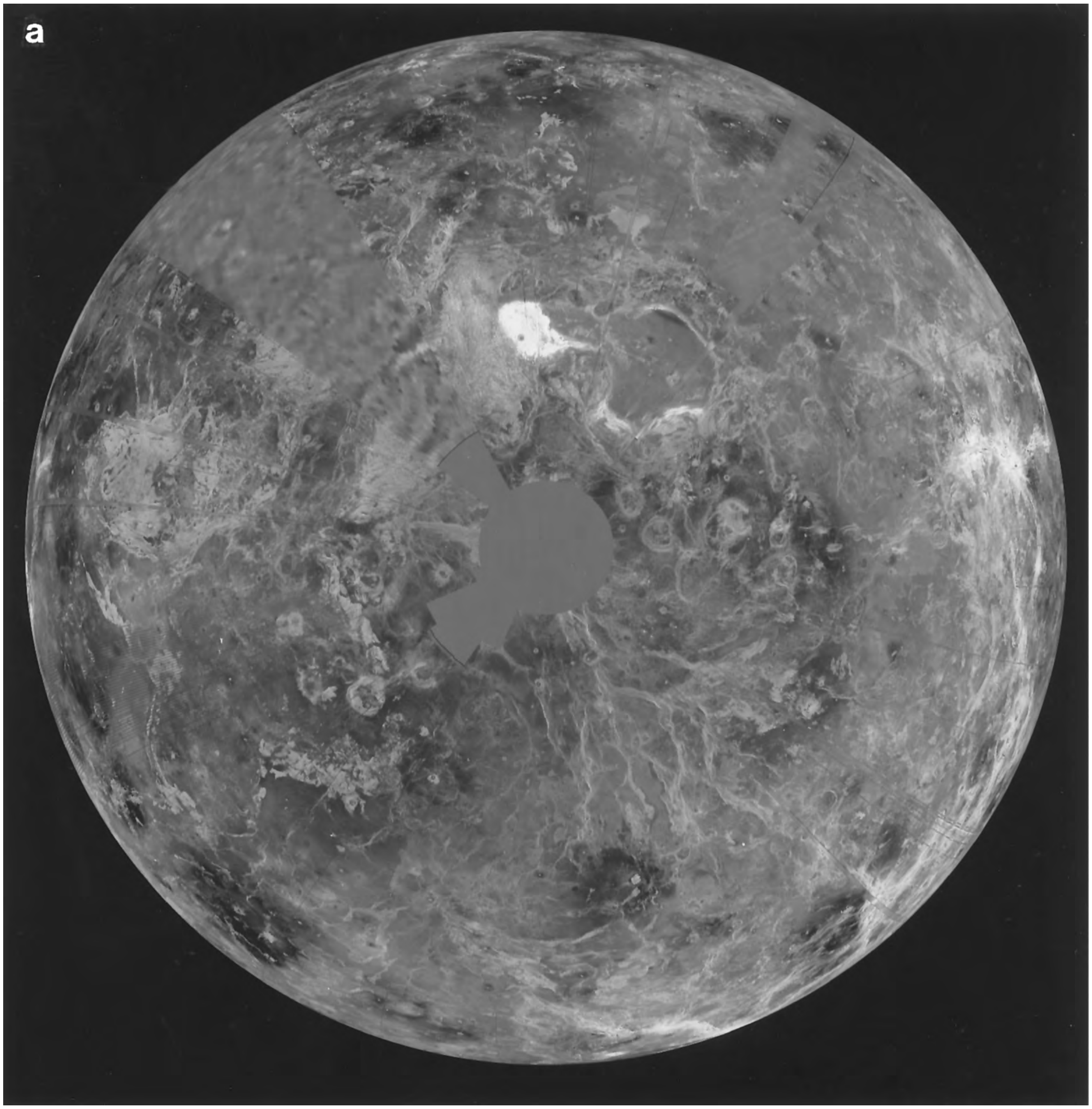


FIGURE 17 *Magellan* radar maps of Venus. (a) Northern-hemisphere projection of mosaics. The north pole is at the center of the image, with 0° and 90°E longitudes at the six and three o'clock positions. Gaps use Pioneer Venus data or interpolations. The bright, porkchop-shaped feature is Maxwell Montes, a tectonically produced mountain range first seen in ground-based images. (b) A 120-m-resolution map of the crater Cleopatra on the eastern slopes of Maxwell Montes. Cleopatra is a double-ringed impact basin that resembles such features seen on the Moon, Mercury, and Mars. (Courtesy of JPL/NASA.) (*continues*)

Pallas, μ_c is only $\sim$0.05 and, as noted above, surface slopes exceed 20°. For asteroids in the 200-km-diameter range, the echoes provide evidence for large-scale topographic irregularities. For example, brightness spikes within narrow ranges of rotation phase suggest large, flat regions on 7 Iris (Fig. 20), 9 Metis, and 654 Zelinda, and delay-Doppler images of 216 Kleopatra reveal a dumbbell shape.

There is a 10-fold variation in the radar albedos of MBAs, implying substantial variations in these objects' surface porosities or metal concentrations or both. The lowest MBA albedo estimate, 0.04 for Ceres, indicates a lower surface bulk density than that on the Moon. The highest MBA albedo estimates, 0.31 for 16 Psyche and 0.44 for Kleopatra, are consistent with metal

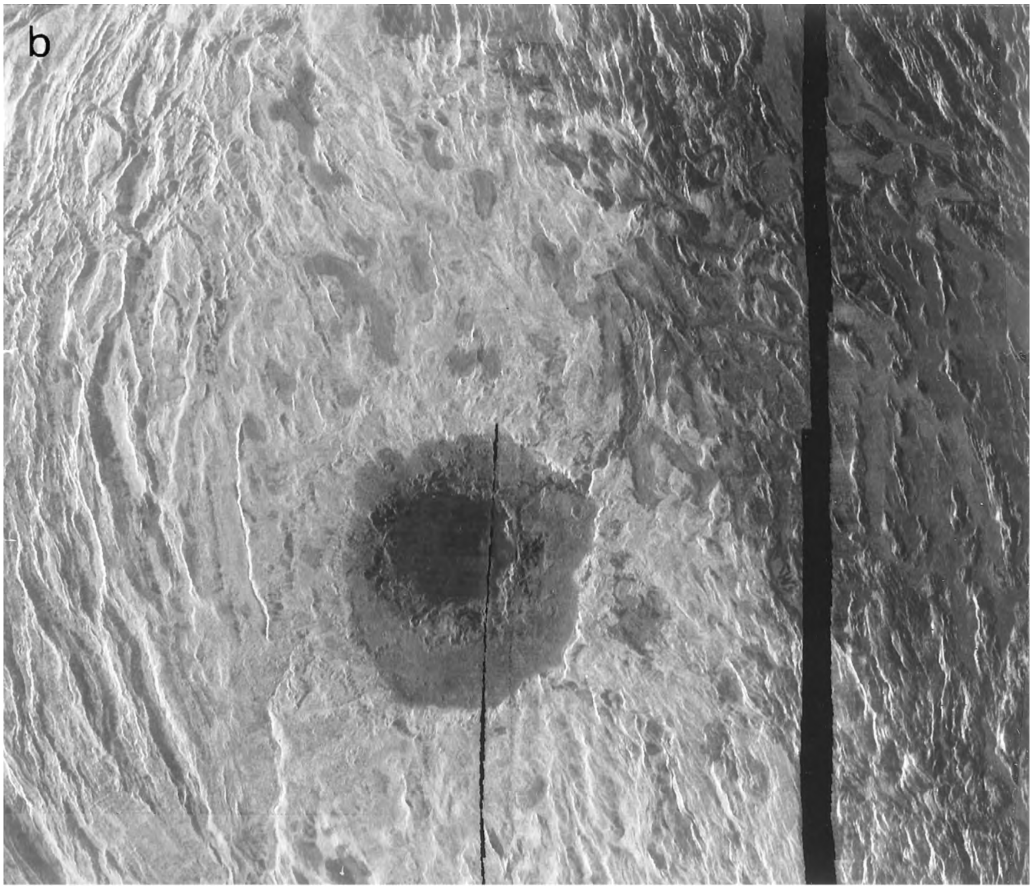


FIGURE 17 (*Continued*)

concentrations near unity and lunar porosities. These objects might be the collisionally stripped cores of differentiated asteroids and by far the largest pieces of refined metal in the solar system. Surprisingly, there is no reason to believe that the two largest classes of radar-observed asteroids (C and S) have different radar-albedo distributions.

The diversity of NEA radar signatures is extreme (see Table II). Some small NEAs are much rougher at decimeter scales than MBAs, comets, or the terrestrial planets. The radar albedo of the 2-km object 6178 (1986DA), 0.58, strongly suggests that this Earth-approacher is a regolith-free metallic fragment, presumably derived from the interior of a much larger object that melted, differentiated, cooled, and subsequently was disrupted in a catastrophic collision. This asteroid, which appears extremely irregular at 10- to 100-m scales and shows hints of being bifurcated, might be (or have been a part of) the parent body of some iron meteorites. At the other extreme, an interval estimate for 1986JK's radar albedo (0.005 to 0.07) suggests a surface bulk density within a factor of 2 of 0.9 g cm^{-3}. Similarly, the distribution of NEA circular polarization ratios runs from near zero to near unity. The highest values, for 2101 Adonis, 1992QN, 3103 Eger, and 3980 1980PA, indicate extreme near-surface structural complexity, but we cannot distinguish between multiple scattering from subsurface heterogeneities (see Section III.K) and single scattering from complex structure on the surface.

The MBAs 951 Gaspra and 243 Ida, imaged by the *Galileo* spacecraft, probably are marginally detectable with the upgraded Arecibo. Both Goldstone and Arecibo have investigated the Gaspra-sized Martian moon Phobos, whose radar properties differ from those of most small, Earth-approaching objects but resemble those of large (∼100-km), C-class, main-belt asteroids. Phobos's surface characteristics may be more representative of Ceres and Pallas than most NEAs. The upper limit on the radar cross section of Deimos, which has defied radar detection, argues for a surface bulk density no greater than about 1 g cm^{-3}.

During the past decade, delay-Doppler imaging of asteroids has produced spatial resolutions as fine as a few decameters. The images generally can be "north–south" ambiguous; that is, they constitute a two-to-one (or even

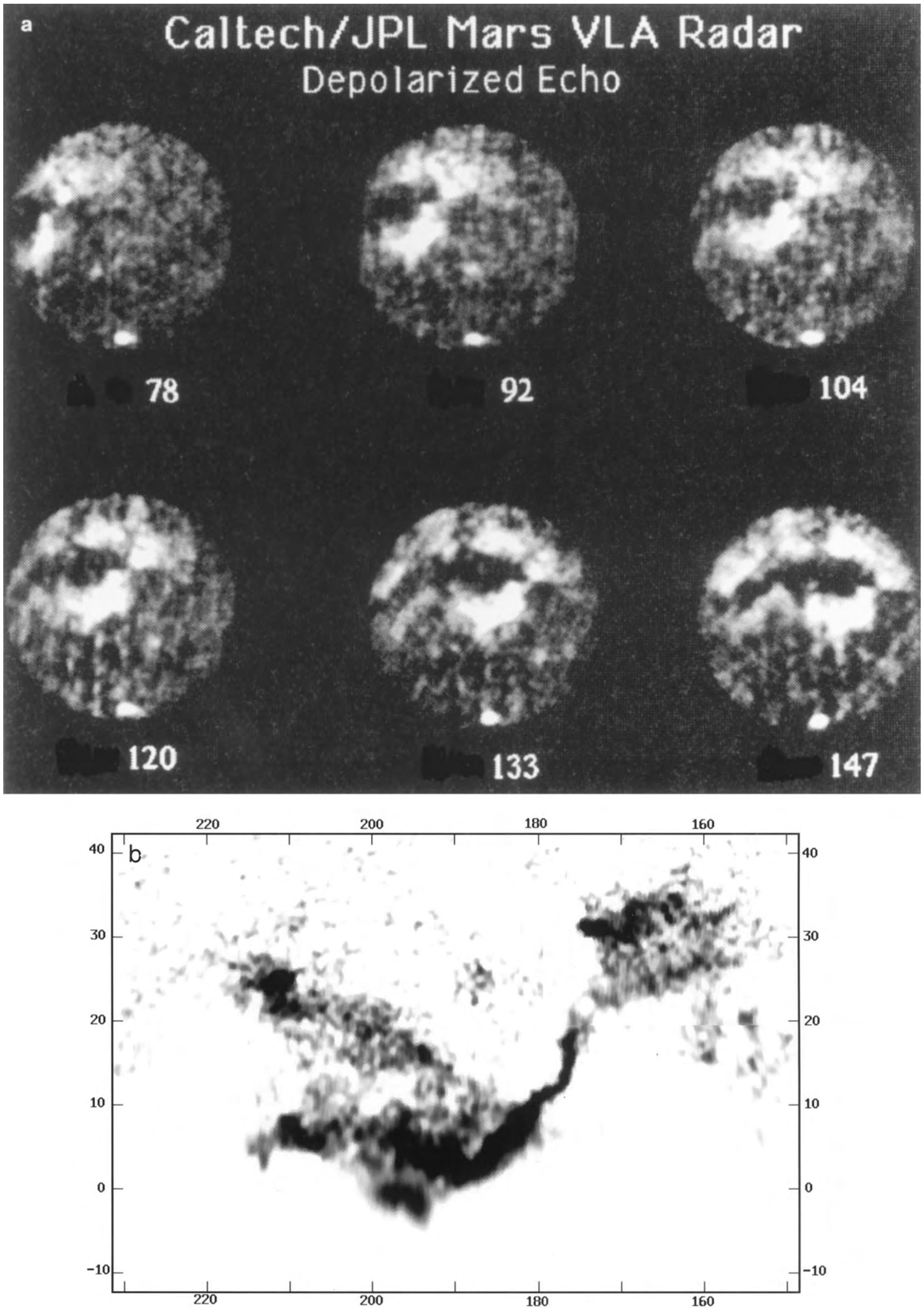


FIGURE 18 Mars radar maps. (a) Goldstone–VLA 3.5-cm, SC radar images of Mars at six longitudes. In the northern hemisphere, the brightest features are in Tharsis, which is traversed by the low-albedo Stealth region. Note the very bright residual south polar ice cap. [From Muhleman, D. O., Butler, B. J., Grossman, A. W., and Slade, M. A. (1991). *Science* **253,** 1508–1513. Copyright 1991 American Association for the Advancement of Science.] (b) Arecibo 13-cm, SC reflectivity map of the Elysium region of Mars, obtained from random-code observations made with the subradar latitude 10°S. The map is north–south ambiguous, but more northerly observations confirm that all of the strong features come from the north. The radar-bright regions (shown here as dark) correspond to Elysium Mons (at ∼214°W longitude, 25°N latitude) and the Elysium flood basin and outflow channel. The SC brightness of these regions is probably caused by extremely rough lava flows. [From Harmon, J. K., Sulzer, M. P., Perillat, P. J., and Chandler, J. F. (1992). *Icarus* **95,** 153–156.]

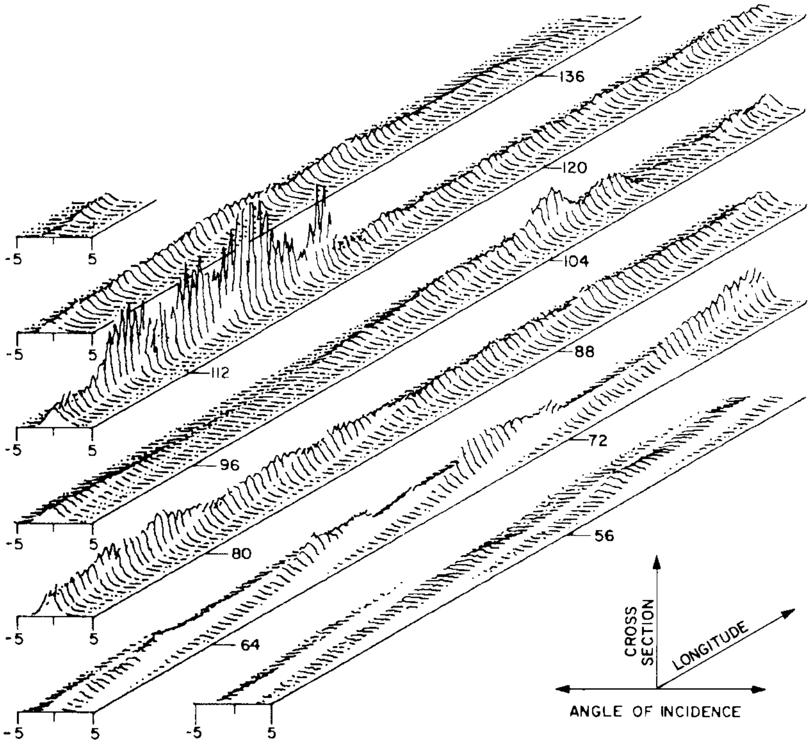


FIGURE 19 Mars echo power spectra as a function of longitude obtained along a subradar track at 16° south latitude. The most sharply peaked spectra correspond to the smoothest regions (i.e., the smallest rms slopes). (Courtesy of G. S. Downs, P. E. Reichley, and R. R. Green.)

many-to-one) mapping from the surface to the image. However, if the radar is not in the target's equatorial plane, then the delay-Doppler trajectory of any surface point is unique. Hence images that provide adequate orientational coverage can be inverted, and in principle one can reconstruct the target's three-dimensional shape as well as its spin state, the radar-scattering properties of the surface, and the motion of the center of mass through the delay-Doppler ephemerides.

The first asteroid radar data set suitable for reconstruction of the target's shape was a 2.5-hr sequence of 64 delay-Doppler images of 4769 Castalia (1989PB) (Fig. 21a), obtained two weeks after its August 1989 discovery. The images, which were taken at a subradar latitude of about 35°, show a bimodal distribution of echo power over the full range of sampled rotation phases, and least-squares estimation of Castalia's three-dimensional shape (Fig. 21b) reveals it to consist of two kilometer-sized lobes in contact. Castalia apparently is a contact-binary asteroid formed from a gentle collision of the two lobes.

If the radar view is equatorial, unique reconstruction of the asteroid's three-dimensional shape is ruled out, but a sequence of images that thoroughly samples rotation phase can allow unambiguous reconstruction of the asteroid's pole-on silhouette. For example, observations of 1620 Geographos yield ~400 images with ~100-m resolution. The pole-on silhouette's extreme dimensions are in a ratio, 2.76 ± 0.21, that establishes Geographos as the most elongated solar system object imaged so far (see Fig. 8). The images show craters as well as indications of other sorts of large-scale topographic relief, including a prominent central indentation. Protuberances at the asteroid's ends may be related to the pattern of ejecta removal and deposition caused by the asteroid's gravity field.

Delay-Doppler imaging of 4179 Toutatis in 1992 and 1996 achieved resolutions as fine as 125 nsec (19 m in range) and 8.3 mHz (0.15 mm sec^{-1} in radial velocity), placing hundreds to thousands of pixels on the asteroid. This data set provides physical and dynamical information that is unprecedented for an Earth-crossing object. The images (Fig. 22) reveal this asteroid to be in a highly unusual, non-principal-axis (NPA) spin state with several-day characteristic time scales. Extraction of the information in this imaging data set required inversion with a much more comprehensive physical model than in the analysis of Castalia images; free parameters included the asteroid's shape and inertia matrix, initial conditions for the asteroid's spin and orientation, the radar-scattering properties of the surface, and the delay-Doppler trajectory of the center of mass. The shape (Fig. 23) reconstructed from

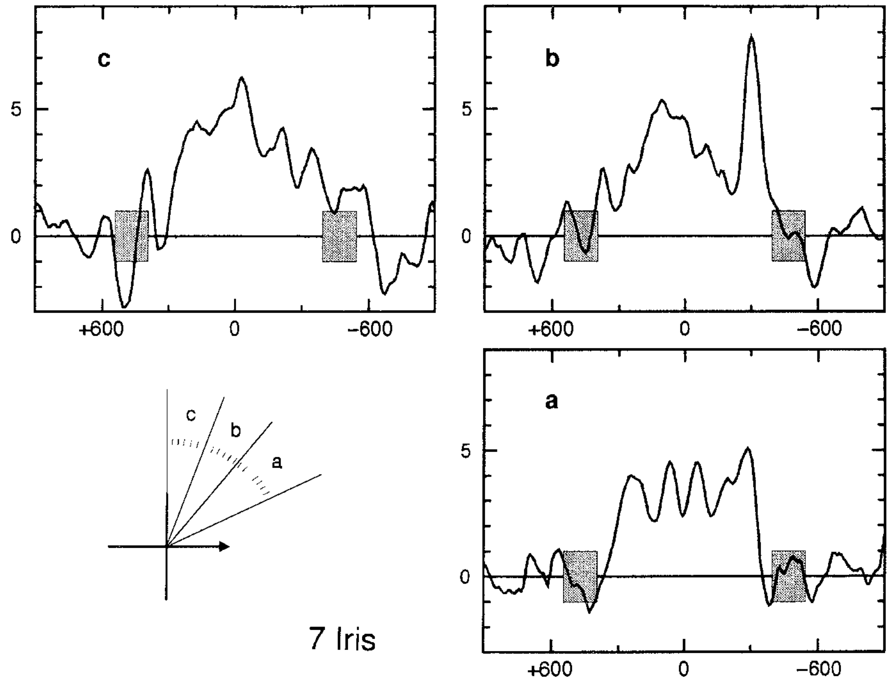


FIGURE 20 Thirteen-centimeter echo spectra of the ~200-km-diameter asteroid 7 Iris, obtained within three narrow rotation-phase intervals. OC echo power in standard deviations is plotted versus Doppler frequency. The shaded boxes show frequency intervals thought to contain the echo edges. A radar "spike" appears in (b) at −305 Hz, but not in spectra at adjacent phases, so it is probably not due to a reflectivity feature but rather to a temporary surge in radar-facing surface area, perhaps a flat facet ~20 km wide. [From Mitchell, D. L., *et al.* (1995). *Icarus* **118,** 105–131.]

the low-resolution images of Toutatis has shallow craters, linear ridges, and a deep topographic "neck" whose geologic origin is not known. It may have been sculpted by impacts into a single, coherent body, or Toutatis might actually consist of two separate objects that came together in a gentle collision. Toutatis is rotating in a long-axis mode (see Fig. 23) characterized by periods of 5.4 days (rotation about the long axis) and 7.4 days (average for long-axis precession about the angular momentum vector). The asteroid's principal moments of inertia are in ratios within 1% of 3.22 and 3.09, and the inertia matrix is indistinguishable from that of a homogeneous body. Such information has yet to be determined for any other asteroid or comet, and probably is impossible to acquire in a fast spacecraft flyby. Higher resolution images (e.g., Fig. 22b) from the 1992 and 1996 experiments are now being used to refine the Toutatis model.

Accurate shape models of near-Earth asteroids open the door to a wide variety of theoretical investigations that previously have been impossible or have used simplistic models (spheres or ellipsoids). For example, the Castalia and Toutatis models are being used to explore the stability and evolution of close orbits, with direct application to the design of robotic and piloted spacecraft missions, to studies of retention and redistribution of impact ejecta and to questions about plausible origins and lifetimes of asteroidal satellites. Accurate models also allow realistic investigations of the effects of collisions in various energy regimes on the object's rotation state, surface topography, regolith, and internal structure. Simulations of impacts into Castalia using smooth-particle hydrodynamics code have begun to suggest how surface and interior damage depends on impact energy, impact location, and the equation of state of the asteroidal material. These computer investigations have clear ramifications for our understanding of asteroid collisional history, for exploitation of asteroid resources, and eventually for deflection/destruction of objects found to be on a collision course with Earth.

K. Jupiter's Icy Galilean Satellites

Among all the radar-detected planetary bodies in the solar system, Europa, Ganymede, and Callisto have the most bizarre radar properties. Their reflectivities are enormous compared with those of the Moon and inner planets (see Table II). Europa is the extreme example (Fig. 24), with an OC radar albedo (1.0) as high as that of a metal sphere. Since the radar and optical albedos and estimates of fractional water frost coverage increase by satellite in the order Callisto/Ganymede/Europa, the presence of water ice has long been suspected of playing a critical role in determining the unusually high reflectivities even though ice

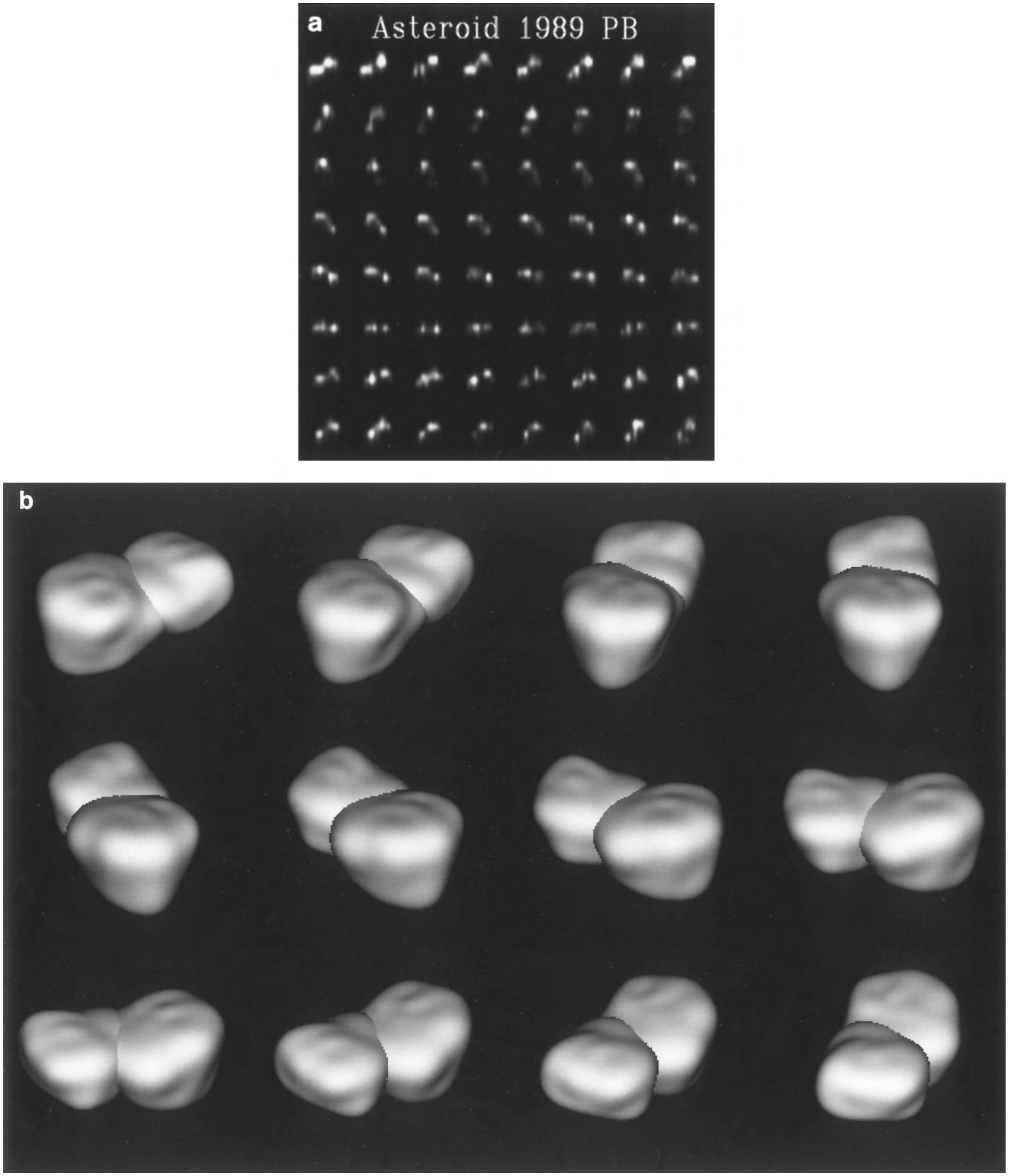


FIGURE 21 Radar results for near-Earth asteroid 4769 Castalia (1989PB). (a) Arecibo radar images. This 64-frame "movie" is to be read like a book (left to right in the top row, etc.). The radar lies toward the top of the page, in the image plane, which probably is about 35° from the asteroid's equatorial plane. In each frame, OC echo power (i.e., the brightness seen by the radar) is plotted versus time delay (increasing from top to bottom) and frequency (increasing from left to right). The object is seen rotating through about 220° during the 2.5-hr sequence. [From Ostro, S. J., Chandler, J. F., Hine, A. A., Shapiro, I. I., Rosema, K. D., and Yeomans, D. K. (1990). *Science* **248,** 1523–1528. Copyright 1990 AAAS.] (b) Three-dimensional computer model of Castalia from inversion of the images in (a). The reconstruction uses 167 shape parameters and has a resolution of about 100 m. This contact-binary asteroid is about 1.8 km long. [From Hudson, R. S., and Ostro, S. J. (1994). *Science* **263,** 940–943. Copyright 1994 AAAS.]

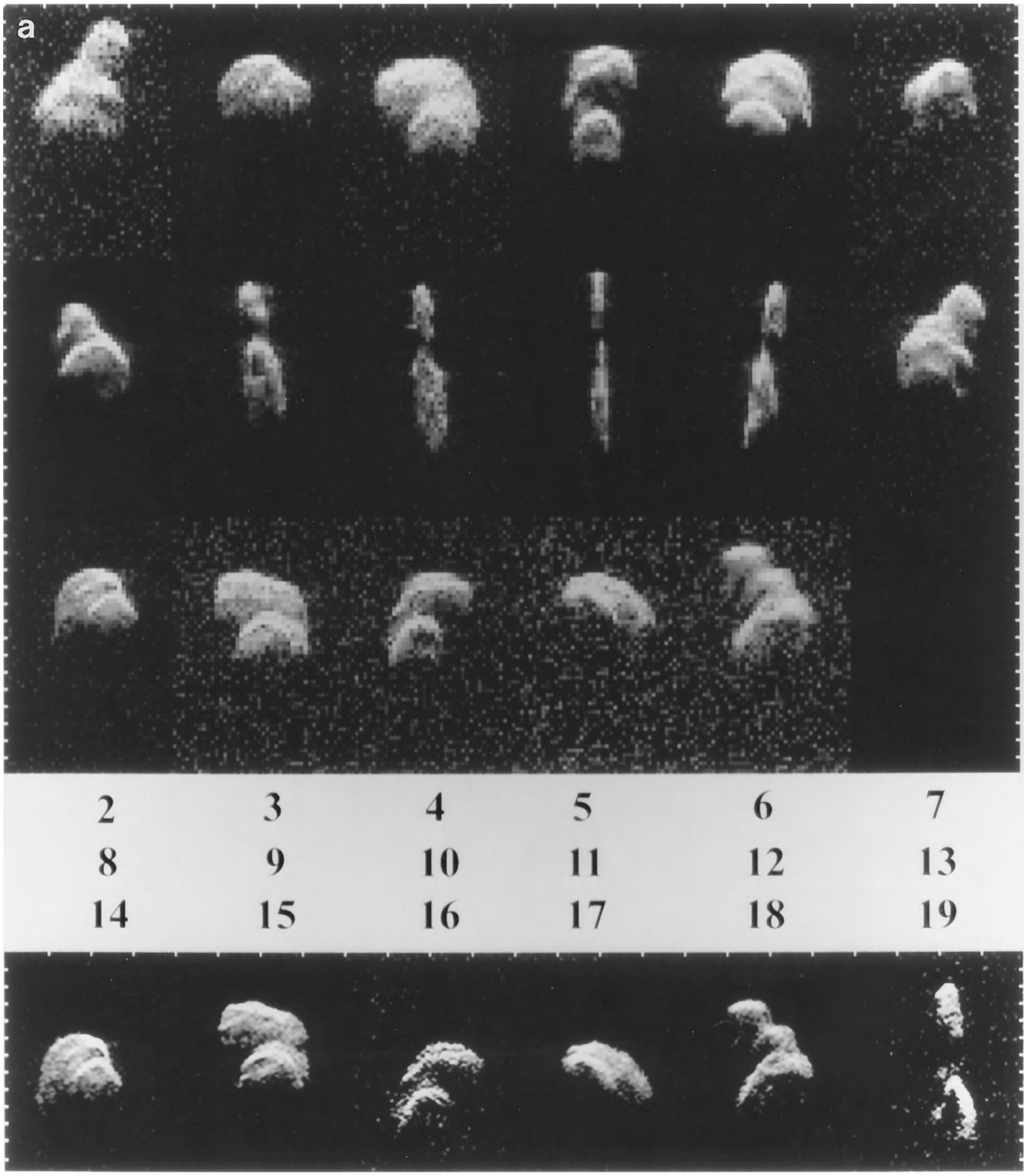


FIGURE 22 Radar images of near-Earth asteroid 4179 Toutatis. (a) Goldstone low-resolution images (top three rows) and Arecibo images (bottom row) obtained on the indicated dates in December 1992, plotted with time delay increasing toward the bottom and Doppler frequency increasing toward the left. On the vertical sides, ticks are 2 msec (300 m) apart. Two horizontal sides have ticks separated by 1 Hz for Goldstone and 0.28 Hz for Arecibo; those intervals correspond to a radial velocity difference of 18 mm sec^{-1}. (b) A high-resolution (125 nsec × 33 mHz) Goldstone image obtained with Toutatis 3.6 million km (10 lunar distances) from Earth. The spatial resolution is 19 × 46 m. [From Ostro, S. J., *et al.* (1995). *Science* **270,** 80–83. Copyright 1995 AAAS.] (*continues*)

is less radar reflective than silicates. Despite the satellites' smooth appearances at the several-kilometer scales of *Voyager's* high-resolution images, a diffuse scattering process and hence a high degree of near-surface structure at centimeter to meter scales is indicated by broad spectral shapes and large linear polarization ratios ($\mu_L \sim 0.5$).

The most peculiar aspect of the satellites' echoes is their circular polarization ratios, which exceed unity. That is,

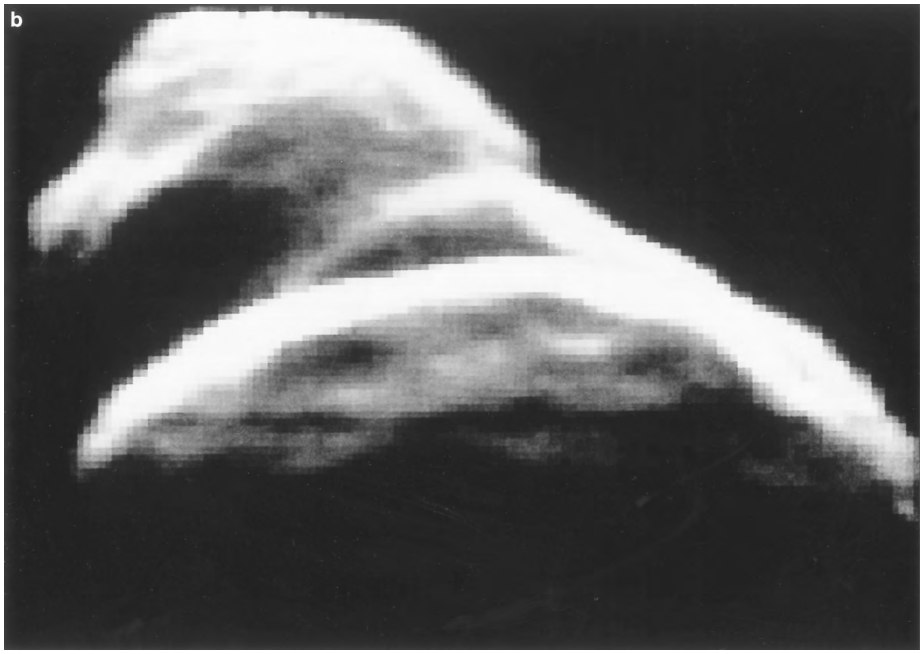


FIGURE 22 (*Continued*)

in contrast to the situation with other planetary targets, the scattering largely preserves the handedness, or helicity, of the transmission. Mean values of μ_C for Europa, Ganymede, and Callisto are about 1.5, 1.4, and 1.2, respectively. Wavelength dependence is negligible from 3.5 to 13 cm, but dramatic from 13 to 70 cm (Fig. 25). Significant polarization and/or albedo features are present in the echo spectra and in a few cases correspond to geologic features in Voyager images.

The icy satellites' echoes are due not to external surface reflections but to subsurface "volume" scattering. The high radar transparency of ice compared with that of silicates permits deeper radar sounding, longer photon path lengths, and higher order scattering from regolith heterogeneities—radar is seeing. Europa, Ganymede, and Callisto in a manner in which the Moon has never been seen. The satellites' radar behavior apparently involves the coherent backscatter effect, which accompanies any multiple-scattering process; occurs for particles of any size, shape, and refractive index; and was first discovered in laboratory studies of the scattering of electrons and of light. Coherent backscatter yields strong echoes and $\mu_C > 1$ because the incident, circularly polarized wave's direction is randomized before its helicity is randomized and also before its power is absorbed. The vector-wave theory of coherent backscatter accounts for the unusual radar signatures in terms of high-order, multiple anisotropic scattering from within the upper few decameters of the regoliths, which the radar sees as an extremely low-loss, disordered random medium. Inter- and intrasatellite albedo variations show much more dynamic range than μ_C variations and are probably due to variations in ice purity.

As sketched in Fig. 25, there are similarities between the icy Galilean satellites' radar properties and those of the radar-bright polar caps on Mars, features inside perpetually shadowed craters at the poles of Mercury (see Fig. 15), and the percolation zone in the Greenland ice sheet. However, the subsurface configuration in the Greenland zone, where the scattering heterogeneities are "ice pipes" produced by seasonal melting and refreezing, are unlikely to resemble those on the satellites. Therefore, unique models of subsurface structure cannot be deduced from the radar signatures of any of these terrains.

L. Comets

Since a cometary coma is nearly transparent at radio wavelengths, radar is much more capable of unambiguous detection of a cometary nucleus than are optical and infrared methods, and radar observations of several comets (see Table I) have provided useful constraints on nuclear dimensions. The radar signature of one particular comet (IRAS-Araki-Alcock, which came within 0.03 AU of Earth in May 1983) revolutionized our concepts of the physical nature of these intriguing objects. Echoes obtained at both Arecibo (Fig. 26) and Goldstone have a narrowband component from the nucleus as well as a much weaker broadband component from large particles ejected mostly from the sunlit side of the nucleus. Models of the

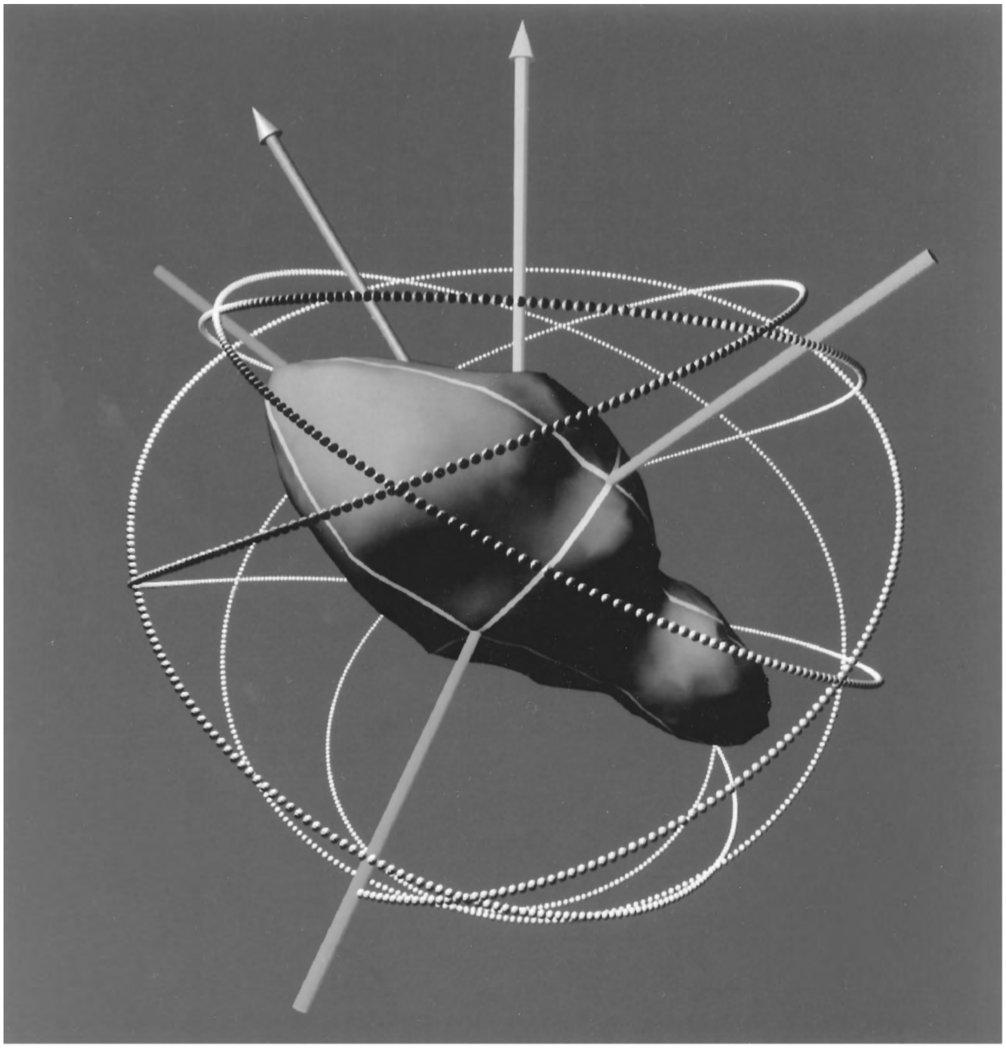


FIGURE 23 Toutatis's shape and non-principal-axis spin state from inversion of the images in Fig. 21a. The axes with no arrow tips are the asteroid's principal axes of inertia and the vertical arrow is its angular momentum vector; the direction of the spin vector (the arrow pointing toward 11 o'clock) relative to the principal axes is a (5.41-day) periodic function. A flashlamp attached to the short axis of inertia and flashed every 15 min for 20 days would trace out the intricate path indicated by the small spheres stacked end-to-end; the path never repeats. Toutatis's spin state differs radically from those of the vast majority of solar system bodies that have been studied, which are in principal-axis spin states. For those objects, the spin vector and angular momentum vector point in the same direction and the flashlamp's path would be a circle.

echoes suggest that the nucleus is very rough on scales larger than a meter, that its maximum overall dimension is within a factor of 2 of 10 km, and that its spin period is 2–3 days. The particles are probably several centimeters in size and account for a significant fraction of the particulate mass loss from the nucleus. Most of them appear to be distributed within ~1000 km of the nucleus, that is, in the volume filled by particles ejected at several meters per second over a few days. The typical particle lifetime may have been this short, or the particle ejection rate may have been highly variable.

In late 1985, radar observations of comet Halley, which was much more active than IRAS-Araki-Alcock, yielded echoes with a substantial broadband component presumed to be from a large-particle swarm, but no narrowband component, a negative result consistent with the hypothesis that the surface of the nucleus has an extremely low bulk density. In 1996, Goldstone obtained 3.5-cm echoes from the nucleus and coma of comet Hyakutake (C/1996 B2). The coma-to-nucleus ratio of radar cross section is about 12 for Hyakutake versus about 0.3 for IAA. The radar signatures of these three comets strengthen impressions

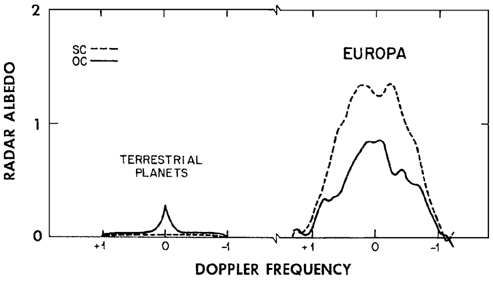


FIGURE 24 Typical 13-cm echo spectra for the terrestrial planets are compared to echo spectra for Jupiter's icy moon Europa. The abscissa has units of half the echo bandwidth.

about the diversity, and unpredictability, of comet physical properties and have obvious implications for spacecraft operations close to comets.

M. Saturn's Rings and Titan

The only radar-detected ring system is quite unlike other planetary targets in terms of both the experimental techniques employed and the physical considerations involved. For example, the relation between ringplane location and delay-Doppler coordinates for a system of particles traveling in Keplerian orbits is different from the geometry portrayed in Fig. 5. The rings are grossly overspread (see Table II), requiring the use of frequency-stepped waveforms in delay-Doppler mapping experiments.

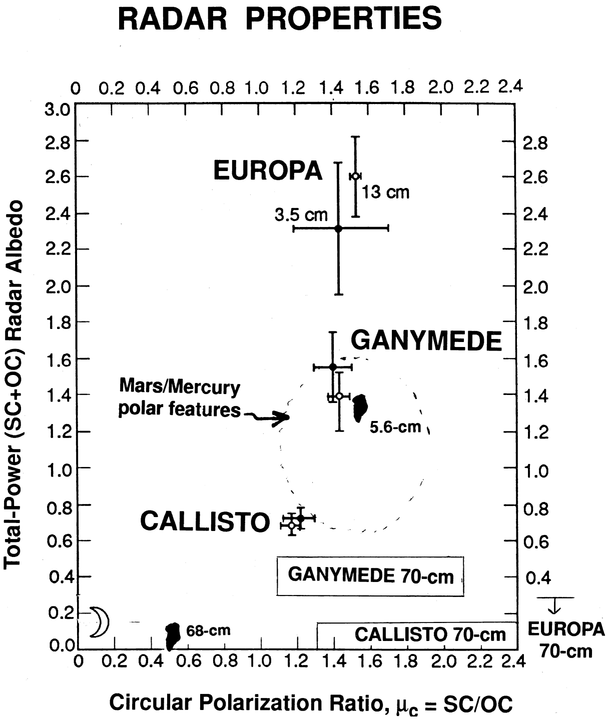


FIGURE 25 Radar properties of Europa, Ganymede, and Callisto compared to those of some other targets. The icy Galilean satellites' total-power radar albedos do not depend on wavelength between 3.5 and 13 cm, but plummet at 70 cm. There are large uncertainties in those objects' μ_C at 70 cm. The solid symbols shaped like Greenland indicate properties of that island's percolation zone at 5.6 and 68 cm. The domain of most of the bright polar features on Mars and Mercury is sketched.

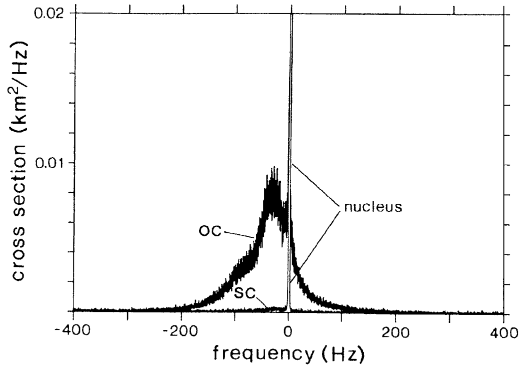


FIGURE 26 OC and SC echo spectra obtained at 13 cm for comet IRAS-Araki-Alcock, truncated at 2% of the maximum OC amplitude. The narrowband echo from the nucleus is flanked by a broadband echo from large (1-cm) particles in a 1000-km-radius cloud surrounding the nucleus. [From Harmon, J. K., Campbell, D. B., Hine, A. A., Shapiro, I. I., and Marsden, B. G. (1989). *Astrophys. J.* **338,** 1071–1093.]

Radar determinations of the rings' backscattering properties complement results of the *Voyager* spacecraft radio occultation experiment (which measured the rings' forward scattering efficiency at identical wavelengths) in constraining the size and spatial distributions of ring particles. The rings' circular polarization ratio is ~1.0 at 3.5 cm and ~0.5 at 13 cm, more or less independent of the inclination angle δ between the ring plane and the line of sight. Whereas multiple scattering between particles might cause some of the depolarization, the lack of strong dependence of μ_C on δ suggests that the particles are intrinsically rougher at the scale of the smaller wavelength. The rings' total-power radar albedo shows only modest dependence on δ, a result that seems to favor many-particle-thick models of the rings over monolayer models. Delay-Doppler resolution of ring echoes indicates that the portions of the ring system that are brightest optically (the A and B rings) also return most of the radar echoes. The C ring has a very low radar reflectivity, presumably because of either a low particle density in that region or compositions or particle sizes that lead to inefficient scattering.

Apart from landing a spacecraft on Titan, radar provides the most direct means to study the cloud-covered surface of Saturn's largest moon. *Voyager* and ground-based data indicate a surface temperature and pressure of 94 K and 1.5 bar and show that the atmosphere is mostly N_2 with traces of hydrocarbons and nitriles. Thermodynamic considerations imply a nearsurface reservoir of liquid hydrocarbons, possibly consisting of a kilometer-deep global ocean. However, that configuration which would give a radar albedo of only several percentages, has been ruled out by Goldstone–VLA and Arecibo detections that yield OC albedo estimates that are significatly higher, at least for part of the object. More work is needed to elucidate Titan's radar properties.

IV. PROSPECTS FOR PLANETARY RADAR

The 1993–1999 upgrading of the Arecibo telescope increased the instrument's average radar sensitivity by a factor of 20, more than doubling its range and reducing by nearly an order of magnitude the diameter of the smallest object detectable at any given distance. The impact of the Arecibo upgrade on planetary science is expected to be fundamental and far-reaching, especially for studies of small bodies and planetary satellites. The quality, in terms of signal-to-noise ratio and spatial resolution, of radar measurements has jumped by an order of magnitude. Several short-period comets will become easy targets. The preupgrade Arecibo could barely skim the inner edge of the main asteroid belt, but the upgraded telescope has access to asteroids throughout the belt. The instrument is expected to provide high-resolution images of dozens of asteroids per year.

Efforts are underway to increase the near-Earth asteroid discovery rate by one to two orders of magnitude. Most of the optically discoverable NEAs traverse the detectability windows of the upgraded Arecibo and/or Goldstone telescopes at least once during any given several-decade interval. In view of the utility of radar observations for orbit refinement and physical characterization, there is considerable motivation to do radar observations of newly discovered NEAs whenever possible. The initial radar reconnaissance of a new NEA might eventually become an almost-daily opportunity. Interferometric radar techniques using the 10-antenna Very Long Baseline Array (VLBA) to receive echoes from Arecibo or Goldstone transmissions should synthesize unambiguous, high-resolution images of NEAs. Also, observations using transcontinental and intercontinental baselines should permit direct measurement of the shapes of these objects.

Radar investigations of natural satellites will rap enormous benefits from ground-based and space-borne radar reconnaissance. The near-surface physical properties of Deimos, Io, and Titan will be readily discernible with Arecibo. Doppler images of Titan may furnish a coarse-resolution, nearly global albedo map, while the *Cassini* spacecraft, with its high-resolution, 13.8-GHz (2.2-cm) radar instrument, is journeying toward its arrival at Saturn in 2004. That instrument, which will function as a synthetic-aperture radar imager, an altimeter, and a passive radiometer, is designed to determine whether oceans exist on Titan and, if so, to determine their distribution. There is growing interest in the possibility of a subsurface ocean on Europa and in the feasibility of using an orbiting, long-wavelength (∼6-m) radar sounder to probe many kilometers below that object's fractured crust, perhaps by 2010. In summary, planetary radar astronomy appears to be on the verge of producing an enormously valuable body of new information about asteroids, comets, and the satellites of Mars, Jupiter, and Saturn.

SEE ALSO THE FOLLOWING ARTICLES

Gamma-Ray Astronomy • Gravitational Wave Astronomy • Millimeter Astronomy • Primitive Solar System Objects: Asteroids and Comets • Radar • Radio Astronomy, Planetary • Solar System, General • Ultraviolet Space Astronomy • X-Ray Astronomy

BIBLIOGRAPHY

Benner, L. A. M., Ostro, S. J., Giorgini, J. D., Jurgens, R. F., Mitchell, D. L., Rose, R., Rosema, K. D., Slade, M. A., Winkler, R., Yeomans, D. K., Campbell, D. B., Chandler, J. F., Shapiro, I. I. (1997). "Radar detection of near-Earth asteroids 2062 Aten, 2101 Adonis, 3103 Eger, 4544 Xanthus, and 1992 QN," *Icarus* **130,** 296–312.

Butrica, A. J. (1996). "To See the Unseen: A History of Planetary Radar Astronomy," NASA History Series No. SP-4218. NASA, Houston.

Harmon, J. K., Slade, M. A., Velez, R. A., Crespo, A., Dryer, M. J., and Johnson, J. M. (1994). "Radar mapping of Mercury's polar anomalies," *Nature* **369,** 213–215.

Hudson, R. S., and Ostro, S. J. (1995). "Shape and non-principal axis spin state of asteroid 4179 Toutatis," *Science* **270,** 84–86.

Magri, C., Ostro, S. J., Rosema, K. D., Thomas, M. L., Mitchell, D. L., Campbell, D. B., Chandler, J. F., Shapiro, I. I., Giorgini, J. D., and Yeomans, D. K. (1999). "Mainbelt Asteroids: Results of Arecibo and Goldstone radar observations of 37 objects during 1980–1995," *Icarus* **140,** 379–407.

Margot, J. M., Campbell, D. B., Jurgens, R. F., and Slade, M. A. (1999). "Topography of the lunar poles from radar interferometry: A survey of cold trap locations," *Science* **284,** 1658–1660.

Mitchell, D. L., Ostro, S. J., Hudson, R. S., Rosema, K. D., Campbell, D. B., Velez, R., Chandler, J. F., Shapiro, I. I., Giorgini, J. D., and Yeomans, D. K. (1996). "Radar observations of asteroids 1 Ceres, 2 Pallas, and 4 Vesta," *Icarus* **124,** 113–133.

Muhleman, D. O., Grossman, A. W., and Butler, B. J. (1995). "Radar investigation of Mars, Mercury, and Titan," *Annu. Rev. Earth Planet Sci.* **23,** 337–374.

Ostro, S. J. (1993). "Planetary radar astronomy," *Rev. Modern Physics* **65,** 1235–1279.

Pettengill, G. H., Ford, P. G., Johnson, W. T. K., Raney, R. K., and Soderblom, L. A. (1991). "*Magellan*: Radar performance and data products," *Science* **252,** 260–265.

Shapiro, I. I., Chandler, J. F., Campbell, D. B., Hine, A. A., and Stacy, N. J. S. (1990). "The spin vector of Venus," *Astron. J.* **100,** 1363–1368.

Simpson, R. A., Harmon, J. K., Zisk, S. H., Thompson, T. W., and Muhleman, D. O. (1992). *In* "Mars" (H. Kieffer, B. Jakosky, C. Snyder, and M. Matthews, eds.), pp. 652–685, Tucson, Univ. of Arizona Press.

Slade, M. A., Butler, B. J., and Muhleman, D. O. (1992). "Mercury radar imaging: Evidence for polar ice," *Science* **258,** 635–640.

Stacy, N. J. S., Campbell, D. B., and Ford, P. G. (1997). "Arecibo radar mapping of the lunar poles: A search for ice deposits," *Science* **276,** 1527–1530.

Tyler, G. L., Ford, P. G., Campbell, D. B., Elachi, C., Pettengill, G. H., and Simpson, R. A. (1991). "*Magellan*: Electrical and physical properties of Venus' surface," *Science* **252,** 265–270.

Yeomans, D. K., Chodas, P. W., Keesey, M. S., Ostro, S. J., Chandler, J. F., and Shapiro, I. I. (1992). "Asteroid and comet orbits using radar data," *Astron. J.* **103,** 303–317.

Radio Astronomy, Planetary

Samuel Gulkis
Jet Propulsion Laboratory

Imke de Pater
University of California, Berkeley

GLOSSARY

Antenna temperature Measure of the noise power collected by the antenna and delivered to the radio receiver. Specifically, the temperature at which a resistor, substituted for the antenna, would have to be maintained in order to deliver the same noise power to the receiver in the same frequency bandwidth.

Blackbody Idealized object that absorbs all electromagnetic radiation that is incident on it. The radiation properties of blackbody radiators are described by the Planck function. Planetary radio astronomers use the properties of blackbody radiators to describe the radiation from planets.

Brightness temperature The definition is not unique; great care is needed to decipher the intention of a given author. The temperature at which a blackbody radiator would radiate an intensity of electromagnetic radiation identical to that of the planet for a specific frequency, frequency bandwidth, and polarization under consideration is one definition of brightness temperature. A second definition is that it is the intensity of radiation under consideration divided (normalized) by the factor $(\lambda^2/2k)$. The normalization factor dimensionally scales the intensity to have units of temperature. The two definitions show the largest departures at low temperatures and high frequencies.

Effective area Equivalent cross section or collecting area of a radio telescope to an incident radio wave; a measure of a radio telescope's capability to detect weak radio signals.

Effective temperature Temperature at which a blackbody radiator would radiate over all frequencies an intensity of electromagnetic radiation identical to that radiated from a planet.

Equivalent blackbody disk temperature Temperature of a blackbody radiator with the same solid angle as the planet that gives the same radiation intensity at the Earth as observed from the planet at a specified frequency and bandwidth.

Flux density Power per unit area and per unit frequency of an electromagnetic wave crossing an imaginary plane surface from one side to the other. In observational radio astronomy, the mks system of units is generally used, and the units of flux density are watts per square meter per hertz.

Flux unit or jansky Commonly used unit of flux density equal to 1×10^{-26} W m^{-2} Hz^{-1}. The size of the unit is suited to planetary radio emissions, which are very weak.

Nonthermal radio emission Radio emission produced by processes other than thermal emission. Cyclotron and synchrotron radio emission are two examples of nonthermal radio emission.

Optical depth Atmospheric attenuation is usually expressed by giving the dimensionless quantity "optical depth" along a specified path. A signal that passes through an atmosphere whose optical depth is τ is attenuated by the factor $e^{-\tau}$.

Thermal radio emission Continuous radio emission from an object that results from the object's temperature. Blackbody radiation is a form of thermal radio emission.

PLANETARY RADIO ASTRONOMY is the study of the physical characteristics of the planets in the solar system by means of the electromagnetic radio radiation emitted by these objects. The term is also used more generally to include the study of planetary ring systems, the moon, asteroids, satellites, and comets in the solar system. Radio astronomy generally refers to the (vacuum) wavelength range from about 0.5 mm (600 GHz $= 600 \times 10^9$ Hz) to 1 km (300 kHz) and longward. There is no wide acceptance of the upper limit frequency. Observations from the ground cannot be carried out below a few megahertz because of the Earth's ionosphere, which is opaque to very low frequency radio waves. At millimeter and submillimeter wavelengths, the terrestrial atmosphere defines windows where radio emissions from cosmic sources can reach the ground. Radio emissions have been measured from all of the planets and some satellites, asteroids, and comets. The observed continuum emissions from the planets can be broadly classified as quasi-thermal (having the same general shape as a blackbody emitter) and nonthermal (i.e., cyclotron, synchrotron). Narrow spectral lines from molecules have been observed in the atmospheres of planets, satellites, and comets. Planetary radio emissions originate in the solid mantles, atmospheres, and magnetospheres of the planets. A number of solar system spacecraft have carried radio astronomical instrumentation.

I. INTRODUCTION

A. Brief History

The science of radio astronomy began with the pioneering work of Karl G. Jansky, who discovered radio emission from the Milky Way galaxy while studying the direction of arrival of radio bursts associated with thunderstorms. Ten years later, in 1942, while trying to track down the source of radio interference on a military antiaircraft radar system in England, J. S. Hey established the occurrence of radio emission from the sun.

R. H. Dicke and R. Beringer working in the United States made the first intentional measurement of a solar system object in 1945. They observed radio emission from the Moon at a wavelength of 1.25 cm, thus beginning the first scientific studies at radio wavelengths of the planets and satellites of the solar system. Subsequent observations revealed that the microwave emission from the Moon varies with lunar phase but the amplitude of these variations is much smaller than that observed at infrared wavelengths. This result was interpreted in terms of emission originating below the surface of the Moon, where the temperature variations are smaller than at the surface. Within a few years, it was widely recognized that the long wavelengths provided by radio measurements offered a new and important tool for solar system studies, namely the capability of probing into and beneath cloud layers and surfaces of the planets. However, thermal radio emissions from the planets are exceedingly weak and nearly a decade elapsed before system sensitivity was sufficiently improved to enable the detection of planetary thermal emission.

Meanwhile, another unanticipated discovery was made. In June 1954, when the angular separation between the Sun and a supernova remnant known as the Crab Nebula was small, astronomers B. Burke and K. Franklin of the Carnegie Institution were attempting to study the effect of the solar corona on radio waves from the Crab Nebula. Occasionally, they observed bursts of radio interference, which they initially thought were due to the Sun. That hypothesis was discarded when it was discovered that the origin of the interference bursts was nearly fixed with respect to the background stars. Further observations and examinations of the data revealed that the emissions were in fact originating from Jupiter, which happened to be located in the same region of the sky as the Sun and the Crab Nebula. Because the intensity of the emissions was much too strong to be of thermal origin, Franklin and Burke concluded that Jupiter was a source of nonthermal radio emission.

The first successful measurements of thermal emission from the planets were made in 1956 at the Naval Research Laboratory in Washington, D. C. C. H. Mayer, T. P. McCullough, and R. M. Sloanaker scanned Venus, Mars, and Jupiter with a 15-m parabolic antenna equipped with a new 3-cm-wavelength radio receiver. They detected weak thermal emission from these three planets when each was observed at its closest distance to the Earth.

In the intervening years, thermal emission has been measured from all of the planets in the solar system.

Because of the faintness of its radio emission, Pluto was the last planet to be detected. A few asteroids, satellites, and comets have been measured as well. Nonthermal radio emission has been measured from Jupiter, Saturn, Uranus, Neptune, and Earth. In this article we give an overview of the techniques used by planetary radio astronomers and discuss what has been learned from the measurements and what can be done in the future. In the interest of brevity, we do not discuss specific observations of asteroids and satellites, although they rightfully belong in any discussion of planetary radio astronomy.

B. Measurement Objectives

The primary objective of planetary exploration is to determine the physical characteristics of the planets, satellites, asteroids, and comets in order to obtain an understanding of the origin and evolution of the solar system, including the origin of life on the planet Earth. One objective associated with this goal is to determine the composition and physical characteristics of these bodies and their atmospheres. Studies of the energy budget and redistribution of energy within solid surfaces and atmospheres are part of this work. Another objective is to investigate the magnetic fields and ionized plasmas that surround some of the bodies in the solar system and to understand the interaction of the magnetic fields with the solar wind and the cosmic environment.

Planetary research involves many scientific disciplines and requires a variety of instruments and techniques, including astronomical studies from the Earth, planetary spacecraft flybys, orbiters, and probes, and eventually manned landings. Each of the various approaches used has a particular strength that experimenters try to exploit. Thus far, most planetary radio astronomy has been carried out from the ground, but the techniques carry over to spacecraft as well.

Planetary radio astronomy measurements provide complementary data to other observational techniques. They also provide some unique data. For example, radio measurements can be used to provide information about planetary atmospheres and planetary subsurface materials to much greater depth than other remote-sensing techniques. The greater penetration is a result of neutral gases and solids being more transparent to radio waves than higher frequency waves such as infrared or visible light. Also, the scattering from particulate materials in planetary atmospheres is generally less at radio wavelengths than at shorter wavelengths. The relative transparency of atmospheres, clouds, and surfaces to radio waves allows the planetary radio astronomer to measure thermal profiles of planetary atmospheres beneath the cloud layers in the atmosphere and to measure temperatures beneath the solid surface of the planet. Taking advantage of this property, radio astronomers were the first to measure the very high surface temperature of cloud-covered Venus.

Very high resolution spectroscopy is another area in which radio astronomy provides a unique capability. It is possible to achieve very high spectral resolution in the radio and submillimeter spectral region by translating the frequency of the radio signal under study to a convenient place in the frequency spectrum where spectrum analysis is easier to achieve with either digital or analog techniques. This process takes place without disturbing the relation of the sidebands to the carrier frequency. The process of frequency translation is referred to by such names as heterodyne, mixing, or frequency conversion. The heterodyne technique makes it possible to measure absorption and emission line shapes in greater detail than has been possible at shorter wavelengths. Both temperature–altitude and composition–altitude profile distributions of absorbing chemical species (e.g., NH_3, H_2O, CO) can be deduced from spectroscopic measurements.

Finally, we note that both synchrotron and cyclotron radiation from (solar system) planets is confined to the radio region of the spectrum. This circumstance is due to the range of planetary magnetic field strengths and particle energies found in the solar system. The existence of Jupiter's strong magnetic field was first deduced from earth-based measurements of its polarized radio emission.

C. Physical Properties of the Planets

The physical characteristics of the planets are required in order to make qualitative estimates of the radio power flux densities expected from them. Table I presents physical data for the planets.

II. BASIC CONCEPTS

A. Thermal (Blackbody) Radiation

Any object in thermodynamic equilibrium with its suroundings (having a temperature above absolute zero) emits a continuous spectrum of electromagnetic radiation at all wavelengths, including the radio region. This emission is referred to as thermal emission. The concept of a "blackbody" radiator is frequently used as an idealized standard which can be compared with the absorption and emission properties of real materials. A blackbody radiator is defined as an object that absorbs all electromagnetic radiation that falls on it at all frequencies over all angles of incidence. No radiation is reflected from such an object. According to thermodynamic arguments

TABLE I Physical Data for the Planets

	Mean distance (AU)[a]	Mass (Earth = 1)	Radius (equator) (km)	Obl.[b]	Density (g/cm^3)	Bond albedo[c]	Diameter (arc sec)[d]	
							Min.	Max.
Mercury	0.387	0.055	2,440	Small	5.427	0.06	4.7	12.2
Venus	0.723	0.815	6,052	Small	5.204	0.77	9.9	62.2
Earth	1.000	1	6,371	1/298.2	5.515	0.39	—	—
Mars	1.524	0.107	3,390	1/156.6	3.934	0.16	3.5	24.6
Jupiter	5.203	317.9	69,911	1/16.7	1.326	0.45	30.5	49.8
Saturn	9.537	95.2	58,232	1/9.3	0.687	0.61	14.7	20.5
Uranus	19.191	14.6	25,362	1/100	1.318	0.42	3.4	4.2
Neptune	30.069	17.2	24,624	1/38.5	1.638	0.42	2.2	2.4
Pluto	39.482	0.0017	1,151	Unknown	1.1	0.55	0.06	0.1

[a] AU, Astronomical unit = 149.6×10^6 km.

[b] Obl., Oblateness of planet = 1 − (polar radius/equatorial radius).

[c] Bond albedo, ratio of reflected solar radiation to incident solar radiation. Spectral range should be specified, for example, visual bond albedo.

[d] Diameter, angular extent of disk when planet–Earth distance is greatest (min.) and least (max).

embodied in Kirchhoff's law, a good absorber is also a good emitter. The blackbody radiator emits the maximum amount of thermal radiation possible for an object at a given temperature. The radiative properties of a blackbody radiator have been well studied and verified by experiments.

A blackbody radiator is an idealized concept rather than a description of an actual radiator. Only a few surfaces, such as carbon black, carborundum, platinum black, and gold black, approach a blackbody in their ability to absorb incident radiant energy over a broad wavelength range. Many materials are spectrally selective in their ability to absorb and emit radiation, and hence they resemble blackbody radiators over some wavelength ranges and not over others. Over large ranges of the radio and infrared spectrum, planets behave as imperfect blackbodies. Later, we will see how the deviations from the blackbody spectrum contain information about physical and chemical properties of these distant objects.

An important property of a blackbody radiator is that its total radiant energy is a function only of its temperature; that is, the temperature of a blackbody radiator uniquely determines the amount of energy that is radiated into any frequency band. Planetary radio astronomers make use of this property by expressing the amount of radio energy received from a planet in terms of the temperature of a blackbody of equivalent angular size. This concept is developed more fully in the following paragraphs.

The German physicist Max Planck first formulated the theory that describes the wavelength dependence of the radiation emitted from a blackbody radiator in 1901. Planck's theory was revolutionary in its time, requiring assumptions about the quantized nature of radiation. Planck's radiation law states that the brightness of a blackbody radiator at temperature T and frequency ν is expressed by

$$B = (2h\nu^3/c^2)(e^{h\nu/kT} - 1)^{-1}, \tag{1}$$

where B is the brightness in watts per square meter per hertz per radian; h is Planck's constant (6.63×10^{-34} J sec); ν is the frequency in hertz; $\lambda = c/\nu$ is the wavelength in meters; c is the velocity of light (3×10^{-8} m/sec); k is Boltzmann's constant (1.38×10^{-23} J/K); and T is the temperature in kelvins.

Equation (1) describes how much power a blackbody radiates per unit area of surface, per unit frequency, into a unit solid angle. The curves in Fig. 1 show the brightness for three blackbody objects at temperatures of 6000, 600, and 60 K. The radiation curve for the undisturbed Sun is closely represented by the 6000 K curve over a wide frequency range. The other two curves are representative of the range of thermal temperatures encountered on the planets. It should be noted in Fig. 1 that the brightness curve that represents the Sun peaks in the optical wavelength range, while representative curves for the planets peak in the infrared. This means that most of the energy received by the planets from the Sun is in the visible wavelength range, while that emitted by the planets is radiated in the infrared. Radio emissions are expected to play only a small role in the overall energy balance of the planets because the vast majority of the power that enters and leaves the planets is contained within the visible and infrared region of the spectrum.

A useful approximation to the Planck radiation law can be obtained in the low-frequency limit where $h\nu$ is small compared with kt ($h\nu \ll kT$). This condition is generally met over the full range of planetary temperatures and at radio wavelengths. It leads to the

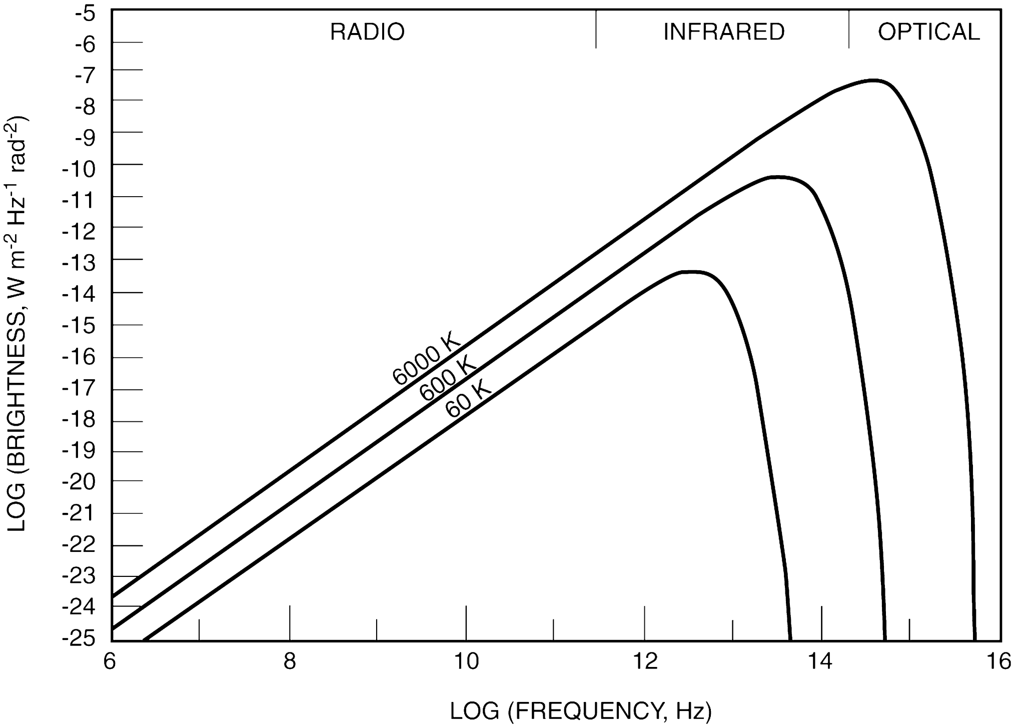


FIGURE 1 Blackbody radiation curves at 6000, 600, and 60 K. The 6000 K curve is representative of the solar spectrum.

Rayleigh–Jeans approximation of the Planck law, given by

$$B = 2\nu^2 kT/c^2 = 2kT/\lambda^2. \qquad (2)$$

The Rayleigh–Jeans approximation shows a linear relationship between physical temperature and the Planck brightness B. The brightness is also seen to decrease as the inverse square of the wavelength, approaching infinity as the wavelength gets shorter and shorter. The Planck brightness, on the other, hand reaches a maximum value at some wavelength and decreases at longer and shorter wavelengths. The Rayleigh–Jeans approximation matches the Planck law at wavelengths considerably longer than the wavelength of peak brightness. However, for shorter wavelengths, the approximation gets progressively worse. At a temperature of 100 K and a wavelength of 1 mm, the error is ~8%.

Planetary radio astronomers estimate the radio power emitted by the planets by measuring with a radio telescope the power flux density received at the Earth. Figure 2 illustrates the geometry involved in the measurement of power from an ideal blackbody radiator. The spectral power (per unit frequency) emitted by an elemental surface element of the blackbody of area dA into a solid angle $d\Omega$ is given by $B\cos(\theta)\,d\Omega\,dA$, where θ is the angle between the normal to the surface and the direction of the solid angle $d\Omega$. The total power (per unit frequency interval) radiated by a blackbody radiator is obtained by integrating the brightness over the surface area and over the solid angle into which each surface element radiates. The total spectral power density produced by a spherical blackbody radiator of radius r at a distance d from the blackbody is given by

$$S = \frac{1}{4\pi d^2}\iint B\cos(\theta)\,ds\,d\Omega \qquad (3a)$$

$$= 2\pi kT(r/d)^2/\lambda^2. \qquad (3b)$$

The double integral represents integration over the surface area of the emitting body and over the hemisphere into which each surface element radiates. The quantity S is called "flux density." Flux density has units of power per unit area per unit frequency. A common unit of flux density is the flux unit (f.u.) or jansky (Jy), which has the value 10^{-26} W m^{-2} Hz^{-1}.

If a planet radiates like a blackbody and subtends a solid angle of Ω steradians at the observer's distance, then the flux density produced by the planet is given by (using the Rayleigh–Jeans approximation)

$$S = 2kT\Omega/\lambda^2 \qquad (4a)$$

$$= B\Omega \qquad (4b)$$

The convention generally adopted for calculating Ω for a planet is to use the polar (PSD) and equatorial (ESD) semidiameter values in the expression

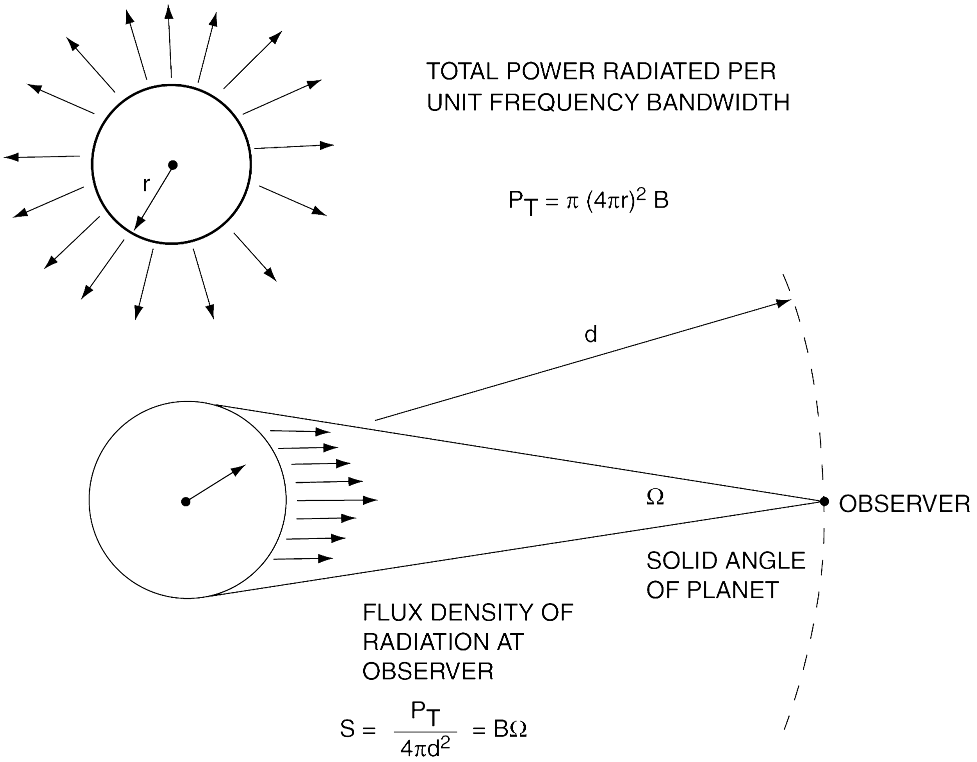

TOTAL POWER RADIATED PER UNIT FREQUENCY BANDWIDTH
$P_T = \pi\,(4\pi r)^2\,B$
r
d
Ω
OBSERVER
SOLID ANGLE OF PLANET
FLUX DENSITY OF RADIATION AT OBSERVER
$S = \frac{P_T}{4\pi d^2} = B\Omega$

FIGURE 2 Relationship between the brightness and power radiated by a blackbody spherical radiator of radius *r* and the flux density at a distance *d* from the blackbody.

$$\Omega = \pi \times \text{PSD} \times \text{ESD}. \tag{5}$$

The American Ephemeris and Nautical Almanac (AENA) provides values for PSD and ESD. A web site (http://ssd.jpl.nasa.gov) operated by the JPL Solar System Dynamics Group also provides these data. Equations (4) and (5) can be combined to yield the expression

$$S = 5.1 \times 10^{-34} T\theta_{\rm E}\theta_{\rm P}/\lambda^2 \quad \text{W m}^{-2}\,\text{Hz}^{-1} \tag{6a}$$

$$= 5.1 \times 10^{-8} T\theta_{\rm E}\theta_{\rm P}/\lambda^2 \quad \text{Jy} \tag{6b}$$

where $\theta_{\rm E}$ and $\theta_{\rm P}$ are the apparent equatorial and polar diameters of the planets in seconds of arc and λ is in meters.

Even though the planets do not radiate like a blackbody, planetary radio astronomers express the observed brightness in terms of the temperature of an equivalent blackbody that would produce the same brightness. This temperature is called the brightness temperature $T_{\rm B}$, defined as follows [from Eq. (4)]:

$$T_{\rm B} = B\lambda^2/2k = (S/\Omega)\lambda^2/2k. \tag{7}$$

The brightness temperature for a planet can be calculated once the flux density S and solid angle Ω are known. The brightness temperature approximates the physical temperature the more the planet behaves like a blackbody radiator.

For problems that deal with the energy budget of the planets, it is necessary to know the total amount of power radiated over all frequencies. Once again the concept of the blackbody is useful. The Planck radiation law can be integrated over all frequencies and solid angles to obtain the relationship known as the Stefan–Boltzmann law, given by

$$R = \sigma T^4, \tag{8}$$

where R is the rate of emission, expressed in units of watts per square meter in the mks system. The constant σ has a numerical value of 5.67×10^{-8} W m^{-2} K^{-4}. Using the blackbody concept, we can now estimate the energy balance of planets that absorb visible light and radiate energy into the infrared.

B. Thermal Emission from Atmospheres and Surfaces

1. Effective Temperatures of the Planets

The amount of thermal radiation expected from a given planet depends in detail on the physical characteristics of the planet's atmosphere and surface. A starting point for understanding the observed flux densities of the planets is to assume that the planets are blackbodies in equilibrium with the energy they receive from the Sun and that which is radiated into free space. The radiation energy incident from the Sun on a unit area per unit time is 1.39×10^3 W m^{-2} sec^{-1} at the Earth. This quantity is called the solar constant. The incident solar flux available to heat a planet is given by

$$(1 - A)S_0 \pi R^2/d^2, \tag{9}$$

where A is the fraction of the incident solar flux that is not absorbed, πR^2 is the cross-sectional area of the planet, S_0 is the solar constant at astronomical unit (AU), and d is the mean distance of the planet from the Sun in astronomical units. The quantity A is known as the Bond albedo or Russel–Bond albedo of the planet. Disregarding any significant internal heat sources or heating from charged particles, the total flux of absorbed radiation must equal the total flux of outgoing radiation when the planet is in equilibrium. The Stefan–Boltzmann law provides the relationship between effective temperature T_{E} and the absorbed flux:

$$\int \sigma T_{\mathrm{E}}^4 \, ds = \pi R^2(1 - A)S_0/d^2. \tag{10}$$

If the planet rotates rapidly, equilibrium will be reached between the insulation and the radiation from the entire planetary surface area, $4\pi R^2$. This leads to the estimate

$$T_{\mathrm{E}} = 277(1 - A)^{1/4} d^{-1/2} \quad \mathrm{K}. \tag{11}$$

If a planet did not rotate and its emitted radiation came only from the sunlit hemisphere, the effective temperature of the sunlit hemisphere would increase by the factor $2^{1/4}$ because of the reduction in the emission surface area. The equilibrium temperature for this case would become

$$T_{\mathrm{E}} = 330(1 - A)^{1/4} d^{-1/2} \quad \mathrm{K}. \tag{12}$$

Figure 3 shows the calculated effective temperatures of the planets for the rapidly rotating and nonrotating cases. The albedos and distances used are those given in Table I. Having obtained the effective temperatures, it is possible to predict flux densities for the planets by using the effective temperatures from Eq. (11) or (12) and the angular diameter data for the planets in Table I in Eq. (6a) or (6b).

Thus far, we have discussed the ideal model in which the planets behave like blackbody radiators. This gives planetary astronomers a crude model from which they can estimate flux densities and search for departures. The planets would not be very interesting to study if they behaved like blackbodies since a single parameter, namely the temperature, could be used to define their radiation properties. More important, it is the departures from the simple model that allows radio astronomers to deduce the physical properties of the planets.

The observed temperatures of the planets depart markedly from this ideal model for a number of different reasons. The presence of atmospheres on the planets produces strong perturbations from the ideal model. Atmospheres modify the amount of heat that enters and leaves the planets over the entire electromagnetic spectrum. Strong "greenhouse" effects can raise the temperatures considerably over that calculated from the ideal models. The presence of internal sources of energy within a planet can modify its effective temperature and affect the thermal profile of the atmosphere. Surface emissivity effects modify the apparent temperature of the planets. Nonuniform heating of the planets by the Sun due to orbit eccentricity and rotational effects also produces departures from the ideal model. Nonthermal emission from planetary magnetospheres produces the largest departures from the blackbody model. The role of the planetary astronomer is to sort out the various effects that take place and to measure and infer the actual physical characteristics of the planets.

2. Radiative Transfer in Planetary Atmospheres

The apparent brightness temperature of a deep atmosphere is related to the physical parameters of the atmosphere, such as pressure, temperature, and composition, through the equation of radiative transfer. To a good approximation at radio wavelengths, the equation of radiative transfer for a ray making an angle $\cos^{-1}\mu$ with the vertical in a lossy medium is

$$\begin{aligned} T_{\mathrm{B}}(\nu, \mu) = \int T(z) \exp\left[-\int \alpha(z', \mu)\mu^{-1} dz' \right] \\ \times \alpha(z, \nu)\mu^{-1} \, dz, \end{aligned} \tag{13}$$

where $\alpha(z, \nu)$ is the absorption per unit length of the atmosphere at frequency ν and depth z, and $T(z)$ is the physical temperature along the line of sight. This equation states that the brightness temperature in any given direction is the sum of the radiation emitted at each point along the trajectory, each component being attenuated by the intervening medium. The equation neglects scattering and variations of the index of refraction. These effects are important in certain wavelength ranges and for certain ray trajectories. Sometimes it is necessary to add another term to Eq. (13) to account for the presence of a solid surface.

Measurements with single-dish antennas generally have insufficient angular resolution to determine the brightness distribution across the planetary disk. The mean disk brightness temperature T_{D} can be calculated by integration of Eq. (13) over all angles of incidence. This yields

$$T_{\mathrm{D}}(\nu) = 2 \int_0^1 T_{\mathrm{B}}(\nu, \mu)\mu \, d\mu. \tag{14}$$

This equation is usually used to calculate the average disk brightness temperature for a model atmosphere for comparison with observations. Unlike the blackbody radiation model, this model predicts a wavelength dependence of

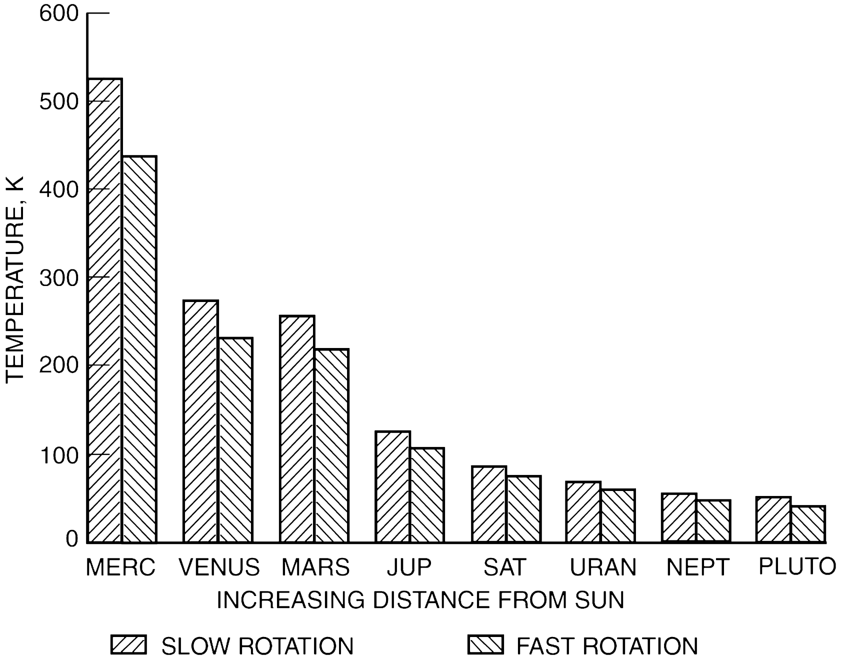


FIGURE 3 Theoretical effective temperatures of the planets for two models. The higher temperature model is for the case where the planet is not rotating [Eq. (12)]. The lower temperature model is for the case where the planet is rapidly rotating [Eq. (11)].

the brightness temperature. The frequency dependence is introduced by the frequency dependence of the absorption and the thermal gradients in the atmosphere.

The absorption may either vary slowly with frequency or exhibit an abrupt change over a narrow frequency range. The two kinds of absorption are referred to as nonresonant and resonant absorption, respectively. The standard classical theory of nonresonant molecular absorption is due to Debye. Resonant absorption is produced by the discrete transitions from one energy level to another in a molecule that cause the molecule to absorb or emit at particular frequencies. The study of resonant absorption lines in planetary atmospheres is referred to as planetary spectroscopy.

A brief discussion of the energy levels in the molecule CO is helpful for understanding resonant absorption and the usefulness of planetary spectroscopy as a tool for studying planetary atmospheres. The rotational energy levels in CO are quantized according to the relation

$$E_\mathrm{r} = J(J+1)B, \tag{15}$$

where J ($J = 0, 1, 2, \ldots$) is the rotational quantum number and B the rotational constant. The rotational constant for a diatomic molecule is defined by

$$B = h^2 8\pi^2 I, \tag{16}$$

where I is the moment of inertia of the molecule about the axis perpendicular to the line connecting the nuclei and h is Planck's constant. The transitions between the various energy levels are limited by the exclusion principle to $\Delta J = \pm 1$. The energy of a quantum emitted or absorbed during a transition is given by

$$h\nu = J(J+1)B - (J-1)(J)B = 2JB. \tag{17}$$

When B is expressed in megahertz, the transition frequency ν in megahertz is expressed directly as $2JB$. The value of B for CO is 57,897.5 MHz. Thus, CO has two rotational transitions in the millimeter spectrum, a ground state (∼115 GHz) and the first excited state (∼230 GHz). The ground-state transition corresponds to a transition between $J = 0$ and 1; the first excited state corresponds to a transition between $J = 1$ and 2. Both of these transitions have been observed in the atmospheres of Venus and Mars.

A number of factors cause the energy levels in a molecule to vary slightly and cause the molecule to emit or absorb over a range of frequencies. The range of frequencies or "width" of a spectral line is determined primarily by three factors: (1) natural attenuation, (2) pressure broadening, and (3) Doppler broadening. Pressure broadening and Doppler broadening are the dominant mechanisms in most planetary astronomy applications. Natural attenuation is interpreted as the disturbance of the molecule by zero-point vibration of electromagnetic fields, which are always present in free space.

Doppler broadening is caused by the frequency shift introduced by the molecule's motion. The statistical effect of the simultaneous observation of a large number of molecules moving at various velocities is to spread the frequency of the spectral line over a range of frequencies. The spectral line produced by thermal motions in a gas (in thermodynamic equilibrium) is symmetric and has a full-width at half-maximum of

$$\Delta\nu = 7.2 \times 10^{-7}(T/M)^{1/2}\nu, \tag{18}$$

where T is the temperature of the gas, M is the molecular weight of the molecule, and ν is the frequency of the spectral line without Doppler shift. The Doppler linewidth of the ground-state transition ($J = 0$ to 1) of CO at 200 K is $\sim$200 kHz. A measure of the Doppler width can be used to estimate the temperature of the gas.

Pressure broadening can be interpreted as the effect of collisions disrupting the processes of emission and absorption in a molecule. A number of different theoretical line shapes have been derived to explain pressure broadening. The one most widely used is the Van Vleck–Weisskopf line shape. Linewidths due to pressure broadening are proportional to pressure, the proportionality constant depending on the particular molecules involved in the collisions. For CO broadened by CO_2, the linewidth of the ground-state transition of CO is approximately

$$\Delta\nu = 3.3p(300/T)^{0.75}, \tag{19}$$

where p is the pressure in millibars and $\Delta\nu$ is the linewidth in megahertz. At a pressure of 1 mbar and temperature of 200 K, the pressure-broadened linewidth of CO is $\sim$4.5 MHz. Observations of pressure-broadened spectral lines can be used to determine the altitude distribution of a molecule in a planetary atmosphere.

3. Radiative Transfer in Planetary Subsurfaces

While some planets have deep atmospheres, others, such as Mercury and Mars, have relatively tenuous atmospheres. For these planets (and the Moon and satellites) the atmospheres are nearly transparent at radio wavelengths, except possibly in narrow wavelength ranges, where resonant absorption lines can produce strong absorption. Thermal emission from the surfaces of these planets is easily observed at radio wavelengths; it is possible to interpret the measurements in terms of the physical properties of the near-surface materials.

Observations of the thermal radiation at radio wavelengths provide measurements that are complementary to infrared measurements of the subsurface materials. Thermal emission in the infrared is emitted very close to the surface of the planet because the opacity of most minerals is high at infrared wavelengths. Consequently, infrared thermal emission reflects the physical characteristics of the near-surface materials. The opacity of typical planetary materials is considerably less at radio wavelengths, and the observed thermal emission originates at greater depth.

The interpretation of thermal emission data from a planetary surface begins with an analysis of heat transfer in solids. The solid surface receives heat from the incoming solar radiation and transports it downward, mainly by conduction and radiation. Boundary conditions are set by the heating of the surface by the Sun, the nighttime cooling, and the internal sources of heat, if any. The surface temperature responds to the heating and cooling, being controlled by the "thermal inertia" of the near-surface material. Because the planets are heated and cooled periodically due to their spin, a thermal wave is set up in the surface layers. The equation of radiative transfer is used to relate the temperature structure in the subsurface layers to the observed thermal emission.

The formal solutions of the equations of heat transfer in solids and radiative transfer depend on the following properties of the near-surface material: the complex dielectric constant $\varepsilon(\lambda)$, thermal conductivity k (ergs per centimeter per second per Kelvin), specific heat c (ergs per gram per Kelvin), and density ρ (grams per cubic centimeter). The analytic theory of heat transfer at planetary surfaces begins by assuming that the temperature at any point on the surface can be expanded in a Fourier series in time:

$$T_s(t) = T_0 + \sum_{n=1}^{\infty} T_n \cos[(n\omega t) - \Phi_n], \tag{20}$$

where ω (radians per second) is the fundamental heating frequency (i.e., rotation rate of the planet as seen from the Sun) and t is time. Assuming that the planet is a semiinfinite homogeneous slab with constant thermal properties, the equilibrium solution for the subsurface temperature distribution is

$$T(x,t) = T_0 + \sum_{n=1}^{\infty} T_n \exp(-x\beta_n)\cos[(\omega t - \beta_n x - \Phi_n)], \tag{21}$$

where x is the depth beneath the surface and β_n is given by

$$\beta_n = (n\omega\rho c/2k)^{1/2} = \beta_1 n^{1/2}. \tag{22}$$

Equation (21) represents a series of thermal waves propagating into the surface and attenuating with distance. The higher harmonics are attenuated more rapidly than the lower harmonics since β_n increases as the square root of the harmonic number n. The attenuation and phase of each harmonic depend on the quantity β_1, which is termed the "thermal absorption coefficient" of the planetary material. The reciprocal of β ($L_\text{t} = 1/\beta_1$) is termed the "thermal skin depth." At a distance of 3–4 thermal skin depths, the fluctuations in subsurface temperature are practically zero.

Given the thermal absorption coefficient and the boundary conditions on the heating, it is possible to determine the constants of temperature (T_0 and T_n) and phase (Φ_n) in Eq. (21). The inverse problem is faced by the radio astronomer, namely to determine the thermal absorption coefficient from measurements of the thermal emission. This is done in the following manner. The temperature distribution given by Eq. (21) is used in the equation of radiative transfer

$$T_{\mathrm{B}} = [1 - R_P(\nu, \theta_0)] \times \int_0^\infty T(x)\exp(-k_\nu x/\cos\theta_i)\,(k_\nu/\cos\theta_i)\,dx \tag{23}$$

to compute the radio brightness temperatures at any wavelength for comparison with observations. In this expression k_ν is the power absorption coefficient at frequency ν, θ_i is the angle of incidence of radiation just below the surface, and θ_0 is the angle of incidence of the observation. The function $R_P\,(\nu, \theta_0)$ is the Fresnel reflection coefficient of polarization P emerging at angle θ_0.

After substituting (the series) Eq. (21) for $T(x, t)$ in Eq. (23), the integral can be evaluated as a series of integrals. Each integral can be put in the form of a standard Laplace transform and integrated directly. The resulting brightness temperature at time t at a specified point on the surface is given by

$$T_{\mathrm{B}}(\nu, p, t) = [1 - R_P]\sum_{n=1}^{\infty}\frac{T_n\cos[n\omega t - \Phi_n - \Psi_n(\theta_i)]}{[1 + 2\delta_n(\theta_i) + 2\delta_n 2(\theta_i)]^{1/2}}, \tag{24}$$

where

$$\delta_n(\theta_i) = n^{1/2}\beta_1\cos(\theta_i)/k_\nu, \tag{25}$$

$$\Psi_n = \tan^{-1}\{\delta_n[\theta_i/(1 + \delta_n(\theta_i)]\}. \tag{26}$$

Defining l/K_ν as the radio absorption length (L_{e}), we have that δ_1 reduces to the ratio of the radio absorption length to the thermal absorption length at normal incidence. Equations (24)–(26) relate the observational data to the physical parameters of the surface materials. Only the first few terms of Eq. (24) are normally important because the higher order terms are small. The most important terms determined from the observations are δ_1 and Ψ_1. The δ_1 term defines the reduction in amplitude of the fundamental diurnal wave component from its surface value. It is best determined from observations at several different wavelengths spanning a wavelength range of 2:1 or more. The Ψ_1 term defines the phase shift in the diurnal wave at depth. It also can be deduced from measurements at several wavelengths, but usually with somewhat less precision than δ_1.

The parameter δ_1 can be written in terms of the physical characteristics of the surface material as follows:

$$\delta_1 = L_{\mathrm{e}}L_{\mathrm{t}} = (\Omega\rho c/2k)^{1/2}\lambda[2\pi\sqrt{\varepsilon}\tan(\Delta)]^{-1}, \tag{27}$$

where $\tan(\Delta) = 2\sigma/\varepsilon_{\mathrm{r}}\nu$ is the loss tangent, ε_{r} is the real part of the dielectric constant, and σ is the electrical conductivity (mhos per meter) of the medium.

Radio observations by themselves do not permit separation of the physical parameters contained in δ_1 and Ψ_1. Nevertheless, the radio data, when combined with infrared data, radar data, and laboratory data for real materials, constrain the material properties and in some cases allow one to exclude certain classes of materials in favor of others.

C. Nonthermal Radio Emission

Thermal radio emission in solids and neutral gases arises from the emission of quanta from individual atoms and molecules in thermodynamic equilibrium with each other. Random collisions between ions and electrons in thermal equilibrium with each other in an ionized gas also produce thermal emission. The electrons and ions in this case have a Maxwellian velocity distribution. When energy sources are present that produce particles having a non-Maxwellian velocity distribution, the system is not in thermodynamic equilibrium. Efficient processes can arise under these conditions that produce large amounts of radio energy. Nonequilibrium conditions frequently arise in an ionized gas or plasma typical of those found in space. In a fully ionized gas, nonequilibrium conditions can lead to coherent and incoherent plasma emissions. The radiation that arises from these mechanisms is called nonthermal emission. Cyclotron emission and synchrotron emission are examples of nonthermal emission. In the case of thermal emission, the blackbody radiation laws limit the radiation to an amount corresponding to the temperature of the body. For nonthermal radiation, this limit does not exist. The brightness temperature of nonthermal radiation sometimes exceeds millions of degrees, even though the effective temperature of a planet does not exceed several hundred degrees.

The classic sources of nonthermal radio emission within the solar system are Jupiter's magnetosphere and the solar corona. Nonthermal radio emissions have also been observed from the Earth's magnetosphere and the other three giant planets, Saturn, Uranus, and Neptune. The emissions from the latter planets can only be observed from spacecraft near the planets. Cyclotron and coherent plasma emission account for much of the low-frequency (<10 MHz) nonthermal emission from these planets and the solar corona. This emission is highly variable and the details of the generation processes involved are not clearly understood.

Synchrotron radiation is the dominant source of emission from Jupiter from about 50 MHz to 5 GHz; it also accounts for continuum bursts of type IV from the sun. The theory of synchrotron radiation is well developed. A review article on magnetospheric radio emissions by Carr, Desch, and Alexander, contained in the book edited by Dessler (1983), lists a number of references.

High-energy electrons moving in a magnetic field produce synchrotron radiation. Although the observed synchrotron radiation from a planet or the sun is the integrated emission from many electrons, an understanding of the radiation characteristics of a single electron in a magnetic field is sufficient to understand the qualitative aspects of the observations.

A single charged electron moving in a magnetic field is accelerated unless its velocity is solely in the direction of the magnetic field. This causes the electron to emit electromagnetic waves. The nature of these waves depends on whether the electron is nonrelativistic (velocity $\ll 3 \times 10^{10}$ cm/sec) or relativistic (velocity $\sim 3 \times 10^{10}$ cm/sec). The radiation emitted by nonrelativistic electrons is referred to as cyclotron radiation; radiation emitted by relativistic electrons is referred to as synchrotron emission.

A nonrelativistic electron with mass m and charge e, in the presence of a magnetic field (of magnitude B), moves in a helical path with the sense of rotation of a right-hand screw advancing in the direction of the magnetic field. The frequency of rotation about the magnetic field, sometimes called the electron cyclotron frequency or the gyrofrequency, is given by (where B is given in gauss)

$$f_c = Be/2nm = 2.8B(\mathrm{G}) \quad \mathrm{MHz}. \tag{28}$$

Cyclotron radiation is emitted in all directions and has a frequency equal to the gyrofrequency. The radiation is polarized with the polarization depending on the direction of propagation. The polarization is circular when viewed along the direction of the magnetic field and linear when viewed in the plane of the orbit. At intermediate angles the polarization is elliptical.

Relativistic electrons radiate not only at the gyrofrequency, but also at the harmonics. The relativistic mass increase with energy causes the harmonic spacing to decrease with increasing energy until the synchrotron spectrum is essentially smeared into a continuum. The radiation from a relativistic electron is highly nonisotropic. The emitted radiation is concentrated within a narrow cone about the instantaneous direction of the velocity vector with an approximate half-cone-width given by

$$\theta \cong mc^2/E \quad \mathrm{rad}, \tag{29}$$

where E is the electron energy. When E is expressed in millions of electron volts (MeV) and θ in degrees, the expression becomes

$$\theta \cong 29/E(\mathrm{MeV}) \quad \mathrm{deg}. \tag{30}$$

An observer situated in the plane of the electron orbit would see one pulse per revolution of the electron. These pulses recur at the relativistic gyrofrequency of the electron.

A single electron of energy E radiates synchrotron emission with an intensity spectrum that varies as $\nu^{1/3}$ (ν is frequency) up to a critical frequency $0.29\nu_c$ and decreases exponentially at higher frequencies. The critical frequency is defined as

$$\begin{aligned}\nu_c &= (3e/4nmc)(E/mc^2)B\\ &= 16.08B(\mathrm{G})E^2(\mathrm{MeV}) \quad \mathrm{MHz}.\end{aligned} \tag{31}$$

Thus a 10-MeV electron in a 1-G magnetic field will radiate a maximum intensity near 0.29×1608 MHz. The spectral density of the radiation near the frequency of the maximum intensity is

$$I(\nu = 0.29\nu_c) \cong 2.16 \times 10^{-29} B(\mathrm{G}) \quad \mathrm{W/Hz}. \tag{32}$$

The total energy radiated per second (in watts) is given by

$$P = 6 \times 10^{-22} B^2(\mathrm{G})E^2(\mathrm{MeV}) \sin^2\alpha \quad \mathrm{W} \tag{33}$$

where α is the pitch angle of the electrons ($\alpha = 90^\circ$ for electrons with no motion in the direction of the magnetic field).

As in the case of nonrelativistic electrons, the polarization is linear when the electron orbit is seen edge-on and elliptical or circular elsewhere. Since the intensity of the radiated power is beamed in the plane of the orbit, synchrotron radiation is predominantly linearly polarized.

It is possible to deduce many properties of Jupiter's magnetic field and of the high-energy particle environment from radio measurements of Jupiter by judicious use of the equations given above. Detailed calculations of synchrotron emission are complex since they involve integrals over the volume of the emitting electrons and over the electron energy spectrum while taking into account the complex geometry of the magnetic field and polarization properties of the radiation.

III. INSTRUMENTATION FOR PLANETARY RADIO ASTRONOMY

A radio telescope is a device for receiving and measuring radio noise power from planets, satellites, asteroids, and comets as well as from galactic and extragalactic radio sources. A simple radio telescope consists of an antenna for collecting the noise power (in a specified bandwidth, polarization, and from a limited range of directions) and a sensitive receiver–recorder for detecting and recording the power. The antenna is analogous to the objective

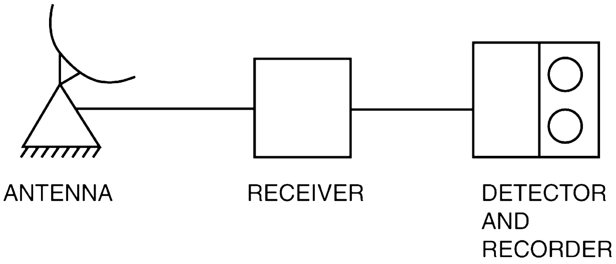


FIGURE 4 Basic components of a simple radio telescope.

lens or primary mirror of an optical telescope; the receiver is analogous to the recording medium of the optical telescope (i.e., photographic plate, photodetectors, etc.). Figure 4 shows the basic components and configuration of a simple radio telescope. Single antennas may be connected together electrically to form an "array" radio telescope which has greater sensitivity and directivity than a single antenna. There are many different types of radio telescopes in use today and their capabilities and visual appearances show much diversity.

Three important properties of an antenna are its effective area $A_e(\theta, \phi)$, normalized antenna power pattern $P_N(\theta, \phi)$, and gain G. The effective area is a measure of the wavefront area from which the antenna can extract energy from a wave arriving at the antenna from different directions. It can be thought of as the equivalent cross section of the antenna to the incident wave front. For most radio telescopes, the effective area is less than the physical area. A working definition of the effective area is given by

$$P = \frac{1}{2} S A_e \, d\nu, \tag{34}$$

where P is the power the antenna can deliver to a matched load in bandwidth $d\nu$ when flux density S is incident on the antenna. The factor 1/2 is introduced because it is assumed that the radiation is unpolarized and that the antenna is responsive to only one polarization component.

The effective area of an antenna is a function of the direction of arrival of the waves. The effective area to a signal from a distant transmitter, as a function of direction, is called the antenna power pattern. The effective area normalized to unity in the direction of the maximum effective area $A_{\text{e-max}}$ is the normalized power pattern of the antenna:

$$P_N(\theta, \phi) = A_e(\theta, \phi)/A_{\text{e-max}}. \tag{35}$$

By reciprocity arguments, the normalized power pattern is the same for both transmitting and receiving.

A general expression for the power delivered to a radio receiver from a planet is obtained by integrating the brightness over the solid angle of the planet and over the bandwidth of the receiver as follows:

$$P = \frac{1}{2} \iiint B(\theta, \phi, \nu) A_e(\theta, \phi, \nu) \sin(\theta) \, d\theta \, d\phi \, d\nu \tag{36a}$$

$$P = \frac{1}{2} A_{\text{e-max}} \iiint B(\theta, \phi, \nu) \times P_N(\theta, \phi, \nu) \sin(\theta) \, d\theta \, d\phi \, d\nu. \tag{36b}$$

A typical antenna power pattern consists of a large number of lobes with one or a few lobes being much larger than the others, as shown in Fig. 5. The lobe with the largest maximum is called the main lobe, while the remaining lobes are called side or back lobes.

The half-power beamwidth of the main lobe, Θ, is the angle between the two directions in which the received power is half of that in the direction of maximum power. The half-power beamwidth is a measure of the ability of the antenna to separate objects that are close together in angle. In optical systems, this is known as resolving power. An estimate of the beamwidth in radians of the main lobe is given by $\Theta \sim \lambda/D$, where D is the linear dimension of the antenna in the plane in which the beam is measured and λ is the wavelength in the same units as D. The resolving power can be improved by using either a shorter wavelength or a larger diameter telescope.

Two major classifications of large radio telescopes are (1) filled-aperture telescopes and (2) unfilled-aperture telescopes. Filled-aperture radio telescopes generally consist of a single reflecting element that focuses the received radio waves to a point or, in some cases, along a line. Examples of filled-aperture radio telescopes are the 64-m parabolic reflector antenna at Parkes, Australia, the 100-m parabolic reflector at Bonn, Germany, and the ~300-m spherical reflector antenna at Arecibo, Puerto Rico. The newest of the large filled-aperture radio telescopes is the 100-m Green Bank Telescope (GBT). The surface panels of the GBT are supported by motor-driven actuators to compensate for the structural deformations of the antenna. The GBT is expected to operate at frequencies up to 80 GHz under the most favorable conditions available at the site.

The resolving power of most filled-aperture radio telescopes is generally less than is required for adequate resolution of the disks of the planets. The few exceptions are the large telescopes such as the GBT and the 30-m Pico Villeta telescope, which operate at short, millimeter wavelengths. The largest single-dish antennas presently available to radio astronomers are capable of partially resolving the disks of the planets of largest angular diameter, Venus and Jupiter. For example, an antenna with a maximum dimension of 100 m operating at a wavelength of 10 cm would have a main lobe width of approximately 1/1000 rad or 3.44 arc min. The angular diameters

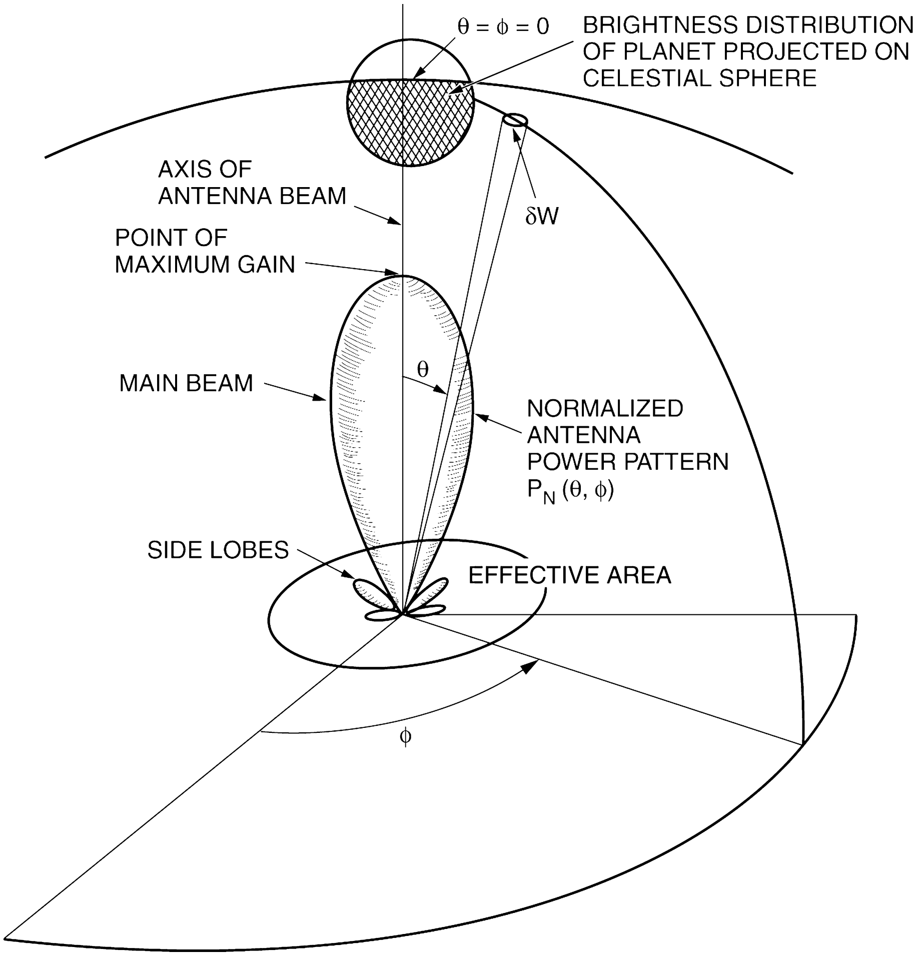

θ = φ = 0
BRIGHTNESS DISTRIBUTION OF PLANET PROJECTED ON CELESTIAL SPHERE
AXIS OF ANTENNA BEAM
δW
POINT OF MAXIMUM GAIN
MAIN BEAM
θ
NORMALIZED ANTENNA POWER PATTERN $P_N(\theta, \phi)$
SIDE LOBES
EFFECTIVE AREA
φ

FIGURE 5 Antenna power pattern and its relation to a measurement of the flux density from a planet.

(Table I) of Venus and Jupiter are approximately 1 arc min; thus an antenna with an aperture of ∼344 is required just to "fill the main lobe." Adequate resolution of even the largest planets (at 10-cm wavelength) requires apertures at least 10 times larger.

Higher resolving power can be achieved by connecting together electrically the outputs of two or more filled-aperture antennas to the input of a common radio receiver. The resolving power of such a system depends on the maximum separation between the individual elements, even though the space between the elements is unfilled. Multiple-element radio telescopes (arrays) with spaced separations between the elements to achieve high resolving power are termed unfilled-aperture radio telescopes. The simplest unfilled-aperture radio telescope is a total power interferometer with two identical elements, as shown in Fig. 6. The output from the two-element interferometer modifies the power pattern of a single element with an angular modulation of scale λ/D superimposed.

The Very Large Array (VLA) radio telescope, located on the plains of San Augustin west of Socorro, New Mexico, is a classic example of an unfilled-aperture radio telescope. This instrument, completed in early 1981, is especially important to planetary radio astronomers. The VLA consists of 27 antennas, each having a diameter of 25 m. The individual antennas are arranged in a huge, Y-shaped pattern. The antennas are mounted on tracks so that they can be moved into four configurations along the Y-shaped baseline. This allows the radio telescope to be custom tailored to a particular kind of measurement. The maximum antenna separation is 36 km across. At the highest frequency (43 GHz), the VLA has a resolution of .043 arc sec. The effective sensitivity is equivalent to a filled-aperture telescope having a diameter of 130 m. Continuum, polarization, and spectral line observations at eight wavelength bands (0.7, 1.3, 2.0, 3.6, 6, 20, 90, and 405 cm) are supported. The VLA has sufficient resolving power to measure the brightness distributions across all of the planets at its shortest wavelengths.

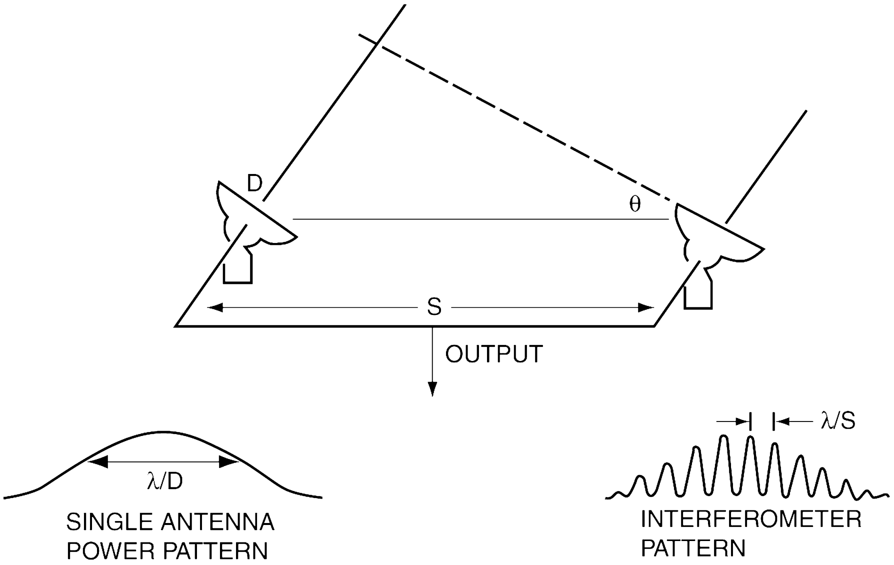


FIGURE 6 Total power interferometer. (Top) Geometry of a two-element interferometer. (Bottom) Antenna response for a single element of the interferometer (left) and response of the interferometer (right) to a completely unresolved planet.

Small arrays at millimeter wavelengths have been operational for over a decade and have returned many interesting continuum and spectroscopic images of the planets. Major millimeter arrays are the BIMA millimeter array, the Owens Valley Radio Observatory (OVRO) millimeter array, the Plateau de Bure Observatory, and the Nobeyama Millimeter Array (NMA). The BIMA and OVRO millimeter arrays are likely to be combined into a single large millimeter array in the future. The Submillimeter Array (SMA) is an exploratory instrument for high-resolution observations at submillimeter wavelengths. The SMA will initially consist of eight 6-m antennas sited on Mauna Kea at an elevation of 4050 m. A major step in millimeter solar system research will be provided by the construction of very large millimeter arrays. Two large arrays are currently in their early phases. The Atacama Large Millimeter Array (ALMA) will consist of 64 (or more) 12-m antennas located at an elevation of 16,400 ft in Llano de Chajnantor, Chile. This array will provide imaging in all atmospheric windows between 1 cm and 350 μm. The array configuration will provide baselines ranging from 150 m to 10 km. Spatial resolution of 10 marcsec, 10 times better than the VLA, will be possible. A Large Millimeter and Submillimeter Array (LMSA) has been proposed in Japan. This array will consist of 50 10-m antennas and will cover frequencies from 80 to 800 GHz. These new very large arrays will open up many possibilities for planetary radio astronomers. Imaging of planetary atmospheres will be greatly improved and general circulation studies will be greatly enhanced. Planetary satellites will be easily resolved making temperature and wind measurements possible. Imaging and spectroscopy of comets will be greatly improved and many new molecules can be studied.

The performance of a filled-aperture antenna is sometimes expressed by specifying the gain of the antenna. The antenna gain is a measure of the ability of the antenna to concentrate radiation in a particular direction. Gain is defined as the ratio of the flux density produced in direction (θ, ϕ) by an antenna when transmitting with an input power P to the flux density produced by the same transmitter feeding an antenna that radiates equally in all directions. The antenna gain, effective area, and beam solid angle Ω are interrelated quantities. The relationship between them is given by

$$G = 4\pi A_{\text{e-max}}/\lambda^2 = 4\pi/\Omega. \qquad (37)$$

Radio astronomers generally specify the received power P delivered to the receiver in the frequency bandwidth $d\nu$ as the antenna temperature T_{a}, defined by

$$T_{\text{a}} = P/k\,d\nu, \qquad (38)$$

where k is Boltzmann's constant. This definition follows from the Nyquist theorem, which states that the noise power delivered from a resistor at temperature T into a matched load in bandwidth $d\nu$ is

$$P_{\text{N}} = kT\,d\nu. \qquad (39)$$

The antenna temperature can be thought of as the physical temperature at which a resistor would have to be maintained to deliver the same power to the receiver as the antenna.

The antenna temperature can be thought of as the physical temperature at which a resistor would have to be maintained to deliver the same power to the receiver as the antenna.

The antenna temperature of an unresolved radio source can be calculated directly from the definition of the effective area and the definition of antenna temperature if its flux density is known. For a planet of known solid angle and temperature, the antenna temperature can be calculated by substituting the Rayleigh–Jeans approximation for the brightness, $B = 2kT_\mathrm{p}/\lambda^2$, and the relationship between effective area and beam solid angle,

$$A_{\text{e-max}} = \lambda^2/\Omega, \tag{40}$$

into Eq. (36). When the planet is unresolved by the main beam of the telescope, the antenna temperature is given by

$$T_\mathrm{a} = (\Omega_\mathrm{p}/\Omega)T_\mathrm{p}, \tag{41}$$

where Ω_p and T_p are the solid angle and temperature of the planet, respectively, and Ω is the beam solid angle of the telescope. The relationship states that the antenna temperature of a planet is proportional to the brightness temperature of the planet, with the constant of proportionality being the ratio of the solid angle of the planet to the solid angle of the beam. This approximation holds when $\Omega_\mathrm{p} \ll \Omega$; consequently, the antenna temperature is always less than the brightness temperature.

The concept of specifying the received power in terms of an equivalent temperature is useful for estimating the signal-to-noise ratio for a particular measurement since random noise in the receiving equipment is easily expressed as a temperature. A discussion of signal-to-noise ratio follows.

A. Radio Receivers

Radio astronomy receivers, like ratio telescopes, are highly varied, depending on the type of measurement to be performed. The function of the receiver is to detect and measure the radio emission with as much sensitivity as possible. The receiver also defines the frequency range or ranges of the measurement. Most modern receivers consist of a low-noise amplifier to boost the power of the incoming signal (without adding significant noise), followed by a heterodyne mixer and square law detector. The heterodyne mixer transforms the signal frequency to a convenient (often lower) frequency for detection or further processing. For high-resolution spectroscopic observations, the square law detector may be preceded by spectrometer, such as a filter bank, digital autocorrelator, or acousto-optical spectrometer.

The inherent noise fluctuations of a radio receiver usually determine the weakest signal strength that can be measured with a radio telescope. The statistical nature of noise radiation is such that statistical fluctuations are proportional to the noise power itself. Furthermore, the average of N independent measurements of the noise power is $\sqrt{N}$ times more accurate than a single measurement. Noting that a single independent measurement can be made in the minimum time interval $1/d\nu$, we see that the maximum number of independent measurements that can be made in time t is $t\, d\nu$. Thus, the sensitivity equation for an ideal receiver is given by

$$\Delta T = \text{rms noise power} \approx T_\mathrm{S}/(t\, d\nu)^{1/2}, \tag{42}$$

where the "system temperature" T_S is a measure of the noise power from the receiver and t is the integration time of the measurement. The rms noise power is expressed in kelvins and can be directly compared with the antenna temperature to determine the signal-to-noise ratio (SNR) of a particular measurement.

A modern radio telescope may have a system temperature of 20 K or less in the frequency range 1–10 GHz, where the radiation from the terrestrial atmosphere and galaxy are both low. Noise temperatures in the millimeter and submillimeter bands are substantially higher. For measurements of the radio continuum, $d\nu$ may be chosen to be 10 MHz or larger, depending on the characteristics of the radio receiver being used. If we adopt a value of 100 MHz for $d\nu$ and 20 K for T_S, then the rms noise power obtained in 1 sec of integration is 0.002 K. For planetary spectroscopy, $d\nu$ would have to be reduced to 1 MHz or less and the rms noise power would increase to 0.02 K. These noise fluctuations can be further reduced by increasing the integration times; however, systematic effects within the receiving equipment prevent ΔT from being pushed to zero.

B. Spacecraft

Earth-based radio observations of the planets have limitations which affect certain types of measurements. These limitations include (1) the inability to obtain spatial resolution on scales of a few meters or less, (2) restrictions on the viewing geometry of the planets, (3) limitations set by the opaqueness and variability of the terrestrial atmosphere, (4) the intrinsic faintness of the radio emissions from planetary bodies, and (5) interference from man-made radio noise. The opacity of the terrestrial atmosphere varies with frequency. The atmosphere is opaque at frequencies lower than about 5 MHz due to the terrestrial ionosphere. Attenuation due to the atmospheric gases, water vapor, and oxygen affects the centimeter, millimeter, and submillimeter bands, but observations are possible from the ground by working in the transparent "windows" in the spectrum. Rain, fog, and clouds occasionally limit the usefulness of the centimeter and millimeter bands.

To overcome these difficulties, a number of spacecraft radio instruments have been proposed, and several have flown on U.S. spacecraft. The first planetary radio system on a U.S. spacecraft was a two-channel microwave radiometer that operated at wavelengths of 13.5 and 19.0 mm, flown on the *Mariner II* spacecraft to Venus in 1962. The microwave radiometer system weighed ~10 kg and used an average power of 4 W. This early system was designed to take advantage of the high spatial resolution and sensitivity that could be achieved from a spacecraft. Another radio astronomy experiment was placed on the Voyager spacecraft that was launched in the late 1970s to the outer planets and targeted to fly by Neptune in 1989. This experiment measures the radio spectra of planetary emissions in the range from 1.2 kHz to 40.4 MHz. The system was designed to measure planetary spectra below the frequency range that is cut off by the Earth's ionosphere and to take advantage of the unique viewing geometry provided by the spacecraft. The Magellan Venus orbiter spacecraft carried a 12.6-cm radio receiver which allowed over 91% of the surface of Venus to be measured.

In the future, we expect to see many more radio astronomy spacecraft experiments. Spacecraft experiments will allow the submillimeter spectral range to be observed without hindrance from the terrestrial atmosphere. Planetary spectroscopy in the submillimeter spectral range is expected to reveal new information about the upper atmospheres of the planets. The ESA ROSETTA spacecraft will carry a millimeter- and submillimeter-wave spectroscopic instrument (MIRO) to investigate the nucleus and coma of a comet.

IV. RESULTS—PLANETS AND COMETS

A. Mercury

Radio measurements of Mercury show no evidence of an atmosphere. Estimates based on other techniques including spacecraft flyby instruments show the atmosphere on Mercury to be extremely tenuous, with a surface pressure of $\sim 5 \times 10^{-15}$ bar. For the purpose of interpreting the radio data, it can safely be assumed that Mercury is an atmosphere-less planet. The radio emission from Mercury is thermal in character, strongly controlled by the high eccentricity of Mercury's orbit and the synchronism between Mercury's spin period and its period of revolution. Solar tidal effects have caused the period of axial rotation of Mercury to be 58.642 days, precisely two-thirds of its orbital period of 87.97 days. One solar day on Mercury is equal to 3 stellar days or 2 Mercurian years. This period equals 176 mean Earth solar days. Because of the synchronism between spin period and revolution period, the Sun takes a curious diurnal path in the sky as seen from the surface of Mercury (Fig. 7). At some longitudes, the Sun rises and sets twice a Mercury day. At perihelion the insolation is approximately twice its value at aphelion. The insolation reaches a maximum value of $\sim 14 \times 10^3$ W/m^2, 10 times the value for Earth. The visual albedo of Mercury is similar to that of the Moon. The spin–orbit coupling and

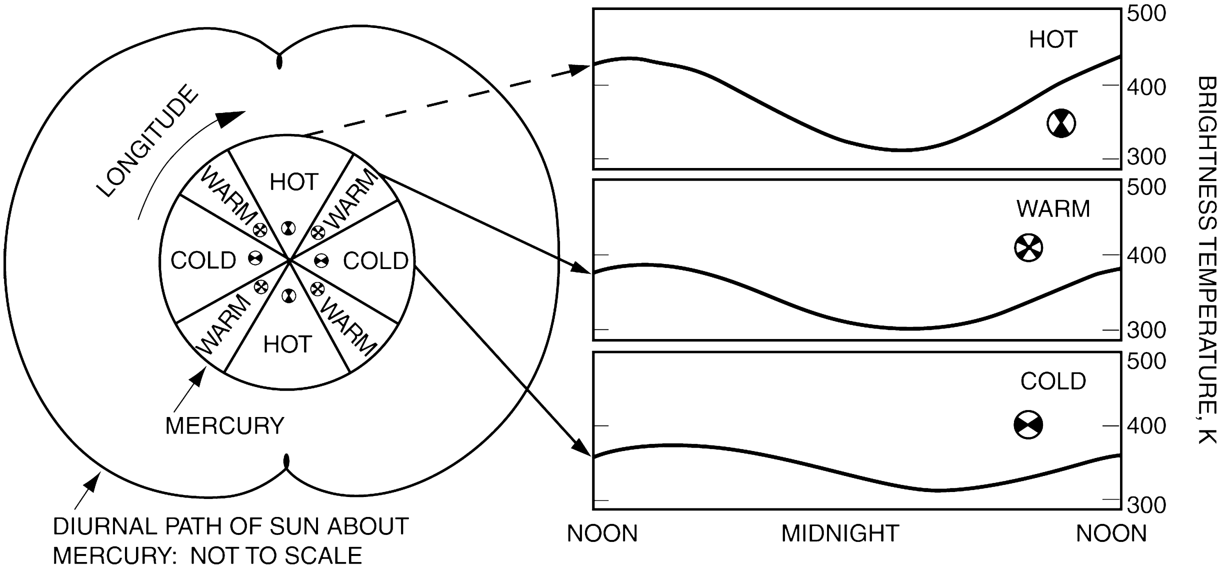


FIGURE 7 (Left) Diurnal path of the Sun about Mercury and (right) representative brightness temperature curves near 4-cm wavelengths for "hot," "warm," and "cold" regions on Mercury.

eccentricity combine to cause the surface of Mercury to be heated very nonuniformly in longitude. A pair of longitudes 180° apart alternatively faces the Sun at perihelion. These longitudes are preferentially heated because their midday insolation occurs when the Sun is nearly stationary on the meridian and the Sun–Mercury distance is at a minimum. At longitudes 90° away from these hot longitudes, the heating is identical but much less than that received at the hot longitudes. The longitudinal temperature variations on Mercury are indicated schematically by the "hot," "warm," and "cold" regions shown on the left in Fig. 7. This pattern of temperature variations is fixed on Mercury because the spin–orbit coupling causes the heating to be cyclic at each longitude (i.e., the same pair of longitudes faces the Sun at perihelion).

The solar heating cycle on Mercury suggests that the two longitudes that see the Sun directly overhead at perihelion (receiving more than twice as much energy as the longitudes 90° away) will be hotter than those 90° away. This is borne out by the radio observations. The right-hand portion of Fig. 7 shows a schematic representation of the variation of temperature at "hot," "warm," and "cold" regions on Mercury at a wavelength near 4 cm. Most of the older, single-dish microwave observations of Mercury do not have sufficient resolving power to resolve the disk of Mercury, so the reported temperatures are averages over the entire visible disk.

Radio images of the planet have been obtained with the VLA and BIMA arrays. Such images clearly show the brightness variation across the disk of Mercury. At short wavelengths, where shallow layers are probed, the temperature is usually highest at local noon. However, when deeper layers are probed, the diurnal heating pattern is less obvious, and one sees the two hot regions. Figure 8 (left side) shows a radio image at 3.6 cm. At this wavelength one probes ~70 cm into the crust. The hot region at longitude 0° is clearly visible on the night side, while the high temperature on the day side is caused both by the 180° hot longitude and solar insolation. Mercury's hot and cold longitudes have been modeled in detail. The image on the right shows residuals after a thermal model was subtracted from the image on the left. Most remarkable here are the negative temperature differences near the poles and along the terminator, suggesting that the poles and terminator are colder than predicted by the model. This is likely caused by surface topography, which causes a permanent shadowing effect at high latitudes and transient effects in the equatorial regions, where crater floors and hillsides are alternately in shadow and sunlight as the day progresses. Some crater floors near the poles are permanently shadowed, and radar observations have revealed evidence for the existence of water ice at such crater floors.

Radio spectra and images have been together with Mariner 10 IR data of the planet used to infer Mercury's surface properties. At first glance, Mercury's surface is quite similar to that of the Moon. Due to the fact that Mercury, like the Moon, is continuously bombarded by small meteorites, one would expect the top few centimeters to have a very low density, while deeper layers are more compact. Observations indeed show that this is the case on both bodies. Researchers found that the microwave opacity on Mercury is roughly a factor of 2–3 smaller than that of most lunar samples. This suggests that the ilmenite content, which is the most common titanium-bearing mineral on the Moon [(Fe, Mg)TiO], is much less abundant on Mercury than on the Moon. Ilmenite is also opaque at optical wavelengths and is largely responsible for the dark appearance of the moon's maria compared to its highlands. Its absence would explain why Mercury is brighter than the Moon at visible wavelengths.

B. Venus

Venus has the densest atmosphere of all the terrestrial planets. The principal atmospheric constituents are carbon dioxide and nitrogen (N_2); their mixing ratios are approximately 96.5% and 3.5%, respectively, below, 100-km altitude. Trace constituents below 100-km altitude are in the range of 0.1%. The Venus atmosphere is covered with thick clouds composed primarily of sulfuric acid and contaminants, making the surface invisible from above. The total pressure at the bottom of the cloud layer (~47 km) is approximately 1.3 atm. Water is highly depleted throughout the atmosphere. The mean physical structure of the atmosphere (pressure and temperature profile) is reasonably well known from the data returned by a number of space probes. The surface pressure and temperature (on a mean surface) are approximately 94 atm and 737 K, respectively.

The effective temperature of Venus deduced from measurements in the infrared is about 240 ± 8 K, corresponding to an altitude of approximately 60 km. Below this level, the temperature distribution generally follows that for an atmosphere that is in convective equilibrium. Convective equilibrium implies that the temperature gradient in the atmosphere is close to the adiabatic value. The temperature gradient is approximately 8.6 K/km. The high surface temperature is believed to be due to the greenhouse effect. The physical basis for this is that the visible light from the sun is only partially absorbed by the clouds and atmosphere. Some of the light reaches the surface and warms it. The heated surface reradiates in the infrared. The atmosphere is highly absorbing in the infrared spectral region to CO_2 and perhaps H_2O. The atmospheric opacity traps the infrared radiation, thereby raising the surface temperature.

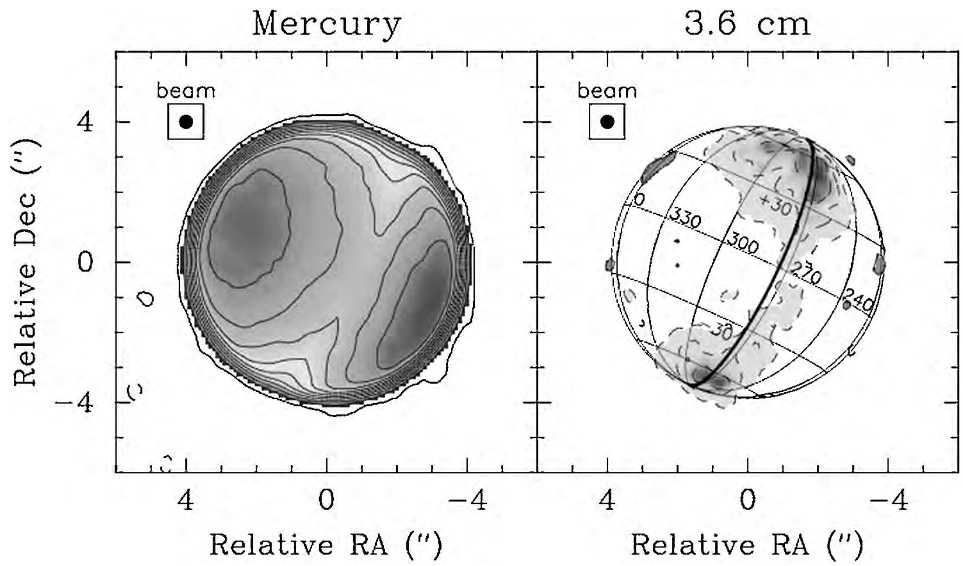


FIGURE 8 Image on the right shows a 3.6-cm thermal emission map of Mercury observed with the VLA. The beam size is 0.4 in. (1/10 of a Mercury radius). The direction to the Sun at the time of the observations and the morning terminator (dashed line) are superimposed on the image. The image shows thermal depressions at both poles and along the sunlit side of the morning terminator. Contours are at 42 K intervals except for the lowest contour, which is at 8 K (dashed contours are negative). The figure on the right shows the residuals after subtracting a model from the observed map. [From Mitchell, D. L., and de Pater, I. (1994). "Microwave imaging of Mercury's thermal emission: observations and models," *Icarus* **110,** 2–32.]

The atmosphere of Venus is opaque at millimeter and short centimeter wavelengths, gradually becoming transparent at longer wavelengths. Individual spectral lines are not observable below the clouds because of pressure broadening. The total vertical optical depth of the atmosphere of Venus at a wavelength of 1 cm is estimated to be slightly less than 20 and to vary approximately as λ^{-2} (optical depth = 1 at ~4 cm). A little more than half of the total opacity is due to collision-induced, nonresonant absorption in CO_2; the remaining opacity is produced by the minor constituents in the atmosphere. Other known or suspected microwave absorbers in the atmosphere are H_2O, SO_2, H_2SO_4, and the sulfuric acid particles in the clouds. Near wavelengths of 6 cm and longward, the atmosphere is sufficiently transparent that it is possible to measure the surface temperature of Venus using radio astronomical methods. The right side of Fig. 8 shows the continuum spectrum of Venus from a few millimeters wavelength out to approximately 6 cm. The left side of Fig. 9 shows the temperature versus altitude profile of Venus. The brightness temperature is seen to rise from about 225 K at 3 mm to about 700 K near 6 cm. This increase in brightness temperature is due to the decreasing opacity of the atmosphere with increasing wavelength. The decreasing opacity allows radio waves to escape from deeper regions in the atmosphere, where it is warmer due to the adiabatic lapse rate. At wavelengths longer than ~15 cm the brightness temperature decreases to ~600 K.

Radio interferometer data, radar data, and spacecraft data have been used to study the surface of Venus, in particular to determine the dielectric constant. Radar reflectivity data place the dielectric constant in the range 4–5. The Magellan radiometer experiment observed the 12.6-cm-wavelength radio emissivity of more than 91% of the Venus surface. With its 2-deg beam width, Magellan observations achieved surface resolutions that varied from 15 by 23 km at periapsis (10°N latitude) to about 85 km at the north pole. The global mean value of emissivity seen using horizontal linear polarization is 0.845, a value that corresponds to a dielectric permittivity of between 4.0 and 4.5, depending on the surface roughness. These values are considerably greater than the values for Mercury, Mars, and the Moon, which range from 2.0 to 2.5. The higher values are suggestive of a surface composed of dry rock not unlike many rocks of the Earth's surface. These values are consistent with the dry basaltic minerals thought to compose the bulk of the Venus surface. The observations have confirmed earlier findings that a few regions on Venus, primarily located at high elevations, possess unexpectedly low values of radiothermal emissivity, occasionally reaching as low as 0.3.

Carbon monoxide is an important constituent of the upper atmosphere of Venus. It is formed primarily by the dissociation of carbon dioxide by solar ultraviolet radiation and is removed by chemical and transport processes in the atmosphere. The ground and first excited rotational

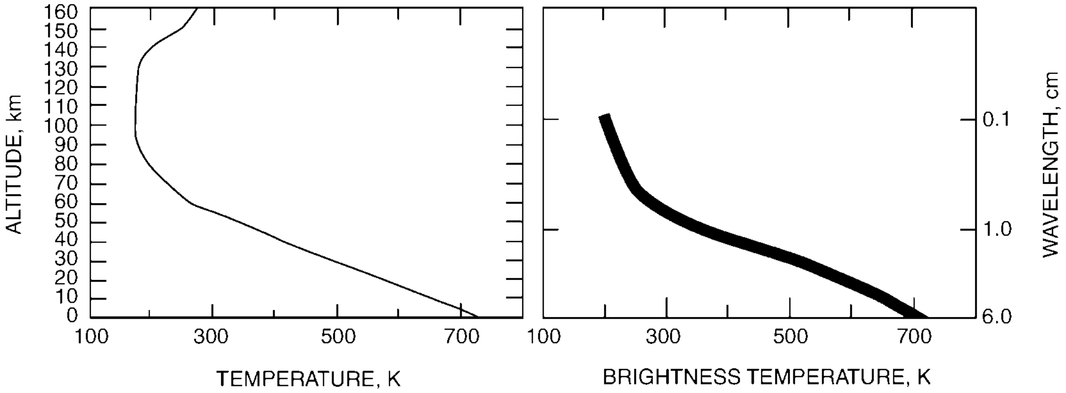


FIGURE 9 Schematic representation of the spectrum of Venus from 1 mm to 6 cm (right) and temperature versus altitude profile (left). The figure illustrates how the atmosphere is probed in altitude by changing the wavelength of the observations.

states of CO (located at very high altitudes in Venus' atmosphere) have been observed to absorb the hot continuous background of the deeper atmosphere. Examples of CO spectra on Venus' day- and night-side hemisphere are shown in Fig. 10. Observed spectral lines have included $^{12}CO(0–1)$, $^{12}CO(1–2)$, $C^{18}O(1–2)$, and $^{13}CO(1–2)$. These observations have been particularly useful for exploring the altitude range from 70 to 110 km.

It has been possible to derive the vertical temperature and mixing ratio profiles of CO in the upper atmosphere of Venus from such observations. As shown in Fig. 9, the spectra reveal considerable variability of CO abundance, which varies with solar phase angle. The variability is believed to be the result of large-scale circulation in the upper atmosphere of the planet. Detailed studies of this variability have been undertaken with arrays of telescopes which operate at millimeter wavelengths. With such arrays the planet can be imaged in the CO lines, so the CO concentration and temperature profile at all solar phase angles can be measured simultaneously. In addition, through measurements of the Doppler shift of the lines at various locations on Venus' disk, the winds in Venus' mesosphere can be studied. For the CO(0–1) transition, a mean wind speed of 100 m/sec in the spectral-line-forming region produces a Doppler shift of 38 kHz. Such Doppler shifts have been observed at altitudes at least as high as 100 km. Continued observations of these spectral lines will lead to a better understanding of the large-scale circulation in Venus' upper atmosphere.

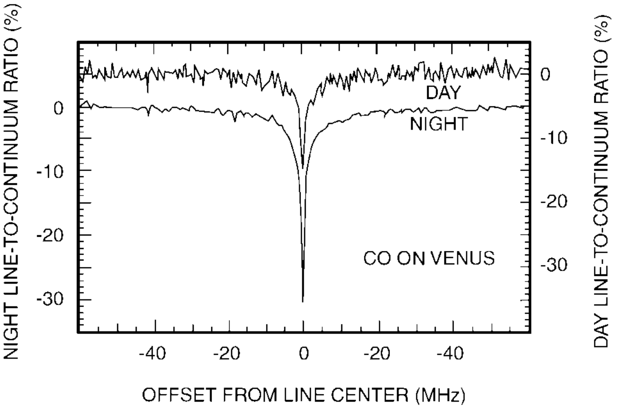


FIGURE 10 Observed absorption spectra of $J = 0 \rightarrow 1$ transition of CO in the atmosphere of Venus. Line center frequency is 115 GHz. [Adapted from Schloerb, P. (1985). *In* "Proceedings ESO-IRAM-ONSALA Workshop on Submillimeter Astronomy."

C. Mars

Mars moves in an orbit slightly larger than the Earth's, always turning its day side toward the Earth as it approaches. Earth-based measurements of the night side are impossible and phase angle coverage is greatly restricted. The axis of rotation is tilted from the perpendicular to the plane of its orbit by 25°, about the same as for the Earth. Radio observations of the disk brightness temperature of Mars show it to have a nearly flat spectrum from about 1 mm to 21 cm. The mean disk brightness temperature is about 215 K. As seen from the Earth, the average surface disk temperature varies by ±15 K as the sub-Earth point moves from afternoon to morning and from midlatitudes to equatorial latitudes. Most observational data used in thermal modeling studies have come from infrared measurements made from spacecraft. The infrared measurements are limited to the near-surface properties. Microwave images obtained with the VLA have extended the thermal models to the subsurface layers of the planet. Microwave images, in contrast to single-dish observations, also yield information on the spatial variations of the brightness temperature. For example, VLA images show the polar regions to be significantly colder than the disk-averaged temperature.

The centimeter radio brightness temperatures of Mars have been found to vary as a function of the central meridian longitude of the planet. Temperature differences as large as 5–10 K are observed over the full range of longitudes. The variations are believed to be due to nonconformity in the Martian surface properties; however, no completely satisfactory explanation of the observations exists.

The Martian atmosphere is very tenuous, having a surface pressure some 200 times less than that of the Earth. The primary constituent of the lower atmosphere is CO_2. Photolysis of CO_2 by solar ultraviolet radiation produces CO and O_2. The diatomic molecule CO plays an important role in determining the millimeter-wave spectrum of Mars.

Both the ground-state and first-excited-state transitions of CO have been observed in the Martian atmosphere. The altitude distribution of CO has been inferred by interpreting the observed line shape in terms of pressure broadening. A column abundance of CO equal to $(2–5) \times 10^{20}$ molecules/cm^2 has been derived from the measurements. This number is very stable over time, even though the CO line profiles show significant variability. This variability has been used to monitor Mars' thermal structure: the atmospheric temperature rises significantly during global dust storms, since the dust grains, heated by the Sun, warm up the atmosphere while at the same time preventing sunlight from penetrating down to the surface. Water vapor (from HDO observations at 226 GHz) has also been observed on Mars and used to determine water vapor distribution and atmospheric behavior. Clancy and collaborators used the VLA at a wavelength near 1.35 cm to image the water vapor in Mars' atmosphere. The water vapor shows up in emission around the planet, where the path length through the atmosphere is largest. The data clearly show the absence of emission over Mars' polar caps, where the atmospheric temperature is so low that all the water has been frozen out. Note that the ground-based detection of water vapor in Mars's atmosphere could only be achieved because of the VLA's high angular resolution, so that the atmosphere on the limb could be separated from the planet itself.

D. Jupiter

Jupiter was the first planet detected at radio wavelengths. The discovery observations occurred in 1955, at the very low frequency of 22.2 MHz. Prediscovery observations of Jupiter were later traced back to 1950. Subsequent observations of Jupiter revealed that its radio spectrum is exceedingly complex, showing both thermal and nonthermal emission mechanisms. Thermal emission from the atmosphere dominates the Jovian spectrum shortward of 7 cm. Nonthermal synchrotron emission dominates the spectrum from $\sim$3 m to 7 cm; brightness temperatures exceed 10^5 K for the synchrotron component. Longward of 7.5 m, Jupiter emits strong and sporadic nonthermal radiation. The radiation exhibits complex frequency, time, and polarization structure. The brightness temperature of this component exceeds 10^{17} K, suggesting a coherent source of emission. The schematic appearance of Jupiter's spectrum is shown in Fig. 11.

Observations of Jupiter at high angular resolution with radio interferometers have been used to map the synchrotron radiation from Jupiter's radiation belts and to separate the thermal from the nonthermal synchrotron components. The nonthermal component is easily identifiable with a radio interferometer because it is greatly extended relative to the optical disk of Jupiter and is strongly linearly polarized.

The thermal component originates in the Jovian atmosphere. The observations are consistent with a deep model atmosphere, composed mostly of hydrogen and helium, in convective equilibrium. The principal source of opacity is ammonia (NH_3), which exhibits very strong absorption in the microwave spectral region.

In December 1995 the Galileo spacecraft released a probe into the atmosphere of Jupiter, which relayed its findings to the spacecraft via radio signals at a frequency of 1.4 GHz. By analyzing the attenuation of the probe radio signal, the ammonia abundance in Jupiter's deep atmosphere (at pressures over 8 bar) was derived to be a factor of $\sim$3.5 larger than the solar N value. Ground-based microwave measurements are most sensitive to layers where the clouds form ($\sim$0.5 bar) down to roughly 10–15 bar. The ground-based microwave measurements do not show as much ammonia as the probe. Apparently, the ammonia abundance in Jupiter's deep atmosphere is significantly decreased at higher altitudes, to roughly half the solar N value just below the upper cloud deck. Scientists do not yet understand why there is so much less ammonia in the upper regions of Jupiter's atmosphere compared to deeper layers.

VLA images resolve the disk of Jupiter and show the familiar zone-belt structure at 2–6 cm. Figure 12 shows a 1.2-arc sec resolution VLA radio image at 2-cm wavelength. The disk diameter of Jupiter was 32 arc sec at the time of the radio observations. Radio images, such as shown in Fig. 12, are usually smeared in longitude, since the observations are integrated over a substantial time interval. The bright (white) beltlike regions are indicative of a higher brightness temperature, which is likely due to a relative depletion of NH_3 gas compared to the darker colored regions. Since the radio waves originate in and below the visible cloud layers, such images contain information complementary to that obtained at IR and optical wavelengths. From the images, the latitudinal variation

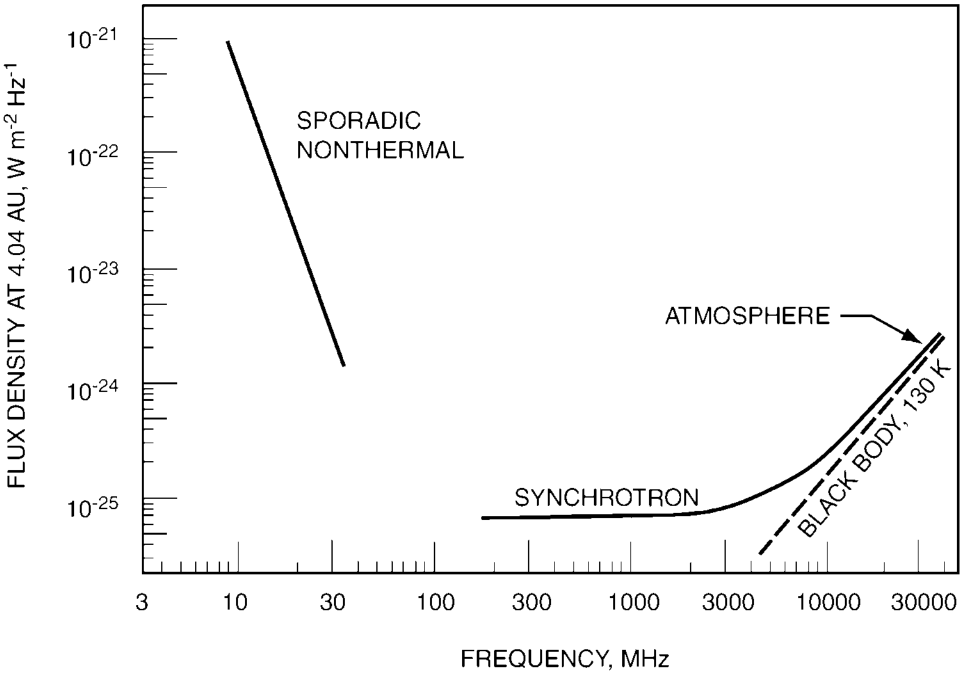


FIGURE 11 Schematic representation of the spectrum of Jupiter, showing the frequency ranges for which atmospheric emission dominates, synchrotron emission dominates, and sporadic nonthermal emission dominates.

of NH_3 gas can be obtained, in addition to the altitude distribution. Such variations must be due to dynamical processes on the planets, for example, zonal winds, upwelling, and subsidence of gas. The zone-belt structure on Jupiter is consistent with upwelling gas in the zones and subsidence in the belts. Images taken in different years show clear variations in the zone-belt structure, indicative of meteorological changes.

Radio interferometric maps of Jupiter's synchrotron emission have been made at a number of different wavelengths. It has been possible to deduce a great deal of information about Jupiter's magnetosphere from the radio measurements. The radio astronomical measurements provided convincing proof that Jupiter has a strong magnetic field, and this information was used to design the first spacecraft sent to Jupiter. The radio measurements show that the magnetic field is primarily dipolar in shape with the dipole axis tilted about 10° with respect to Jupiter's rotational axis. Using the well-developed theory of synchrotron emission (summarized in Section II.C), it has been possible to determine energies and densities of the high-energy electrons that are trapped in Jupiter's magnetic field.

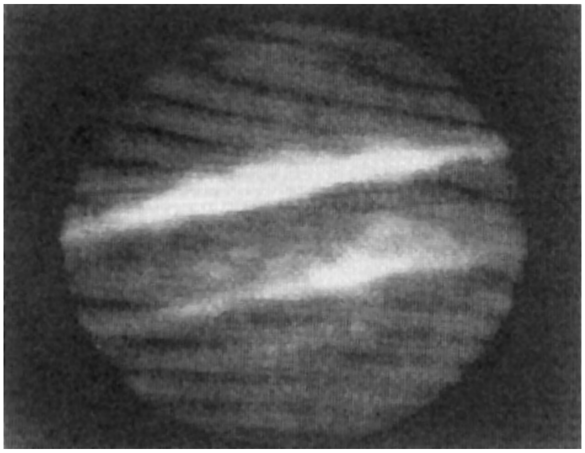


FIGURE 12 Radio photograph of Jupiter at a wavelength of 2.0 cm obtained with the VLA. Resolution is 1.2 arc sec. Equatorial diameter of Jupiter is 32 arc sec.

Figure 13 shows a radio image of Jupiter's synchrotron radiation. The main radiation peaks are indicated by the letters L and R, and the high-latitude emission peaks by Ln, Ls, Rn, and Rs. Magnetic field lines at jovicentric distances of 1.5 and 2.5 (from the O6 magnetic field model) are superimposed. The resolution is 0.3 Jovian radii, roughly the size of the high-latitude emission regions. Thermal emission from Jupiter's atmosphere appears as a disk-shaped region in the center of the figure.

Figure 14 shows a tomographic map of the emission region, obtained by observing Jupiter at all longitudes and reconstructing a three-dimensional map. The emission is seen to be confined to the magnetic equatorial plane out to a distance of ~4 Jovian radii. Several intriguing features are visible. The main radiation peaks (L and R on

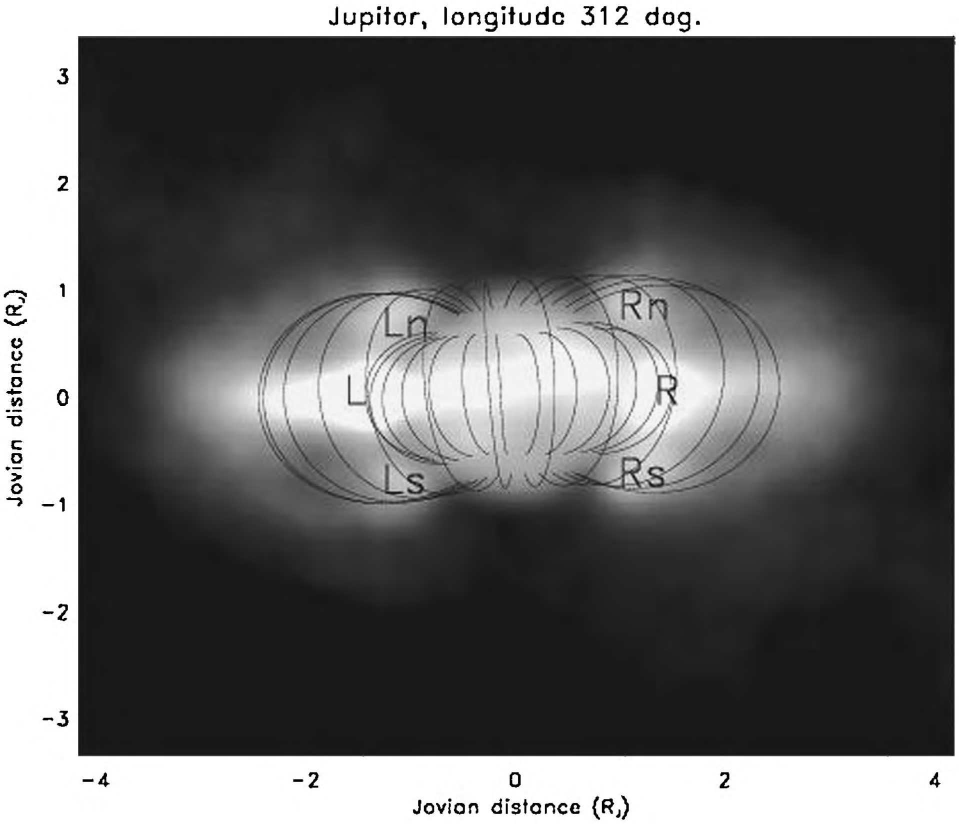


FIGURE 13 Radio photograph of Jupiter's synchrotron emission at a jovicentric longitude of 312°. The image was taken at a wavelength of 20 cm, using the Very Large Array in June 1994. [After de Pater, I., and Sault. (1998). *J. Geophys. Res. Planets* **103**(E9), 19,973–19,984.]

Fig. 13) are usually asymmetric. One of the peaks appears to be brighter than the other peak. The asymmetry of these main radiation peaks is caused by deviations in Jupiter's magnetic field from a pure dipole configuration. These deviations are evident in Fig. 14, where the main ring of radiation is clearly warped like the surface of a potato chip. If Jupiter's field were a dipole field, this ring would be flat, and the radiation peaks (Fig. 13) would always be equal in intensity, though the intensity would vary with jovian rotation, such that it would be smallest when one of the magnetic poles is directed toward us. The secondary emission peaks (Ln, Ls, Rs, Rn in Fig. 13) become high-latitude rings when viewed in a three-dimensional image. These peaks are produced by electrons at their mirror points, and they reveal the presence of a relatively large number of electrons which bounce up and down the field lines at a Jovian distance of 2–2.5 Jovian radii. It is believed that these electrons may have been scattered out of the magnetic equatorial plane, perhaps by the Moon Amalthea, which orbits Jupiter at a distance of 2.5 Jovian radii.

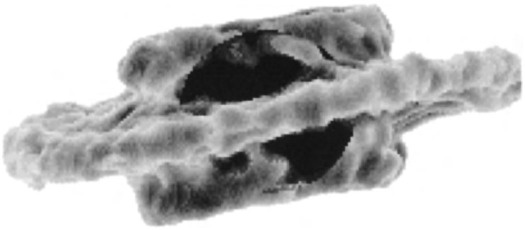


FIGURE 14 Three-dimensional tomographic reconstruction of Jupiter's nonthermal radio emissivity. The planet itself is shown as a black sphere in this visualization. [After de Pater, I., and Sault. (1998). *J. Geophys. Res. Planets* **103**(E9), 19,973–19,984.]

The total radio intensity of Jupiter varies significantly over time (years), and appears to be correlated with solar wind parameters. The impact of comet D/Shoemaker-Levy 9 with Jupiter in July 1994 caused a sudden sharp increase in Jupiter's total flux density, by ~20%. At the same time the brightness distribution of the radio flux density changed drastically. These observations suggest that the impact and associated phenomena significantly modified the electron distribution and possibly the magnetic field as well.

At frequencies below 40 MHz, Jupiter is a strong emitter of sporadic nonthermal radiation. A decameter-wavelength (DAM) component observable from the ground is characterized by complex, highly organized structure in the frequency–time domain and dependent on the observer's position relative to Jupiter. The satellite Io modulates the DAM emission. The Voyager spacecraft added significantly to our knowledge of this low-frequency component when it flew by Jupiter in 1979. A kilometric-wavelength (KOM) component was discovered at frequencies below 1 MHz and the observations of the DAM component were significantly improved. The extremely high brightness temperatures ($>10^{17}$ K), narrow-bandwidth emissions, and sporadic nature all suggest that the very low frequency emissions from Jupiter are generated by energetic particles acting coherently and interacting with the plasma that surrounds Jupiter. The details of the emission process are not well understood.

E. Saturn

Radio emission from Saturn has been observed from the Earth over a wavelength range from 1 mm to approximately 70 cm. The emission is thermal throughout this band, arising in both the atmosphere and the rings. The atmospheric emission is similar to that observed from Jupiter. Model studies indicate that Saturn, like Jupiter, has a deep convective atmosphere. Hydrogen and helium form the bulk of the atmosphere, whereas ammonia in trace amounts provides most of the microwave opacity. High-resolution radio images of Saturn also give information on the latitudinal distribution of NH_3 gas. Figure 15 shows a VLA image at 6-cm wavelength: the resolution is 1.5 arc sec and the disk diameter is 16.8 arc sec. A bright band can be distinguished at midlatitudes, indicative of an average lack of NH_3 gas over the altitude region probed at this wavelength. The region at midlatitudes is likely a region of subsiding gas, just like the bright belts seen on Jupiter. In the 1990s the zonal patterns on Saturn changed drastically compared to what was seen in the 1980s, indicative of strong dynamical interactions.

As shown in Figs. 15 and 16, interferometer observations of Saturn also detect the thermal emission from the ring particles. Most of the radio emission is due to scattering of Saturn's emission off the ring particles. In front of the planet, the rings are visible as an absorption feature; they block out Saturn's radio emission. The scattering characteristics of the rings contain information on the ring particle sizes and composition.

Interferometer observations of Saturn at centimeter and millimeter wavelengths have detected thermal emission from the ring particles. The rings have a low brightness temperature, approximately 10 K. The presence of ring particle sizes larger than a few centimeters is suggested by the radio observations. The observations are consistent with the bulk properties of the ring particles being those of water ice.

The Voyager spacecraft detected two distinct classes of nonthermal emissions from Saturn at frequencies below

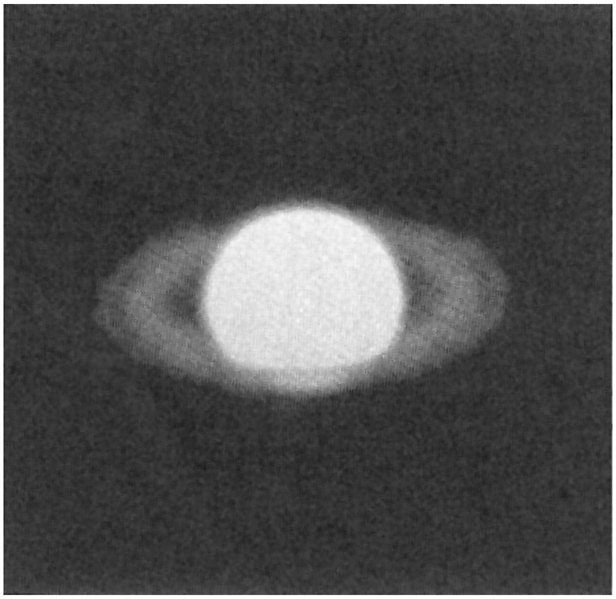


FIGURE 15 Radio photograph of Saturn at a wavelength of 6.14 cm obtained with the VLA. Resolution is 1.5 arc sec. Equatorial diameter of Saturn is 16.83 arc sec. [After de Pater, I., and Dickel, J. R. (1991). *Icarus* **94,** 474–492.]

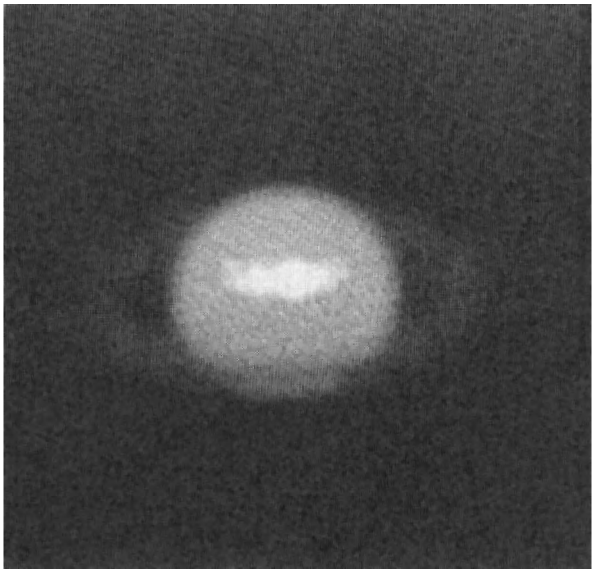


FIGURE 16 Radio photograph of Saturn at a wavelength of 2.0 cm obtained with the VLA. Resolution is 1.5 arc sec. Equatorial diameter of Saturn is 16.83 arc sec. [After de Pater, I., and Dickel, J. R. (1991). *Icarus* **94,** 474–492.]

1 MHz. These emissions are not observable from the Earth because of the opaqueness of the Earth's ionosphere at frequencies below a few megacycles. The first class, called Saturn kilometric radiation, is a relatively narrow band polarized emission. The second class is a broadband, impulsive emission called the Saturn electrostatic discharge.

F. Uranus

Uranus is unique among the planets in having its rotation axis tilted close to the plane of the ecliptic. The north pole of Uranus is inclined $\sim 98^\circ$ to the ecliptic plane (8° south of the plane), and the seasons on Uranus average 21 terrestrial years in length. The effect of this geometry on the large-scale circulation of the Uranium atmosphere is not fully understood, but it is expected to be significant.

As in the case of both Jupiter and Saturn, the disk brightness temperatures of Uranus significantly exceed the expected equilibrium temperature. The observed temperatures are greater than 100 K at wavelengths greater than a few millimeters and longward, whereas the predicted effective temperature is only about 55 K. The radio emission from Uranus is unpolarized within the measurement uncertainties of a few percent. Interferometer observations of Uranus show the emission to be confined to the solid angle of the visible disk, providing evidence that the excess emission is from the atmosphere and is not synchrotron emission. It is believed that the emission from Uranus is thermal, originating in the atmosphere of the planet.

The radio emissions from Uranus arise from sufficient depths that collision-induced absorption by hydrogen is an important source of opacity at millimeter wavelengths. Ammonia is severely depleted in Uranus' atmosphere, at least at pressure levels less than 25 bar. Since, based upon planet formation theories, nitrogen must be present in at least solar proportions, it is believed that ammonia gas is abundant at deeper levels, but reacts with H_2S to form a cloud of NH_4SH. If indeed this process accounts for the observed depletion in NH_3, hydrogen sulfide should be enriched in Uranus' atmosphere by about an order of magnitude over solar S. Such an abundance of H_2S itself will contribute to the radio opacity in Uranus' atmosphere and actually help reconcile observed spectra with models.

Another interesting aspect of the radio emission from Uranus is that its total intensity varies slowly over time. With the help of interferometric (VLA) observations, it is well established that the Uranus pole is warmer than its equator. This will certainly lead to time variations during Uranus' orbit around the Sun. It is not known, however, whether this indeed fully explains the observed time variability.

The Voyager spacecraft detected a wide variety of radio emissions from Uranus during its encounter in January 1986. Most of the emissions were polarized and probably due to maser-cyclotron emission. The emissions range in frequency from about 20 kHz to 800 kHz, well below the frequency range that is observable from the Earth. The emissions suggest a magnetosphere rich in magnetohydrodynamic phenomena.

G. Neptune

Disk brightness temperature measurements of Neptune at centimeter and millimeter wavelengths are sparse, but suggest a spectrum very similar to that of Uranus. The disk brightness temperatures exceed the predicted equilibrium temperature by 50 K or more. High-angular-resolution measurements obtained with the VLA show that the excess emission is not due to synchrotron emission. As in the case of Uranus, model studies suggest that ammonia must be depleted on Neptune (as on Uranus) by roughly two orders of magnitude compared to solar nitrogen values to explain the high brightness temperatures that are observed.

This depletion may be caused by a nearly complete removal of NH_3 gas in the upper atmosphere through the formation of NH_4SH, the same process which depletes ammonia in the Uranus atmosphere. On Neptune the H_2S abundance must be even larger than on Uranus, about 30 times the solar sulfur/hydrogen ratio. Although ammonia gas may be close to the solar N value in Uranus' deep atmosphere, it must be substantial subsolar throughout Neptune. Here a problem arises: A subsolar value for nitrogen gas is inconsistent with theories on planet formation. If a planet forms directly from the primordial solar nebula, the elemental abundances must be equal to that measured on the Sun. Condensable materials accrete as solids, sublime in the protoplanet's atmosphere and therefore enhance the elemental abundances above solar values. Carbon, sulfur, and oxygen compounds, present as CH_4, H_2S, and H_2O in the giant planets, are therefore enhanced above solar values. Similarly, nitrogen is expected to be enhanced above solar N. A subsolar value on Neptune is simply not possible. This dilemma was solved with the detection of emission lines in the 1- to 3-mm band of CO and HCN in Neptune's upper atmosphere.

Since both lines are seen in emission, CO and HCN must be located above the tropopause, in Neptune's stratosphere. Although CO may be brought up from deep depths through rapid convection, HCN must be formed in the stratosphere from nitrogen and carbon products, since HCN, if it existed in Neptune's deep atmosphere, would freeze out on its way up, well before it could reach the stratosphere. If, however, much of the nitrogen in Neptune's atmosphere were present in the form of N_2 rather than NH_3, the N_2 might convect rapidly upward and form HCN in the stratosphere through chemical reactions with carbon products. Moreover, if nitrogen existed in the form of N_2 rather than NH_3, the total nitrogen abundance in Neptune's atmosphere would be consistent with planet formation theories.

The Voyager spacecraft detected a variety of low-frequency radio emissions from Neptune during its encounter in 1989. The emissions were similar to those observed from the other major planets.

H. Comets

Prior to the 1960s, comets were investigated primarily using visible-wavelength observations. The first infrared detection of a comet was achieved in the mid-1960s. Although radio astronomers had attempted to detect cometary emissions since the 1950s, the first widely accepted radio detection of a comet is that of the 18-cm OH transitions in Comet Kohoutek (1973). Radio observations of comets have been used to study all of the major components of comets: nucleus, dust, neutral gas, and the plasma.

Continuum observations of thermal emission from comets provide information on the nucleus temperature and on the large dust grains which surround the nucleus. The thermal emission from comets is very weak, and sensitive receiving systems are needed to make a detection. Short centimeter, millimeter, and submillimeter observations have been the most successful for the detection of thermal emission. In the 1990s sensitive detectors at millimeter and submillimeter wavelengths came on line, just in time to observe two very bright comets: Comet Hyakutake in 1996 and Comet Hale-Bopp in 1997. Radio spectra at submillimeter to millimeter wavelengths were obtained for both comets. The thermal emission is clearly dominated by emission from dust grains rather than the cometary nucleus, and the data have been used to derive the dust mass and dust mass production rate for each comet. Continuum observations of comets will be greatly improved when the new large millimeter and submillimeter arrays become available and when thermal emission measurements from spacecraft are possible.

Radio astronomy of comets has made its most significant advances in the area of spectroscopy. Many so-called parent molecules (molecules which sublimate directly off the nucleus, in contrast to daughter molecules, which are products of the parents) have been observed at radio wavelengths through their rotation lines in the millimeter and submillimeter bands. Observations of these spectral lines give valuable information on the icy composition of the cometary nucleus, gas production rates, physical conditions in coma, and variation of these in time and heliocentric distance.

The timely appearance of the bright comets Hyakutake and Hale-Bopp greatly expanded our knowledge of cometary volatiles. The number of identified parent molecules increased by more than a factor of two through observations of these comets; most of these new species were detected in the millimeter and submillimeter regions. Spectroscopic observations of Comet Hale-Bopp, which could be followed in its orbit for several years, yielded invaluable data on composition and gas production rates. Gas production rates of different molecules could be followed over large heliocentric distances (Fig. 17). At distances greater than 3 AU, the sublimation was dominated by gases more volatile than water. Carbon monoxide, the most volatile species observed, could be detected at a heliocentric distance of 7 AU. It displayed a steady increase in brightness while the comet moved in toward the Sun. This gas clearly must have been key to the

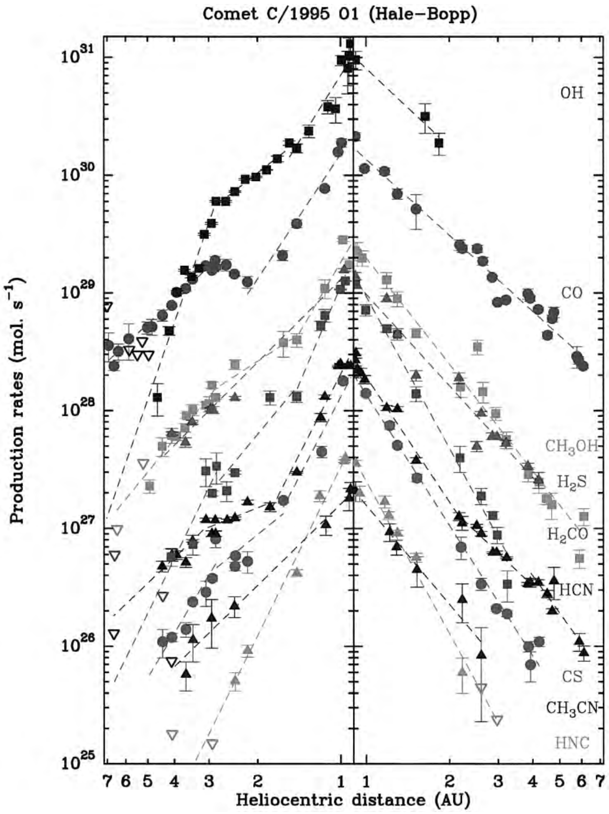


FIGURE 17 Evolution of molecular production rates of comet Hale-Bopp as a function of heliocentric distance. *Left*: Preperihelion data; *right*: postperihelion data. [After Biver, N., Bockelée-Morvan, D., Colom, P., Crovisier, J., Germain, B., Lellouch, E., Davies, J. K., Dent, W. R. F., Moreno, R., Paubert, G., Wink, J., Despois, D., Lis, D. C., Mehringer, D., Benford, D., Gardner, M., Phillips, T. G., Gunnarsson, M., Rickman, H., Bergman, P., Johansson, L. E. B., Winnberg, A., and Rauer, H. (1999). *Earth Moon, Planets* **78,** 5–11.]

observed visible brightness of the comet at these large heliocentric distances. At a heliocentric distance of 3 AU, OH became the most abundant species, confirming that the composition of the comet is dominated by water-ice ($H_2O \rightarrow OH + H$). Inside a heliocentric distance of 1.5 AU all species showed a dramatic increase in production rates. Radio images of various species have been obtained and have been used to derive the relative importance of sublimation directly from the nucleus or from grains in the coma. This varies from gas to gas, and probably differs from comet to comet.

As mentioned above, the first radio detection of a comet was that of the 18-cm OH line. The study of this line has been an important component of cometary research for over 25 years. The 18-cm line is split into four transitions at 1612, 1665, 1667, and 1721 MHz. The observed intensity of the OH line is quite variable. The line is seen in emission or absorption depending on the comet's heliocentric velocity. The explanation of this effect is that it is due to pumping of the OH molecules by solar UV photons. If the comet's heliocentric velocity is such that Fraunhofer absorption lines from the Sun shift into the excitation frequency, the OH molecules are not excited. At all other times they are. This variation in excitation is known as the Swings effect. When OH is not excited, it is seen in absorption against the galactic background. If it is excited, the emission is actually maser emission rather than simple thermal emission.

V. PROSPECTS FOR THE FUTURE

Planetary radio astronomical observations were initially limited to measurements of the disk-averaged brightness temperatures and to the strong nonthermal radio emission from Jupiter. With improved spatial resolution and high-sensitivity receiving systems, it has been possible to map the planets (and satellites and comets) and carry out high-resolution spectroscopic observations. With the rapidly improving instrumentation now on the horizon, eventually it should be possible to study the planets using ground-based and spacecraft instruments in nearly the same detail in which the Earth is being studied with combined sensors on weather satellites and ground stations. As angular resolution improves with time, it should be possible to map the vertical and horizontal distributions of certain chemical species within planetary atmospheres, measure wind speeds, and search for local variations in subsurface properties. Both ground-based radio telescopes and spacecraft radio receivers will play a role in future observations. Large ground-based telescopes will have sufficient resolution to examine the global properties of the planets out to Neptune and to study satellites, asteroids, and comets. The advent of lighter weight and lower power radio receivers will enhance the possibilities of placing radio experiments on spacecraft. Radio telescopes working at submillimeter wavelengths are gradually becoming a reality as a few mountain-top observatories are nearing completion and plans for a submillimeter space telescope are being discussed in a number of countries. The shorter wavelength region of the spectrum will provide new opportunities to study the upper atmospheres of the planets with spectroscopic techniques capable of detecting heretofore unobserved transitions that occur in the submillimeter and infrared regions. In addition, the shorter wavelengths will provide high angular resolution with only moderate-size telescopes. In summary, the future prospects for planetary radio astronomy are bright. This optimistic outlook is based on radio astronomical systems currently under development. The evolution of radio technology will undoubtedly lead to an even brighter future for planetary radio astronomy.

ACKNOWLEDGMENT

The portion of the research described in this publication was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

SEE ALSO THE FOLLOWING ARTICLES

Planetary Atmospheres • Radiation Sources • Radio-Astronomy Interferometry • Radiometry and Photometry • Solar System, General

BIBLIOGRAPHY

Berge, G. L., and Gulkis, S. (1976). "Earth-based radio observations of Jupiter: Millimeter to meter wavelengths," *In* "Jupiter" (T. Gehrels, ed.), University of Arizona Press, Tuscon, AZ.

de Pater, I. (1990). "Radio images of the planets," *Annu. Rev. Astron. Astrophys.* **28,** 347–399.

de Pater, I. (1999). "The solar system at radio wavelengths," *In* "Encyclopedia of the Solar System" (P. Weissman, L. McFadden, and T. V. Johnson, eds.), pp. 735–772, Academic Press, New York.

Dessler, A. J. (ed.). (1983). "Physics of the Jovian Magnetosphere," Cambridge University Press, Cambridge.

Janssen, M. A. (ed.). (1993). "Atmospheric Remote Sensing by Microwave Radiometry," Wiley, New York.

Lewis, J. S., and Prinn, R. G. (1984). "Planets and Their Atmospheres," Academic Press, Orlando, FL.

Morrison, D. (1970). "Thermophysics of the planet Mercury," *Space Sci. Rev.* **11,** 271–307.

Muhleman, D. O., Orton, G. S., and Berge, G. L. (1979). "A model of the Venus atmosphere from radio, radar, and occultation observations," *Astrophys. J.* **234,** 733–745.

Sullivan, W. T., III (ed.). (1984). "The Early Years of Radio Astronomy," Cambridge University Press, Cambridge.

Radio-Astronomy Interferometry

T. Joseph W. Lazio
Naval Research Laboratory

GLOSSARY

Closure phase Sum of the measured coherence phase around a triad of array elements; invariant to all effects that factorize per element, e.g., extra path lengths introduced into the signal path by the atmosphere.

Coherence Time-averaged product of the electric field at two different points in space. For a distant incoherent source the measured coherence is a Fourier transform of the source brightness distribution.

Correlator Computer for calculating the coherence between two antennas, usually realized with digital logic.

Deconvolution Correction for the effects of limited sampling of the coherence function. Accomplished by utilizing knowledge of the coherence sampling function and *a priori* information about the source brightness distribution, such as positivity and confinement.

Heterodyne Shift in the frequency of a spectrum accomplished by adding a carrier wave at an offset frequency, and then passing the signal into a nonlinear device.

Interferometer Device for measuring coherence of the electromagnetic field.

Self-calibration Calibration of interferometric array using the source of emission being imaged in addition to an external calibrator; relies on factorization of systematic effects by antenna.

INTERFEROMETERS are used widely in radio astronomy for high angular resolution imaging of celestial objects. An interferometer measures the time-averaged product or coherence of the electromagnetic field at two separated points. Each measurement corresponds to a sample of the Fourier transform of the brightness distribution of an object. From arrays of such radio interferometers and by exploiting the rotation of the Earth to change the aspect of a single interferometer, many such samples can be accumulated, and an image formed in a computer. Interferometers have been formed with antenna separations exceeding the Earth's diameter, providing angular resolutions of better than a milliarcsecond ($\sim$5 nanoradians), and recent developments enable ever shorter and ever longer wavelengths to be used.

I. THEORY OF INTERFEROMETRY

At optical wavelengths astronomical imaging is performed almost exclusively by monolithic telescopes, which either reflect or refract light to a focal plane. Diffraction limits the highest resolution obtained to approximately λ/D, where λ is the observing wavelength and D is the diameter of the aperture. At radio wavelengths ($\lambda \approx 0.001$–10 m) resolutions comparable to those obtained in the optical would require telescopes with aperture diameters of tens or hundreds of kilometers. In optical astronomy the desire for higher angular resolution led to the development of the Michelson interferometer, in which light beams from two (widely) separated mirrors are combined to form interference fringes. After the initial experiments, optical interferometry was not pursued until recently because of the numerous practical difficulties including stabilization of the light paths, atmospheric turbulence, and low signal levels. At radio wavelengths these problems are much simplified, especially at moderate resolution of an arcsecond (5 microradians) or more: stabilization of the signal path is possible for antenna separations up to a few tens of kilometers, the effects of the Earth's atmosphere may be removed through calibration, and photon-rich radio signals can be divided and amplified many times with a negligible impact on the signal-to-noise ratio, unlike typical optical signals which consist of relatively few photons and cannot be amplified coherently. Furthermore, radio frequencies are low enough that electronic equipment is capable of measuring the electric field itself and preserving phase information, in contrast to optical interferometers which generally are limited to measuring the second moment of the electric field (the intensity).

As a result radio interferometry has been developed vigorously over the past 40 years since Ryle and Vonberg built the first simple Michelson adding radio interferometer. Many variants of the Michelson interferometer were tried and discarded, though some turned out to be of importance in other fields. (For example, Hanbury Brown and Twiss developed an intensity interferometer in which interference fringes were formed from the electric field intensity. The utility of the intensity interferometer is limited by poor sensitivity, but it did contribute to the development of quantum optics, because intensity correlations implied the existence of photon bunching, then unknown.) A Michelson interferometer forms fringes by direct addition of the signals, but systematic errors are troublesome. For example, a strong, spatially extended background source (e.g., the Galactic radio emission) will magnify the effects of amplifier gain drifts.

Because the electric field can be measured directly, an alternative design—a multiplying interferometer—is usually preferred in order to reduce the contribution of uncorrelated components of the signal. A multiplying interferometer forms the electric field coherence or time-averaged product of the electric field, $E(\boldsymbol{r}, t)$, at two separate points in space and time,

$$\Gamma(\boldsymbol{r}_i, \boldsymbol{r}_j, t_i, t_j) = \left\langle E(\boldsymbol{r}_i, t_i) E^*(\boldsymbol{r}_j, t_j) \right\rangle_t$$

where $\langle \cdots \rangle_t$ denotes an average over time. Because the electric field can be sampled directly and digitized, the multiplication and averaging can be realized in a special-purpose digital computer.

Given the considerable spatial extent of most celestial objects, it might seem surprising that measurements of the electric field coherence at the Earth could yield any information about their structures. Indeed, most radio sources emit incoherent radiation: At the source $\Gamma = 0$ unless $\boldsymbol{r}_i = \boldsymbol{r}_j$ and $t_i = t_j$. Figure 1 illustrates how coherence arises in a distant radiation field. The two point sources P_1 and P_2 are incoherent. However, at the distant points Q_1 and Q_2, the electric field contains contributions from both P_1 and P_2 and is therefore partially coherent. The generalization of this idea leads to the van Cittert-Zernike theorem. A simple version of this theorem produces a Fourier transform relation between the sky brightness $B(x, y)$ and the measured coherence function

$$\Gamma(u, v) = \int B(x, y)\, e^{2\pi i(ux+vy)}\, dx\, dy,$$

where position on (strictly: tangent to) the celestial sphere is measured in radians in a Cartesian system (x, y) centered on the source and antenna separations are measured in wavelengths in a Cartesian system (u, v) aligned with (x, y). (More general versions of this expression include

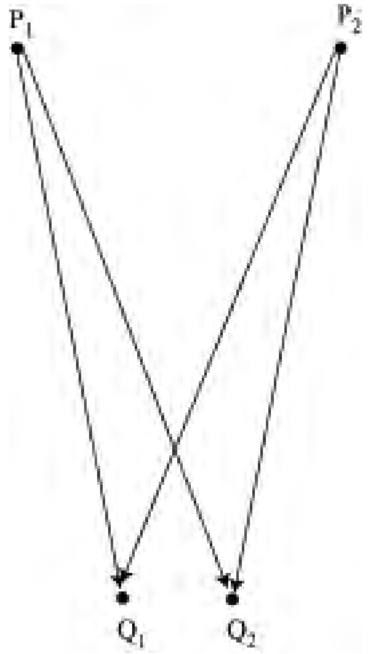


FIGURE 1 After propagation through space, a radiation field emitted by two mutually incoherent objects, P_1 and P_2, becomes partially coherent at two distant points, Q_1 and Q_2. The coherence function $\Gamma(Q_1, Q_2)$ contains information about the source of the radiation field.

temporal lags in Γ from which spectral information can be recovered.)

This Fourier relationship between brightness and coherence is central to interferometry. It has several important features. First the coherence measured between a given pair of points on the ground is a Fourier coefficient of the sky brightness, with a spatial frequency given by the separation of the points (in wavelengths) as seen from the object. Second, the coherence depends *only* on the separation of the two points, not their absolute position. Third, a physical brightness distribution must be real, implying that $\Gamma(\boldsymbol{r}_i, \boldsymbol{r}_j) = \Gamma(\boldsymbol{r}_j, \boldsymbol{r}_i)$ or $\Gamma(u, v) = \Gamma(-u, -v)$. These last two facts mean that $N(N-1)/2$ independent measurements of the coherence can be formed from samples of the radiation field from N antennas, which is important both for economy of measurement and for understanding systematic errors.

In principle, image formation by simple Fourier inversion of the measured $\Gamma(u, v)$ is possible only if the coherence is measured at all separations, u and v. In practice, limited sampling is acceptable provided that some correction is made in the processing. Furthermore, the Earth's rotation usually can be exploited to enhance the sampling produced by an array. As seen from the object being imaged, an array will rotate in the u-v plane, each pair of elements following an ellipse. Over a period of up to 12 hr, a large area in the u-v plane can be sampled without moving the antennas physically.

A unique aspect of interferometers is that the sampling of the u-v plane limits both the smallest *and largest* angular scales that can be detected. Like monolithic telescopes, interferometers have an angular resolution of approximately λ/D, where D is the largest separation between two antennas. Objects smaller than this scale appear to have an angular size of λ/D. Unlike monolithic telescopes, interferometers also have a largest angular scale, approximately λ/d where d is the smallest separation between the antennas. As the interferometer measures no spatial frequencies smaller than λ/d, any angular structures on these scales is largely invisible. Information on largest angular scales can be recovered by *mosaicing*. This technique incorporates the coherence across the individual antenna elements (which are necessarily smaller than the shortest antenna spacing!) into the measured Γ, but at the cost of requiring the celestial source to be observed multiple times at offset pointings.

II. DESIGN OF A RADIO-INTERFEROMETRIC ARRAY

The objective of a radio-interferometric array is to sample the coherence function over as much of the u-v plane as possible. At most wavelengths, the radiation is collected using parabolic reflectors which focus the radiation to a focal plane; typical antenna diameters are 5–100 m with 25 m being a popular compromise between construction cost and sensitivity. The ability to point the antennas and the surface accuracy limits the highest practical observing frequency. At wavelengths longer than about 50 cm, dipoles become favored because of the large antenna diameter that would be required to produce reasonable antenna gains.

Amplification of the radiation is vital at most wavelengths. Over the years receiver technology has advanced steadily to decrease the noise introduced at this stage and to increase the bandwidths feasible. Current state-of-the-art amplifiers use cryogenically cooled field-effect transistors (FETs) or high-electron-mobility transistors (HEMTs) for centimeter wavelengths, and either Schottky devices or SIS devices at millimeter wavelengths. At meter wavelengths, cryogenically cooled amplifiers are not required because the noise introduced by the sky typically dominates that introduced by the receiver.

Once the signals have been collected and amplified, they are relayed to a central location for estimation of the coherence. Often this is more easily done at low frequencies, so heterodyne techniques are used to mix the signal down to an intermediate frequency (IF) for transmission to the central location. For this conversion an accurate frequency standard is required at each antenna in order to maintain coherence throughout the array. This frequency standard may be either distributed from a central location (via cables, waveguides, radio links, or optical fibers) or derived from an extremely stable clock such as a hydrogen maser for widely separated antennas. The IF signal is then transmitted to the central location; for short distances the same route as the frequency standard can be used, whereas for long distances (>100 km) the signals are commonly digitally sampled and recorded on magnetic tapes which are then shipped to the central location for later processing. Bandwidths allowed by current technology exceed 100 MHz.

At the central location the signals must be multiplied and integrated (i.e., correlated) to form estimates of the coherence function. As digital correlators are preferred almost universally for their lower levels of systematic errors, the signals must be digitized before correlation. One-bit digitization has been popular for many years and leads to only a small loss in the signal-to-noise ratio. The advent of cheap, fast samplers has spurred more elaborate digitization schemes, such as three-level encoding, which usually are preferred for their lower loss of signal to noise. Before the correlation is performed, the signals must be synchronized to eliminate the continually changing geometric delay of one antenna relative to another due to the Earth's rotation. For digitized signals large buffers of high-speed memory are used to delay the signals.

Even after crude digitization the processing rates required in the correlation step far exceed those feasible even with the fastest supercomputers. This problem can be exacerbated by a need to correlate for a number of different temporal lags, which is necessary if either the geometry of the interferometry is poorly known or spectral information is desired. For typical sample rates of 10^7 samples/sec for each of 10–30 antennas and up to 512 different temporal lags, the required multiplication rates can approach 10^{14}/sec. Special-purpose hardware is vital, and much effort has been expended in this direction. Most correlators use custom VLSI chips for the crucial multiplier-accumulator unit. For arrays of many antennas, the conventional correlator design of many multipliers and accumulators, one per lag per antenna pair, is extremely inefficient. An FFT chip to transform the signals to the frequency domain before multiplication is often preferable as it saves in digital logical and can also allow elimination of interfering signals.

Earth rotation, while helpful in sampling regions of the u-v plane, complicates correlation. First, as a given pair of antennas traces out its ellipse in the u-v plane, fine structure in the coherence function will be smeared out unless the integration time is short ($\sim$10 sec). Second, the relative motion between the antennas introduces a differential Doppler shift in the received radiation, which must be cancelled. The measured coherence samples are usually averaged for as long as possible, within the limits posed by tolerable smearing of the coherence function or by uncertainties in the interferometry geometry.

The Earth's atmosphere can be a major source of error in the final coherence samples. At centimeter wavelengths the effect of the troposphere can be ignored for resolutions poorer than about 1 arcsec. For higher resolutions the averaged coherence samples must be calibrated by interleaving observations of an object of known strength and position nearby in the sky. In most cases even this procedure is not sufficient to allow high-quality imaging because the atmospheric variations will be partially decorrelated over even small antenna separations. The best remedy is then to "self-calibrate" on the source of interest itself (see Section III). At long wavelengths ($>$0.1 m) the ionosphere is the dominant cause of phase errors, but self-calibration can also remove these effects.

High-quality imaging requires good sampling of the coherence function. The largest separation of the antennas fixes the highest resolution possible, while the distribution of the samples over the u-v plane determines the complexity of the structure than can be imaged. As the information collected scales as N^2, it is desirable to have as many elements as possible. With a fixed budget one must include consideration of a number of factors including the number and diameter of antennas, cost and feasibility of the correlator, and ancillary computing costs. Practical constraints on the placement of antennas include the availability of land, the signal distribution system, and the mobility of antennas.

Table I summarizes salient details about radio-astronomical arrays currently in operation around the world.

III. IMAGING WITH RADIO-INTERFEROMETRIC ARRAYS

Once a correlator has produced the coherence samples, a Fourier inversion is required to produce an image. For this process modern computers have proven to be

TABLE I

Name	Location	Number of antennas	Maximum baseline (km)	Observing wavelength (cm)
Cambridge low-frequency synthesis telescope	Cambridge, UK	60	4.6	200
Very Large Array (VLA)	New Mexico, USA	27	35	400, 90, 21–18, 6, 3.6, 2, 1.3, 0.7
MERLIN	Jodrell Bank, UK	8	217	200, 75, 18, 6, 2
Westerbork Synthesis Radio Telescope	The Netherlands	14	2.7	120–65, 50, 40–25, 21, 18, 13, 6, 3.6
Australia Telescope Compact Array	Narrabri, Australia	6	6	20, 13, 6, 3
Molongo Synthesis Telescope	Australia	88	1.6	75
Caltech Millimeter Array	California, USA	6	0.48	0.3, 0.1
BIMA Millimeter Array	California, USA	10	2	0.3, 0.1
Plateau de Bure Interferometer	France	5	0.47	0.3, 0.1
Nobeyama Millimeter Array	Japan	6	0.35	0.3, 0.2, 0.1
Giant Metrewave Radio Telescope	India	30	25	200, 125, 100, 50, 20
Very Long Baseline Array (VLBA)	USA	10	8610	90, 50, 21–18, 13, 6, 3.6, 2, 1.3, 0.7, 0.3

revolutionary. Before electronic computers, man-weeks could be consumed turning a day's observations into a small image of the object. Besides facilitating the Fourier inversion, high-speed computers have driven the development of various algorithms to correct for defects in the coherence samples, such as poor coverage of the u-v plane and systematic errors due to the time-varying atmosphere above each antenna. Figure 2 illustrates the advances in fidelity and dynamic range in radio-interferometric imaging over the past 30 years, using images of the powerful radio galaxy Cygnus A. As can be seen, imaging has progressed from the first crude structural determination in radio interferometry (Fig. 2a), really little more than simple models, to being able to produce high-resolution images capable of distinguishing fine-scale filamentary structure within the lobes and a "jet" tracing the supply of energy from the central core to one of the lobes (Fig. 2d). While the 1974 image (Fig. 2c) was made from well-sampled coherence measurements processed by simple Fourier inversion, the 1984 image (Fig. 2d) relied on two new techniques: (1) a nonlinear deconvolution algorithm, the maximum entropy method, was used to correct for the limited sampling of the u-v plane, and (2) a self-calibration algorithm corrected for the effects of the Earth's atmosphere on the measured coherence functions.

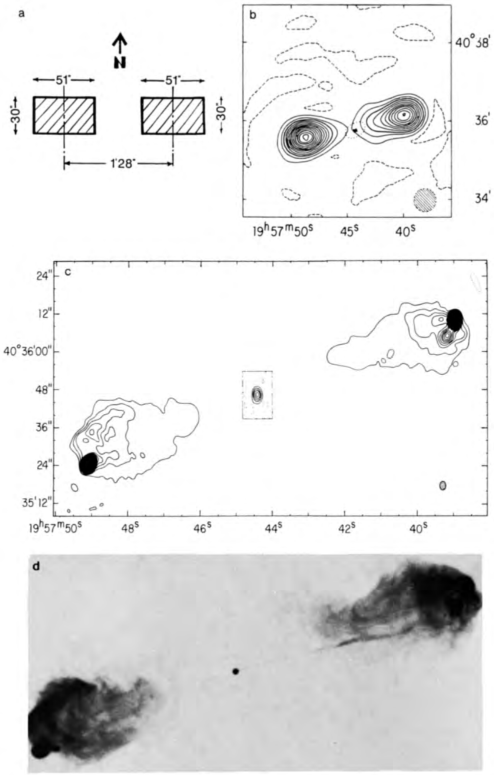
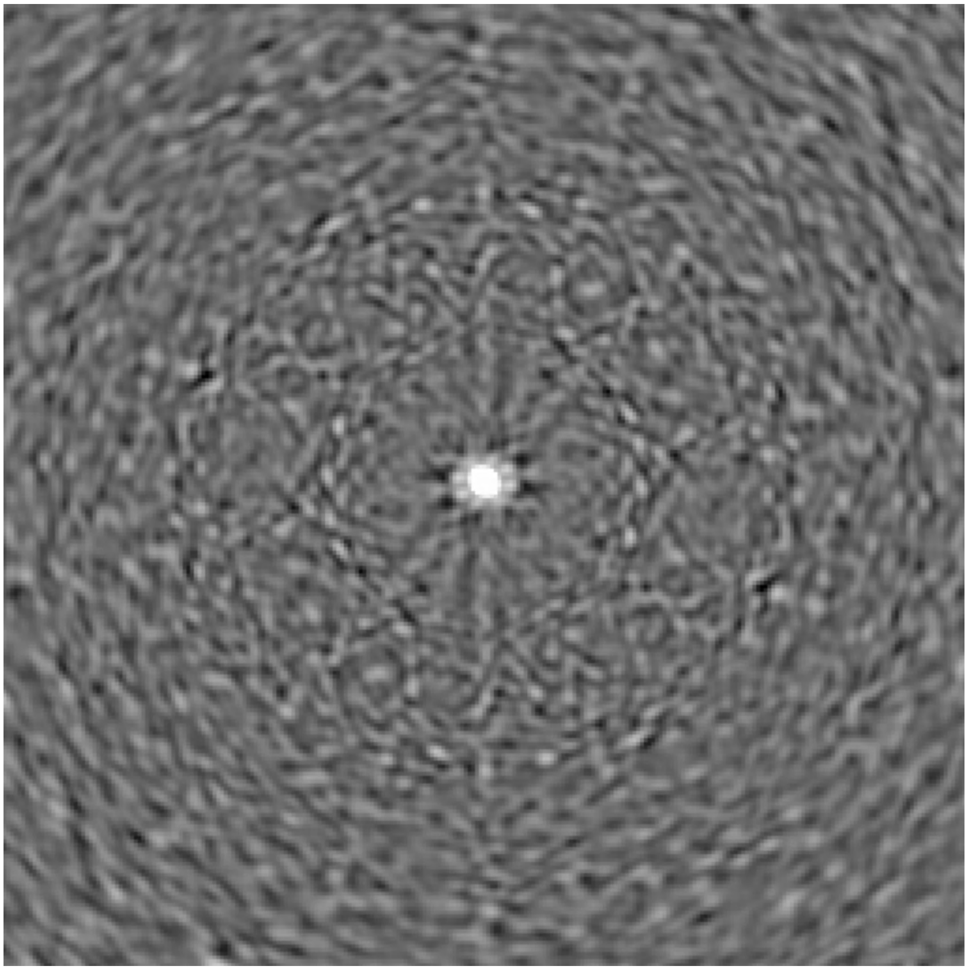


FIGURE 2 The evolution of imaging capability as demonstrated on the strong radio galaxy Cygnus A. (a) Jennison and Das Gupta (1953) first demonstrated double structure. (b) Ryle, Elsmore, and Neville (1965) showed that the radio "lobes" drop off in brightness toward the center of the radio galaxy. (c) Hargrave and Ryle (1974) resolved two "hot spots" of bright emission in the lobes, and also detected emission from the central galaxy. (d) Perley, Dreher, and Cowan (1984) found a jet of emission linking the core and the lobes and also found filaments in the lobes. The energy source at the center of the galaxy is likely to be a supermassive black hole accreting material from its surroundings. [(a) Reprinted by permission from *Nature* (*London*) **172,** 996; (b) *Nature* (*London*) **205,** 1259. Copyright ©1953 and 1965, respectively, Macmillan Magazines Ltd. (c) Reprinted by permission from *Mon. Not. R. Astron. Soc.* **166,** 305 (1974). (d) Reprinted by permission from *Astrophys. J.* **285,** L35 (1984).]

Deconvolution algorithms rely on *a priori* information about the sky brightness to interpolate into portions of the u-v plane that were not sampled. The two most common methods are CLEAN and maximum entropy method (MEM). The former assumes that the sky is mostly dark and that sources can be described as a combination of point sources, while the latter assumes positivity for the sky brightness of sources. Figure 3 illustrates deconvolution with the CLEAN algorithm.

Self-calibration also relies on *a priori* information about the sky brightness, but it mainly exploits the fact that the number of systematic errors introduced at each antenna (N) is usually much less than the number of sampled coherences ($\sim N^2$). The errors can be atmospheric in origin, but any source of error that can be considered to affect each antenna separately can also be removed. Thus, self-calibration can also correct for slow frequency drifts in the frequency standard used at each antenna and has thereby allowed true imaging in very-long-baseline interferometry (VLBI), in which the frequency standards are completely independent from antenna to antenna.

Self-calibration is predicated on the concept of phase closure developed by Jennison in the earliest days of radio interferometry. The observed coherence on the i-j baseline, V_{ij}, is assumed to be related to the "true" visibility as $V_{ij} = g_i g_j{}^* \Gamma_{ij}$ where g_i is the complex antenna gain for the ith antenna. The closure phase is the sum of the observed coherence phases around a triad of array elements. Any phase effect that is specific to a specific element occurs twice in the sum but with opposite sign and vanishes, and the closure phase is therefore invariant to antenna-based phase errors. An analogous closure amplitude that is invariant to gain amplitude errors exists for a quartet of antennas. Self-calibration is closely related to the use of adaptive optics systems in optical astronomy to counteract the effects of atmospheric turbulence. [Analogous concepts have also been developed by Hauptmann in X-ray crystallography (structure invariants) and by Weigelt in optical imaging (speckle masks).] Often multiple iterations of self-calibration and deconvolution are performed, in order to correct for phase and/or gain amplitude errors



(a)

FIGURE 3 Deconvolution and self-calibration of an image of the radio galaxy 3C 390.3 at 327 MHz: (a) point-source response; (b) dirty image corrupted by effects of limited sampling of coherence function; (c) CLEAN deconvolved image showing residual phase errors; and (d) the image after a single iteration of self-calibration. Additional self-calibration could be used to reduce further the effects of residual phase errors. All images are shown as negatives, with black being regions of most intense radio emission and white being regions of less intense radio emission, with transfer functions that are linear between black and white. For images shown in (b), (c), and (d) the intensity levels corresponding to black and white are held fixed. In order to emphasize low-level intensities, the images are saturated.

on shorter and shorter time scales. Figure 3 illustrates the additional improvement obtained by combining deconvolution and self-calibration.

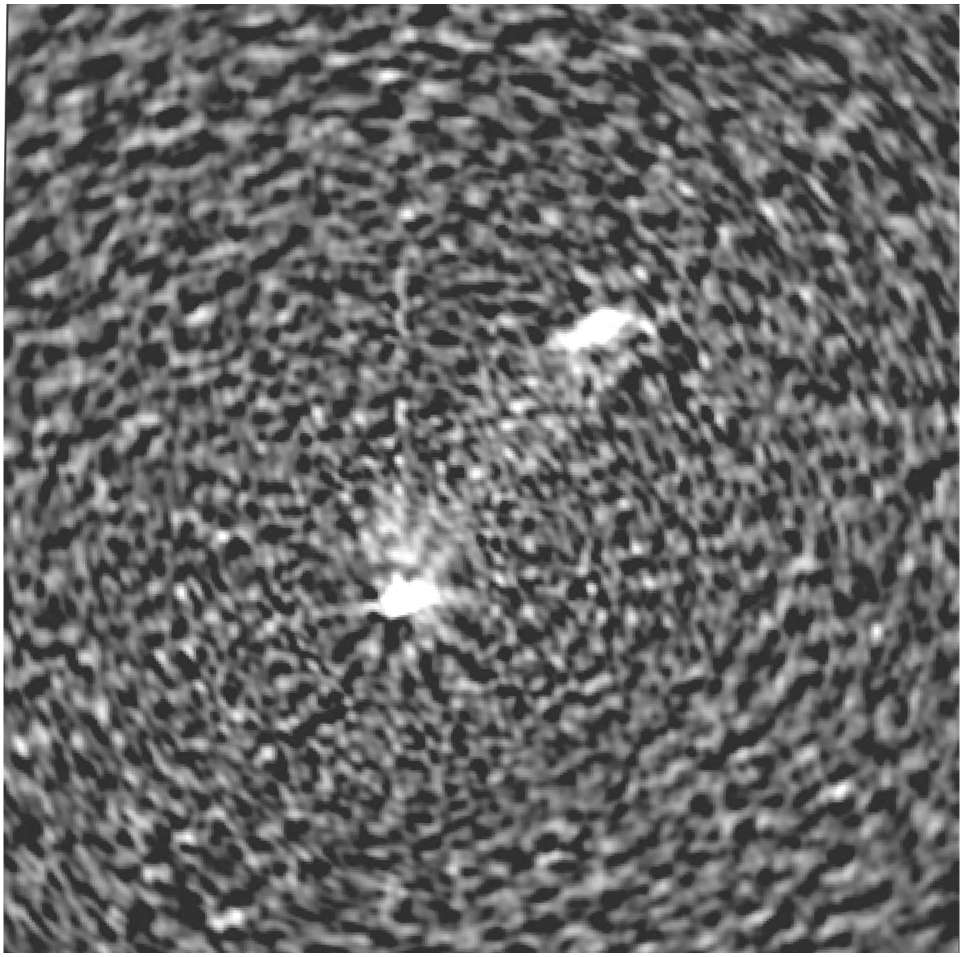
The net effect of these improvements has been to augment considerably the capabilities of radiointerferometric arrays at the cost of increased computing requirements.

IV. FUTURE PROSPECTS

The driving forces behind radio interferometry for the past 50 years have been the quest for higher sensitivity and angular resolution. These principles will continue to drive radio interferometry over the next decade.

A. Sensitivity, Radio Frequency Interference, and Imaging Algorithms

Improved sensitivity can be achieved either by improving receiver characteristics or by increasing the number of interferometer elements (total collecting area of the array). Receiver characteristics that can be improved include lowering the noise introduced by the amplifiers or increasing the bandwidth accepted by the receivers. Unfortunately, receiver noise is often not the limiting factor at short or long wavelengths and has been decreased generally to an irreducible physical minimum at intermediate wavelengths. Increasing the receiver bandwidth can improve the sensitivity to continuum sources of emission



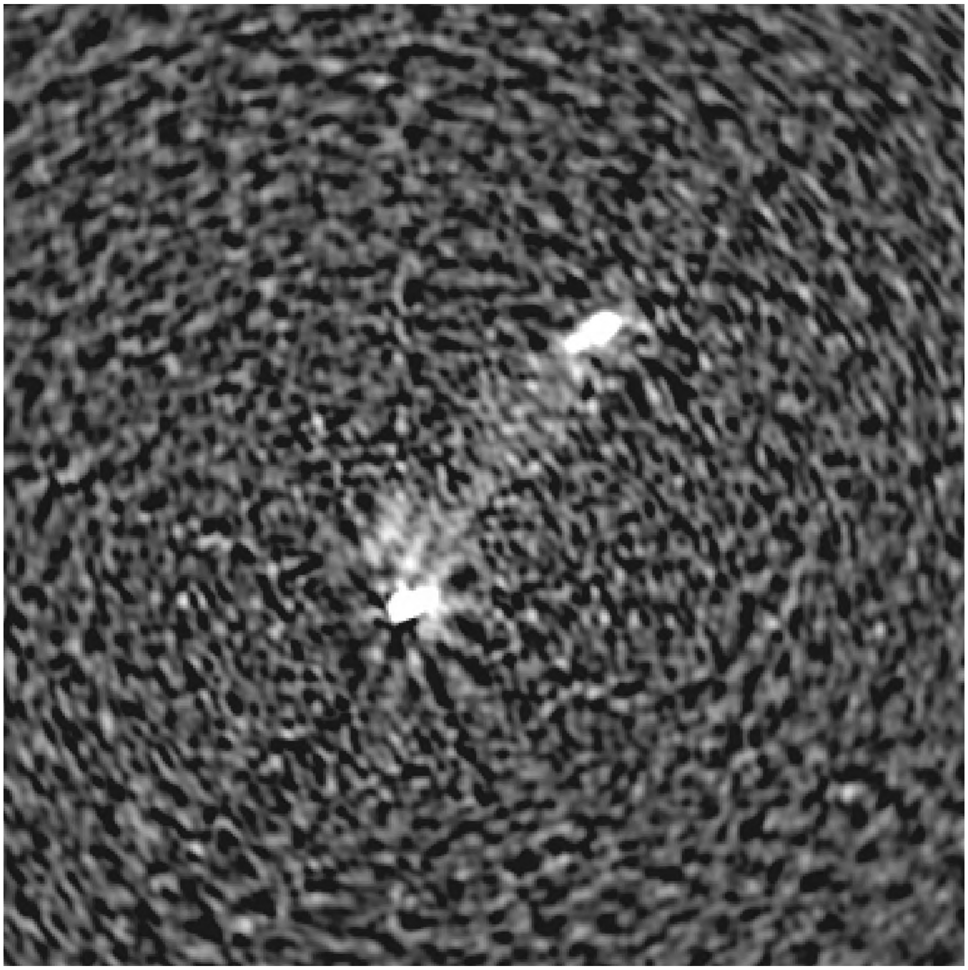
(b)

FIGURE 3 (*Continued*)

but does not improve the sensitivity for sources of line emission.

Efforts to increase receiver bandwidths must be coupled with strategies for mitigating radio frequency interference (RFI). Radio astronomy (and other passive users of the radio spectrum) has only a limited portion of the spectrum reserved exclusively for its use. Observations outside of these reserved spectral regions can be required either in an effort to obtain a wider continuum bandwidth or because an astrophysically interesting atomic or molecular transition line has been Doppler shifted by the motion of the emitting source. Previous RFI mitigation techniques have relied on the remote siting of interferometers and the use of filters. The advent of widespread satellite communications and wide-bandwidth, spread-spectrum broadcasts require additional measures. In addition to increased use of regulatory avenues, measures being investigated by radio astronomers include adaptive nulling. Various techniques fall under this rubric, but all make use of tracking the source(s) of RFI so as to remove its effects. For instance, the power reception pattern (beam) of an array could be altered continuously so that a null in the sidelobe pattern is directed toward the RFI source.

The alternate possibility for increasing the sensitivity of an interferometer is to increase the number of collecting elements and thereby the total collecting area. Doing so improves the array's sensitivity to both continuum and line sources of emission and is at least part of the motivation behind proposed new or expanded arrays across the entire radio spectrum. However, increasing the number of array elements also increases the complexity of the central correlator which scales as N^2. Proposals for arrays containing 100 or more elements, compared to a typical number of 10–30 elements in present-day arrays, have been made. Perhaps the most ambitious of various proposals is that for the Square Kilometer Array (SKA) which would have 10^6 m^2 of collecting area and operate at least over the wavelength range 1 cm–1 m. Key to the SKA's ultimate success will be whether such a large amount of collecting area can be constructed cheaply. A variety of



(c)

FIGURE 3 (*Continued*)

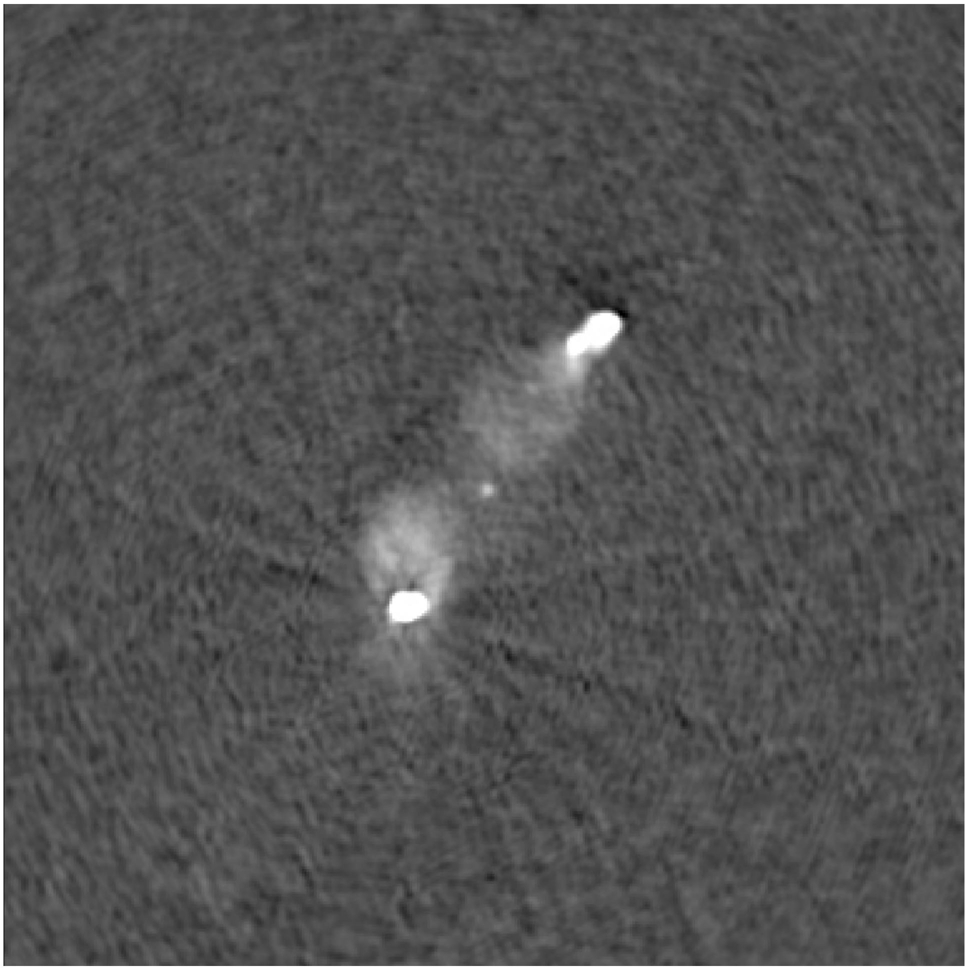
novel possibilities have been suggested to collect and focus the radiation, many of which forego traditional parabolic antennas (e.g., Luneberg lenses).

Fully exploiting the capabilities of new interferometer arrays will also require the development of new imaging algorithms, particularly deconvolution algorithms. Even from existing arrays, the images produced can be limited as much by errors in existing deconvolution algorithms as by residual (self-)calibration errors. Some recent progress has been the development of a nonnegative least squares (NNLS) deconvolution algorithm, in which a series of linear equations are inverted with constraints on the positivity of the solutions applied.

B. Millimeter Interferometry

At meter and centimeter wavelengths the primary emission mechanisms of celestial sources are either synchrotron or thermal bremsstrahlung, both continuum processes, with the 21-cm line of neutral hydrogen and the transitions from the OH^- radical near 18 cm serving as important exceptions. At wavelengths shorter than about 1 cm, rotational and vibrational transitions of various molecules become alternate, and often dominant, emission processes. Thus, in addition to the possibility of higher resolution than at centimeter and meter wavelengths, there is the possibility of three-dimensional structure information—via the Doppler shift of molecular lines observed from a source—and chemical information.

Challenges posed by millimeter interferometers include both the mechanical stability and tolerances of the antennas and the effect of the troposphere. Like antennas at any other wavelength, the surfaces of the antennas must be smooth enough that most of the incident radiation is reflected to the focus rather than scattered. Mechanical tolerances are particularly demanding given that the typical millimeter-wave antenna is only a factor of 2–3 smaller in diameter than a centimeter-wavelength antenna,



(d)

FIGURE 3 (*Continued*)

but the observation wavelength is an order of magnitude smaller.

The troposphere introduces two primary difficulties—phase fluctuations and opacity—both of which result primarily from atmospheric H_2O and O_2 molecules. Phase fluctuations can be mitigated by self-calibration, as at other wavelengths, or by fast calibration switching. The typical initial phase calibration scheme described above (§II) involves interleaved observations of a calibrator and a target source. If the calibrator can be observed frequently enough to track the tropospheric phase fluctuations, this fast-switching calibration scheme will remove most of the phase fluctuations. At millimeter wavelengths the troposphere can vary on time scales as short as 10 sec, requiring source switching to be on comparable time scales as well. Thus, the individual antennas must be mechanically stable enough to be driven from source to source this quickly without exciting any resonances, and motions must damp quickly enough so that the antennas can point accurately.

The opacity of the troposphere at millimeter wavelengths is unlike the situation at longer wavelengths (or optical wavelengths) in that it both attenuates celestial signals and increases the system noise. The latter effect occurs because the atmosphere contributes an effective noise temperature of $T_{atm}(1 - \exp[-\tau_{atm}])$ with $\tau_{atm} \sim 0.1$–0.5. One technique for reducing the effects of the troposphere is to place millimeter arrays at as high an altitude as possible so as to be above as much of the atmospheric water vapor as possible.

A crucial imaging problem for millimeter interferometers is that the sizes of the celestial sources being imaged are commonly larger than the smallest spatial frequency they sample (λ/d). Routine use of the mosaicing technique will be required to recover full information about the coherence function.

The most ambitious plan for millimeter interferometry is the Atacama Large Millimeter Array (ALMA, Fig. 4a), a US-European (and potentially Japanese) collaboration

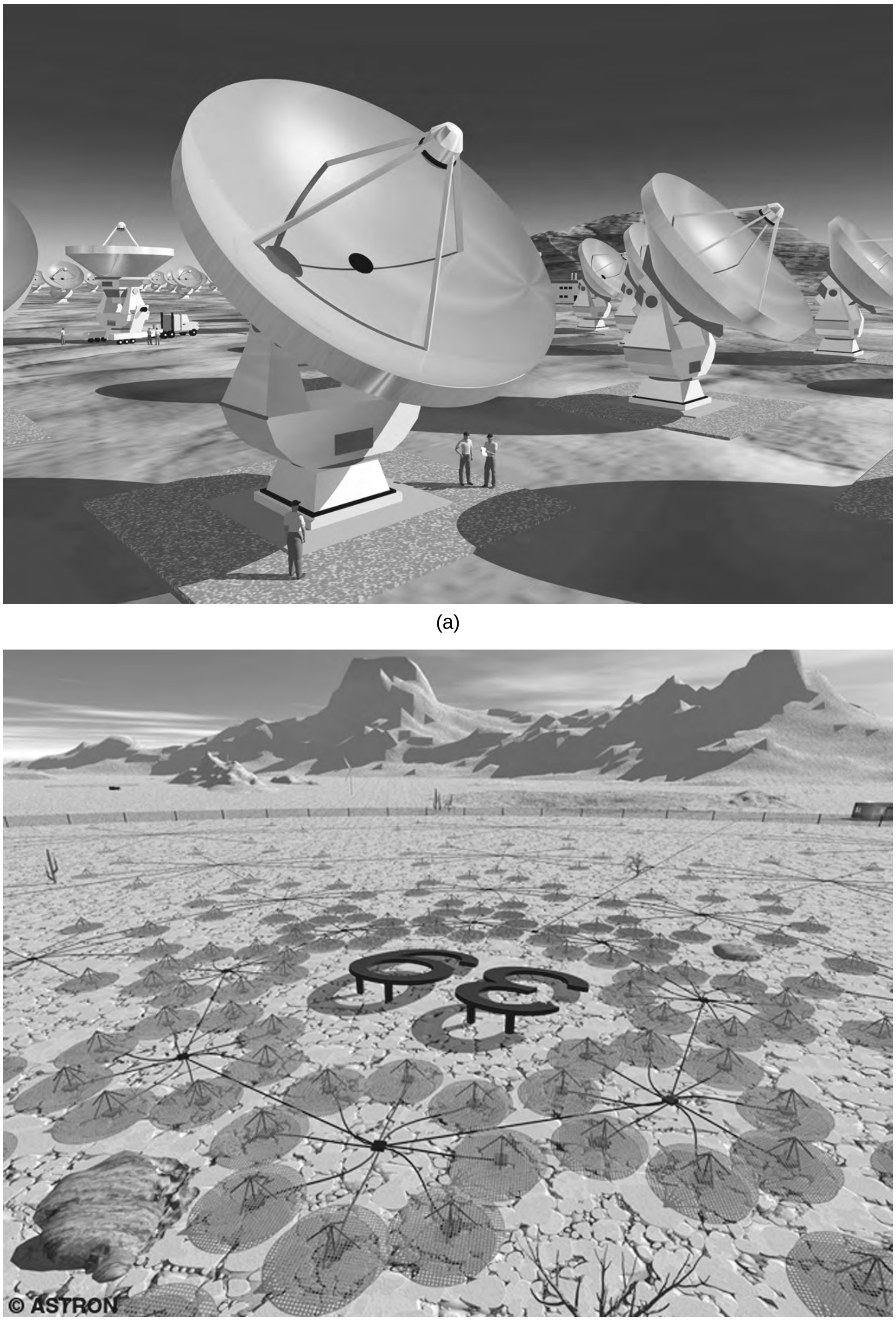


FIGURE 4 Future interferometers. (A) An artist's conception of the Atacama Large Millimeter Array (ALMA) proposed to be constructed on the (5000-m altitude) Chilean Atacama desert. The array will consist of at least Sixty four 12-m antennas operating between 22 and 350 GHz with the elements capable of being moved to produce baselines ranging from 15 m to 10 km. The highest angular resolutions will range from 0.3 arcsec at 22 GHz to 0.02 arcsec at 350 GHz. Image courtesy of the European Southern Observatory. (B) An artist's conception of a Low Frequency Array (LOFAR) "station." The array will consist of at least 40 stations of phased-array dipoles with baselines approaching 500 km and operating between 15 and 200 MHz. The highest angular resolutions will range from 8 arcsec at 15 MHz to 0.6 arcseconds at 200 MHz. Image courtesy of the Netherlands Foundation for Research in Astronomy (NFRA/ASTRON).

that aims to construct a millimeter interferometer in the Chilean Atacama desert at 5000 m altitude. ALMA will consist of at least sixty four 12-m antennas arrayed in movable configurations with baseline lengths between 150 m and 10 km and operating over the wavelength range 0.35–10 mm.

C. Long-Wavelength Interferometry

Radio astronomy was discovered at 20 MHz, but the quest for higher angular resolution with the single-dish telescopes then available, led quickly to most observations being conducted at higher frequencies. Interferometers were constructed to operate at frequencies below 100 MHz, but they had short baselines (<5 km). Short baselines were motivated by the belief that ionospherically induced phase fluctuations would destroy coherence on longer baselines.

The short baselines, and correspondingly limited angular resolution, also contributed to limited sensitivity. Observed at low angular resolution, sources overlap leading to *source confusion* and sensitivities far worse than expected from receiver noise alone. A crucial assumption of a two-dimensional Fourier relation between the observed coherences and the sky brightness is that the interferometer's extent along the line of sight to the source is negligible. The large field of view of meter and decameter instruments ($>1^\circ$) invalidates this assumption. Approaches to dealing with the large field of view include performing the full three-dimensional Fourier transform or tessellating the sky into small facets over which the two-dimensional assumption is appropriate. Both approaches are computationally intensive. Most work has focused on the latter approach because, if observed at sufficiently high resolution, the sky remains mostly dark at these wavelengths, and the entire field of view does not need to be tessellated: Facets need to be placed only on sources. This approach is also amenable to parallel computer algorithms as the Fourier transforms required to produce the facets can be done independently of each other.

The development of self-calibration algorithms in the 1980s led to the realization that they could be employed to remove ionospheric phase fluctuations. Subsequently, interferometer baselines have been extended to tens of kilometers, and the success of these relatively long baseline interferometers (particularly the 74 MHz system on the Very Large Array, cf. Table I) have spurred plans for a Low Frequency Array (LOFAR, Fig. 4B). This instrument will have baselines of order 100 km and a collecting area of order 10^6 m^2 at 15 MHz, leading to a sensitivity that will be competitive with centimeter-wavelength interferometers.

D. Space-Based Very-Long-Baseline Interferometry

The diameter of the Earth presents a limit in improving the resolution of centimeter- and meter-wavelength interferometers. In order to obtain resolutions better than 0.02λ arcsec, with λ in meters, requires placing one or more antennas in space.

The first successful demonstration of space-based interferometry used a NASA Tracking and Data Relay Satellite System (TDRSS) satellite. Although not designed for observing celestial objects, the TDRSS antenna, when combined with ground-based antennas, nevertheless resulted in the first interferometer that included baselines larger than the Earth's diameter. The first dedicated space VLBI antenna was onboard the VLBI Space Observing Programme (VSOP) satellite (later renamed to the Highly Advanced Laboratory for Communications and Astronomy, HALCA).

Space-based antennas present a number of difficulties in forming an interferometer. The high launch cost per kilogram of payload means that antennas must be either fairly small or made of extremely lightweight materials. To date, space-based antennas have been small, requiring large ground-based antennas to be used in order to have adequate sensitivity. In contrast to well-known positions and stable clocks of ground-based antennas, space-based antennas have poorly known locations and clock signals transmitted by uplinks subject to atmospheric fluctuations. These differences entail considerable effort to find and maintain coherence on the baselines to the space-based antenna. Furthermore, although long, ground–space baselines are often fairly similar. Rather than covering a wide range of (instantaneous) u-v values, as is typical of ground–ground baselines, ground–space baselines are concentrated in a narrow range at any given instant.

One compensating factor, at least for antennas in low-Earth orbit, is that long tracks in the u-v plane can be obtained in relatively short intervals due to the rapid satellite motion. There is also a natural limit to the length of the baselines that can be used. Plasma density fluctuations scatter radio waves propagating through the interplanetary and interstellar media causing point sources seen through these media to be angularly broadened, not unlike the broadening of stellar images at optical wavelengths because of neutral atmospheric turbulence ("seeing"). The magnitude of the broadening is both direction and frequency dependent, but roughly a baseline of D Earth radii will be limited by interstellar scattering at wavelengths below 1 m$/D$.

The success of HALCA has stimulated more ambitious plans, though all still contemplate placing only one antenna in space. VSOP2 would be a follow-on mission to HALCA with ten times its sensitivity and operating at frequencies as high as 43 GHz. RadioASTRON is a long-delayed Soviet/Russian radio telescope slightly larger than HALCA and operating over a larger wavelength range. ARISE is a proposed NASA mission that would place a 25 m diameter antenna in orbit with operating frequencies as high as 86 GHz.

ACKNOWLEDGMENTS

I thank T. Cornwell, R. Ekers, and K. Weiler for helpful comments. Basic research in radio astronomy at the NRL is supported by the Office of Naval Research.

SEE ALSO THE FOLLOWING ARTICLES

ASTROPHYSICS • FIELD-EFFECT TRANSISTORS • GRAVITATIONAL WAVE DETECTORS • IMAGE RESTORATION, MAXIMUM ENTROPY METHOD • MILLIMETER ASTRONOMY • RADIOCARBON DATING • SIGNAL PROCESSING, DIGITAL

BIBLIOGRAPHY

Hanbury Brown, R. (1974). “The Intensity Interferometer: Its Application to Astronomy,” Taylor and Francis, London.

Kellerman, K. I., and Thompson, A. R. (1988). “The very long baseline array,” *Sci. Am.* **258,** 54–64.

Napier, P. J., Thompson, R. T., and Ekers, R. D. (1983). “The very large array: Design and performance of a modern synthesis radio telescope,” *Proc. IEEE* **71,** 1295–1320.

Sullivan, W. T. (1982). “Classics in Radio Astronomy,” Reidel Publ., Boston, MA.

Taylor, G. B., Carilli, C. L., and Perley, R. A. (1999). “Synthesis Imaging in Radio Astronomy II,” Astron. Soc. of the Pacific, San Francisco, CA.

Thompson, A. R., Moran, J. M., and Swenson, G. W. (1986). “Interferometry and Synthesis in Radio Astronomy,” Wiley (Interscience), New York.

Vershuur, G. (1987). “The Invisible Universe Revealed—The Story of Radio Astronomy,” Springer-Verlag, Berlin.

Telescopes, Optical

L. D. Barr
National Optical Astronomy Observatories (retired)

GLOSSARY

Airy disc Central portion of the diffracted image formed by a circular aperture. Contains 84% of the total energy in the diffracted image formed by an unobstructed aperture. Angular diameter $= 2.44\,\lambda/D$, where λ is wavelength and D is the unobstructed aperture diameter. First determined by G. B. Airy in 1835.

Aperture stop Physical element, usually circular, that limits the light bundle or cone of radiation that an optical system will accept on-axis from the object.

Coherency Condition existing between two beams of light when their fluctuations are closely correlated.

Diameter-to-thickness ratio Diameter of the mirror divided by its thickness. Term is generally used to denote the relative stiffness of a mirror blank: 6:1 is considered stiff, and greater than 15:1 is regarded as flexible.

Diamond turning Precision-machining process used to shape surfaces in a manner similar to lathe turning. Material is removed from the surface with a shaped diamond tool, hence the name. Accuracies to one microinch ($\frac{1}{40}$th μm) are achievable. Size is limited by the machine, currently about 2 m.

Diffraction-limited Term applied to a telescope when the size of the Airy disc formed by the telescope exceeds the limit of seeing imposed by the atmosphere or the apparent size of the object itself.

Effective focal length (EFL) Product of the aperture diameter and the focal ratio of the converging light beam at the focal position. For a single optic, the effective focal length and the focal length are the same.

Electromagnetic radiation Energy emitted by matter with wavelike characteristics over a range of frequencies known as the electromagnetic spectrum. The shortest waves are gamma rays (<1 Å) and the longest are radio waves (>40 μm). Visible light is in the intermediate range.

Field of view Widest angular span measured on the sky that can be imaged distinctly by the optics.

Focal ratio (f/ratio) In a converging light beam, the reciprocal of the convergence angle expressed in radians. The focal length of the focusing optic divided by its aperture size, usually its diameter.

Image quality Apparent central core size of the observed image, often expressed as an angular image diameter that contains a given percentage of the available energy. Sometimes taken to be the full width at half maximum (FWHM) value of the intensity versus angular radius function. A complete definition of image quality would include measures of all image distortions present, not just its size, but this is frequently difficult to do, hence the approximations.

Infrared For purposes of this article, wavelength region from about 0.8 to 40 μm.

Optical path distance Distance traveled by light passing through an optical system between two points along the optical path.

Seeing Measure of disturbance in the image seen through the atmosphere. Ordinarily expressed as the angular size, in arc-seconds, of a point source (a distant star) seen through the atmosphere, that is, the angular size of the blurred source. Seeing disturbances arise from air density fluctuations due to temperature variations along the line of sight.

Ultraviolet For purposes of ground-based telescopes, the wavelength region from about 3000 to 4000 Å. The remaining UV region down to about 100 Å is blocked by the earth's atmosphere.

OPTICAL TELESCOPES were devised by European spectacle makers around the year 1608. Within two years, Galileo's prominent usage of the telescope marked the beginning of a new era for astronomy and a proliferation of increasingly powerful telescopes that continues unabated today. Because it extends what the human eye can see, the optical telescope in its most restricted sense is an artificial eye. However, telescopes are not subject to the size limitation, wavelength sensitivities, or storage capabilities of the human eye and have been extended vastly beyond what even the most sensitive eye can accomplish. In this article, the word light is used to mean all electromagnetic radiation collected by telescopes but the properties of astronomical telescopes operating on the ground in the optical/IR spectral wavelength range from 3000 to about 40,000 Å (0.3–40 μm) are the principal subject. The atmosphere transmits radiation throughout much of this range. Telescopes designed for shorter wavelengths are either UV or X-ray telescopes and must operate in space. The Hubble Space Telescope is the best-known example of a UV (and optical) telescope and is discussed in Section VII. Telescopes operating at wavelengths longer than 40 μm are in the radio-telescope category and are not considered in this article. Emphasis is placed on technical aspects of present-day telescopes rather than history.

I. TELESCOPE SIZE CONSIDERATIONS AND LIGHT-GATHERING POWER

An astronomical telescope works by capturing a sample of light emitted or reflected from a distant source and then converging that light by means of optical elements into an image resembling the original source, but appropriately sized to fit onto a light-sensitive detector (e.g., the human eye, a photographic plate, or a phototube). Figure 1 illustrates the basic telescope elements. It is customary to assume that light from a distant object on the optical axis arrives as a beam of parallel rays sufficiently large to fill the telescope entrance, as shown.

The primary light collector can be a lens, as in Fig. 1, or a curved mirror, in which case the light would be shown arriving from the opposite direction and converging after reflection. The auxiliary optics may take the form of eyepieces or additional lenses and mirrors designed to correct the image or modify the light beam. The nature and arrangement of the optical elements set limits on how efficiently the light is preserved and how faithfully the image resembles the source, both being issues of prime concern for telescope designers.

The sampled light may have traveled at light speed for a short time or for billions of years after leaving the source, which makes the telescope a unique tool for studying how the universe was in both the recent and the distant past. Images may be studied to reveal what the light source looked like, its chemistry, location, relative motion, temperature, mass, and other properties. Collecting light and forming images is usually regarded as a telescope function. Analyzing the images is done by various instruments designed for that purpose and attached to the telescope. Detectors are normally part of the instrumentation. The following discussion deals with telescopes.

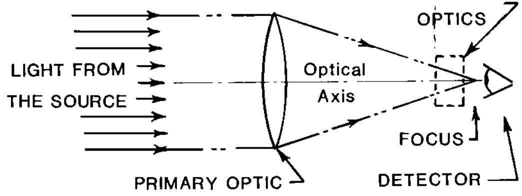


FIGURE 1 Basic telescope elements. Refractive lens could be replaced with a curved reflective mirror.

A. Telescope Size and its Effect on Images

The size of a telescope ordinarily refers to the diameter, or its approximate equivalent, for the area of the first (primary) image-forming optical element surface illuminated by the source. Thus, a 4M telescope usually signifies one with a 4-m diameter primary optic. This diameter sets a maximum limit on the instantaneous photon flux passing through the image-forming optical train. Some telescopes use flat mirrors to direct light into the telescope (e.g., solar heliostats); however, it is the size of the illuminated portion of the primary imaging optic that sets the size.

The size of a telescope determines its ability to resolve small objects. The Airy disc diameter, generally taken to be the resolution limit for images produced by a telescope, varies inversely with size. The Airy disc also increases linearly with wavelength, which means that one must use larger sized telescopes to obtain equivalent imaging resolution at longer wavelengths. This is a concern for astronomers wishing to observe objects at infrared (IR) wavelengths and also explains in part why radio telescopes, operating at even longer wavelengths, are so much larger than optical telescopes. (Radio telescopes are more easily built larger because radio wavelengths are much longer and tolerances on the “optics” are easier to meet.)

Telescopes may be used in an interferometric mode to form interference fringes from different portions of the incoming light beam. Considerable information about the source can be derived from these fringes. The separation between portions of the primary optic forming the image (fringes) is referred to as the baseline and sets a limit on fringe resolution. For a telescope with a single, round primary optic, size and maximum baseline are the same. For two telescopes directing their beams together to form a coherent image, the maximum baseline is equal to the maximum distance between light-collecting areas on the two primaries. More commonly, the center-to-center distance would be defined as the baseline, but the distance between any two image-forming areas is also a baseline. Thus, multiple-aperture systems have many baselines.

As telescope size D increases so does the physical size of the image, unless the final focal ratio F_{f} in the converging beam can be reduced proportionately, that is,

$$\text{final image size} = F_{\mathrm{f}} \cdot D \cdot \theta,$$

where θ is the angular size of the source measured on the sky in radians. With large telescopes this can be a matter of importance when trying to match the image to a particular detector or instrument. Even for small optics, achieving focal ratios below about f/1.0 is difficult, which sets a practical limit on image size reduction for a given situation.

Another size-related effect is that larger telescopes look through wider patches of the atmosphere which usually contain light-perturbing thermally turbulent (varying air density) regions that effectively set limits on seeing. Scintillation (twinkling) and image motion are caused by the turbulence. However, within a turbulent region, slowly varying isotropic subregions (also called isoplanatic patches) exist that affect the light more or less uniformly. When a telescope is sized about the same as, or smaller than, a subregion and looks through such a subregion, the instantaneous image improves because it is not affected by turbulence outside the subregion. As the subregions sweep through the telescope’s field of view, the image changes in shape and position. Larger telescopes looking through many subregions integrate or combine the effects, which enlarges the combined image and effectively worsens the seeing. However, these effects diminish with increasing wavelength, which means that larger telescopes observing at IR wavelengths may have better seeing than smaller ones observing in the visible region.

Studies of atmospheric turbulence effects have given rise to the development of special devices to make optical corrections. These are sometimes called rubber mirrors or adaptive optics. An image formed from incoming light is sensed and analyzed for its apparent distortion. That information is used to control an optical element (usually a mirror) that produces an offsetting image distortion in the image-forming optical train. By controlling on a star in the isoplanatic patch with the object to be observed (so that both experience similar turbulence effects), one can, in principle, form corrected images with a large ground-based telescope that are limited only by the telescope, not the atmosphere. In practice, low light levels from stars and the relatively small isoplanatic patch sizes (typically a few arc-seconds across) have hampered usage of adaptive optics on stellar telescopes. One technique, successfully used in large telescopes, is to split off the visible wavelength portion of the light beam for image distrotion analysis while allowing the longer wavelength (IR) portion of the beam to pass through to the adaptive optic and to form the final corrected image.

B. Telescope Characteristics Related to Size

At least three general, overlapping categories related to telescope size may be defined:

1. Telescopes small enough to be portable. Sizes usually less than 1 m.

2. Mounted telescopes with monolithic primary optics. Sizes presently range up to 6 m for existing telescopes, with many 8 m telescopes either planned or under construction. Virtually all ground-based telescopes used by professional astronomers are in this category.

3. Very large telescopes with multielement primary optics. Proposed sizes range up to 25 m for ground-based telescopes. Only a few multielement telescopes have actually been built.

A fourth category could include telescopes small enough to be launched into Earth's orbit, but the possibility of an in-space assembly of components makes this distinction unimportant.

One cannot, in a short space, describe all of the telescope styles and features. Nevertheless, as one considers larger and larger telescopes, differences become apparent and a few generalizations can be postulated.

In the category of small telescopes, less than 1 m, one finds an almost unlimited variety of telescope configurations. There are few major size limitations on materials for optics. Polishing of optics can often be done manually or with the aid of simple machinery. Mechanical requirements for strength or stiffness are easily met. Adjustments and pointing can be manually performed or motorized. Weights are modest. Opportunities for uniqueness abound and are often highly prized. Single-focus operation is typical. Most of the telescopes used by amateur and professional astronomers are in this size range. Figure 2 illustrates a 40-cm telescope used by professional astronomers.

In the 1–2 m size range, a number of differences and limitations arise. Obtaining high-quality refractive optics is expensive in this range and not practical beyond. Simple three-point mechanical supports no longer suffice for the optics. The greater resolving-power potential demands higher quality optics and good star-tracking precision. Telescope components are typically produced on large machine tools. Instrumentation is likely to be used at more than one focus position. Because of cost, the domain of the professional astronomer has been reached.

As size goes above 2 m, new issues arise. The need to compensate for self-weight deflections of the telescope becomes increasingly important to maintaining optical alignment. Flexure in the structure may affect the bearings and drive gears. Bearing journals become large enough to require special bearing designs, often of the hydrostatic oil variety. The observer may now be supported by the telescope instead of the other way around. Support of the primary optics is more complex and obtaining primary mirror blanks becomes a special, expensive task. Automated operation is typical at several focal positions. Star-tracking automatic guiders may be used to control the telescope drives, augmented by computer-based pointing correction tables. Figure 3 illustrates a 4-m telescope with all of these features.

At 5 m, the Hale Telescope on Mount Palomar is regarded as near the practical limit for equatorial-style mountings (see Section V). Altitude–azimuth (alt–az) mountings are better suited for bearing heavy rotating loads and are more compact. With computers, the variable drive speeds required with an alt–az telescope can be managed. Mounting size and the length of the telescope are basic factors in setting the size of the enclosing building. For technical reasons and lower cost, the present trend in large telescopes is toward shorter primary

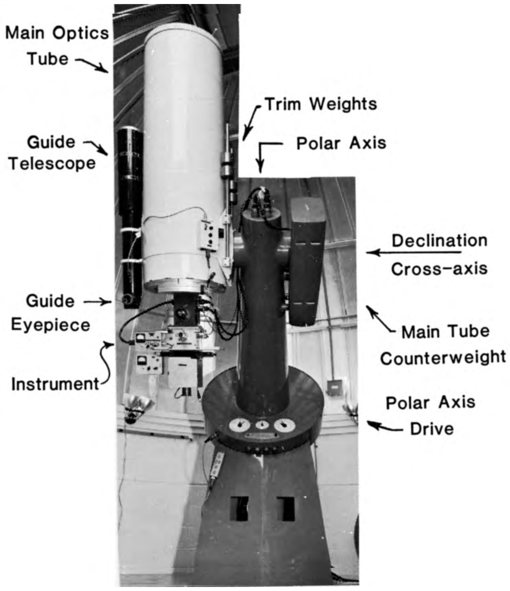


FIGURE 2 40-cm telescope on an off-axis equatorial mount. [Courtesy National Optical Astronomy Observatories, Kitt Peak.]

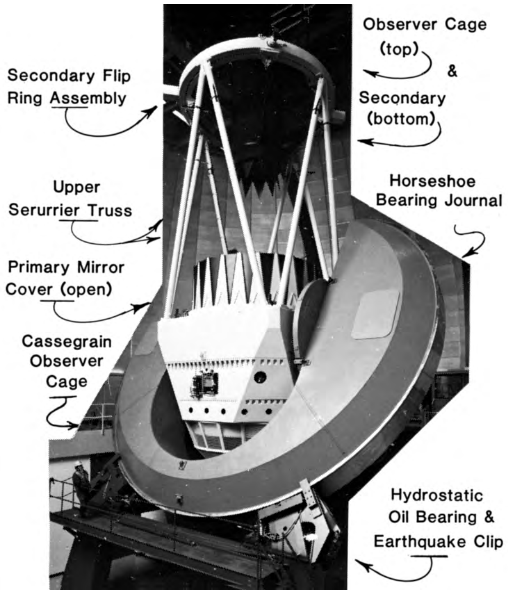

Observer Cage (top) & Secondary (bottom)
Secondary Flip Ring Assembly
Horseshoe Bearing Journal
Upper Serurrier Truss
Primary Mirror Cover (open)
Cassegrain Observer Cage
Hydrostatic Oil Bearing & Earthquake Clip

FIGURE 3 Mayall 4-m telescope, with equatorial horse-shoe yoke mountings. [Courtesy National Optical Astronomy Observatories, Kitt Peak.]

TABLE I Telescopes 3 Meters or Larger Built Since 1950

Date completed (projected)	Telescope and/or institution	Primary mirror size (m)	Primary focal ratio	Mounting style
1950	Hale Telescope, Palomar Observatory, California	5	3.3	Equatorial horseshoe yoke
1959	Lick Observatory, California	3	5.0	Equatorial fork
1973	Mayall Telescope, Kitt Peak National Observatory (KPNO), Arizona	4.0	2.7	Equatorial horseshoe yoke
1974	Cerro Tololo International Observatory (CTIO), Chile	4.0	2.7	Equatorial horseshoe yoke
1975	Anglo-Australian Telescope (AAT), Australia	3.9	3.3	Equatorial horseshoe yoke
1976	European Southern Observatory (ESO), Chile	3.6	3.0	Equatorial horseshoe yoke
1976	Large Altazimuth Telescope (BTA), Special Astrophysical Observatory, Caucusus Mtns, Russia	6.0	4.0	Alt–az
1979	Infrared Telescope Facility (IRTF), Hawaii	3.0	2.5	Equatorial English yoke
1979	Canada-France-Hawaii Telescope (CFHT), Hawaii	3.6	3.8	Equatorial horseshoe yoke
1979	United Kingdom Infrared Telescope (UKIRT), Hawaii	3.8	2.5	Equatorial English yoke
1979	Multiple Mirror Telescope (MMT) Observatory, Mt. Hopkins, Arizona	4.5[a]	Six 1.8 m[b]	Alt–az at f/2.7
1983	German-Spanish Astronomical Center, Calar Alto, Spain	3.5	3.5	Equatorial horseshoe fork
1986	Wm. Herschel Telescope, La Palma, Canary Islands	4.2	2.5	Alt–az
1989	New Technology Telescope (NTT), ESO, Chile	3.5	2.2	Alt–az
(1993)	Wisconsin-Indiana-Yale-NOAO (WIYN) Telescope, Kitt Peak, Arizona	3.5	1.75	Alt–az
(1992)	Apache Point Observatory, New Mexico	3.5	1.75	Alt–az
(1992)	Keck Ten Meter Telescope (TMT), Hawaii, Segmented Mirror Telescope (SMT). 36-element (hexagons) paraboloidal primary	10[a]	1.75	Alt–az
Proposed for Hawaii (1987) & CTIO (1999)	National Optical Astronomy Observatories (NOAO), Arizona	8	1.8	Alt–az Disc
Proposed for Chile (1997)	Magellan Project Telescope, Las Campanas, Chile	8	1.2	Alt–az Disc
(1998)	Japanese National Large Telescope (JNLT), Hawaii	8.3	1.9	Alt–az
(1998)	Columbus Project Telescope, Mt. Graham, Arizona, MMT style	11.3[a]	Two 8 m at f/1.2[b]	Alt–az “C-Ring”
(1998)	The Very Large Telescope (VLT), ESO, Chile. An array of telescopes	16[a]	Four 8.2 m at f/2[b]	Alt–az
Proposed	Russia, SMT style with a 400-element (hexagons) spherical primary.	25[a]	2.7	Alt–az

[a] Equivalent circular mirror diameter with equal light gathering area.
[b] Number, size, and f/ratio of individual primary mirror.

focal lengths and alt–az mounts. This trend is evident from Table I, which lists the telescopes 3 m in size or larger that have been built since about 1950. Also listed are the major large telescopes proposed for construction in the late 1980s and the 1990s, which will be discussed in the next section. The largest optical telescope in operation today is the Russian 6 m, which incorporates a solid, relatively thick (650 mm) primary mirror that had to be made three times in borosilicate glass and finally in a low-expansion material before it was successful. Such difficulty indicates that 6 m may be a practical limit for that style of mirror. New approaches are needed to go beyond.

C. The New Giant Telescopes

The desire for greater light-gathering power and image resolution, especially at IR wave-lengths, continues to press astronomers to build telescopes with larger effective apertures. Costs for a given telescope style and imaging performance have historically risen nearly as the primary aperture diameter to the 2.5 power. These factors have given impetus to a number of new technology telescope designs (see Proposed Projects in Table I) that are based on one or more of the approaches discussed in the following. Computer technology plays a strong part in all of these approaches.

1. Extending the Techniques for Making Lightweight Monolithic Mirror Blanks

Sizes up to about 8 m are considered feasible, although the Russian 6 m is the largest telescope mirror produced before 1985. Further discussion on blank fabrication methods is provided in Section III. Supporting such large mirrors to form good images will be difficult without some active control of the surface figure and thermal conditions in the mirror blank.

Several American universities and the national observatories in the United States and Japan are planning telescopes of this variety.

2. Making a Large Mirror from Smaller Segments

This is also known as segmented mirror telescope, or SMT. In principle, no limit exists for the size of a mosaic of mirror segments that functions optically as a close approximation to a monolithic mirror. For coherency each segment must be precisely and continuously positioned with respect to its neighbors by means of position sensors and actuators built into its support. The segments may be hexagonal, wedge shaped, or other to avoid large gaps between segments. Practical limits arise from support structure resonances and cumulative errors of the segment positioning system. Manufacturing and testing the segments require special methods since each is likely to be a different off-axis optic that lacks a local axis of symmetry but must have a common focus with all the other segments.

The University of California and the California Institute of Technology have adopted this approach for their Keck Observatory 10-m telescope (see Fig. 4), completed in 1992. Russia also has announced plans for a 25-m SMT utilizing a spherical primary to avoid the problems of making aspheric segments.

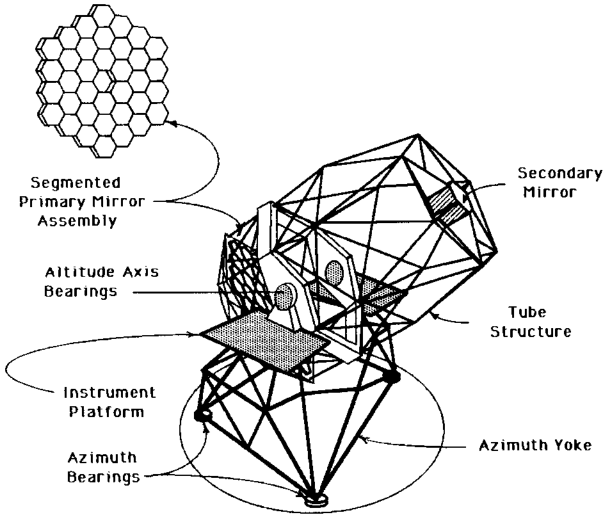


FIGURE 4 Conceptual sketch of a segmented mirror telescope (SMT) using the arrangement adopted for the Keck Ten Meter Telescope, which uses 36 hexagonal mirrors in a primary mirror mosaic to achieve light gathering power equivalent to a single 10-m diameter circular mirror. Each mirror segment must be positioned accurately with respect to its neighbors to achieve the overall optical effect of an unsegmented parabola. The gaps between the segments cause unwanted diffraction effects in the image in proportion to the area they occupy in the primary aperture; hence, they are minimized. Other segment shapes are possible, and the total telescope size is limited only by its structural stiffness and optics alignment provisions. Thus, very large SMTs are theoretically possible.

3. Combining the Light from an Array of Telescopes

Several methods may be considered:

1. Electronic combination after the light has been received by detectors at separate telescopes. Image properties will be those due to the separate telescopes, and coherent combining is not presently possible. Strictly speaking, this is an instrumental technique and will not be considered further.

2. Optical combination at a single, final focus of light received at separately mounted telescopes. To maintain coherency between separate light beams, one must equalize the optical path distance (OPD) between the source and the final focus for all telescopes, a difficult condition to meet if telescopes are widely separated.

3. Placing the array of telescopes on a common mounting with a means for optically combining the separate light beams. All OPDs can be equal (theoretically), thus requiring only modest error correction to obtain coherency between telescopes. This approach is known as the multiple mirror telescope (MMT).

The simplest array of separately mounted individual telescopes is an arrangement of two on a north–south (NS) baseline with an adjustable, combined focus between them (the OPD changes occur slowly with this arrangement when observing at or near the meridian). Labeyrie pioneered this design in the 1970s at Centre d'Etudes et de Recherches Géodynamiques et Astronomiques (CERGA) in France, where he used two 25-cm telescopes on a NS variable baseline of up to 35 m to measure successfully numerous stellar diameters and binary star separations, thereby showing that coherent beam combination and the angular resolution corresponding to a long telescope baseline could be obtained. Other schemes for using arrays of separate telescopes on different baselines all require movable optics in the optical path between the telescopes to satisfy the coherency conditions, and so far none has been successfully built. However, the European Southern Observatory is building an array of four 8-m telescopes in Chile which will be known as the "Very Large Telescope" (VLT) when completed in the late 1990s. One form of the VLT is an in-line array with an "optical trombone" arrangement for equalizing OPDs. Other arrangements are also under consideration.

The MMT configuration was first used by the Smithsonian Astrophysical Observatory (SAO) and the University of Arizona (UA). The SAO/UA MMT on Mount Hopkins uses six 1.8-m image-forming telescopes arrayed in a circle around a central axis. Six images are brought to a central combined focus on the central axis, where they may be incoherently stacked, coherently combined, or used separately. The effective baseline for angular resolution (i.e., the maximum separation between reflecting areas) is 6.9 m, and the combined light-gathering power is equivalent to a single 4.5-m diameter mirror. A two-element MMT using 8-m mirrors and a 22-m baseline, illustrated in Fig. 5 and called the "Columbus Project," is under construction by the UA in collaboration with the Observatorio Astrofisico de Arcetri, Italy. Appropriately, this telescope has been dubbed the "Big Binocular."

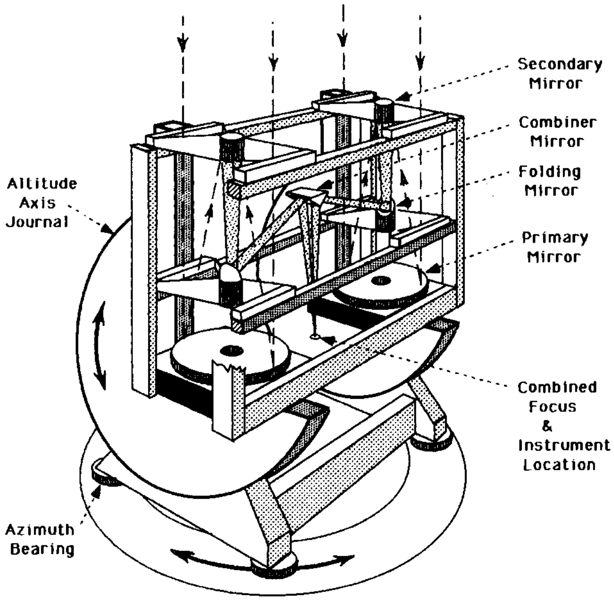


FIGURE 5 Conceptual sketch of a multiple mirror telescope (MMT) using the arrangement planned for the Columbus "Big Binocular" Telescope, which uses two 8-m diameter telescopes mounted on a common structure with provisions to combine their separate output light beams. With careful control of the optical path distances in each telescope, the resolving power of the telescope is equivalent to that of a single-mirror telescope of a size equal to the maximum separation between the primary mirrors of the two telescopes. Total light gathering power is equal to the sum of the individual telescope powers. There is no theoretical limit to the number of individual telescopes in an MMT, but the beam combining and structural provisions become increasingly complex.

II. OPTICAL CONFIGURATIONS

Light entering the telescope is redirected at each optical element surface until it reaches the focal region where the images are most distinct. The light-sensitive detector is customarily located in an instrument mounted at the focal region. By interchanging optics, one can create more than one focus condition; this is commonly done in large telescopes to provide places to mount additional instruments or to produce different image scales. The arrangement of optics and focal positions largely determines the required mechanical support configuration and how the telescope will be used.

The early telescopes depended solely upon the refractive power of curved transparent glass lenses to redirect the light. In general, these telescopes were plagued by chromatic aberration (rainbow images) until the invention in 1752 of achromatic lenses, which are still used today in improved forms. Curved reflective surfaces (i.e., mirrors) were developed after refractors but were not as useful until highly reflective metal coatings could be applied onto glass substrates. Today, mirrors are more widely used than lenses and can generally be used to produce the same optical effects; they can be made in larger sizes, and they are without chromatic aberration. These are still the only two means used to form images in optical telescopes. Accordingly, telescopes may be refractive, reflective, or catadioptric, which is the combination of both.

A. Basic Optical Configurations: Single and Multielement

The telescope designer must specify the type, number, and location of the optical elements needed to form the desired image. The basic choices involve material selections and the shapes of the optical element surfaces. Commonly used surfaces are flats, spheres, paraboloids, ellipsoids, hyperboloids, and toroidal figures of revolution.

The optical axis is the imaginary axis around which the optical figures of revolution are rotated. Light entering the telescope parallel to this axis forms the on-axis (or zero-field) image directly on the optical axis at the focal region. The field of view (FOV) for the telescope is the widest angular span measured on the sky that can be imaged distinctly by the optics.

In principle, a telescope can operate with just one image-forming optic (i.e., at prime focus), but without additional corrector optics, the FOV is quite restricted. If the telescope is a one-element reflector, the prime focus and hence the instrument/observer are in the line of sight. For large telescopes (i.e., >3 m) this may be used to advantage, but more commonly the light beam is diverted to one side (Newtonian) or is reflected back along the line of sight by means of a secondary optic to a more convenient focus position. Figure 6 illustrates the optical configurations most commonly used in reflector telescopes. In principle, the reflectors shown in Fig. 6 could be replaced with refractors to produce the same optical effects. However, the physical arrangement of optics would have to be changed.

In practice, one tries to make the large optics as simple as possible and to form good images with the fewest elements. Other factors influencing the configuration include the following:

1. Simplifying optical fabrication. Spherical surfaces are generally the easiest to make and test. Nonsymmetric aspherics are the opposite extreme.
2. Element-to-element position control, which the telescope structure must provide. Tolerances become tighter as the focal ratio goes down.
3. Access to the focal region for viewing or mounting instrumentation. Trapped foci (e.g., the Schmidt) are more difficult to reach.
4. Compactness, which generally aids mechanical stiffness.
5. Reducing the number of surfaces to minimize light absorption and scattering losses.

Analyzing telescope optical systems requires a choice of method. The geometric optics method treats the incoming light as a bundle of rays that pass through the system while being governed by the laws of refraction and reflection. Ray-tracing methods based on geometric optics are commonly used to generate spot diagrams of the ray positions in the final image, as illustrated in Fig. 7a. More rigorous analysis based on diffraction theory is done by treating the incoming light as a continuous wave and examining its interaction with the optical system. Figure 7b is a computer-generated plot of light intensity in a diffracted image formed by a circular aperture. The central peak represents the Airy disc. For further detail on the use of these methods the reader is referred to texts on optics design.

B. Wide Field Considerations

Modern ground-based telescopes are usually designed to resolve images in the 0.25–1.0-arcsec range and to have FOV from a few arc minutes to about one degree. In general, distortions or aberrations due to the telescope optics exist in the images and are worse for larger field angles. Table II lists the basic types. Space-based telescopes (e.g., the Hubble telescope) can be built to resolve images in the 0.1 arcsec region because atmospheric seeing effects are absent, but compensation must still be made for optical aberrations. Aberrations can also arise from nonideal placement of the optical surfaces (i.e., position errors) and from nonuniform conditions in the line of sight.

To correct distortions produced by optics, the designer frequently tries to cancel aberrations produced at one surface by those produced at another. Extra optical surfaces may be introduced for just this purpose. Ingenious optical corrector designs involving both refractors and reflectors have resulted from this practice. Many of these designs are described in texts on optics under the originators' names (e.g., Ross, Baker, Wynne, Shulte, and Meinel). It is possible, however, to cancel some field aberrations by modifying the principal optical surfaces or by using basic shapes in special combination. For example, the Ritchey–Crétian telescope design for a wide FOV reduces coma by modifying the primary and secondary surfaces of a Cassegrain telescope. The Mersenne telescope cancels aberrations from the primary (a parabola) with the secondary (another parabola).

A spherical mirror with the aperture stop set at its center of curvature has no specific optical axis and forms equally good images everywhere in the field. Images of distant objects have spherical aberrations, however, and the focal region is curved. The Schmidt telescope compensates for most of the spherical aberration by means of an aspheric refractor at the center of curvature. The Maksutov telescope introduces an offsetting spherical aberration by means of a spherical meniscus refractor. Numerous variations of this approach have been devised yielding well-corrected images in field sizes of 10 deg and more. However, large fields may have other problems.

During a long observing period, images formed by a telescope with a very large FOV (i.e., ~1 deg or more) are affected differently across the field by differential refraction effects caused by the atmosphere. Color dispersion effects (i.e., chromatic blurring) effectively enlarge the images as the telescope looks through an increasing amount of atmosphere. Differential image motion also occurs, varying as a function of position in the FOV, length of observation, and telescope pointing angle. Partial chromatic correction can be made by inserting a pair of separately rotatable prisms, called Risley prisms, ahead of the focal position. Even with chromatic correction, however, images are noticeably elongated (>0.5 arcsec) at the edge of a 5-deg field compared with the on-axis image after a continuous observing period of a few hours.

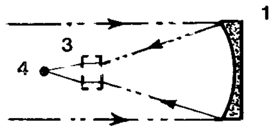
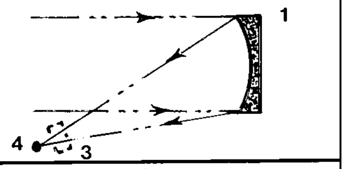
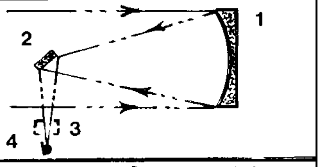
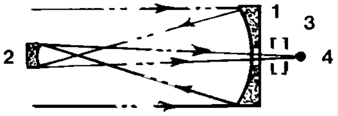
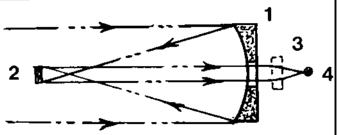
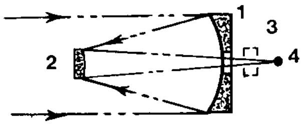
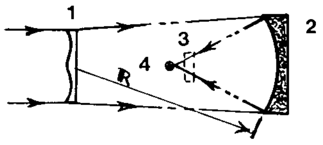
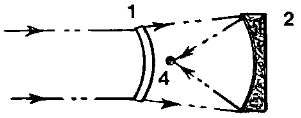
TYPE	PRIMARY OPTIC	SECONDARY OPTIC	CONFIGURATION 1-PRIMARY 2-SECONDARY 3-EYEPIECES/CORRECTORS 4-FOCUS
KEPLERIAN GALILEAN (if refractive)	SPHERE or PARABOLA	NONE	
HERSCHELIAN	OFF-AXIS PARABOLA	NONE	
NEWTONIAN	PARABOLA	DIAGONAL FLAT	
GREGORIAN	PARABOLA	ELLIPSE	
MERSENNE	PARABOLA	PARABOLA	
CASSEGRAIN	PARABOLA	HYPERBOLA	
RITCHEY- CHRÉTIEN	MODIFIED PARABOLA	MODIFIED HYPERBOLA	
DALL-KIRKHAM	ELLIPSE	SPHERE	
SCHMIDT	ASPHERIC REFRACTOR	SPHERE	
BOUWERS- MAKSUTOV	REFRACTIVE MENISCUS	SPHERE	

FIGURE 6 Basic optical configurations for telescopes.

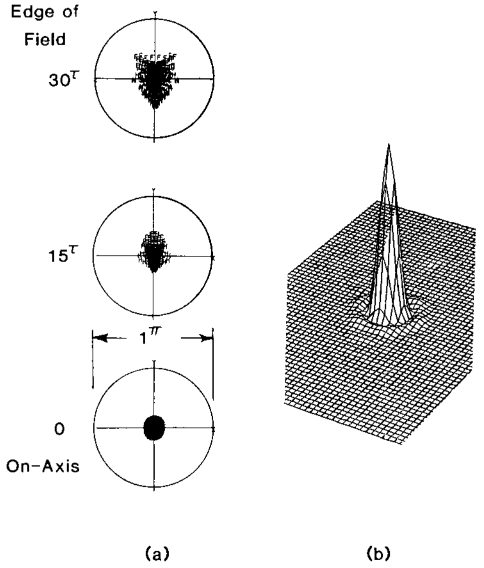


FIGURE 7 Examples of image analysis methods. (a) Typical spot diagrams of well-corrected images in the focal plane of a 1-deg field. Each image represents the ray bundle at that field location. (b) Computer-generated three-dimensional plot of intensity in a perfect image that has been diffracted by a circular aperture.

C. Instrumental Considerations: Detector Matching and Baffling

At any focal position, the distance measured along the optical axis within which the images remain acceptably defined is referred to as the depth of focus. The detector should be adjusted to the most sharply focused position in this region; however, it is more common to adjust the focus (e.g., by moving the secondary) to sharpen the images at the detector. Certain instruments containing reimaging optics (e.g., spectrographs) require only that the image be formed at the entrance to the instrument (e.g., at a slit or aperture plate). There is a practical limit to the amount of allowable focal position movement obtained by moving the secondary because optical aberrations, especially spherical aberration, are introduced when the mirror is displaced from its theoretically perfect position.

TABLE II Basic Image Aberrations Occurring in Telescopes

Type	Condition
Spherical aberration[a]	Light focuses at different places along the optical axis as a function of radial position in the aperture.
Coma[a]	Image size (magnification) varies with radial position in the focal region. Off-axis flaring.
Field curvature[a]	Off-axis images are not focused on the ideal surface, usually a plane.
Astigmatism[a]	Light focuses at different places along the optical axis as a function of angular position in the aperture.
Distortion[a]	Focused off-axis image is closer to or further from the optical axis than intended.
Chromatic aberration	Shift in the focused image position as a function of wavelength.

[a] Also known as Seidel aberrations.

To view all or part of the focused FOV, it is common to insert a field mirror into the beam at 45 deg to divert the desired portion of the field out to an eyepiece or a TV monitoring camera. The undiverted light to be observed passes on to the instrument. The diverted light can be used for guiding purposes or to make different observations.

The angular size of the focused image (i.e., the image scale) and its sensitivity to defocusing changes is governed by the effective focal length (EFL) of the optical system that formed the image:

$$\begin{aligned}\text{image scale} &= \frac{1}{\text{EFL}}\\ &= \frac{\text{radians on the sky}}{\text{length in the focal plane}}.\end{aligned}$$

Large EFL values produce comfortable focal depths, but the image sizes are relatively large. The physical size of the detector may place a limit on the FOV that can be accommodated for a given scale, hence a potential need for a different focal position or reimaging optics. This issue has become quite important with the advent of solid-state image detectors, known as CCDs (Charge Coupled Devices), that have much higher quantum efficiency than photographic plates but are difficult to make in large sizes (i.e., approximately 50 × 50 mm is considered large in 1990).

Light baffles and aperture stops are important but frequently neglected aspects of telescope design. It is common to place a light baffle just ahead of the primary to block out unwanted edge effects, in which case the baffle may become the aperture stop. In some cases, the primary surface may be reimaged further along the optical path where a light baffle can be located. This may be done either to reduce the size of the required baffle or to locate it advantageously in a controlled environment (e.g., at cryogenic temperature to reduce thermal radiation effects). In other cases, the aperture may be set by undersizing one of the optical elements that is further along the optical path. An example is an IR-optimized telescope, where the secondary is made undersized to ensure that the detector cannot see past the edge of the primary mirror. The effective light-gathering power of the telescope can be significantly reduced under these conditions.

Obstructions in the light path further reduce light-gathering power. The most common obstructions are the secondary and auxiliary optics along with the mechanical struts that support them. It is common for 10–20% of the aperture to be obstructed in this manner. Depending upon the telescope style and the instrument location, it may be necessary to use a light baffle to prevent the detector from seeing unwanted radiation. For example, a detector at the Cassegrain focus of a two-mirror telescope can see unfocused light from stars directly past the perimeter of the secondary mirror unless obstructing baffles are provided. The size and location of these baffles are generally determined empirically by ray-tracing, and it is a fact that larger FOV requires larger baffles that mean more obstruction. Thus, one cannot truly determine telescope light-gathering power until the usage is considered.

III. TELESCOPE OPTICS

Telescope performance depends fundamentally upon the quality of optics, especially the surface figures. That quality in most telescopes is a compromise between what the optical designer has specified, what the glassmakers and opticians can make, and how well the mechanical supporting structures perform. Especially for larger sizes, one should specify the optics and their supports at the same time. How the optics are to be tested should also be considered since the final figure corrections are almost always guided by test results. Optical figuring methods today range from simple manual lapping processes to sophisticated computer-controlled polishers (CCP) and direct machining using diamond tools. The remarks to follow are only a summary of a complex technical field.

A. Refractive Optics

Light passing through the surface of a refractive optic is changed in direction according to Snell's law of refraction:

$$\eta_1 \sin\theta_1 = \eta_2 \sin\theta_2,$$

where η_1 and η_2 are the refractive indices of the materials on either side of the surface and θ_1 and θ_2 the angles with respect to the surface normal of incidence and refraction. Producing a satisfactory refractor is, therefore, done by obtaining transparent glass or another transmitting material with acceptable physical uniformity and accurately shaping the surfaces through which light passes. Sometimes the optician can alter the surface to compensate for nonuniform refractive properties. Losses ordinarily occur at the surface due to scattering and also to the change in refractive index. Antireflection coatings can be applied to reduce surface transmission losses, but at the cost of restricting transmission to a specific wavelength range. Further losses occur internally by additional scattering and absorption.

Manufacturing methods for refractive optics are similar to those for reflectors.

B. Reflective Optics

Light reflecting off a surface is governed by the law of reflection:

$$\theta_1 = -\theta_2,$$

where θ_1 and θ_2 are the angles of incidence and reflection, respectively. Controlling the slope at all points on the reflecting surface is, therefore, the means for controlling where the light is directed. Furthermore, any part of the surface that is out of its proper position, measured along the light path, introduces a change at that part of the reflected wavefront (i.e., a phase change) equal to twice the magnitude of the position error. Surface accuracy is thus the prime consideration in making reflective optics. Achieving high reflectivity after that is usually done by applying a reflective coating.

The most common manufacturing method is to rough-machine or grind the mirror blank surface and progressively refine the surface with abrasive laps. Certain kinds of mirrors, especially metal ones, may be diamond turned. In this case, the accuracy of the optical surface may be governed by the turning machine, whereas ordinarily the accuracy limits are imposed principally by the optical test methods and the skill of the optician. High-quality telescope mirrors are typically polished so that most of the reflected light (>80%) is concentrated in an image that is equivalent to the Airy disc or the seeing limit, whichever is larger.

C. Materials for Optics

Essential material requirements for all optical elements are related to their surfaces. One must be able to polish or machine the surfaces accurately, and afterward the blank should not distort uncontrollably. Residual stresses and unstable alloys are sources of dimensional instability to be avoided. Stresses also cause birefringence in refractors.

1. Refractors

Refractors should transmit light efficiently and uniformly throughout the operating wavelength region. However, there is no single material that transmits efficiently from 0.3 to 40 μm. One typically chooses different materials for the UV, near-IR (to $\simeq$2 μm), and far-IR regions. Glassmakers can control the index of refraction to about one part in 10^6, but only in relatively small blanks (<50-cm diameter). In larger sizes, index variations and inclusions limit the availability of good-quality refractor blanks to sizes less than 2-m diameter.

2. Reflectors

The working part of a reflector optic is usually the thin metallic layer, 1000–2000 Å thick, that reflects the light. An evaporated layer of aluminum, silver, or gold is most commonly used for this purpose, and it obviously must be uniform and adhere well to the substrate. Most of the work in making a reflector, however, is in producing the uncoated substrate or mirror blank. The choice of material and the substrate configuration are critically important to the ultimate reflector performance.

Reflector blanks, especially large ones, require special measures to maintain dimensional (surface) stability. The blank must be adequately supported to retain its shape under varying gravitational loads. It must also be stable during normal temperature cycles, and it should not heat the air in front of the mirror because that causes thermal turbulence and worsens the seeing. A mirror-to-air temperature difference less than about 0.5°C is usually acceptable. The support problem is a mechanical design consideration. Thermal stability may be approached in any of several ways:

1. Use materials with low coefficients of thermal expansion (CTE).
2. Lightweight the blank to reduce the mass and enhance its ability to reach thermal equilibrium quickly (i.e., by using thin sections, pocketed blanks, etc.). Machineability or formability of the material is important for this purpose.
3. Use materials with high thermal conductivity (e.g., aluminum, copper, or steel).
4. Use active controls for temperature or to correct for thermally induced distortion. Elastic materials with repeatable flexure characteristics are desirable. Provisions in the blank for good ventilation may be necessary.

Materials with low CTE values include borosilicate glass, fused silica, quartz, and ceramic composites. Multiple-phase (also called binary) materials have been developed that exhibit near-zero CTE values, obtained by offsetting the positive CTE contribution of one phase with the negative CTE contribution from another. Zerodur made by Schott Glaswerke, ULE by Corning Glass Works, and epoxy–carbon fiber composites are examples of multiple-phase materials having near-zero CTE over some range of operating temperature. Fiber composites usually require a fiber-free overlayer that can be polished satisfactorily.

One usually considers a lightweight mirror blank to reduce costs or to improve thermal control. This is particularly true for large telescopes. Reducing mirror weight often produces a net savings in overall telescope cost even if the lightweighted mirror blank is more expensive than a corresponding solid blank. The important initial step is to make the reflecting substrate or faceplate as thin as possible, allowing for polishing tool pressures and other external forces. One has three hypothetical design options:

1. Devise a way to support just the thin monolithic faceplate. This approach works best for small blanks. Large, thin blanks require complex supports and, possibly, a means to monitor the surface shape for active control purposes. The European Southern Observatory's 3.5 m New Technology Telescope (NTT), commissioned in 1990, is a good example of this approach. The additional complexity of the mirror support is made more worthwhile by using it to correct for image aberrations due to other causes. The NTT has resolved images of 0.33 arcsec (containing 80% of the focused light and all of the aberrations due to the telescope and atmosphere), perhaps, the best performance to date for a large ground-based telescope.
2. Divide the faceplate into small, relatively rigid segments and devise a support for each segment. Segment position sensing and control is required: a sophisticated technical task. The Keck Ten Meter Telescope uses this approach.
3. Reinforce the faceplate with a gridwork of ribs or struts, possibly connected to a backplate, to create a sandwichlike structure. One may create this kind of structure by fusing or bonding smaller pieces, by casting into a mold, or by machining away material from a solid block.

All of these approaches have been used to make lightweight reflector blanks. Making thin glass faceplates up to about 8 m is considered feasible by fusing together smaller pieces or direct casting. Titanium silicate, fused silica, quartz, and borosilicate glass are candidate materials. Castings of borosilicate glass up to 6 m (e.g., the Palomar 5-m mirror) have been produced; 8 m is considered feasible. Structured (i.e., ribbed) fused silica and titanium silicate mirrors up to 2.3 m have been produced (e.g., the Hubble Space Telescope mirror); up to 4 m is considered feasible.

Metals may also be used for lightweight mirrors but have not been widely used for large telescopes because of long-term dimensional (surface) changes. The advent of active surface control technology may alter this situation in the future. Most metals polish poorly, but this can be overcome by depositing a nickel layer on the surface to be polished. Most nonferrous metals can also be figured on diamond-turning machines; however, size is limited presently to about 2 m.

IV. SPECTRAL REGION OPTIMIZATION: GROUND-BASED TELESCOPES

The optical/IR window of the atmosphere from 0.3 to 35 μm is sufficiently broad that special telescope features are needed for good performance in certain wavelength regions. Notably, these are needed for the UV region less than about 0.4 μm and the thermal IR region centered around 10 μm. In the UV, it is difficult to maintain high efficiency because of absorption losses in the optics. When observing in the IR region, one must cope with blackbody radiation emitted at IR wavelengths by parts of the telescope in the light path, as well as by the atmosphere. Distinguishing a faint distant IR source from this nearby unwanted background radiation requires special techniques. Using such techniques, it is common for astronomers to use ground-based telescopes to observe IR sources that are more than a million times fainter than the IR emission of the atmosphere through which the source must be discerned.

A. UV Region Optimization

The obvious optimizing step for the UV region is to put the telescope into space. If the telescope is ground-based, however, one good defense against UV light losses is to use freshly coated aluminum mirror surfaces. Reflectivity values in excess of 90% can be obtained from freshly coated aluminum, but this value rapidly diminishes as the surface oxide layer develops. Protective coatings such as sapphire (Al_2O_3) or magnesium fluoride (MgF_2) can be used to inhibit oxidations. Also, multilayer coatings can be applied to the surfaces of all optics to maximize UV throughput. However, these coatings greatly diminish throughput at longer wavelengths, which leads to the tactic of mounting two or more sets of optics on turrets, each set being coated for a particular wavelength region. The desired set is rotated into place when needed. This tactic obviously works best for small optics, not the primary.

Many refractive optics materials absorb strongly in the UV. Fused silica is good low-absorbing material. If optics are cemented together, the spectral transmission of the cement should be tested. Balsam cements are to be avoided.

B. IR Region Optimizations

One cardinal rule for IR optimization is to minimize the number and sizes of emitting sources that can be seen by the detector. This includes mechanical hardware such as secondary support structures and baffles, as well as seemingly empty spaces such as the central hole in the primary mirror. All of these emit black body radiation corresponding to their temperatures.

Since the detector obviously must see the optical surfaces, another cardinal rule is to reduce the emissivity of these surfaces with a highly reflective coating. If a coating is 98% reflective, it emits only 2% of the blackbody radiation that would otherwise occur if the surface were totally nonreflective. If possible, the detector should see only reflective surfaces, and these should be receiving radiation only from the sky or other reflective surfaces in the optical train. Objects that must remain in the line of sight can also be advantageously reflective provided that they are not looking at other IR-emitting objects that could send the reflected radiation into the main beam.

Achieving these goals may require one or more of the following special telescope features:

1. Using an exchangeable secondary support structure. This enables elimination of oversize secondaries and baffles that might be needed for other kinds of observation.
2. Using the secondary mirror as the aperture stop and making it sufficiently undersized that it cannot see past the rim of the primary mirror.
3. Putting all of the secondary support (except the struts) behind the mirror so that none of the hardware is visible to the detector.
4. Placing a specially shaped (e.g., conical) reflective plug at the center of the secondary to disperse radiation emitted from the central hole region of the primary mirror.

A basic technique for ground-based IR observing is that of background subtraction. This involves alternating the pointing direction of the telescope between the object (thus generating an object-plus-background signal from the detector) and a nearby patch of sky that has no apparent object (thus generating a background-only signal). Signal subtraction then eliminates the background signal common to both sky regions. Methods for alternating between positions include (1) driving the telescope between the two positions, usually at rates below 0.1 Hz, (2) wobbling the secondary mirror at rates between 10 and 50 Hz (called chopping), and (3) using focal plane modulators such as rotating aperture plates (called focal plane chopping). With solid state CCD detectors, one may simultaneously observe the object and a nearby background patch and electronically remove the measured background through appropriate processing of the data from individual pixels on the detector.

V. MECHANICAL CONFIGURATIONS

The mechanical portion of a mounted, ground-based telescope must support the optics to the required precision, point to and track the object being observed, and support

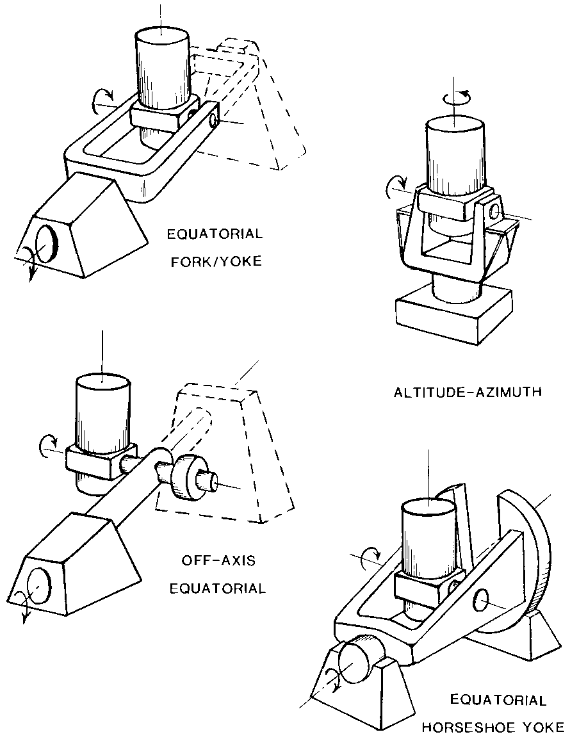


FIGURE 8 Basic telescope mounting styles in popular use. Numerous variants on each style are in existence.

the instrumentation in accessible positions. It is customary to distinguish between the telescope (or tube), which usually supports the imaging optics, and the mounting, which points the telescope and includes the drives and bearings. Tracking motion is usually accomplished by the mounting.

Most telescope mountings incorporate two, and occasionally three, axes of rotation to enable pointing the telescope at the object to be observed and tracking it to keep it centered steadily in the FOV. In general, the rotating mass is carefully balanced around each axis to minimize driving forces and the location of each rotation axis is chosen to minimize the need for extra counterweights.

Figure 8 illustrates the basic mounting styles that are discussed in the next section.

A. Mounting Designs

A hand-held telescope is supported and pointed by the user. The user is the mounting in this case. The mechanical mounting for a telescope performs essentially the same function, except that a mechanical mounting can support heavier loads and track the object more smoothly. As a rule of thumb, short-term tracking errors in high-quality telescopes are less than 10% of the smallest resolved object that can be observed with the telescope. Smoothness of rotation is important for long-term observations (i.e., no sudden movements) which mandates the use of high-quality bearings. Pressurized oil-film bearings (hydrostatics) are used in large telescopes for this reason.

Telescope drives range widely in style. The chief requirements are smoothness, accuracy, and the ability to move the telescope rapidly for pointing purposes (i.e., slewing) or slowly for tracking (i.e., at one revolution per day or less). Electric motor driven traction rollers, worm gears, or variants on spur gears are most commonly used. Position measuring devices (encoders) are often used to sense telescope pointing and to provide input data for automatic drive controls. Adjustments in tracking rates or pointing are accomplished either by manual control from the observer or, possibly, by star-tracking automatic guiding devices. Pointing corrections may also be based on data stored in a computer from mounting flexure and driving-error calibrations done at an earlier time. Telescope pointing accuracies to about 1 arcsec are currently possible with such corrections. Once the object is located in the FOV, the ability to track accurately is the most important consideration.

1. Equatorial Mounts

Astronomical telescopes ordinarily are used to observe stars and other objects at such great distances that they would appear stationary during an observation period if the earth did not rotate. Accordingly, the simplest telescope tracking motion is one that offsets the earth's rotation with respect to "fixed" stars (i.e., sidereal rate) and is done about a single axis parallel to the earth's north–south (N–S) polar axis. Equatorial mountings are those that have one axis of rotation (i.e., the polar or right ascension axis) set parallel to the earth's N–S axis. This axis is tilted toward the local horizontal plane (i.e., the ground) at an angle equal to local latitude. A rotatable cross-axis (also called declination axis) is needed for initial pointing and guiding corrections, but the telescope does not rotate continuously around this axis while tracking.

The varieties of equatorial mountings are limited only by the designer's imagination. The basic varieties, however, are the following:

1. Those that mount the tube to one side of the polar axis and use a counterweight on the opposite side to maintain balance. For an example, see Fig. 2. These are sometimes called off-axis or asymmetric mounts. It is also possible to mount a second telescope in place of the counterweight.

2. Those that support the tube on two sides in a balanced way to eliminate the need for a heavy counterweight. Yokes and forks are most commonly used, especially for larger telescopes. Theses are sometimes called symmetric

mounts. For an example, see Fig. 3 which shows a horseshoe yoke mount.

2. Other Mounting Styles

The alt–az mounting is configured around a vertical (azimuth) axis of rotation and a horizontal cross-axis (the altitude or elevation axis). The altitude–altitude (alt–alt) mounting, not widely used, operates around a horizontal axis and a cross-axis that is horizontal when the telescope points at the meridian and is tilted otherwise (similar to an English yoke with its polar axis made horizontal). With either of these styles, because neither axis is parallel to the earth's rotation, it is necessary to drive both axes at variable rates to track a distant object. Furthermore, the FOV appears to rotate at the focal region, which often necessitates a derotating instrument mounting mechanism, also moving at a variable rate. These factors inhibited the use of these mountings, except for manually guided telescopes, until the advent of computers on telescopes. Computers enable second-to-second calculation of the drive rates, which is required for accurate tracking. The alt–az configuration cannot track an object passing through the local zenith because the azimuth drive rate theoretically becomes infinite at that point. In practice, alt–az telescopes are operated to within about 1 deg of zenith.

The famous Herschel 20-ft telescope, built in England in 1783, was the first large alt–az telescope. Very few were built after that, but the trend today is toward alt–az mounts (see Section I.B and Table I). The ability to support the main azimuth bearing with a solid horizontal foundation is advantageous, as is the fact that the altitude axis bearings do not change in gravity orientation. These are important considerations when bearing loads of hundreds of tons must be accommodated. The alt–alt mounting is not as suitable for carrying heavy loads because the cross-axis is usually tilted with respect to gravity.

B. Telescope Tubes and Instrument Considerations

Design of the tube begins with the optical configuration. Tube structures are designed to maintain the optics in alignment, either by being stiff enough to prevent excessive deflections or by deflecting in ways that maintain the optics in the correct relative position. The well-known Serrurier truss first used on the Palomar 5-m telescope is a much-copied example of the latter (see Fig. 9). The tube structure is normally used to support the instruments at the focal positions, sometimes along with automatic guiders, field-viewing TV monitors, calibration devices, and field derotators. Large telescopes often do not use the Serrurier truss design specifically, but the flexure of the structure is usually designed to perform the same function. Modern computer-based structural analysis programs, developed since about 1960 and thus not available to Serrurier, provide the tools to create lighter, stiffer structures without ambiguity about the positions of the optics.

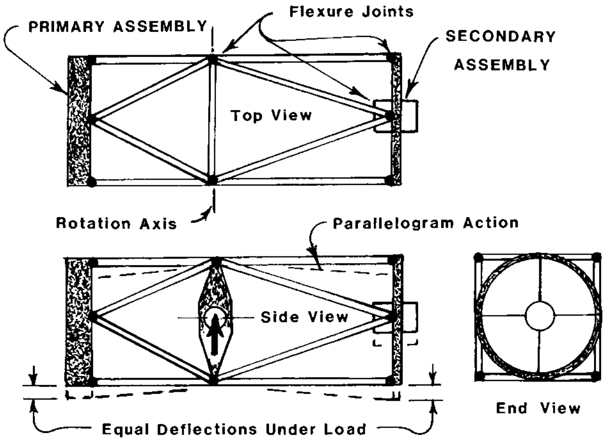


FIGURE 9 Serrurier truss used to maintain primary-to-secondary alignment as the tube rotates. Equal deflections and parallelogram action at both ends keep the optics parallel and equidistant from the original optical axis. Similar flexure is designed into most large telescope tube structures.

Focal positions (i.e., instrument mount locations) on the tube obviously move as the telescope points and tracks, which can be a problem for instruments at those locations that work poorly in a varying gravity environment. In those cases, one can divert the optical beam out of the telescope tube along the cross-axis to a position on the mounting or even outside the mounting. Flat mirrors are normally used for this purpose. To reach a constant-gravity position with an equatorial mounting, one must use several mirrors to bring the converging beam out: first along the declination axis, then the polar axis, and finally to a focus off the mounting. This is known as the coudé focus and is commonly used to bring light to spectrographs that are too large to mount on the telescope tube.

One can reach a constant-gravity focus (instrument location) on an alt–az telescope by simply diverting the beam along the altitude cross-axis to the mounting structure that supports the tube. This is called the Nasmyth focus after its Scottish inventor. The instrument rides the mounting as it rotates in azimuth but does not experience a change in gravity direction.

VI. CONSIDERATIONS OF USAGE AND LOCATION

Considering the precision built into most optical telescopes, one would expect them to be sheltered carefully. In practice, most telescopes must operate on high mountains, in the dark, and in unheated enclosures opened wide to the night sky and the prevailing wind. Under these conditions, it is not unusual to find dust or dew on the optics, a certain amount of wind-induced telescope oscillation, and insects crawling into the equipment. Certain insects flying through the light path can produce a noticeable amount of IR radiation. Observer comforts at the telescope are minimal.

In designing a telescope, one should consider its usage and its environment. A few general remarks in this direction are provided in the following sections.

A. Seeing Conditions

The seeing allowed by the atmosphere above the telescope is beyond ordinary control. Compensation may be possible as discussed in Section I.A. but the choice of site largely determines how good the imaging is. Seeing conditions in the region of 0.25 arcsec or less have been measured at certain locations, but more typically, good seeing is in the 0.5–1.0 arcsec range. Beyond 2–3 arcsec, seeing is considered poor. To the extent possible, one should build the telescope to produce images equal to or better than the best anticipated seeing conditions.

Locating the telescope at high altitudes usually reduces the amount of atmosphere and water vapor that is in the line of sight (important for IR astronomy), however, the number of clear nights and the locally produced thermal turbulence should also be considered. In many locations, a cool air layer forms at night near the ground which can be disturbed by the wind and blown through the line of sight. In other cases, warm air from nearby sources can be blown through the line of sight. In either case, telescope seeing is worsened.

Other seeing disturbances can originate inside the telescope enclosure. Any source of heat (including observers) is a potential seeing disturbance. Also, any surface that looks at the night sky, and hence is cooled by radiative exchange, may be a source of cooled air that can disturb seeing if it falls through the line of sight. If possible, it is desirable to allow the telescope enclosure to be flushed out by the wind to eliminate layers and pockets of air of different temperatures. Some telescope buildings have been equipped with air blowers to aid in the process, but dumping the air well away from the building has not always been possible even though it should be done.

The study of atmospheric seeing has become a relatively advanced science, and the telescope builder is well advised to consult the experts in choosing a site or designing an enclosure. Having chosen a site, one may be guided by the truism that seeing seldom improves by disturbing Mother Nature.

B. Nighttime versus Daytime Usage

With the advent of IR astronomy, optical telescopes began to be used both day and night because the sky radiation background is only slightly worse at IR wavelengths during the day compared with night. During the day it is much harder to find guide stars, and the telescope must often point blindly (and hence, more accurately) at the objects to be observed; but much useful data can be obtained. Some problems arise from this practice, however.

A major purpose of the telescope enclosure, other than windscreening, is to keep the telescope as close as possible to the nighttime temperature during the day so that it can equalize more rapidly to the nighttime temperature at the outset of the next night's observing. Obviously, this cannot be done if the telescope enclosure has been open during the day for observation. The condition is worsened if sunlight has been allowed to fall on the telescope during the day. Accordingly, optical/IR telescopes should be designed for rapid thermal adjustment.

Some of the design options in thermal control are (1) to insulate heavy masses that cannot equalize quickly, (2) to reduce weights and masses, (3) to provide good ventilation (i.e., avoiding dead air spaces that act as insulators), (4) to make surfaces reflective so that radiative coupling to the cold night sky is minimized, and (5) to isolate or eliminate heat sources. One should also consider using parts made from materials with low thermal expansion, but these have limited value if their heating effects are allowed to spoil the telescope seeing.

C. Remote Observing on the Ground

The traditional stereotype of an astronomer is a person perched on a high stool or platform, peering through the eyepiece and carefully guiding the telescope. The modern reality is likely to be quite different. Sophisticated electronic detectors replace the eye. Automatic star-tracking guiders take over the guidance chore. The astronomer sits in a control room sometimes far away from the telescope. A TV monitor shows the FOV or, at least, that part of the field not falling on the detector. A computer logs the data and telescope conditions. The telescope is not even seen by the astronomer: It can be in the next room or even a continent away if the communication link is properly established.

The advent of space-based astronomy clearly marked the time when the astronomer and the telescope were separated. The same separation is taking place in ground-based astronomy, albeit less dramatically. Numerous demonstrations have occurred during the 1970s and 1980s in which astronomers conducted observing runs on telescopes located at distant sites. In one case, the astronomer was in

Edinburgh, Scotland, and the telescope was in Hawaii. The "first light" pictures taken with the European Southern Observatory's 3.5-m NTT in Chile were transmitted to astronomers in Munich, Germany. In both cases the connection was through a communications satellite. This trend is likely to accelerate as the cost for such connections reduces and the data transmission rates increase.

VII. OBSERVING IN SPACE: THE HUBBLE SPACE TELESCOPE

The Hubble Space Telescope, or HST, named in honor of Dr. Edwin P. Hubble, was launched by the National Aeronautics and Space Administration (NASA) into orbit 614 km above the earth on April 24, 1990, from Space Shuttle "Discovery." Launch had been delayed more than 3 years following the loss of the Space Shuttle "Challenger" in 1986. Figure 10 illustrates the HST and its major components. Although many other smaller telescopes have previously been used in space, the HST represents such a large step in observing power and expected lifetime (15 years) that, for many astronomers, it marks the maturation of space-based astronomy. The discussion to follow is about the HST, but many of its operating characteristics are common to all telescopes operating in space.

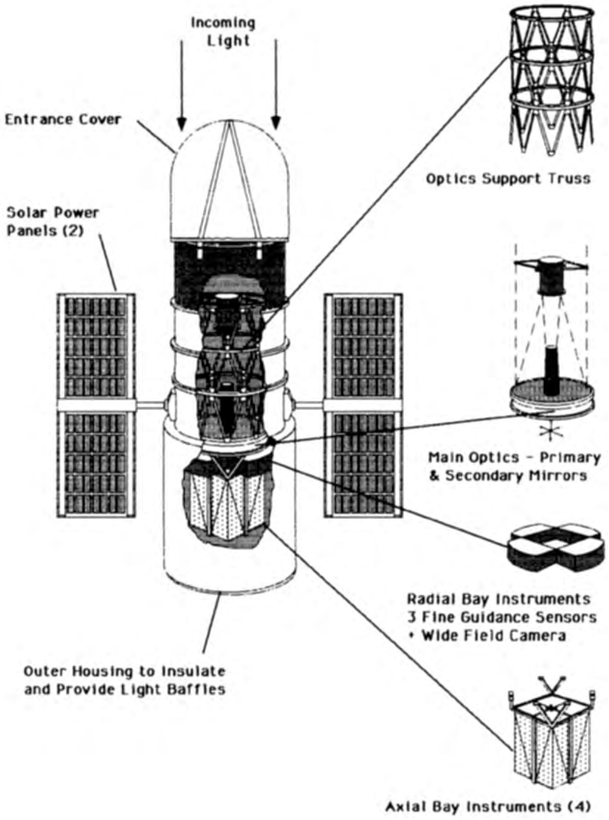

Incoming Light
Entrance Cover
Optics Support Truss
Solar Power Panels (2)
Main Optics – Primary & Secondary Mirrors
Radial Bay Instruments 3 Fine Guidance Sensors + Wide Field Camera
Outer Housing to Insulate and Provide Light Baffles
Axial Bay Instruments (4)

FIGURE 10 Schematic illustration of the Hubble Space Telescope (HST), showing its principal components. Light is focused behind the primary mirror and directed to the radial and axial bay instruments. Data are telemetered to the ground from antennas not shown. HST is optimized for UV region performance down to 0.12 μm wavelength and for observation of faint objects in both the UV and visible wavelength ranges.

The HST differs significantly from ground-based telescopes. It can observe throughout the UV region, its principal scientific justification, and it avoids the image distortions and "sky background" light emission due to the earth's atmosphere. Thus, very faint, small (presumably very distant) objects can be discerned that would be hopelessly lost in sky background seen from the ground. To produce reasonably good images at short UV wavelengths, the optics had to be substantially better than those in even the best telescopes on the ground, which results in excellent optical region images.

The HST can also observe in all regions of the IR, but its optics and structures are not cooled to minimize radiation emission, and its images are diffraction limited at IR wavelengths. In the late 1990s, NASA launched the "Space Infrared Telescope Facility," a 1-m class telescope that operates inside an enclosure cooled by liquid helium to avoid the excess radiation problem.

X-ray telescopes must operate in space like the HST, but their optics consist of near-conical shapes, appearing almost cylindrical in form, with the central axis pointed at the object. Light reflects at such steep incidence angles from these surfaces that it is not absorbed despite the high energy of the rays; hence it can be directed to a focus by properly shaping the optics.

HST operates under remote control in a hostile, high radiation, vacuum environment while producing its own power from solar panels. To avoid artificial light scattering, the nearby region of space must not be polluted with gases or debris from the telescope because no helpful breeze will blow it away. Since instrument changeovers are not possible except by visiting astronauts, HST must carry all of its instruments, essentially ready to work at all times. To maintain its orbital altitude, HST completes one "day-night" orbit of the earth every 96 min. Optical alignment must be maintained despite the rapid hot-to-cold-to-hot temperature cycling. The short observing periods, the multiplicity of instrumentation, the remote location, and the enormous number of objects to study produced a need to schedule astronomers differently than on the ground. For this purpose, a new organization, the Space Telescope Science Institute in Baltimore, Maryland, was formed in 1981 and is responsible for the scientific operation of the telescope.

The optical configuration of HST is "Ritchey-Chrétien" similar to many ground-based telescopes, with an 18 arcmin field of view. The mirrors are held in alignment by a special graphite-epoxy truss structure that is nearly dimensionally invariant under changing temperature conditions. Light from the focal plane can be diverted into four "radial bay" positions which contain three "fine guidance sensors" (FGSs) and the "wide-field camera" (WFC), an instrument for recording at all times a portion of any scene viewed by the telescope. By adjusting the pointing of the telescope, light from a chosen object can be directed into the entrance of one of four "axial bay" instruments directly behind the focal plane. When an object is being observed by an axial bay instrument, the WFC observes an adjacent region which astronomers hope will result in unexpected "serendipity discoveries." The axial bay contains the "faint object camera," two spectrographs (high and low resolution), and a photometer. All of these instruments are equipped with electronic detectors for converting light into transmittable telemetry data received on the ground by NASA's global network of radio receiving antennas.

The primary mirror blank for the HST is a sandwich structure made of Corning's "zero-expansion" ULE with an "egg-crate" style central core. It weighs about 20% as much as an equivalent solid disc. The low expansion properties of ULE enabled thin pieces to be assembled by flame welding and fusion bonding in a furnace without fear of thermal stress fracture and also avoids dimensional changes in the changing temperature environment of space. The primary mirror was later polished by the Perkin–Elmer Corporation to an accuracy enabling 80% of the focused test light (at 0.63 μm wavelength) to be concentrated into an image smaller than 0.1 arcsec across, very near the diffraction limit and an extraordinary achievement for such a large mirror.

When launched, the 2.4 m HST was the largest optical telescope designed for astronomy ever orbited and also the costliest ($2 billion). Built in the period 1977–1985, HST cost 200 times as much as typical ground-based telescopes of equivalent size built during the same period. Annual operating expenses are expected to be nearly 10% of construction costs, which makes it the world's most expensive observatory. Thus, the financial commitment by the U.S. government to HST represents major support of astronomy and was obtained only after a prolonged effort by hundreds of astronomers during the 1960s and 1970s.

One may imagine the dismay of all concerned when the HST was found soon after launch to have flawed imaging ability due to an incorrect curvature of its primary mirror. The images were blurred by spherical aberration 15 times greater than the specified 0.1 arcsec image resolution and no ground-controlled adjustment could eliminate the problem. Despite this enormous initial setback, HST has become operational. Corrective measures were taken in 1993 when a "second generation" instrument package containing a compensating lens was installed by astronauts. Until then, astronomers used a computer to subtract the aberration from the images, a tedious but workable process. Considering the level of quality of the optics when tested individually and the imaging problems discovered later, a few words about the mirror testing are appropriate.

Testing of the HST primary required compensation of the test results for the effects of gravity which the mirror, tested on the ground, would not experience in space. This was done successfully. Additionally, the mirror was optically tested using a "null lens," which is an optic used to make the mirror seem to be a sphere that is focusing light originating at or near its center of curvature, a position that can be located with precision. In the telescope, the mirror functions as a parabola focusing light originating at infinity, much like an automobile headlamp in reverse. Simulating these conditions in an optical shop is very difficult for large mirrors, so a null lens is used. However, any errors existing in the null lens remain in the testing data and cannot easily be separated from errors in the mirror. Accordingly, opticians generally perform several different tests for comparison, searching for commonality in the results. One element of the "most accurate" HST null lens was incorrectly positioned, which caused the opticians to polish the mirror to a different curvature than was needed to function properly with the secondary mirror. The disagreement with other tests (made with other, less accurate null lenses) was discounted, and no shop tests were made with both the primary and secondary mirrors working together (as is typical for most telescopes). Unfortunately, the images formed in space by the two-mirror system for HST did contain considerable spherical aberration (the less accurate tests were right), and correction will require a new secondary mirror or an additional correcting lens. Repolishing the primary mirror would be much more expensive.

Pointing the HST, or any orbiting telescope, to a new object for observation and tracking on it steadily and accurately is challenging. Changing position is accomplished by a "reaction motor" on each of three coordinate axes. Powered acceleration of the motor rotor produces a reactive torque on the telescope, causing it to rotate in the opposite direction. On reaching the desired rotational velocity, the telescope and rotor "coast." Motion is stopped by applying power to the rotor to slow its rotation. This "slewing" action coarsely positions the telescope, enabling a pair of "fixed head" star trackers (small guide telescopes with relatively large fields of view) to identify a pair of preselected, bright, guide stars. Next, the three "fine guidance sensors" (FGS), receiving light from the main focal plane as mentioned earlier, begin searching until two have "locked on" to suitable guide stars, presumably faint, but near the object to be observed and known in position to high accuracy. After that, signals from the FGSs are used to control the reaction motors for tracking. The system is designed to control the 12-ton HST to a pointing accuracy of about 0.01 arcsec, roughly 10 times more accurate than typical ground-based telescope performance and all done without a solid foundation to react against.

Only two FGSs are needed to guide the HST. The third FGS can be used to measure star positions (i.e., to do astrometry) after the other two have locked onto stars with known position coordinates. Thus, HST is equipped to upgrade its own guide-star catalog to a precision greater than is possible from the ground. The process of pointing to a new object and locking onto guide stars requires upward of 30 min, a large penalty in observing time but a toll that must be paid to operate in space. Ground-based telescopes typically perform the same operations in less than 5 min.

The HST is an extraordinary telescope and may indeed prove to be the beginning of the era in which most professional astronomy is done in space. Dark skies are becoming hard to find on the ground, and ingenious corrections for the atmosphere will never be as good as avoiding the problem altogether, not to mention the loss of wavelength coverage on the ground. Many other orbiting telescopes are in progress, and plans already exist at NASA for a 10-m class optical space telescope, plus much larger telescopes as to operate in the millimeter and longer wavelength range.

Nevertheless, most professional astronomy is still done on the ground, as the formidable costs of working in space, exemplified by HST, must be justified, presumably by new discoveries. The location and special abilities of the HST (when its optics are finally corrected) assures this possibility. The worth of new discoveries is impossible to measure in advance, but it seems that a bright future for humanity depends upon them. Galileo would agree.

SEE ALSO THE FOLLOWING ARTICLES

ELECTROMAGNETICS • GAMMA-RAY ASTRONOMY • INFRARED ASTRONOMY • OPTICAL DIFFRACTION • ULTRAVIOLET SPACE ASTRONOMY

BIBLIOGRAPHY

Barlow, B. V. (1975). “The Astronomical Telescope,” Wykeham, London.

Bell, L. (1981). “The Telescope,” Dover, New York.

Burbidge, G., and Hewitt, A. (eds.) (1981) “Telescopes for the 1980s,” Annual Reviews, Palo Alto, CA.

Driscoll, W. G., and Vaughan, W. (eds.) (1978). “Handbook of Optics,” McGraw-Hill, New York.

King, H. C. (1979). “The History of the Telescope,” Dovehhr, New York.

Kingslake, R. (1983). “Optical System Design,” Academic Press, Orlando, Florida.

Kuiper, G., and Middlehurst, B. (eds.) (1960). “Telescopes,” Stars and Stellar Systems, Vol. 1. University of Chicago Press, Chicago.

Learner, R. (1981). “Astronomy through the Telescope,” Van Nostrand Reinhold, New York.

Marx, S., and Pfau, W. (1982). “Observatories of the World,” Van Nostrand Reinhold, New York.

Schroeder, D. J. (1987). “Astronomical Optics,” Academic Press, San Diego.

Ultraviolet Space Astronomy

George R. Carruthers
Naval Research Laboratory

GLOSSARY

Blackbody radiator Object that absorbs or emits radiation with 100% efficiency at all wavelengths and whose emission spectrum follows the laws derived by Max Planck and others.

Extinction Attentuation of radiation by the processes of scattering, pure absorption, or both.

Extreme ultraviolet Wavelength range below 1000 Å, extending to the X-ray wavelength range below 100 Å.

Far ultraviolet Wavelength range below 2000 Å, usually the range 1000–2000 Å.

Flux distribution Radiated energy, in energy units or number of photons per unit area or for the total object, per second vs wavelength in a region of the electromagnetic spectrum.

Ionosphere Region of a planetary atmosphere in which a significant portion of the atoms or molecules are ionized (i.e., stripped of one or more electrons).

Middle ultraviolet Wavelength range 2000–3000 Å.

Resonance transition Spectral transition of an atom, molecule, or ion that is between the ground state (state of lowest energy) and an excited state (of higher energy), with corresponding emission or absorption of a discrete wavelength of electromagnetic radiation.

Subordinate transition Spectral transition between two excited states.

ULTRAVIOLET (UV) SPACE ASTRONOMY is the study of extraterrestrial objects in the UV wavelength range of the electromagnetic spectrum. This portion of the spectrum provides information unavailable in other wavelength ranges, particularly on planetary and stellar atmospheres and the interstellar medium. However, because the UV spectrum is inaccessible to ground-based telescopes, UV astronomy must be conducted with instruments on rockets, Earth satellites, and other space vehicles outside the Earth's atmosphere.

I. SIGNIFICANCE OF THE RESEARCH

One of the primary benefits of doing astronomy in space rather than from ground-based observatories is the much wider range of the electromagnetic spectrum that is accessible to observation (Fig. 1). In particular, the entire UV and X-ray wavelength range below 300 nm (3000 Å) is absorbed by oxygen and ozone in the Earth's atmosphere and, hence, is totally inaccessible to even the largest ground-based telescopes. To observe the sun, stars, and other celestial objects in this wavelength range, the instruments must be carried above the absorbing atmosphere by means of sounding rockets or space vehicles.

The UV spectral range is of importance to astronomy for a number of reasons. Two of the most significant are as follows: (1) the primary, or *resonance*, transitions (those involving the ground state) of the most common atoms, ions, and molecules occur in the UV spectral range below 300 nm; (2) very hot stars, having surface temperatures in excess of 10,000 K (vs ~5800 K for our sun), emit much or most of their radiation in the ground-inaccessible UV.

Hydrogen and helium are the most abundant elements in the universe, accounting for more than 99% of all atoms in stars, the interstellar medium, and the giant planets Jupiter and Saturn. Of the heavier elements, the most abundant in

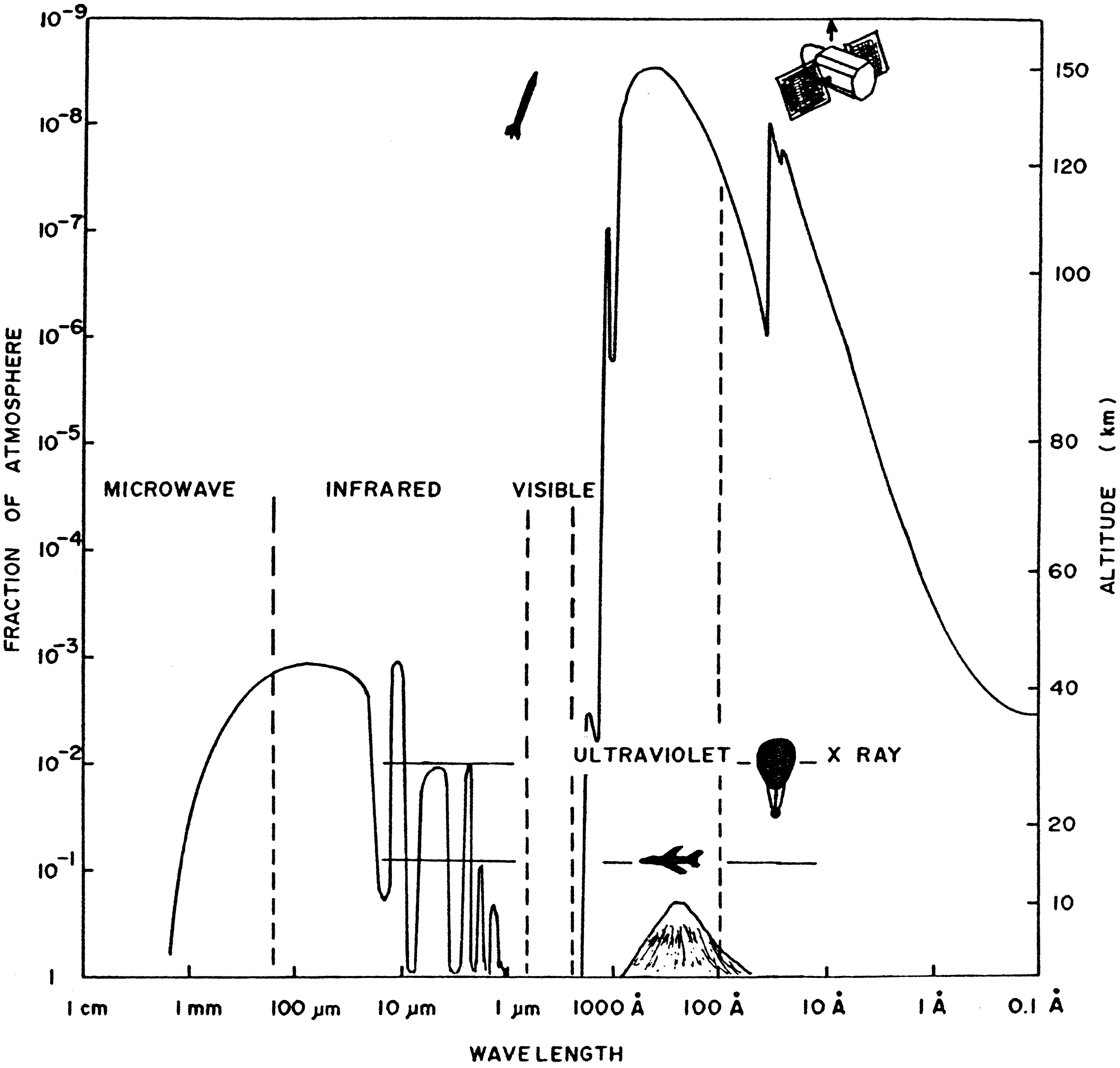


FIGURE 1 Altitude and fraction of atmosphere remaining overhead that must be reached in order to observe one-half of the radiation coming from an extraterrestrial source vs wavelength in the electromagnetic spectrum.

numerical order are oxygen, carbon, neon, and nitrogen. For all of these, in atomic or ionic form or in simple molecules such as H_2, N_2, and CO, the resonance spectral transitions occur in the far UV (below 200 nm). The resonance transitions (to or from the ground state) are usually much stronger and more quantitatively useful for remote sensing of gas composition and physical state than are the subordinate transitions (between two excited states). This is particularly true for cool material, such as in the interstellar medium and in planetary or cometary atmospheres, where most of the atoms or molecules reside in the ground state. Therefore, observations in the far UV provide information on the composition and properties of astronomical objects that cannot be obtained at ground-accessible wavelengths.

Stars that are of much higher temperature than our sun (ranging from ∼10,000 K to more than 200,000 K) emit much or most of their radiation at wavelengths below the 300-nm limit of ground-based observations. Therefore, it is difficult to determine, from ground-based observations alone, the total energy outputs and energy flux distributions of these stars. Although the flux distributions of hot stars have crude resemblances to those of classical blackbody radiators (in that the peak of the distribution shifts to shorter wavelengths as the temperature increases), more detailed theoretical models predict significant differences from blackbody distributions, particularly at the shortest wavelengths. Comparisons of observations with theory and accurate determinations of the "effective" temperatures of hot stars therefore require accurate measurements in the far UV.

Even for cooler stars, such as our sun, UV measurements are important for detecting and characterizing energetic phenomena that occur in their outer atmospheres. For the sun, these include the high-temperature solar corona, solar flares, and other manifestations of solar activity. Also, solar UV radiation is largely responsible for the maintenance of the terrestrial and other planetary ionospheres and for exciting UV emissions observed in planetary atmospheres and in comets.

II. INSTRUMENTATION FOR ULTRAVIOLET SPACE ASTRONOMY

The instrumentation used for UV space astronomy is, in most respects, similar to that used in ground-based astronomy. The differences have to do, primarily, with the following: (1) UV radiation is not transmitted as efficiently by, or by as wide a variety of, transparent materials (used for windows, lenses, and other refractive optical elements) as are visible and near-infrared radiation; and (2) reflective coatings for mirrors and reflection gratings are not as efficient (especially at wavelengths below 120 nm) than at longer wavelengths.

Categories of instruments can be defined by the types of measurements to be made, such as imagery, photometry (measurement of light intensity), spectroscopy, and spectrophotometry (measurement of intensity vs wavelength in a spectrum). There is no sharp distinction between these categories, and they are not mutually exclusive; it is mainly a matter of emphasis. For example, imagery emphasizes high spatial resolution and simultaneous coverage of a wide field of view (in comparison to the size of the resolution element), but photometric information can also be obtained from images. Traditional spectroscopy emphasizes the detection, identification, and wavelength measurement of spectral features, whereas spectrophotometry emphasizes measurement of the intensity of spectral features or intensity vs wavelength in a continuum. Both can be accomplished with the same instrument, if properly designed. Imagery in a very narrow wavelength range (e.g., by the use of interference filters) can be considered a form of spectroscopy or spectrophotometry, and "imaging spectrographs" or "hyperspectral imagers" are those that retain spatial intensity resolution in one dimension while providing spectral intensity resolution in the transverse dimension.

As is true in ground-based astronomy, the type of instrumentation used depends to some extent on the object of study; for example, solar studies require different instrumentation than do observations of stars or nebulae.

Since UV astronomy can be done only at very high altitudes or in space, its progress has been paced by the development of rocket vehicles for carrying the instruments above the atmosphere and other payload equipment such as pointing controls for directing the instruments at the objects of interest with adequate accuracy and stability. The rate of progress in the various subfields of UV astronomy has, therefore, depended on both instrumentation and supporting vehicular developments.

Solar UV astronomy was the first subfield of UV astronomy to be developed; because the sun is so much brighter than any other astronomical object, the requirements regarding both scientific instrumentation and vehicle attitude control systems were less stringent than for observations of other celestial objects. Solar UV astronomy began in the late 1940s with experiments flown by Naval Research Laboratory scientists on captured German V-2 rockets. The initial instruments were unpointed, but nevertheless returned new and useful information on the intensity and spectral distribution of solar UV radiation. In the early 1950s, the development of instrument pointing controls for Aerobee sounding rockets imparted greater sensitivity and higher spectral resolution to UV spectroscopy and photometry.

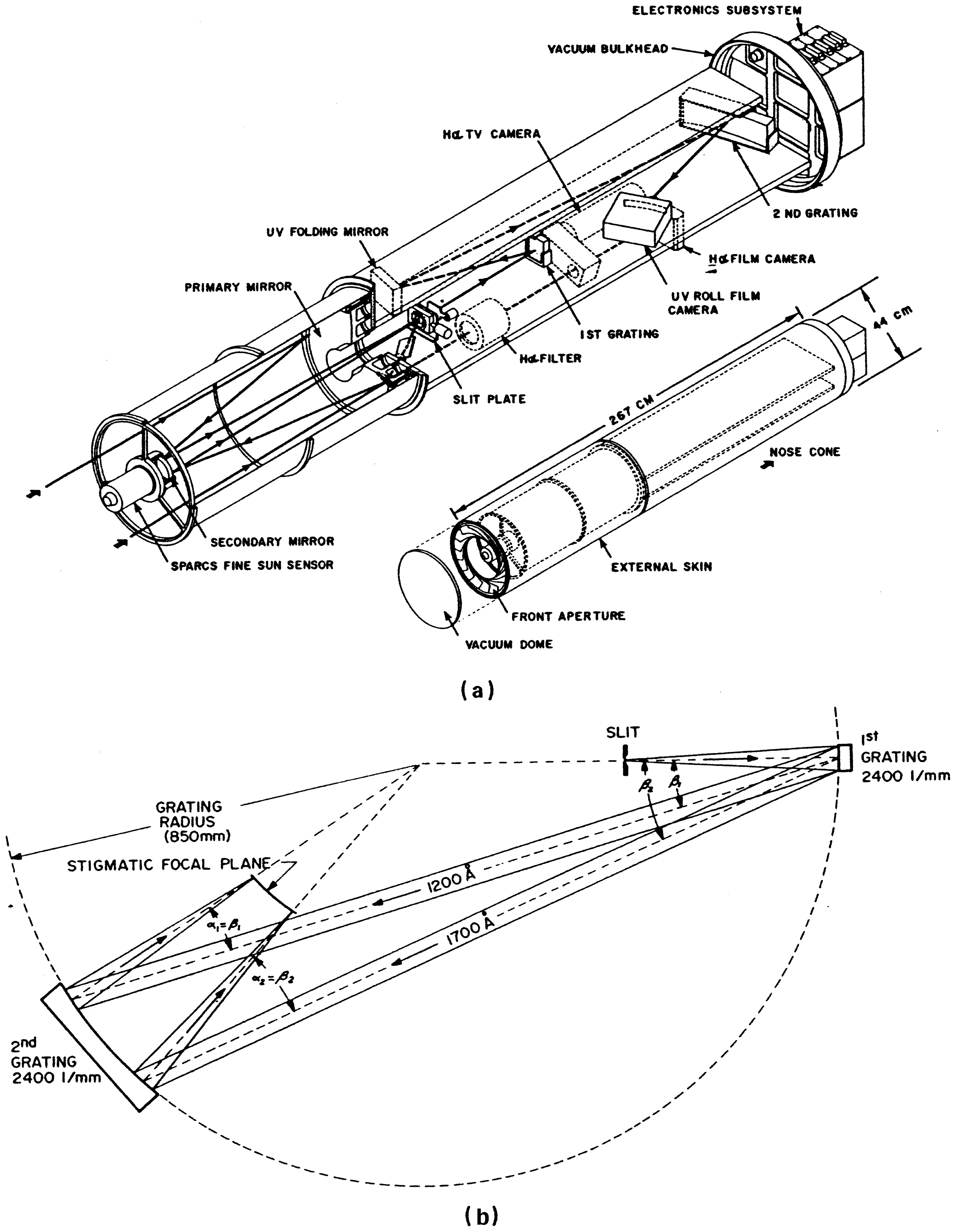


FIGURE 2 Diagram of the Naval Research Laboratory's high-resolution telescope and spectrograph used in solar studies from rockets and *Spacelab 2*. (a) Diagram of the complete telescope and spectrograph. The optical parameters are as follows. Spatial: field of view, $0.5'' \times 16'$; resolution, $<1''$. Spectral: range, 1175–1715 Å and Hα; resolution, 0.05 Å. (b) Simplified schematic of the spectrograph. (Naval Research Laboratory illustrations.)

The first measurements of stellar UV radiation were made from free-spinning Aerobee rockets in the late 1950s; the development of inertial attitude control systems in the early 1960s greatly improved the quality of data return. The first satellite vehicles dedicated to space astronomy were the *Orbiting Solar Observatories* (OSOs), the first of which was launched in 1962, and the *Orbiting Astronomical Observatories* (OAOs), the first successful launch of which was in 1968. The advent of satellite observatories tremendously increased the available observing time, in comparison to a typical 5-min sounding rocket flight. This allowed much more comprehensive studies of the sun, stars, and other celestial objects than possible with sounding rockets. However, the latter still remained useful for carrying out special-purpose investigations not suitable for the long-duration (but technologically more difficult) satellite observatories.

Some of the unmanned satellites primarily dedicated to UV solar or celestial astronomy include the *Solar Maximum Mission* (SMM), the *International Ultraviolet Explorer* (IUE), the *Extreme Ultraviolet Explorer* (EUVE), the *Far Ultraviolet Spectroscopic Explorer* (FUSE), and the *Solar Heliospheric Observatory* (SOHO). Ultraviolet studies of the sun are also a significant part of the *Upper Atmospheric Research Satellite* (UARS) mission, and UV astronomy is an important part of the operations of the *Hubble Space Telescope*.

Ultraviolet astronomical observations have also been carried out in several manned missions, including *Gemini*, *Apollo*, and *Skylab*, as well as many space shuttle missions. These latter include solar observations, beginning with the *Spacelab-2* shuttle flight in 1985, and celestial observations with the *Astro* UV astronomy missions flown in 1990 and 1995. In addition, the space shuttle was used in 1990 as the launch vehicle for the *Hubble Space Telescope* (HST), a much larger and more comprehensive space observatory, which still remains operational in orbit. The HST, unlike previous space astronomical observatories, was designed so that it could be revisited by the space shuttle for repairs and upgrades (in extravehicular activities by the space shuttle crew members); such missions were carried out in 1993, 1997, and 1999. The space shuttle also made upgrades and repairs on the *Solar Maximum Mission* satellite, initially launched on an unmanned *Delta* rocket. In addition to these, numerous other solar and celestial UV astronomy experiments have been conducted in space shuttle flights, using instruments carried in the payload bay of the space shuttle, or as short-duration, free-flying payloads on carriers such as *Spartan* (developed by NASA's Goddard Space Flight Center) and *AstroSPAS* (developed by the European Space Agency).

There are a large variety of instruments used in UV space astronomy, and, in general, they are similar to instrumentation used in ground-based astronomy. In the UV, the number of optical components (mirrors, gratings, etc.) must be kept to a minimum because of the generally lower efficiencies of optical elements in the UV as compared with the visible. However, instruments now in use or planned for near-future missions in UV astronomy are fully comparable in sensitivity, resolution, and other performance capabilities to similar visible-light instruments.

In UV astronomy, as in visible-light astronomy, photographic film was initially a major recording technique for imagery and spectroscopy, but it has now been almost totally replaced by electronic imaging detectors. These include image intensifiers and electrographic detectors (having final images recorded on film) and devices whose final output is an electronic signal, such as charge-coupled devices (CCDs) and detectors based on microchannel-plate (MCP) electron-multiplier arrays. Electronic imaging detectors can provide much higher sensitivity than provided by direct recording on photographic film. They also can be made sensitive *only* to UV wavelengths, which is particularly important for studies of objects like the sun, which are much brighter in the visible than in the far UV. In addition, detectors whose final output is an electronic signal can send their data back to Earth by radiotelemetry, eliminating the need to recover and process film (hence making them applicable to long-duration satellite and deep-space missions). Another potential advantage of electronic detectors vs photography is that the information is acquired in a more quantitative form, making it more readily and accurately processed and calibrated.

We show here some examples of recent or current UV astronomy instruments. Figure 2 shows the Naval Research Laboratory's High-Resolution Telescope and Spectrometer (HRTS), which has flown on several sounding rocket flights and on *Spacelab-2*. An improved version of this instrument uses a CCD detector in place of the photographic film used previously. The *Astro* Spacelab shuttle payload consisted of three separate instruments: the Johns Hopkins University's *Hopkins Ultraviolet Telescope* (HUT), shown in Fig. 3, was used for spectroscopic measurements in the 85- to 185-nm wavelength range; the Goddard Space Flight Center's *Ultraviolet Imaging Telescope* (UIT), shown in Fig. 4, was used for direct imaging of celestial objects in far- and middle-UV wavelength ranges; and the University of Wisconsin's UV photopolarimeter instrument (WUPPE) was used for studies of stars and interstellar material in the 140- to 320-nm wavelength range.

A number of UV imaging and spectrographic instruments have been or will be used on the currently orbiting HST, launched in 1990 (see Fig. 5). The HST has a collecting mirror aperture of 2.4 m diameter, more

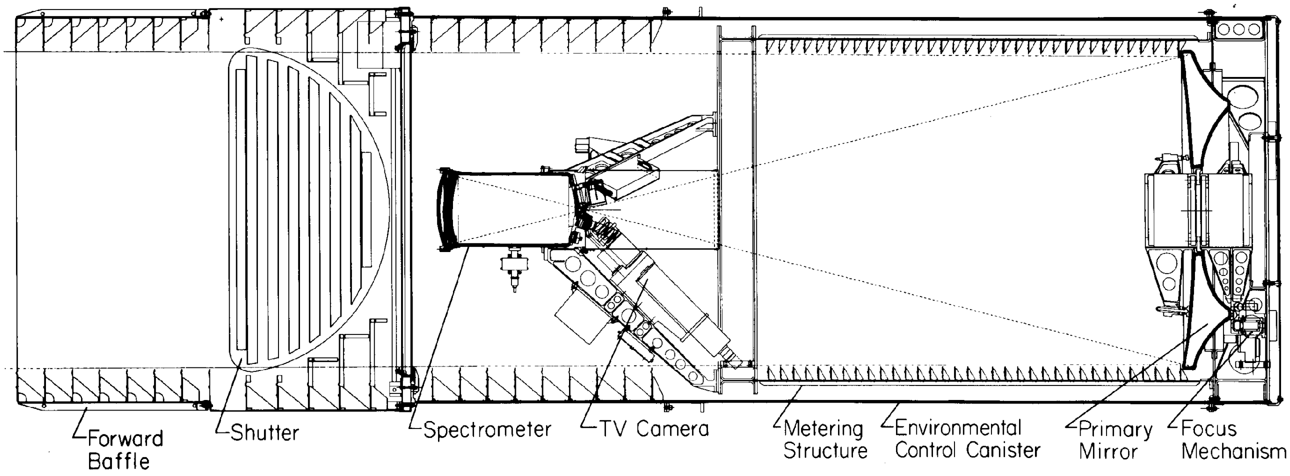


FIGURE 3 Diagram of the *Hopkins Ultraviolet Telescope* (HUT), which was part of the shuttle-based *Astro* Spacelab missions in 1990 and 1995. (Courtesy of A. Davidsen, Johns Hopkins University, Baltimore, MD.)

than twice that of the largest previous space astronomy telescope. It operates in the UV, visible, and near-IR wavelength ranges, and of particular importance is that it has nearly diffraction-limited imaging performance (better than 0.1 arcsec resolution, or about 10 times better than typically achieved, in the visible, with ground-based telescopes). This provides not only improved imaging resolution, but also better senstivity in direct imaging,

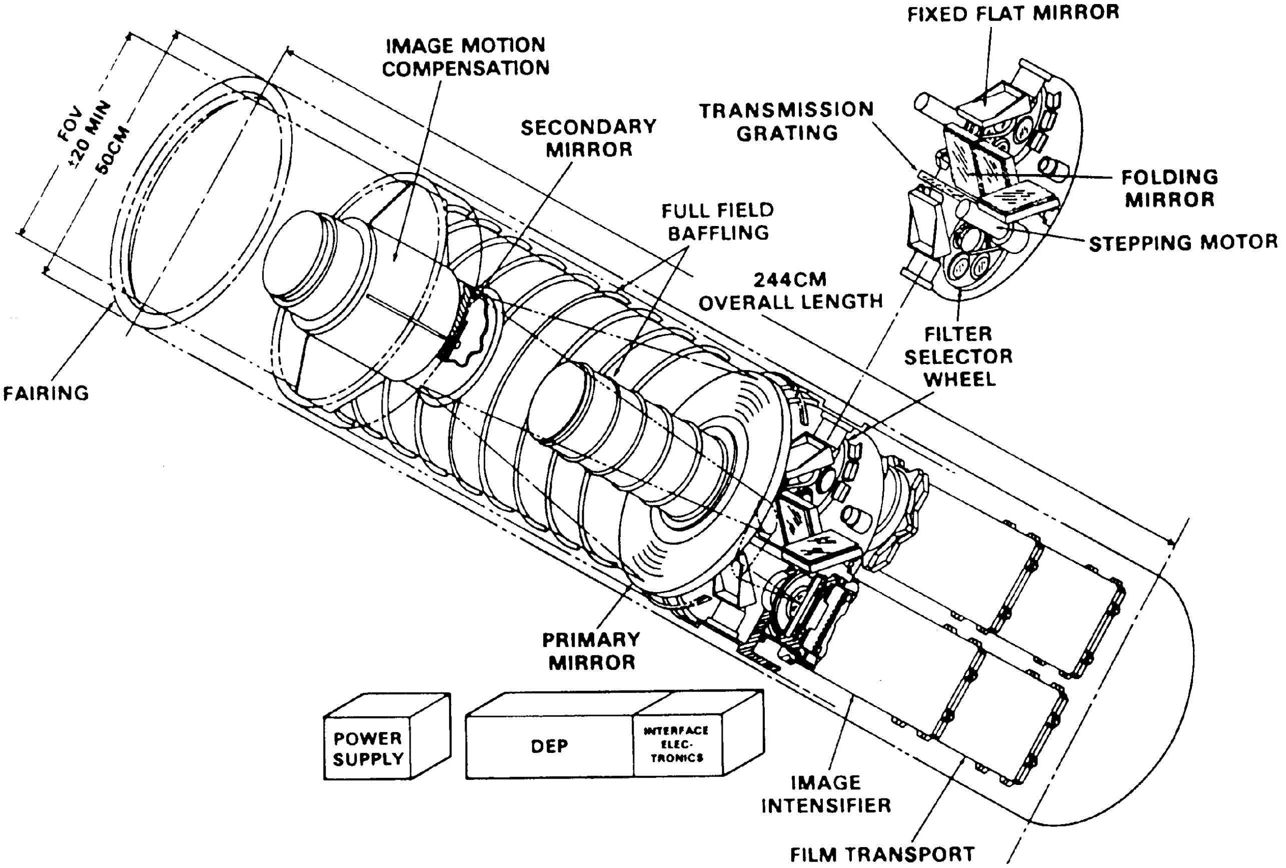


FIGURE 4 Diagram of the *Ultraviolet Imaging Telescope* (UIT), which was part of the shuttle-based *Astro* Spacelab missions. (Courtesy of T. Stecher, NASA GSFC.)

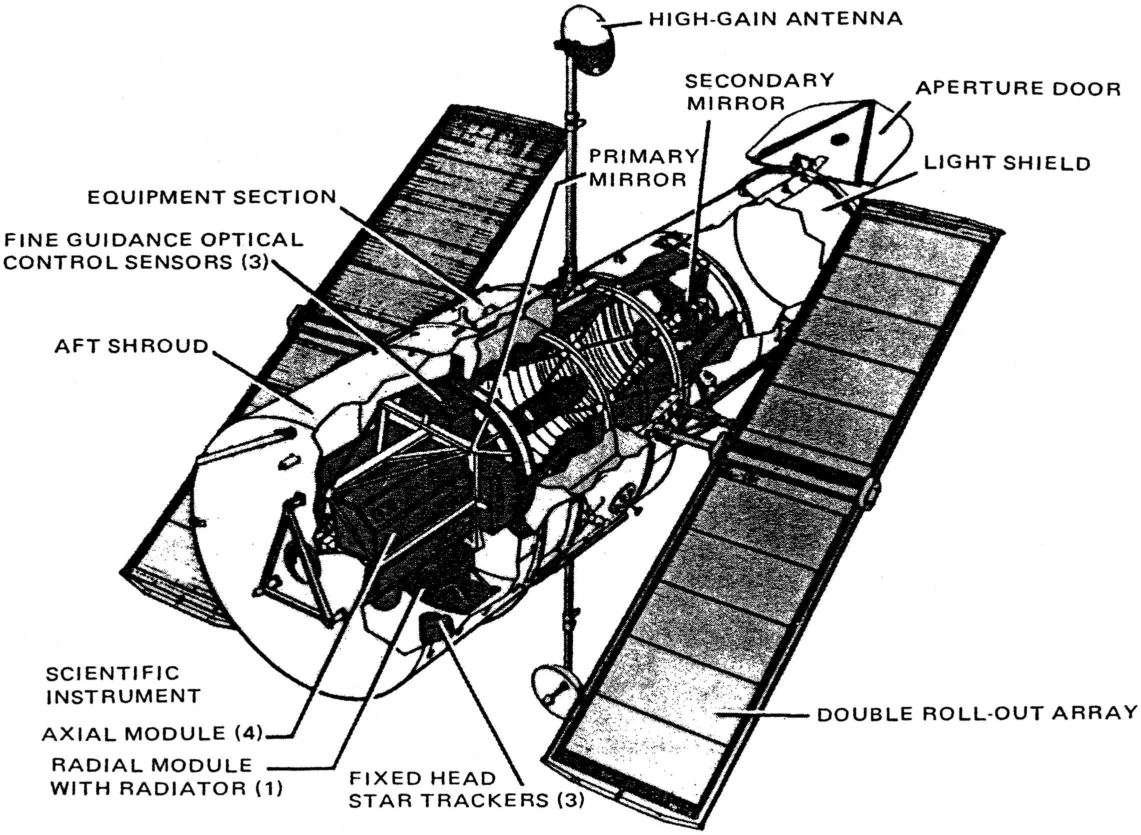


FIGURE 5 Diagram of the 2.4-m-aperture *Hubble Space Telescope*, planned for launch by the space shuttle. The scientific instruments include imaging cameras, a high-speed photometer–polarimeter, a high-resolution spectrograph, and a faint-object spectrograph. (NASA photograph.)

and imaging spectrometry, of compact and point sources (such as stars).

The HST is also unique among long-duration space astronomy missions, in that it was designed so that its instruments could be changed on-orbit, in visits by the space shuttle. This has allowed the "first-generation" instruments included in the original launch to be replaced with improved and more advanced instruments, which have in turn improved and broadened the HST's scientific capabilities. For example, the first-generation *Faint Object Spectrograph* (FOS) and *Goddard High Resolution Spectrograph* (GHRS) were replaced by the *Space Telescope Imaging Spectrograph* (STIS) in a 1997 space shuttle refurbishment mission, and a new, improved far-UV *Cosmic Origins Spectrograph* (COS) is planned for installation in a near-future HST refurbishment mission.

III. THE SUN

The sun is the nearest star, the one of most practical importance to us, and the only one whose surface features can be observed in detail. The visible "surface" of the sun, known as the photosphere, emits a continuous spectrum of radiation whose intensity and spectral distribution crudely resemble those of a classical blackbody radiator having a temperature of ~5800 K. The intensity of solar radiation is a maximum in the visible (near 500 nm wavelength) and decreases rapidly toward shorter wavelengths and less rapidly toward longer wavelengths. Thus, in the UV (particularly in the far UV, below 200 nm) the solar photosphere is far less bright than in the visible.

However, the outer atmospheric layers of the sun, above the photosphere, have much higher gas temperatures than does the photosphere. As seen in Fig. 6, the temperature rises very sharply in a "transition region" between the lower atmosphere (chromosphere) and the upper atmosphere (corona), from ~10,000 K in the upper chromosphere to more than 1,000,000 K in the corona. Highly active regions in the solar atmosphere, such as solar flares, can have much higher temperatures still. However, because of the very low gas densities in these regions, the energy content per unit volume is still very much less than

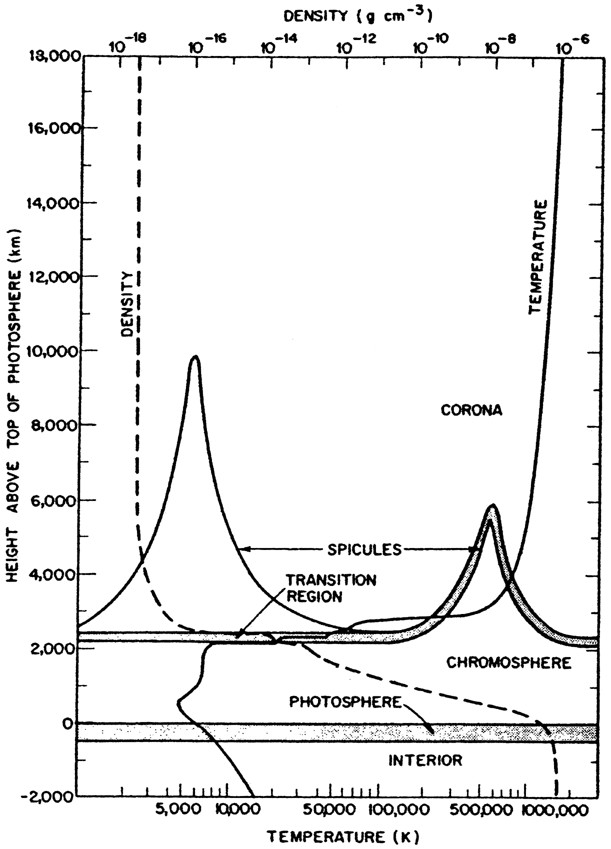


FIGURE 6 Temperature and density in the solar atmosphere vs altitude. Note that a minimum temperature is reached just above the visible-light surface, or photosphere, and temperature rises sharply in the transition region between the chromosphere and corona. [Adapted from Eddy, J. A. (1979). "A New Sun," NASA SP-402. U. S. Govt. Printing Office, Washington, DC.]

in the photosphere, and even the most intense flares contribute very little to the visible brightness of the sun.

One of the most important current problems of solar physics is to explain how the outer atmosphere of the sun can be maintained at million-degree temperatures in contact with a 6000-K photosphere. Related problems include elucidation of the mechanism for the production and acceleration of the solar wind, and of the mechanism of, and physical processes occurring in, solar flares.

In the visible and through the near and middle UV, the solar spectrum is a continuum with superimposed absorption lines (Fraunhofer lines), due to cool gas in the upper photosphere and lower chromosphere (the "temperature minimum" region in Fig. 6). In the far UV, however, the continuum fades away and is replaced by an emission line spectrum, produced by the hotter gas in the transition region and corona.

It is particularly noteworthy that, in the far and extreme UV, the outer atmosphere of the sun is the predominant source of radiation, due to its very high temperature. This constitutes one of the major advantages of short-wavelength observations of the sun: the ability to observe the high-temperature, active regions of the outer atmosphere without interference from the cooler (but, in the visible, far brighter) photosphere. Although (with the exception of sunspots) the solar disk appears almost featureless in white light, observations in far-UV emission lines [and also in narrow wavelength ranges centered on visible absorption lines, such as hydrogen Balmer α (Hα) at 656.3 nm] reveal a great deal of detail and time variability in the structure of the solar atmosphere. Different emission or absorption features are produced by gas at different temperatures and, hence, at different spatial locations in the solar atmosphere. Therefore, observations from the visible to the X-ray wavelength range are needed to characterize the temperature, density, and time variations in the solar atmosphere. For example, imagery of the sun in the light of neutral hydrogen (Lyman α at 121.6 nm) reveals gas having a characteristic temperature of ~10,000 K, whereas the light of ionized helium (He II) at 30.4 nm is emitted by gas with a temperature of ~80,000 K. Five-times-ionized oxygen (O VI) emission at 103.2 and 103.6 nm reveals material at a temperature of ~300,000 K. Nine-times-ionized magnesium (Mg X) emission near 62.5 nm is characteristic of material at a temperature near 1,600,000 K.

Studies of the sun in the UV and X-ray wavelength ranges are also of practical importance, because these radiations significantly influence Earth's upper atmosphere and are primarily responsible for the production and

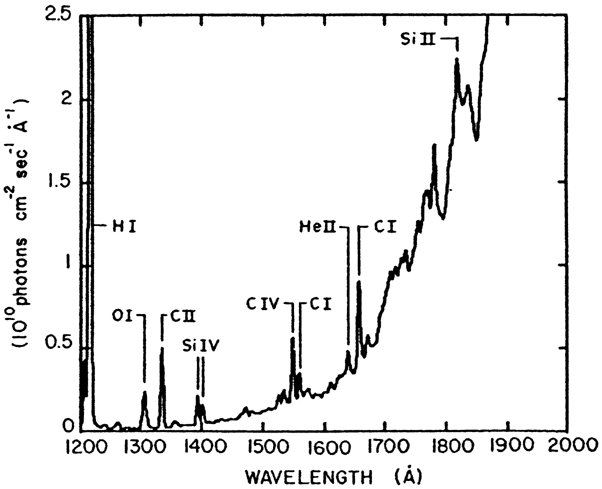


FIGURE 7 Far-UV spectrum of the sun, obtained in a sounding rocket flight, showing the transition to an emission line spectrum below ~1800 Å. The effective resolution is 5 Å. (Courtesy of G. H. Mount, Naval Research Laboratory.)

maintenance of the ionosphere (which is essential to long-distance radio communications). Ultraviolet radiation below 200 nm dissociates molecular oxygen in the upper atmosphere, indirectly resulting in the formation of ozone (O_3), which protects life on Earth from the biologically harmful middle-UV solar radiation. Lyman α radiation ionizes nitric oxide (NO), producing the lower ionosphere, whereas radiations of wavelengths less than 102.6, 91.1, and 79.6 nm ionize molecular oxygen, atomic oxygen, and molecular nitrogen, respectively.

As mentioned previously, the sun was the first extraterrestrial object to be observed in the ground-inaccessible UV, in part because of its brightness (even in the far UV, much greater than that of any other celestial object) and its relative ease of acquisition by primitive rocket pointing control systems. The first rocket observations were exploratory in nature, attempting to define the general nature and intensity distribution of the solar UV spectrum. Successive experiments gradually improved the spectral resolution and photometric accuracy of the measurements and extended them toward shorter wavelengths. Also, once strong emission lines (such as hydrogen Lyman α at 121.6 nm and ionized helium at 30.4 nm) were identified, imaging instruments (such as spectroheliographs) were used to obtain monochromatic images of the entire solar disk in these emissions.

The advent of long-duration satellite, *Skylab,* and space shuttle-based, *Spacelab*, observations has allowed not only more detailed measurements, but also studies of the time variations of the solar UV output and its correlations

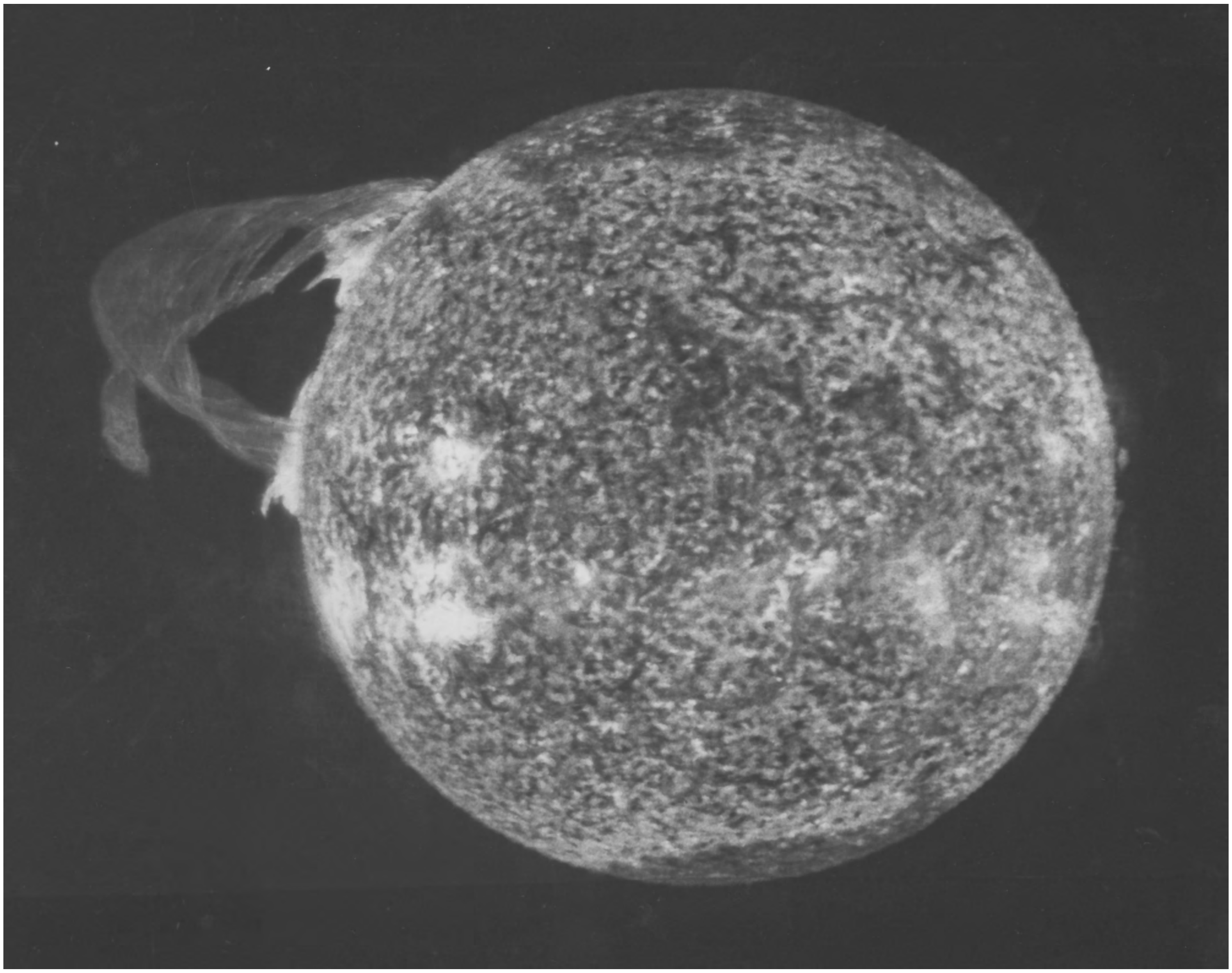


FIGURE 8 Image of the sun in the light of ionized helium (He II) emission at 304-Å wavelength, showing mottled structure of the transition region, active regions associated with sunspots, and a giant eruptive prominence (at left). Obtained with the Naval Research Laboratory extreme-UV spectroheliograph from *Skylab*. (Naval Research Laboratory photograph.)

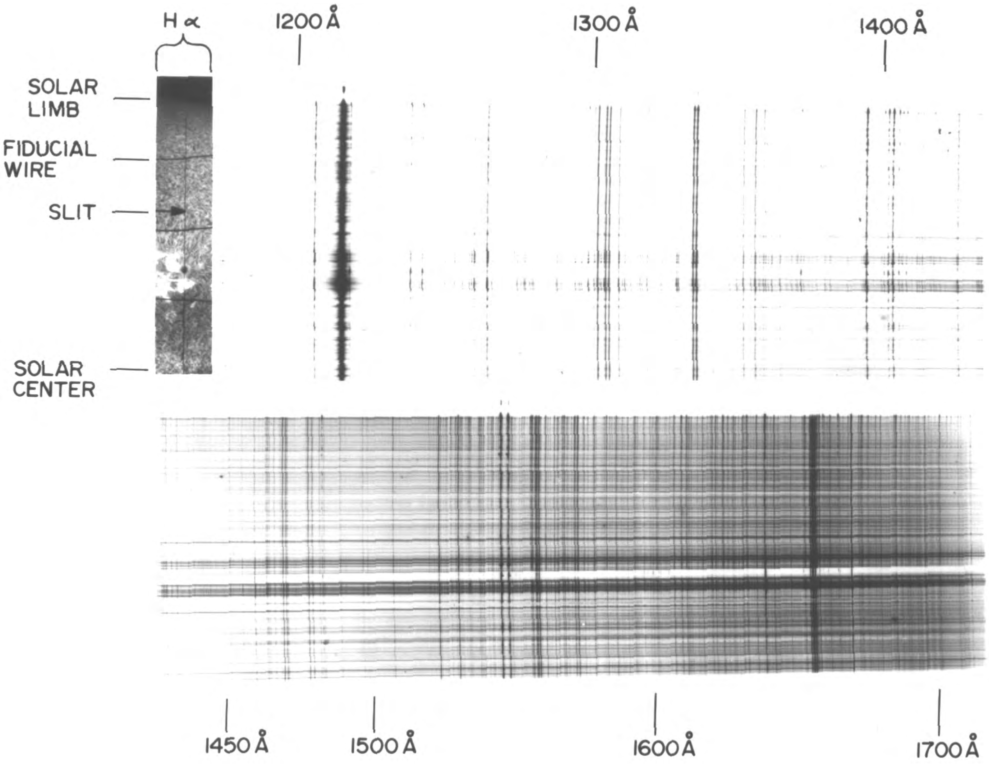


FIGURE 9 Spectrum of the sun obtained with the Naval Research Laboratory high-resolution telescope and spectrograph in a sounding rocket flight. This stigmatic (imaging) spectrograph reveals spatial variations across the disk of the sun (top to bottom; region covered by the spectrograph slit is shown at the left) simultaneously with the spectral variations (left to right). (Naval Research Laboratory photograph.)

with other indications of solar activity, such as sunspots and structure of the white-light corona. Improvements in instrumentation and observing techniques have resulted in improved spatial, spectral, and temporal resolution in the measurements. Also, it has become evident that understanding the physical phenomena taking place in the solar atmosphere requires simultaneous or complementary measurements over a wide range of wavelengths, from the radio through the X-ray and γ-ray ranges.

Figure 7 is a far-UV spectrum of the sun, obtained in a sounding rocket flight, showing the transition of the solar spectrum from a continuum with superimposed absorption lines (Fraunhofer spectrum) to an emission line spectrum below 180 nm. Measurements of this type are useful for measuring the total solar energy output in specific lines and wavelength ranges in the far and extreme UV. Figure 8 is a monochromatic image of the entire sun in the light of ionized helium (He II) at a wavelength of 30.4 nm, taken with the Naval Research Laboratory's extreme-UV spectroheliograph flown on the *Skylab* space station in 1973–1974. Images of this type reveal the structure of the solar atmosphere in various temperature ranges.

Figure 9 is a spectrum of the sun taken in a sounding rocket flight of the Naval Research Laboratory's HRTS instrument. Unlike previous instruments, this combined high spectral resolution with high spatial resolution in the along-slit direction (an imaging spectrograph). Thus, new information on the spatial distributions and velocity structures of solar active regions was obtained, as were variations in the spectral intensity and temperature distributions.

IV. PLANETARY ATMOSPHERES

Ultraviolet measurements are important for studies of planetary atmospheres, including Earth's upper atmosphere. This is due to the fact that the resonance absorption and emission spectral features fall primarily in the UV, making this spectral range much more sensitive for detection and measurement of the most common atmospheric gases than are other wavelength ranges. For example, O_2 absorbs strongly at wavelengths below 200 nm; hence, UV spectroscopy in this wavelength range allows sensitive detection of O_2 in other planetary atmospheres (provided that one observes from above Earth's O_2-rich atmosphere). The hydrogen Lyman α emission line at 121.6 nm, produced by scattering of solar Lyman α radiation by hydrogen atoms, is a very sensitive test for this gas in the outer atmospheres of planets and comets.

In Earth's upper atmosphere (and, in planetary probe missions, the atmospheres of other planets), atmospheric composition can be measured *in situ* with mass spectrometers and related instrumentation. However, UV measurements provide a capability for *remote sensing* of atmospheric composition and its variation with altitude, geographic location, and time, which supplements and extends *in situ* measurements where available, and can be applied to many objects not yet visited by spacecraft (e.g., in observations of other planets by spacecraft in near-Earth orbit). In addition to atmospheric composition, UV measurements can remotely sense the atmospheric temperature structure and its spatial and temporal variations. Also, information on the fluxes, energies, and spatial distributions of incoming energetic particles (such as those that produce Earth's polar auroras) can be obtained.

Three basic types of UV observations applicable to planetary atmosphere studies are (1) observations of far-UV emission features, excited by solar radiation or charged-particle impacts (such as those that cause Earth's polar auroras); (2) observations of middle-UV absorption features, superimposed on the reflected solar spectrum (for example, observations of ozone in Earth's atmosphere or of sulfur dioxide in that of Venus); and (3) observations of middle-UV or far-UV absorption features superimposed on the solar spectrum or a stellar UV spectrum when the sun or star is viewed directly, with a line of sight passing through the planetary atmosphere (solar or stellar occultation). All of these have been applied to studies of Earth's upper atmosphere; various combinations of these have been applied (in various wavelength ranges) to studies of the atmospheres of other planets. The pioneering observations of both Earth's upper atmosphere and those of other planets were made using sounding rocket vehicles. More recent observations of the other planets have been made using both Earth orbiting astronomical observatories (such as the IUE and HST satellites) and planetary flyby or orbiter spacecraft (such as the *Mariners*, *Voyagers*, and *Galileo*).

Figure 10 shows UV spectra of three planets: Earth, Venus, and Jupiter. It is seen that Lyman α (121.6 nm) emission is present in all of these. The atmospheres of both Venus and Mars consist mainly of carbon dioxide, and hence, their UV spectra are similar, but there are subtle and significant differences. Atomic oxygen emissions (130.4 and 135.6 nm) are present in all but the spectrum of Jupiter, which shows only atomic and molecular hydrogen features. The Earth spectrum is nearly unique, showing strong features of atomic and molecular nitrogen, although these have also been detected in the UV spectrum of Titan, the largest satellite of Saturn. For our moon, the planet Mercury, and Ganymede (Jupiter's largest satellite), UV observations have set upper limits on atmospheric densities which are far lower than those established by ground-based measurements. As in the case of the sun, imagery

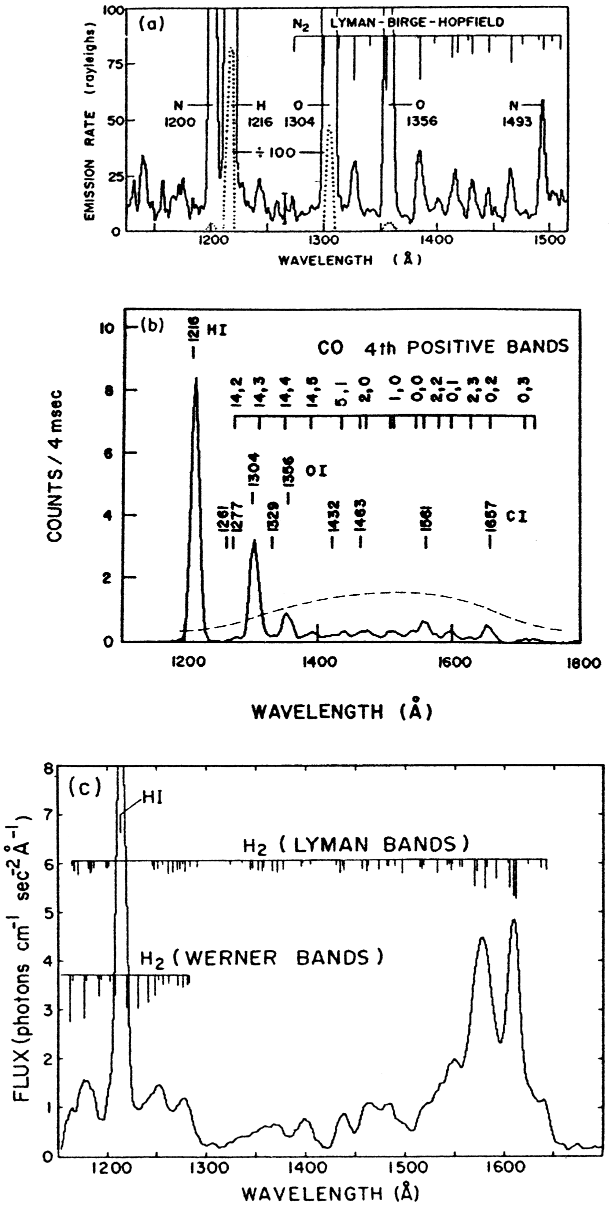


FIGURE 10 Far-UV emission spectra of three planetary atmospheres. (a) Earth's upper atmospheric day air-glow, observed in a sounding rocket flight. [Reprinted with permission from Takacs, P. Z., and Feldman, P. D. (1977). *J. Geophys. Res.* **82**, 5013.] (b) Day airglow of Venus, observed with the spectrometer on the *Pioneer Venus Orbiter*. [Reprinted with permission from Durrance, S. T. (1981). *J. Geophys. Res.* **86**, 9116.] (c) An aurora on Jupiter, observed with the *International Ultraviolet Explorer* satellite. [Reprinted with permission from Durrance, S. T., et al. (1982). *Geophys. Res. Lett.* **9**, 653.]

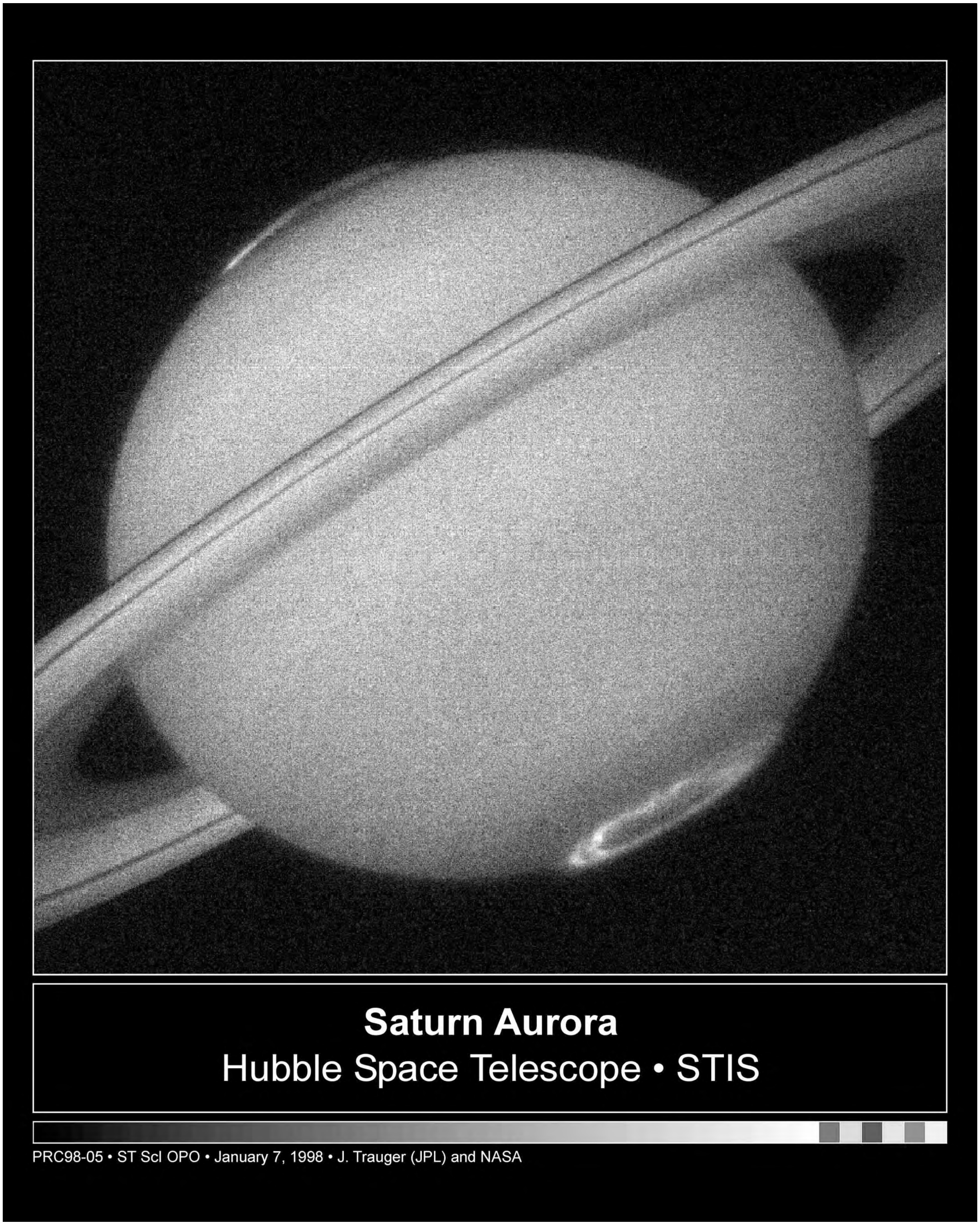

Saturn Aurora
Hubble Space Telescope • STIS
PRC98-05 • ST ScI OPO • January 7, 1998 • J. Trauger (JPL) and NASA

FIGURE 11 Far-UV images of auroras and day airglow on Jupiter and Saturn, obtained with the *Space Telescope Imaging Spectrograph* (STIS) on the *Hubble Space Telescope* (HST). Note that the contrast of the aurora to the dayglow and reflected sunlight background is much greater in the far-UV than in the visible spectral range. (Courtesy of B. Woodgate, NASA GSFC.)

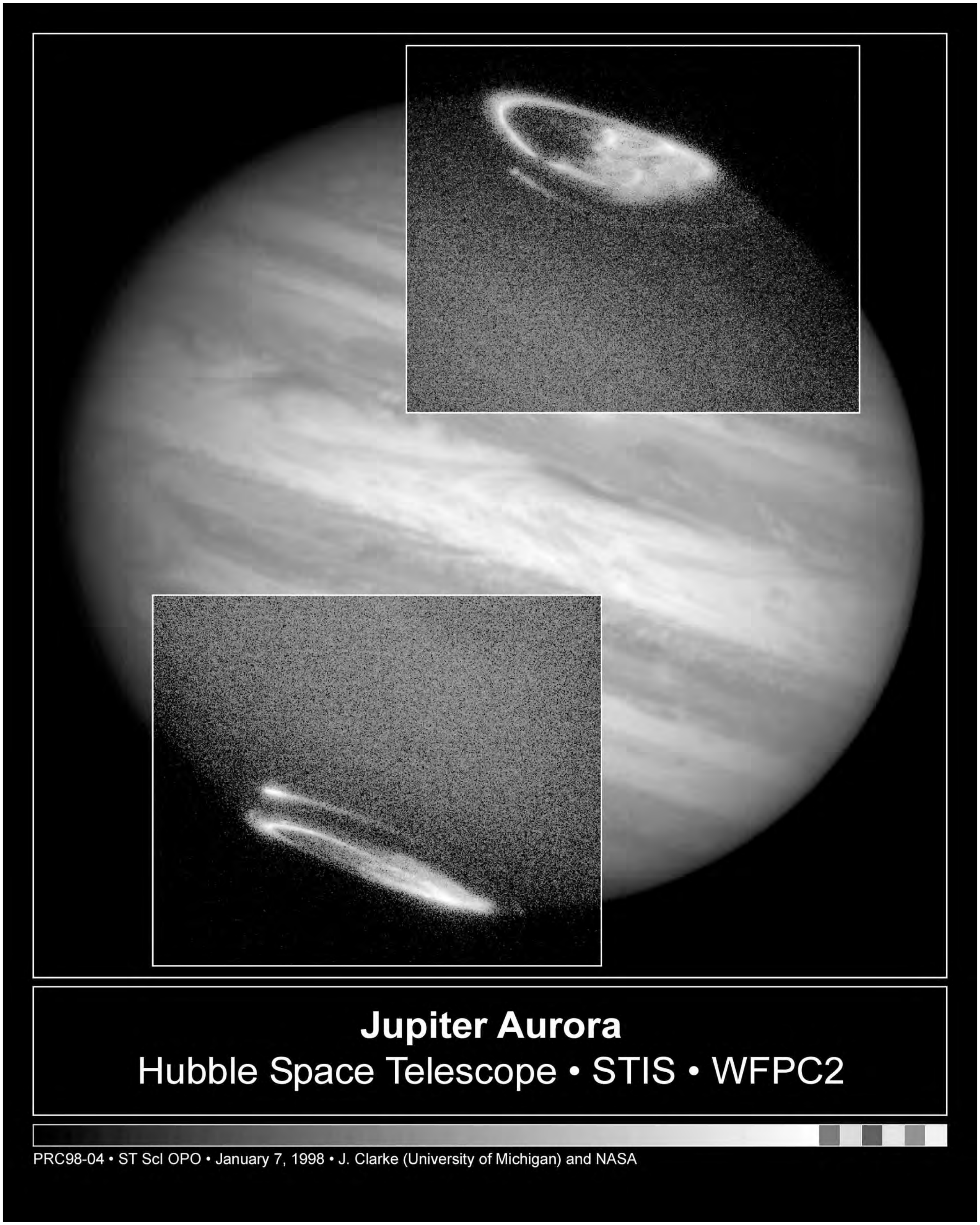

Jupiter Aurora
Hubble Space Telescope • STIS • WFPC2
PRC98-04 • ST ScI OPO • January 7, 1998 • J. Clarke (University of Michigan) and NASA

FIGURE 11 (*continued*)

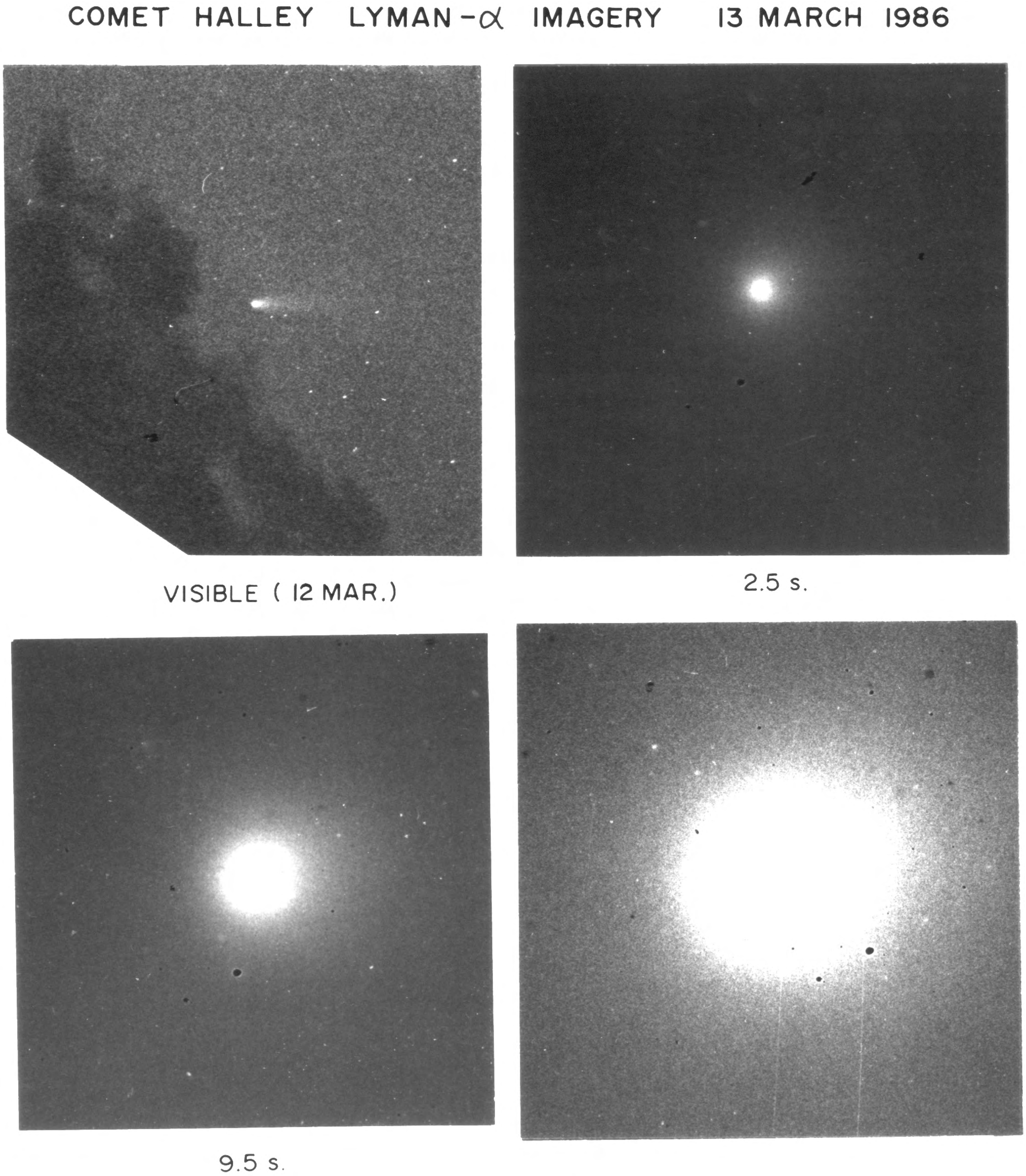


FIGURE 12 Far-UV images of Comet Halley, obtained by the Naval Research Laboratory in a 1986 sounding rocket flight, with exposure times indicated. The images show the great extent of the atomic hydrogen halo surrounding the comet (many times larger than the sun), which scatters Lyman α (121.6 nm) solar radiation.

of planetary atmospheres in individual spectral lines provides information on the altitude distributions of various gases, indirectly yielding information on temperature distributions. In addition, it provides information on the planetographic distributions of localized phenomena, such as auroras (which *Voyager*, *Galileo*, and HST have observed on Jupiter and Saturn). Figure 11 shows HST STIS far-UV images of auroras on Jupiter and Saturn. It is noteworthy that, as is also true of auroras on Earth, the contrast of the aurora against the daylit lower atmosphere is much greater in the far-UV than in the visible spectral range.

V. COMETS

Comets are unique, in comparison with the planets and satellites, in many ways. Of particular significance here is that they are objects of very low mass (comparable to small asteroids), but they are composed largely of volatile materials such as water and ice. Thus, when they approach the sun while traveling along their (typically) highly eccentric orbits, a significant portion of their mass is "boiled off" to produce a gaseous halo, or coma, which is quite prominent in the UV as well as at other wavelengths. As in the case of planetary atmospheres, UV observations can provide important information on the volatile composition of comets. They can also be used to determine the vaporization rates of various cometary materials and to provide information on the physical interactions between the cometary atmosphere and solar UV radiation and the solar wind.

Hydrogen Lyman α and the OH molecular band emission near 310 nm are the two most prominent spectral features in comets. These indicate that water is indeed the dominant volatile constituent of comets; other materials (which are responsible for the ground-accessible features, such as CH, CN, C_2, and NH band emissions) are only minor constituents. Because of the small mass of the hydrogen atom, the cometary hydrogen coma is much larger than the coma revealed in heavier molecular emissions (see Fig. 12); it can be many times larger than the sun. Measurements of the Lyman α brightness distribution in this halo can be used to determine the hydrogen production rate and, hence, the vaporization rate of water and other hydrogen compounds.

Ultraviolet observations of several comets, including West (1975) and Halley (1986) (Fig. 13), revealed that CO is also a prominent cometary constituent. Ultraviolet observations have also resulted in the detection of minor species not previously observed in comets, including atomic and diatomic sulfur, ionized carbon, and the SH radical. Future, more sensitive observations and ones extending to shorter wavelengths are expected to reveal other constituents, such as H_2, N_2, N, N^+, and O^+.

VI. THE STARS

A. Hot Stars

Hot stars, meaning those having surface temperatures in excess of 10,000 K (vs ~5800 K for our sun), emit much or most of their radiation in the UV wavelength range below 300 nm, the limit for ground-based observations. This was predicted theoretically, many years before the firstspace observations, from models of stellar atmospheres. It was also inferred, for the very hottest stars (above 20,000 K), from observations of emission nebulae or ionized hydrogen (H II) regions; in these regions, ionization of hydrogen is produced by stellar radiation in the extreme UV, below 91.2 nm wavelength.

Although the flux distributions for hot stars are rising toward shorter wavelengths in the ground-accessible range, the shape of the distribution (or of the theoretical model curves) is insensitive to temperature for very hot stars. Hence, it is very difficult to determine temperature or total

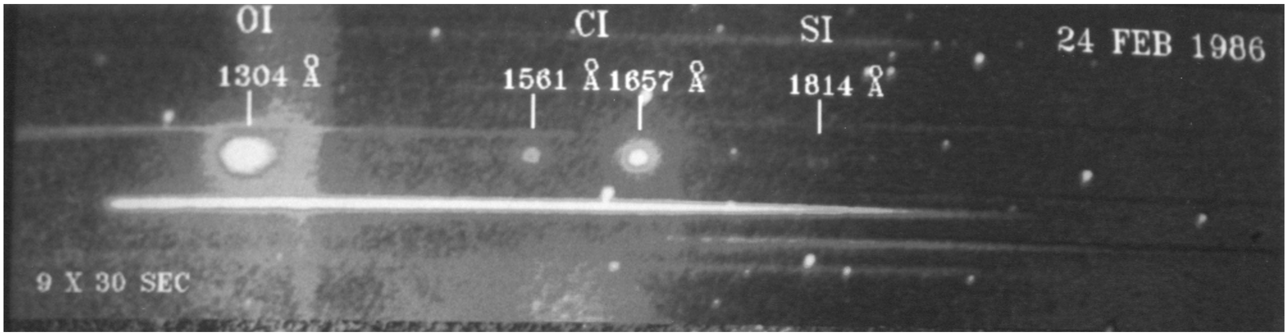


FIGURE 13 Objective-grating far-UV spectrum of Comet Halley, obtained by the Naval Research Laboratory in a 1986 sounding rocket flight, which excludes the hydrogen Lyman α emission. The spectrum reveals far-UV emissions of atomic oxygen, carbon, sulfur, and also carbon monoxide (CO) (unlabeled features between the O I and C I features).

radiation output for very hot stars from ground-based observations of their visible/near-UV spectra.

The earliest sounding rocket observations of hot stars in the UV in the early 1960s indicated much lower UV fluxes than were predicted by the theoretical stellar atmosphere models. However, over the following years, improvements in both the observations and the models resulted in convergence of the two, so that now there is reasonable agreement between theory and observation.

In particular, the improved models include the effects of the far-UV absorption lines (line blanketing), omitted in early models, and predict lower far-UV fluxes for a given temperature. Figure 14 shows a series of model atmosphere flux distributions, computed by E. Avrett of the Harvard-Smithsonian Center for Astrophysics. As can be seen, the flux distributions of hot stars are far more sensitive to temperature in the far-UV range than in the ground-accessible wavelength range.

The first large, self-consistent set of stellar UV spectrophotometric data was that obtained by the University of Wisconsin experiment on the OAO-2 satellite. Figure 15 shows measured flux distributions for four hot stars of different temperatures, a combination of UV observations by OAO-2 with ground-based spectrophotometry. Similar data sets have been obtained with the European TD-1 and ANS satellites and greatly expanded and extended to much fainter stars by the IUE satellite spectrometers.

Current efforts include extensions of measurements to much fainter stars still (including those in nearby galaxies, such as the Magellanic Clouds) using HST; more accurate and complete measurements in the wavelength range below 115 nm, the short-wavelength limit of IUE and HST; using the *Extreme Ultraviolet Explorer* (EUVE) and the *Far Ultraviolet Spectroscopic Explorer* (FUSE); and other, special-purpose UV astronomical instruments.

Most stars are not observable at wavelengths below 91.2 nm, even from space, because this radiation is absorbed by atomic hydrogen in interstellar space. However, because the distribution of atomic hydrogen in space is highly variable, EUVE has been successful in observing a

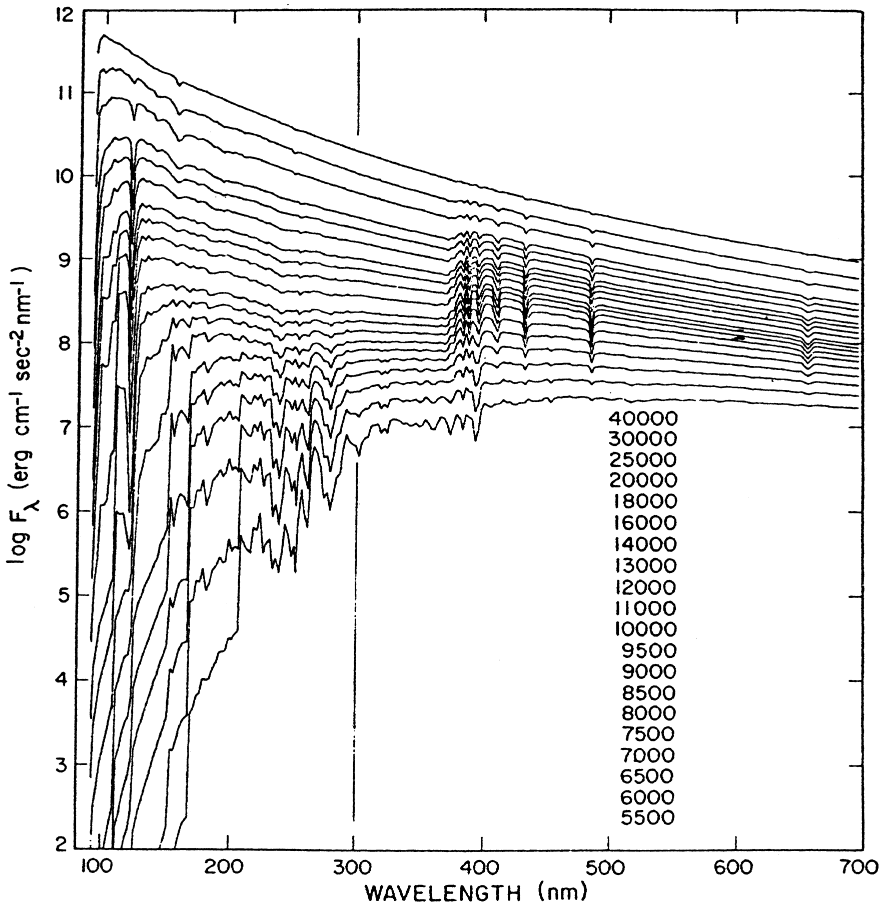


FIGURE 14 Shown are model atmosphere flux distributions, more recent than those in Fig. 13, taking into account absorption line blanketing. Effective temperatures for the models are as indicated, from top to bottom. [Reprinted with permission from Kurucz, R. L. (1979). *Astrophys. J. Suppl.* **40**, 1.]

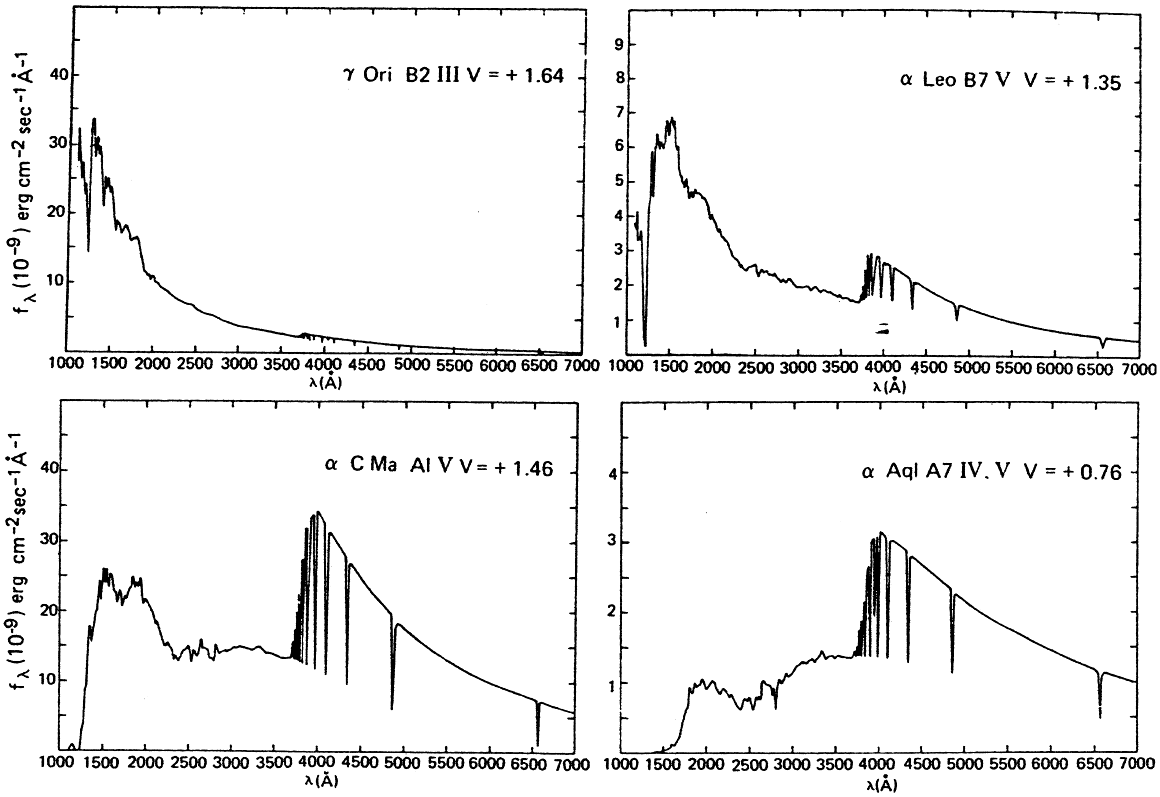


FIGURE 15 Measured flux distributions in the UV and visible for four stars ranging in temperature from about 8000 K (type A7) to 20,000 K (type B2). The UV measurements are from the University of Wisconsin experiment package on OAO-2. (Courtesy of A. Code, University of Wisconsin, Madison.)

number of very hot stars in the extreme-UV range below 91.2 nm, some even at relatively large distances, in regions in space where the atomic hydrogen column density is unusually low.

Spectrometric observations of individual hot stars have been supplemented by imagery of starfields in moderate wavelength ranges in the UV. The advantage of these *imaging surveys* is that much fainter objects can be measured and that a very large number of objects can be recorded in a single exposure. Ultraviolet imagery is particularly useful as a means of surveying large areas of the sky in order to detect and measure hot stars not previously known to exist (or at least not previously known to be hot). Figure 16 compares a far-UV image of the constellation Orion, obtained in a sounding rocket flight, with a visible-light image. The UV image shows, at a glance, the distribution of hot stars without confusion by the far more numerous (and, in some cases, brighter in visible light) stars.

Ultraviolet spectroscopy is useful not only for measuring the energy flux distributions of hot stars, but also for gaining insight into the properties and physical processes occurring in stellar atmospheres. This information is provided by the line features in the spectra. One of the first major discoveries in stellar UV spectroscopy was that the strong line features such as C IV (155 nm) and Si IV (140 nm) had what is known as a P Cygni profile: a very strong, broad absorption feature shifted to shorter wavelengths (relative to its laboratory, or rest, wavelength) combined with an emission feature at, or slightly longer than, the rest wavelength (see Fig. 17). As shown in Fig. 18, the production of such a spectral feature can be explained as being due to a very strong *stellar wind*. The observations indicate that some very hot, luminous stars are losing mass at very high rates—of the order of 10^{-6} solar mass per year. Such high mass-loss rates, many orders of magnitude greater than that associated with the solar wind, can significantly influence the evolution of a massive hot star.

In addition, UV line spectra are useful because they contain the resonance absorption features of most of the common ions expected in stellar atmospheres and provide sensitive means of measuring temperature and relative abundances. Since, as discussed later, starlight can be absorbed or scattered by dust particles in interstellar space, measurements of absorption lines produced in the stellar atmospheres can, in some cases, provide more useful

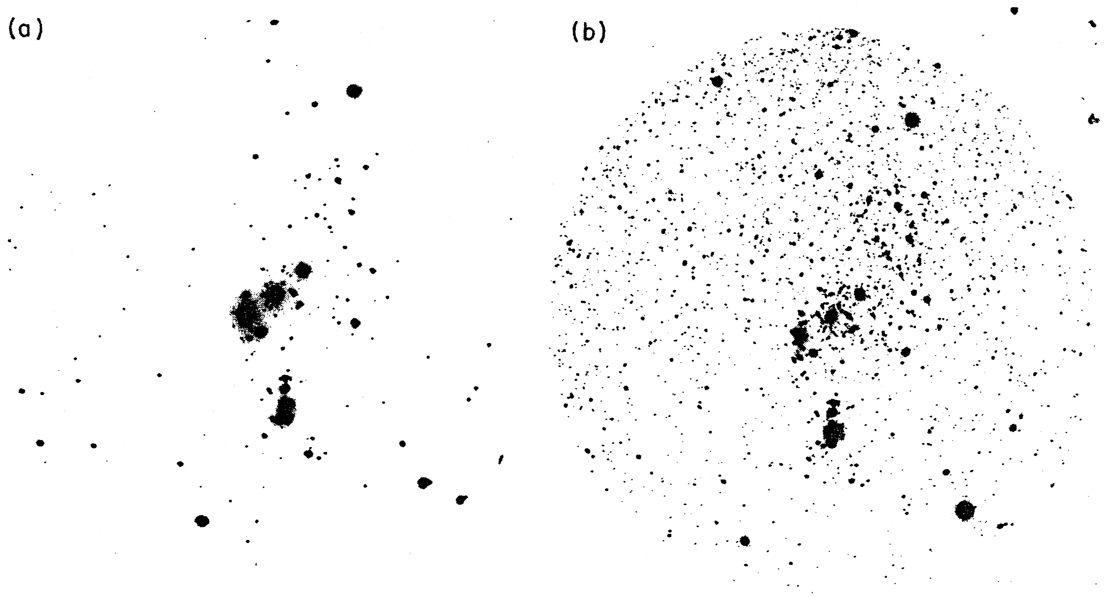


FIGURE 16 Comparison of visible and UV imagery of the constellation Orion. (a) Far-UV (1230–2000 Å) image of Orion obtained by the Naval Research Laboratory in a 1975 sounding rocket flight. (b) Visible-light image, to the same scale. (Hale Observatories photograph.)

temperature information than measurements of the continuum flux distribution.

B. Cool Stars

Cool stars, in the temperature range 3000–9000 K, emit most of their radiation in the wavelength range accessible from the ground (as does our sun). However, as is true for the sun, the UV wavelength range reveals energetic processes in the outer atmospheres of these stars. Objects of similar surface temperatures can exhibit quite different far-UV spectra, indicative of quite different conditions in their outer envelopes.

Although some pioneering observations of cool stars in the UV were obtained with sounding rockets, OAO-2, and *Copernicus*, a truly systematic study covering a wide range of cool star types had to await the higher sensitivity of the IUE satellite. Observations with IUE have revealed that some giant stars, such as Capella, are much more active than our sun (as are also some stars in close binary systems). Cool supergiant stars, on the other hand, do not show evidence of hot coronas; it is presumed that this is due to the presence of strong stellar winds, the energy of which might otherwise go into heating. Cool class M stars, having temperatures of only about 3000 K and being much less luminous than our sun, surprisingly show very active chromospheres and coronas—in some cases, with higher far-UV fluxes per unit area than our sun. These cool stars also sometimes exhibit flare activity far surpassing that of our sun.

In eclipsing binary systems consisting of, for example, a hot star and a cool giant star, UV measurements can provide information on the outer atmosphere of the cool star by observing the hot star as it is eclipsed by the cool star.

These types of observations with IUE have been supplemented by more sensitive observations with the HST spectrographs and extended to shorter wavelengths by the HUT on the *Astro* Spacelab shuttle missions and by the newly operational FUSE. Also, somewhat surprisingly, many late-type stars have been detected by the EUVE (which is most likely seeing EUV emissions from the chromospheres and coronas of these stars, as well as occasional flare activity).

VII. INTERSTELLAR GAS

The space between the stars is not empty, but contains highly rarefied gas and solid particles (dust). The density and temperature, and to some extent the composition, of this interstellar material are highly variable from place to place in our galaxy. On the average, the interstellar gas contains about one hydrogen atom per cubic centimeter, although it can range from 0.01 or less to more than 10^6 (in dense molecular clouds). It is this interstellar material from which new stars are formed.

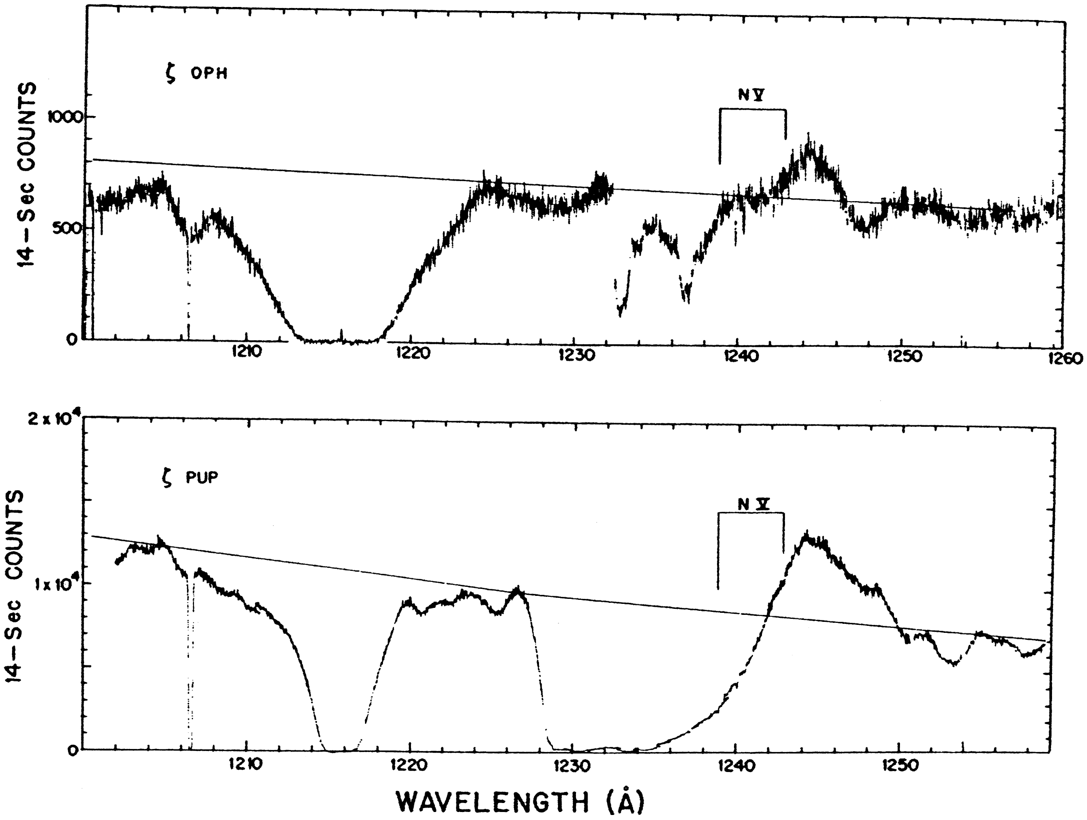


FIGURE 17 Far-UV high-resolution (0.02-nm) spectra of the stars ζ Ophiuchi (spectral type O9.5) and ζ Puppis (type O4) obtained with the Princeton University spectrometer on the *Copernicus* satellite. Of interest here are the absorption features of atomic hydrogen at 121.6 nm (considerably stronger, indicating a larger hydrogen column density, toward ζ Ophiuchi) and the absorption features due to four-times-ionized nitrogen (N V) in the stellar atmospheres. The latter feature exhibits a P-Cygni profile (much stronger in ζ Puppis), indicating high-velocity mass ejection. [Reprinted with permission from Morton, D. C. (1976). *Astrophys. J.* **203**, 386.]

The composition of the interstellar medium is believed to be similar to that of the sun and the stars: Hydrogen accounts for ~90% of all atoms, and helium accounts for most of the remaining 10%. Atoms of all heavier elements make up less than 1%. Much of the heavier element component is believed to be in the form of dust grains rather than in gaseous form, although the relative proportion is variable. Except in the close vicinity of hot and highly luminous stars, this interstellar medium is made evident only by its attenuation of the light of stars seen through it. The dust particles produce continuous attenuation, whereas the gas usually absorbs only in discrete spectral lines. Since the resonance transitions of most of the common elements occur in the far UV, the composition of the interstellar gas was only poorly known before the advent of space UV spectroscopy. Only relatively minor constituents of the interstellar gas, such as sodium and calcium, can be detected and measured by ground-based optical absorption spectroscopy. The far UV allows measurements of the major constituents, such as atomic and molecular hydrogen, and also atomic oxygen, carbon, and nitrogen.

A. Interstellar Atomic and Molecular Hydrogen

The first observations of interstellar atomic hydrogen, by means of its Lyman α absorption line at 121.6 nm, were made from sounding rockets as early as 1966. Molecular hydrogen was first observed in a 1970 rocket flight. Surveys of interstellar atomic hydrogen were made with the OAO-2 satellite spectrometers. However, the instrument that produced the most comprehensive studies of the interstellar gas was the far-UV spectrometer provided by Princeton University, carried on the OAO-3 (*Copernicus*) satellite launched in 1972. This instrument observed in the wavelength range 91.2–300 nm with spectral resolution as good as 0.005 nm. Figure 17 shows spectra of the stars ζ Ophiuchi and ζ Puppis, obtained by *Copernicus*, in the

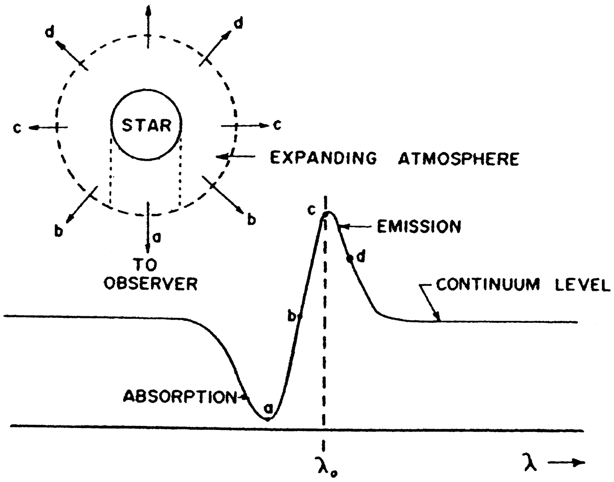


FIGURE 18 Illustration of the formation of a P Cygni profile (emission line near the "rest" wavelength, with blue-shifted absorption) in a spectral feature, resulting from high-velocity outflow of gas from a stellar atmosphere.

region of the Lyman α absorption line. It is apparent that there is much more atomic hydrogen in the line of sight to the former star than toward the latter.

Figure 19 shows a higher resolution spectrum of ζ Ophiuchi, in the region of a molecular hydrogen absorption band. In contrast to atomic absorption spectra, the detailed distribution of absorption intensity in molecular spectra can be used to determine gas temperature, in the low-temperature range existing in molecular clouds. Here, the H_2 absorption features indicate a gas temperature of about 80 K, which is consistent with other determinations.

The *Copernicus* satellite also made the first direct observations of deuterium ("heavy hydrogen") in the interstellar medium, both in atomic form and in the molecule HD (Fig. 19). Accurate knowledge of the ratio of D to H in the interstellar gas is of relevance to theories of the origin and evolution of the universe.

The *Copernicus* observations confirmed theoretical predictions that molecular hydrogen is formed by "three body recombination" of hydrogen atoms on dust particles; here, the dust particle acts as a catalyst and takes up the energy of recombination of the two hydrogen atoms. The hydrogen molecules are broken up, or photodissociated, by UV radiation from hot stars in the interstellar medium outside of dense dust clouds. Thus, molecular hydrogen is the major form of interstellar hydrogen in dense dust clouds, whereas in the general interstellar medium, hydrogen is primarily in atomic form.

Currently, observations of interstellar molecular hydrogen, as well as of atomic hydrogen, deuterium, and HD,

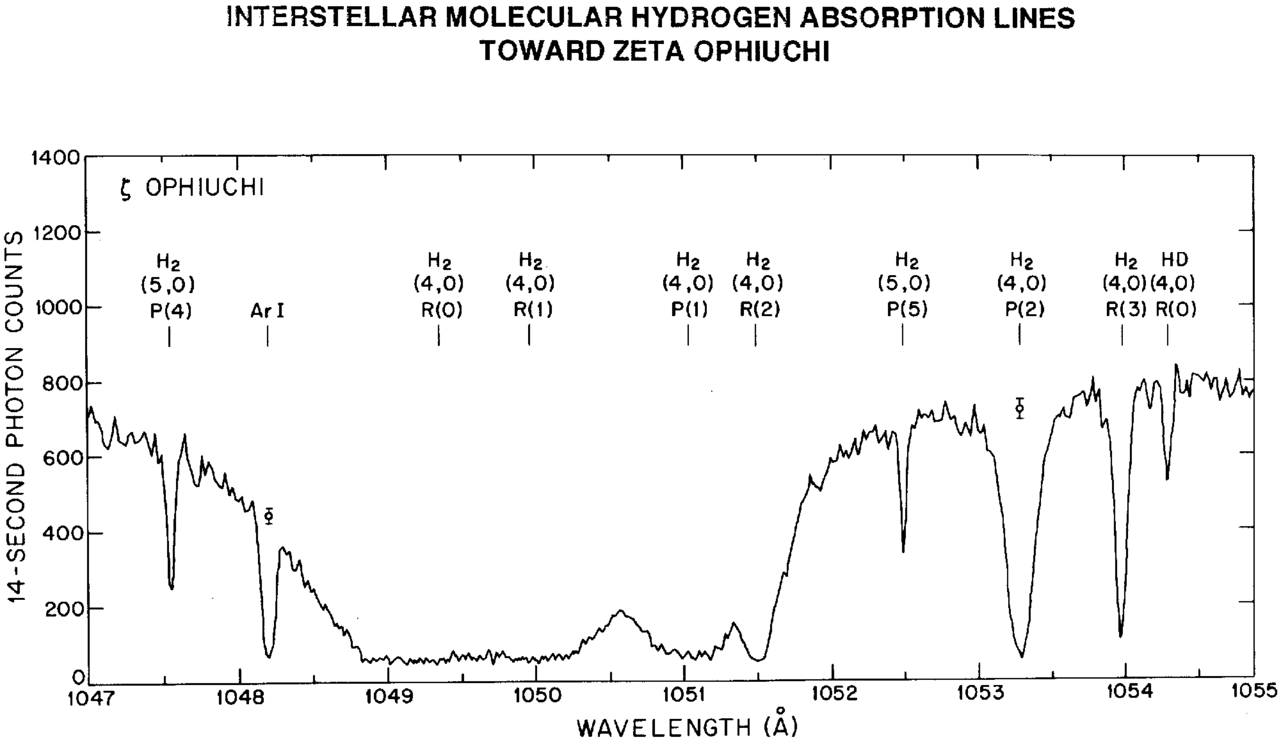


FIGURE 19 *Copernicus* high-resolution (0.005 nm) spectrum of the star ζ Ophiuchi showing absorptions due to interstellar molecular hydrogen (H_2) and HD. [Spitzer, L., and Jenkins, E. B. (1975). *Annu. Rev. Astron. & Astrophys.* **13**, 133.]

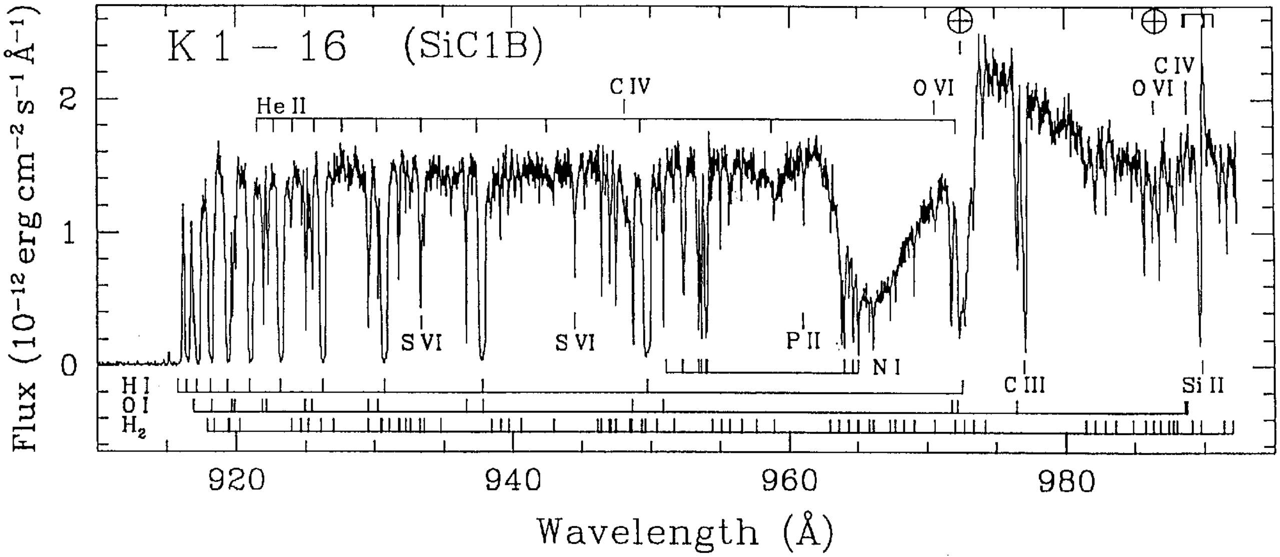


FIGURE 20 FUSE spectrum of the central star of a planetay nebula, with about 0.004 nm resolution, showing absorption features due to interstellar atomic and molecular hydrogen and atomic oxygen (lower scales) in the 91.2–99.2 nm wavelength range. Features due to the atmosphere of the star itself are marked on the upper scale. [Moos, H. W., et al. (2000). *Astrophys. J.* **538**, L1.]

are being extended to much fainter and more distant objects by the FUSE satellite mission. Figure 20 shows a FUSE spectrum of the central star of a planetary nebula. Absorption lines due to interstellar H I, O I, and H_2 are marked below the spectrum.

B. Other Constituents of the Interstellar Gas

The *Copernicus* satellite also observed many other species in the interstellar gas, including neutral and ionized carbon, neutral oxygen, nitrogen, and argon. The molecules CO and OH were also observed toward some stars. It was found that most of the heavier elements were less abundant in the interstellar gas, relative to hydrogen, than expected on the basis of the relative elemental abundances in the sun and stars. It was conjectured that this "depletion" was the result of dust grain formation. Regions of the interstellar medium containing less than the usual proportion of dust showed less depletion than average, whereas the converse was true in lines of sight through dense dust clouds. *Copernicus* also discovered a very hot phase of the interstellar gas, revealing absorption lines of O VI characteristic of temperatures near 300,000 K. This gas is probably heated by shock waves in the interstellar medium resulting from supernova explosions.

The IUE satellite, launched in 1978, did not have as high a spectral resolution as (or reach wavelengths as short as) the *Copernicus* spectrometer. However, in the range of wavelengths longer than 120 nm, IUE was much more sensitive than *Copernicus*. Therefore, it was able to observe fainter and more distant stars (including some in nearby external galaxies, such as the Magellanic Clouds) and stars whose light is more severely attenuated by interstellar dust. It extended studies of high-temperature interstellar gas through observations of C IV, Si IV, and N V absorption lines. In observations of hot stars in the Magellanic Clouds, it detected interstellar absorption lines attributed (because of their Doppler shifts relative to locally produced interstellar absorption lines) to a hot "halo" of gas surrounding our galaxy and also perhaps to similar halos surrounding the Magellanic Clouds.

More recently, the *Goddard High Resolution Spectrograph* (GHRS) and its later replacement, the *Space Telescope Imaging Spectrograph* (STIS) on the HST, have greatly extended the measurements of the interstellar gas by IUE to fainter and more distant stars and to better spectrophotometric accuracy. The recently launched FUSE satellite has also extended the measurements by IUE to shorter wavelengths, and of *Copernicus* to higher sensitivity. In particular, STIS and FUSE have now extended these measurements to very large distances in intergalactic space, using quasi-stellar objects (quasars) as background light sources. Among other results is the finding that oxygen has an unexpectedly high abundance, relative to hydrogen, in intergalactic space.

C. Emission Nebulae

The IUE satellite also provided new information on the composition of interstellar gas through observations of

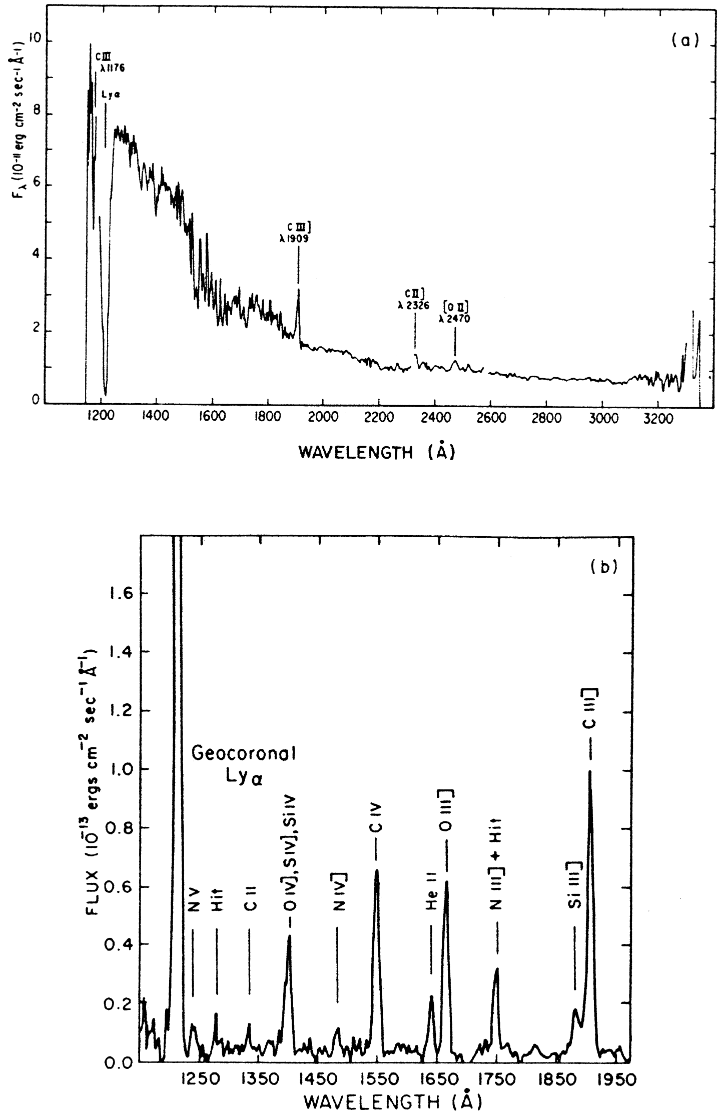


FIGURE 21 Far-UV spectra, obtained with IUE, of diffuse nebulae, (a) Orion nebula, a typical H II region, showing continuum due to dust scattering of UV starlight, with nebular emission lines. [Reprinted with permission from Torres-Peimbert, S., et al. (1980). *Astrophys. J.* **293**, 133.] (b) Cygnus Loop, a supernova remnant, showing emission lines due to shock-heated gas. [Reprinted with permission from Raymond, J. C., et al. (1980). *Astrophys. J.* **238**, 881.]

emission nebulae (H II regions, planetary nebulae, and supernova remnants), in this case by observations of UV emission lines. Figure 21 shows IUE spectra of the Orion nebula (a typical H II region) and of the Cygnus Loop supernova remnant. In particular, the abundance of carbon (difficult to determine from ground-based observations) has been measured by observations of the strong features of C II (232.5 nm), C III (190.9 nm), and (in some cases) C IV (155 nm). In these objects, the electron temperatures (and, in supernova remnants, shock wave velocities) have also been determined from observations in the far UV. Figure 22 is a HUT spectrum of the Cygnus Loop supernova remnant, showing the emission line of O VI (103.2–103.7 nm). The presence of this feature indicates that the temperature of the gas is very high (more than 200,000 K).

The H II regions are representative of gas from which stars have recently formed or are in the process of forming. Planetary nebulae, on the other hand, are formed from the outer envelopes of stars nearing the end points of their evolutionary life cycles. The collapsed, very hot core of a dying star provides the extreme-UV radiation that ionizes the nebula. Since the gas in the planetary nebula has been "processed" by the star, its composition differs from that of an H II region; however, it is the gas ejected by old stars that is recycled to form the next generation of stars.

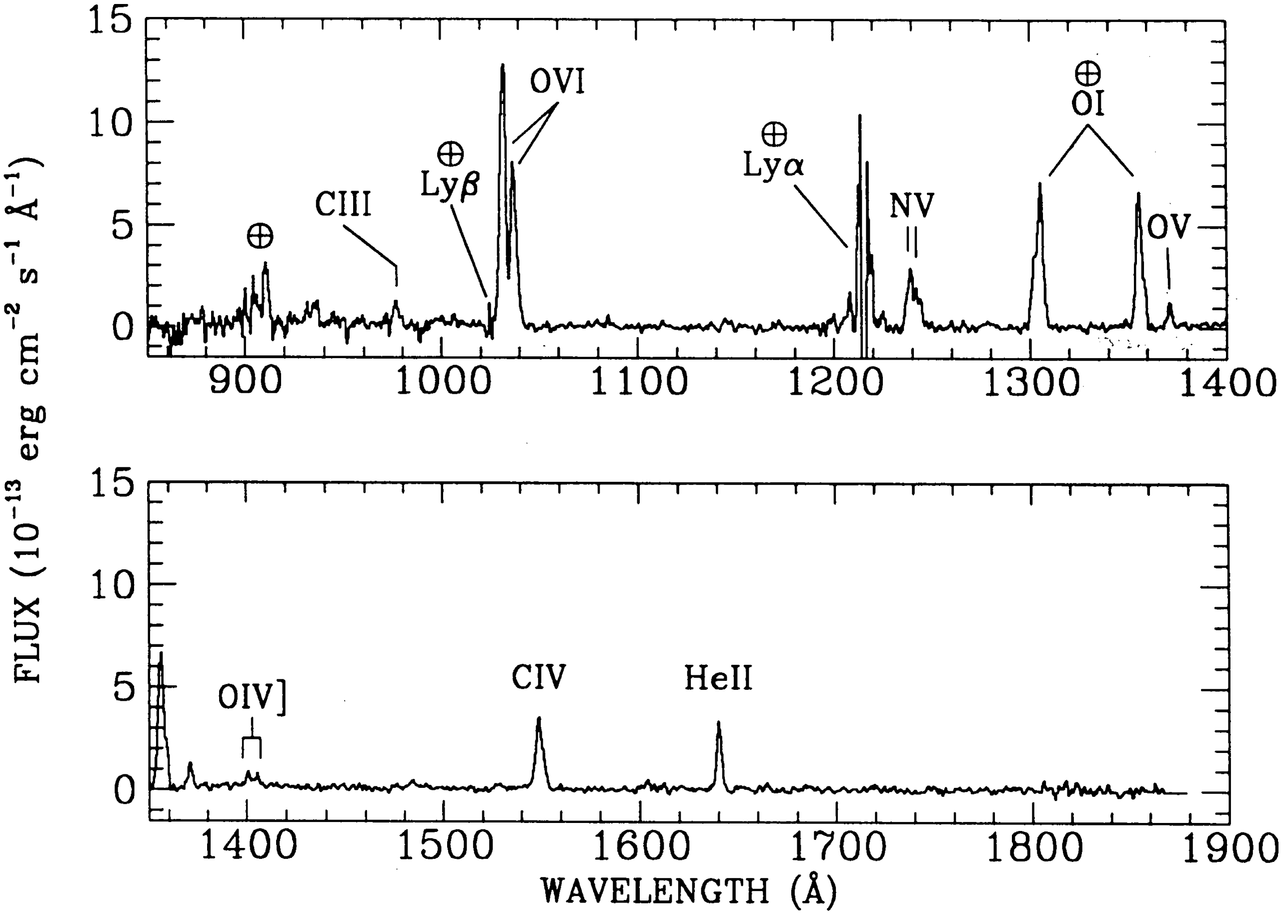


FIGURE 22 Far-UV spectrum of a filament in the Cygnus Loop supernova remnant, obtained with the HUT in the 1990 *Astro-1* shuttle mission. Emission features of O VI and other highly ionized species are revealed. [Blair, W. P., et al. (1991). *Astrophys. J.* **379**, L33.]

Supernova remnants are the result of much more violent disruptions of (more massive) stars; the shock waves produced by the stellar explosion excite emissions from the gas both in the castoff envelope and in the surrounding interstellar gas, which is swept up in the expanding shell. Hence, the elemental compositions of supernova remnants are intermediate between those of H II regions and planetary nebulae.

VIII. INTERSTELLAR DUST

Dust particles in interstellar space make themselves known by both their attenuation and their reflection of starlight. In the visible range, the combined absorption and scattering (extinction) of starlight increases toward shorter wavelengths, so that stars seen through dust clouds appear both dimmer and redder in color than they would otherwise. In the close vicinity of bright stars, dust clouds are illuminated and appear as *reflection nebulae*.

Ultraviolet measurements of interstellar extinction and of reflection nebulae have the potential to provide additional information on the properties and composition of the dust particles. As mentioned, in the visible range, the extinction increases smoothly toward shorter wavelengths. However, early rocket measurements (confirmed by more comprehensive satellite observations), in which the spectra of reddened stars were compared with those of unreddened stars of the same type, showed that this trend does not continue indefinitely into the UV. Instead, the interstellar extinction vs wavelength relation reaches a peak near 220 nm, following which the extinction *decreases* toward shorter wavelengths (to ~160 nm). Below this, the extinction again increases toward shorter wavelengths (see Fig. 23). The peak in the extinction curve near 220 nm matches that expected, theoretically, to be produced by graphite particles. However, it has been found that the

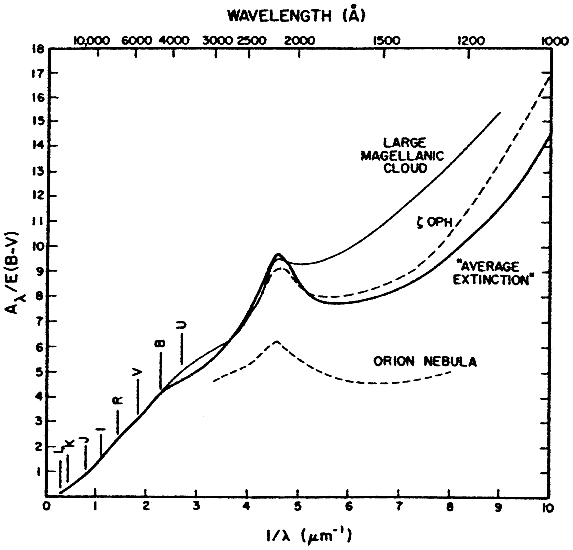


FIGURE 23 Interstellar extinction vs wavelength in the UV and ground-accessible wavelength ranges. The letters indicate central wavelengths of ground-based photometric pass bands. Note large variations in the extinction curve in the UV for different regions of space.

UV extinction curve is highly variable in both magnitude and shape in different directions in our galaxy and in the Magellanic Clouds. Hence, it appears that there must be at least three independent components of the interstellar dust, whose relative abundances vary in different ways with local conditions in the interstellar medium.

Observations of the reflection spectrum of interstellar dust have also yielded some insights concerning the properties of interstellar dust particles; in particular, it has been found that, at least in some reflection nebulae, the particles are very efficient scatterers of far-UV radiation. Except in the vicinity of the 220-nm extinction peak, it appears that most of the interstellar extinction is due to scattering rather than pure absorption.

Most galactic H II regions appear as reflection nebulae in the far UV, as illustrated by the IUE spectrum of the Orion nebula in Fig. 21. This is due to the relatively high dust densities (roughly proportional to gas densities) in these regions, combined with the presence of UV-bright stars (which both illuminate the dust and excite UV and visible emission lines). However, the dust in H II regions is probably not typical of dust in other reflection nebulae (illuminated by somewhat cooler stars) and in the general interstellar medium, since the intense radiation fields in regions such as the Orion nebula can significantly modify the size distributions and compositions of the dust particles. This is evidenced by both their reflection spectra and their extinction curves, which differ from those observed in less excited regions of the interstellar medium.

IX. EXTRAGALACTIC OBJECTS

The external galaxies, like our own, are gigantic clusterings of stars (as many as 10^{11} to 10^{12}) and associated interstellar material. There is a wide variety of shapes

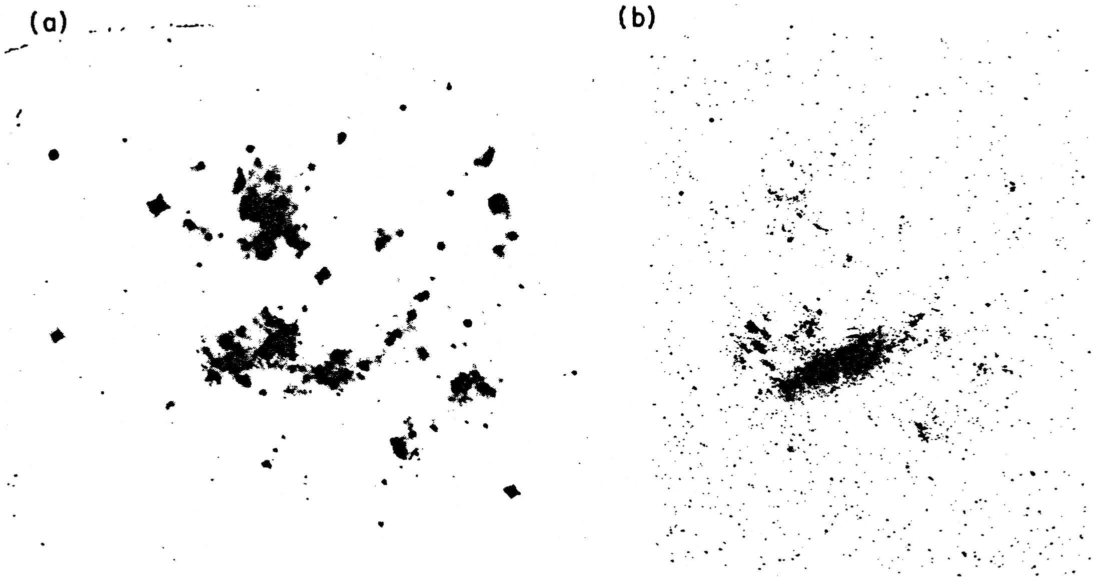


FIGURE 24 Comparison of far-UV and visible imagery of the Large Magellanic Cloud, the nearest external galaxy. (a) Image in the wavelength range 1250–1600 Å (10 min) obtained with the Naval Research Laboratory S201 camera on Apollo 16. (b) Visible-light image, to the same scale. (Lick Observatory photograph.)

and stellar content among external galaxies: many are spiral galaxies resembling our own, whereas others are spherical or elliptical and are devoid of obvious spiral patterns.

As in the case of our galaxy, UV observations are expected mainly to reveal the hot stellar component. However, an advantage in observing other galaxies is that the spatial distribution of the hot stellar population is seen at a glance, whereas this is much more difficult to determine for our own galaxy, which we view from within. The variations in this distribution from one galaxy to the next, and in comparison with the distribution of cooler stars and of interstellar material in these galaxies, provide information on star formation rates and history and on the overall evolution of galaxies. The ability to detect and measure hot stars in the presence of a much larger number of cool stars, by means of UV observations, is even more important in studies of external galaxies than our own, since in the former individual stars often are not individually resolved and hence problems due to image overlap and confusion are more severe. Ultraviolet observations also provide more sensitive measurements of the interstellar material in observations of starlight extinction or reflection.

Some galaxies, notably the Seyfert galaxies and quasi-stellar objects (quasars), exhibit highly energetic activity in their central regions, which far transcends that which can be associated with even the most massive individual stars. The total luminosity of an active galactic nucleus can greatly exceed the total luminosity of the remainder of the galaxy; quasars are, in fact, by far the most luminous objects in the universe. Observations of these objects in the UV can add to the store of knowledge that is needed to acquire an understanding of these objects.

Photometric and spectrophotometric measurements of a number of external galaxies were obtained with OAO-2 and subsequently were supplemented by more sensitive observations with IUE. Ultraviolet images of the Magellanic Clouds and a number of other galaxies have been obtained in sounding rocket flights, the *Apollo 16* mission

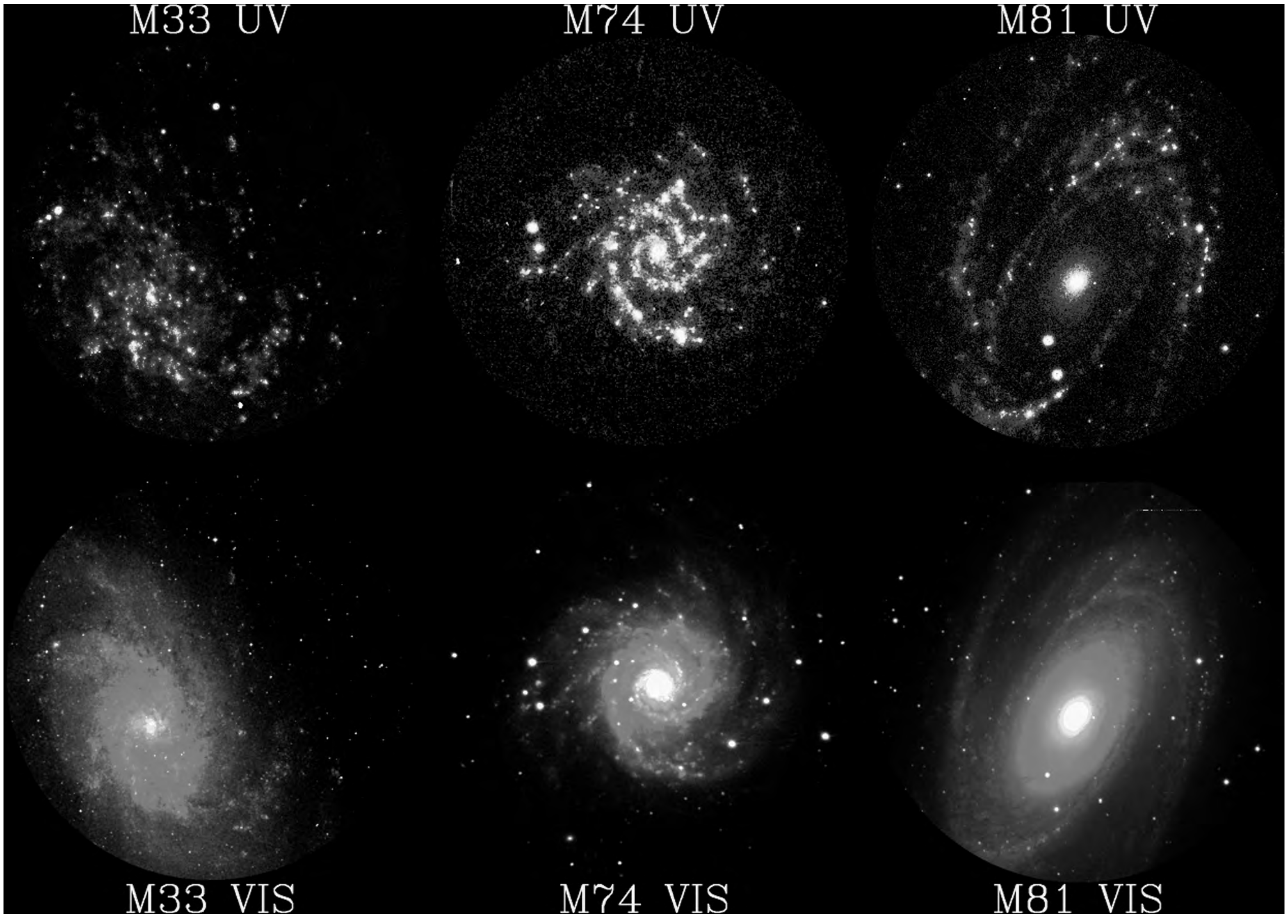


FIGURE 25 Comparison of far-UV images (top), taken with the UIT on the *Astro-2* shuttle mission, and ground-based visible images (bottom) of three spiral galaxies. (Courtesy of T. Stecher, NASA GSFC).

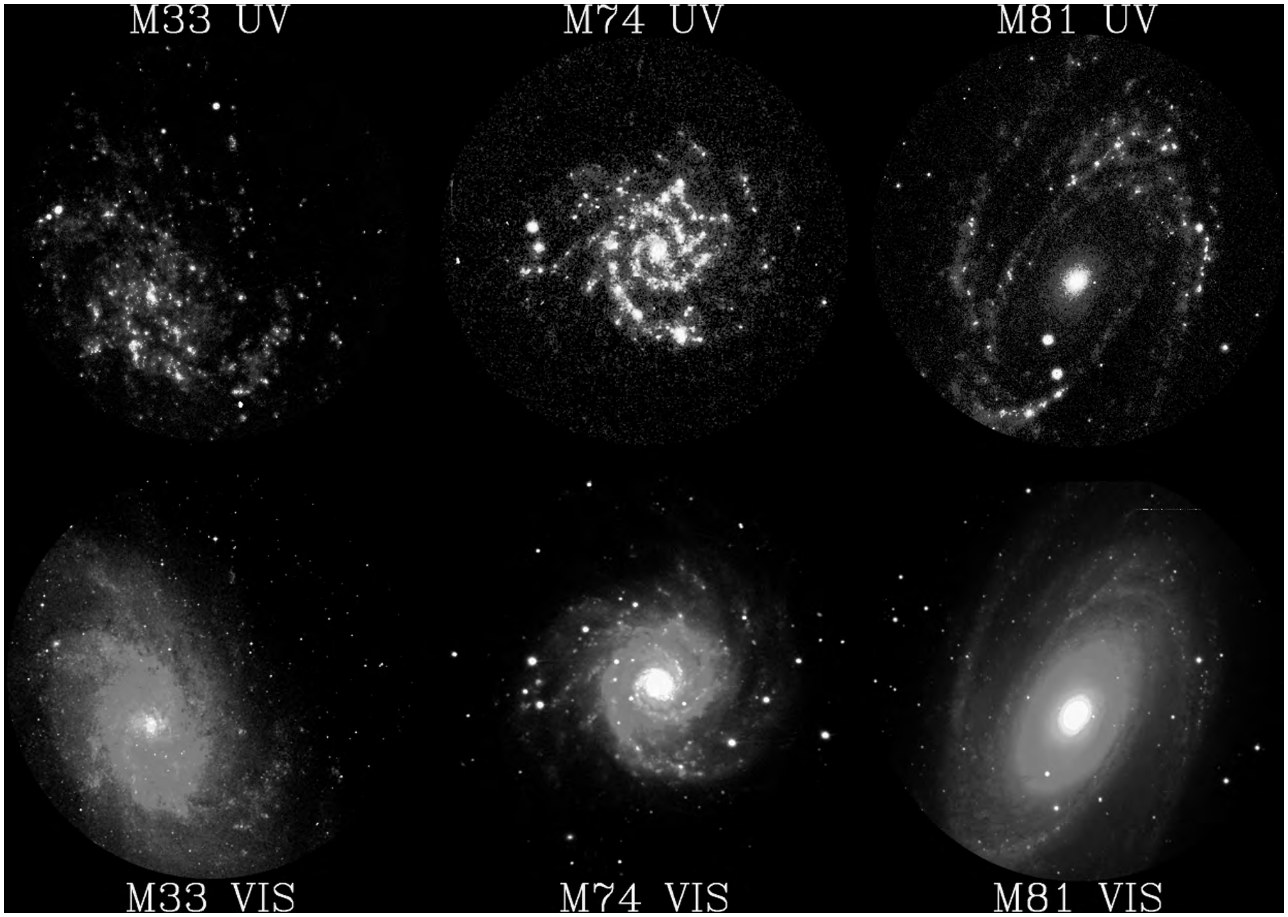


FIGURE 25 (*continued*).

(see Fig. 24), *Astro* space shuttle flights of the Goddard Space Flight Center's UIT (see Fig. 25), and other UV-imaging space experiments. The HUT on the *Astro* space shuttle flights has been used to obtain far-UV spectra of quasars and other extragalactic objects at wavelengths as short as 92 nm.

These observations are being greatly extended using imaging and spectroscopic instruments on the HST, in particular the STIS, and with the recently launched FUSE. Among other things, these new instruments have provided measurements of the gas and dust in external galaxies and in intergalactic space.

One of the most important objectives of the HST is to determine more accurately the distance scale of the universe; it can do this by observing galaxies at much greater distances with the same degree of detail as nearer galaxies are presently observed with ground-based telescopes. Ultraviolet observations are an important aspect of these studies, because the most luminous hot stars are useful distance indicators, and these are much brighter in the UV than in the visible. Also, for very distant objects, the redshift due to the expansion of the universe allows observations of wavelengths below the 91.2 nm short-wavelength limit for nearby objects, set by the absorption due to local interstellar atomic hydrogen.

SEE ALSO THE FOLLOWING ARTICLES

Atomic Spectrometry • Aurora • Galactic Structure and Evolution • Infrared Astronomy • Interstellar Matter • Planetary Atmospheres • Solar Physics • Stellar Spectroscopy • Stellar Structure and Evolution • Telescopes, Optical

BIBLIOGRAPHY

Bowyer, S., and Leinert, C., eds. (1990). "The Galactic and Extragalactic Background Radiation," IAU Symposium 139, Kluwer Academic, Dordrecht.

Bowyer, S., and Malina, R. F., eds. (1996). “Astrophysics in the Extreme Ultraviolet,” Kluwer Academic, Dordrecht.

Chapman, R. D., ed. (1981). “The Universe at Ultraviolet Wavelengths,” NASA CP-2171, U.S. Natl. Aeronaut. Space Admin., Washington, DC.

Code, A. D., ed. (1972). “The Scientific Results from the Orbiting Astronomical Observatory (OAO-2),” U.S. Natl. Aeronaut. Space Admin., Washington, DC.

Cornell, J., and Gorenstein, P., ed. (1983). “Astronomy from Space,” MIT Press, Cambridge, MA.

Cowie, L. L., and Songaila, A. (1986). “High resolution optical and ultraviolet absorption line studies of interstellar gas,” *Annu. Rev. Astron. Astrophys.* 499.

Eddy, J. A. (1979). “A New Sun: The Solar Results from Skylab,” U.S. Natl. Aeronaut. Space Admin., Washington, DC.

Hanle, P. A., and Chambrelain, V. D., eds. (1981). “Space Science Comes of Age,” Smithsonian Institution Press, Washington, DC.

Henbest, N., and Marten, M. (1983). “The New Astronomy,” Cambridge Univ. Press, London and New York.

Kondo, L., Mead, J. M., and Chapman, R. D., eds. (1982). “Advances in Ultraviolet Astronomy,” NASA CP-2238, U.S. Natl., Aeronaut. Space Admin., Washington, DC.

Kondo, Y., ed. (1990). “Observatories in Earth Orbit and Beyond,” Kluwer Academic, Dordrecht.

Kondo, Y., Boggess, A., and Maran, S. P. (1989). “Astrophysical contributions of the international ultraviolet explorer,” *Annu. Rev. Astron. Astrophys.* 397.

Lean, J. (1997). “The sun’s variable radiation and its relevance for Earth,” *Annu. Rev. Astron. Astrophys.* 33.

Lundquist, C. A., ed. (1979). “Skylab’s Astronomy and Space Sciences,” NASA SP-404, U.S. Natl. Aeronaut. Space Admin., Washington, DC.

Noyes, R. W. (1982). “The Sun, Our Star,” Harvard Univ. Press, Cambridge, MA.

O’Connell, R. W. (1999). “Far-ultraviolet radiation from elliptical galaxies,” *Annu. Rev. Astron. Astrophys.* 603.

Petersen, C. C., and Brandt, J. C. (1995). “Hubble Vision: Astronomy with the Hubble Space Telescope,” Cambridge Univ. Press, Cambridge, UK.

Savage, B. D., and Sembach, K. R. (1996). “Interstellar abundances from absorption-line observations with the Hubble Space Telescope,” *Annu. Rev. Astron. Astrophys.* 279.

Spitzer, L., Jr. (1982). “Searching Between the Stars,” Yale Univ. Press, New Haven, CT.

X-Ray Astronomy

M. F. Corcoran
NASA, Goddard Space Flight Center

GLOSSARY

Accretion disk A flattened ring of gas surrounding a normal star or compact object. Gas in the ring spirals onto the central object, which increases the mass of the accreting object and may generate X-rays via the conversion of gravitational potential energy to heat.

Active galaxy A galaxy whose radiation is dominated by the emission from a localized region at the center of the galaxy. This region is thought to consist of a massive torus of gas surrounding a supermassive black hole about a few million times more massive than the sun.

Black hole A compact object produced by the collapse of a massive object such as a star or star cluster from self-gravity. The object collapses to such a small size that not even light can escape from it.

Corona The outermost part of a stellar atmosphere of a star like the sun. The temperature of the corona is typically a few million degrees, and thus the corona is the hottest part of the star that can be directly observed.

Dark matter Matter that does not emit electromagnetic radiation detectable at earth. Dark matter is only detectable from its gravitational effect on luminous matter and photons. Dark matter may comprise 90% or more of the total mass in the Universe.

Electron volt (eV) The change of potential energy experienced by an electron moving from a place where the electric potential has a value of V volts to a place where it has a value of $V + 1$ volts. This is a convenient energy unit when dealing with the motions of electrons and ions in electric fields; the unit is also the one used to describe the energy of X-rays and gamma rays. $1\ \text{eV} = 1.602177 \times 10^{-19}$ joules.

Neutron star Compact object formed from the death of a star of mass greater than about 10 times the mass of the sun, or from the induced collapse of a white dwarf in a binary system due to mass transfer. Typically neutron stars have a radius of about 10 km and a mass near a solar mass.

Planetary nebula The ejected outer atmosphere of a low to intermediate mass star that forms a glowing cloud of gas and dust around the stellar core. This cloud of gas and dust is called a planetary nebula.

Pulsar A rapidly spinning neutron star that emits pulses of electromagnetic radiation.

Solar mass An amount of matter equal to the mass of the sun, i.e., 2×10^{33} g.

Supernova remnant Debris ejected into space as a result of an explosion of a star whose mass is in excess of 10 solar masses.

White dwarf Collapsed remnant of the core of a star whose mass is near 1 solar mass. The mass of a white dwarf is limited to 1.4 solar masses, the so-called Chandrasekhar mass. White dwarfs typically have a radius of about 10^4 km.

X-RAY ASTRONOMY is the study of emission and absorption of high-energy X-radiation produced by cosmic objects, using space-based detectors. Though a relatively young field of astronomical research, over the last 30 years X-ray astronomy has revealed the richness of high energy phenomena in the Universe. X-radiation is produced by nearly all types of cosmic objects, from solar system bodies (planets and comets) to stars and galaxies, to large diffuse nebulae. X-ray emission provides an especially sensitive probe of regions of extreme gravity, extreme temperature, and extreme magnetic field strength in the Milky Way galaxy and in external galaxies.

I. OVERVIEW

The spectrum of electromagnetic radiation is divided into about seven discrete regions based on wavelength or energy. The X-ray energy band comprises radiation with energy in the range a few tenths to a few hundred kilo–electron volts (keV). This range roughly corresponds to the K shell absorption edge of neutral carbon up to the rest mass energy of the electron. Since photon wavelength is inversely proportional to energy, this corresponds to a range in photon wavelength of about one millionth to one ten billionth of a centimeter. For comparison, visible radiation comprises a range of wavelength of about four hundred thousandths to eight hundred thousandths of a centimeter, or energies of about 0.003–0.001 keV. X-ray radiation (first detected by Wilhelm Conrad Röntgen on 8 November 1895 at the University of Würzburg) is thus a form of high energy (short wavelength) radiation usually produced in regions of extreme temperatures or strong magnetic fields, and is often associated with large changes in kinetic energy produced near extremely strong gravitational fields. X-ray astronomy deals with the detection and analysis of X-radiation produced by stars, collapsed stellar remnants, diffuse hot gas, galaxies, and clusters of galaxies. Overall, X-ray emission is an important measure of some of the most energetic physical proceses in the Universe. Absorption of X-rays by material between the observer at earth and the source is also important as it gives astronomers a measure of the amount and composition of the intervening material. As in other branches of astronomy, X-ray astronomy in particular is concerned with understanding the spatial, spectral, and temporal distribution of the emitted radiation. However, because X-rays are so energetic, X-ray astronomy must utilize somewhat specialized detection techniques, though some detectors (like photographic plates and charged-coupled devices) which have been or still are widely used in visible-band astronomy are also useful detectors of X-rays. Because the atmosphere is mostly opaque to X-rays, X-ray astronomy necessarily requires the X-ray telescope to be placed above the bulk of the atmosphere, either carried aloft on a balloon or by a suborbital rocket, or on an orbiting space platform. In the following we discuss how X-rays are produced in the cosmos, and how they are detected and analyzed at earth.

II. PRODUCTION AND ABSORPTION OF X-RAYS

Electromagnetic radiation is the result of the acceleration of charged particles (electron and ions). Cosmic X-rays may be produced by both thermal and nonthermal processes. Thermal processes are those in which the velocity distribution of the relevant charged particles is determined solely by temperature. Hot, optically thick objects (such as planets and stars) will produce “blackbody” radiation in which radiation is emitted as a smooth continuum over a large wavelength range, though the emission peaks near some characteristic wavelength and falls quickly at shorter wavelengths and less rapidly at longer wavelengths. Wien’s law states that the peak of the blackbody continuum occurs at a wavelength λ_{max} given by

$$\lambda_{max} = \frac{0.29}{T}\text{cm} \tag{1}$$

where T is the temperature of the blackbody in kelvins (K). In order for the continuum peak to occur in the X-ray band, the temperature of the blackbody must be $T > 2.3 \times 10^5$ K; thus the presence of thermal blackbody X-ray emission implies high temperatures. For a spherical body of radius R radiating at some temperature T, the blackbody luminosity of the object is given by

$$L_{tot} = 4\pi R^2 S T^4 \tag{2}$$

where S is the Stefan–Boltzmann constant. A cloud of hot, low-density plasma will produce thermal continuum emission due to bremsstrahlung (“braking radiation”). Thermal bremsstrahlung is produced in optically thin plasmas by the acceleration of electrons by positively charged ions in the plasma. Bremsstrahlung is sometimes called “free-free” radiation since the radiation is produced by motions of unbound charged particles. Optically thin plasmas also produce line emission by collisional excitation of electrons to upper atomic bound states and radiative deexcitation to lower states. In general, all optically thin hot plasmas will produce a combination of X-ray continuum bremsstrahlung emission and X-ray line emission.

Nonthermal processes involve the motion of charged particles in the presence of magnetic fields. Examples of nonthermal emission processes are synchrotron and cyclotron emission, in which radiation is produced as electrons and/or ions accelerate around lines of magnetic force and radiate. Another astrophysically important nonthermal process is inverse Compton scattering, in which photons scatter off a nonthermal population of fast-moving electrons, and energy is transferred from the electrons to the photons. For sufficiently energetic electrons (especially those whose velocities approach the speed of light) enough energy may be transferred to boost the photon energy into the X-ray (or, conceivably, even the gamma ray) energy band. Production of X-radiation by nonthermal emission processes generally requires large populations of electrons moving near the speed of light. Gas near bright sources of X-rays can be photoionized, in which the population of the electronic energy levels and ionization state is determined by the incident X-ray spectral flux and not the temperature of the gas. Radiative excitation and deexcitation in photoinized plasmas can produce detectable line emission.

Cosmic X-rays are absorbed by intervening material between the source and Earth, either from material associated with the source (as in the case of an accretion disk, or a stellar wind) or from diffuse nebular gas in our galaxy. Absorption is produced by photoionization of hydrogen and other elements. The photon absorption cross sections of all elements decrease with increasing X-ray energy; thus in general absorption is important for low-energy or "soft" X-rays (X-rays of energy 0.1–1 keV), but less important for higher energy (or "hard") X-rays. Typically absorption is unimportant at energies higher than about 3 keV for all but the most highly absorbed sources. Since absorption is basically produced by photoionization, determining the degree to which the X-ray source emission is absorbed offers clues to the ionization state, the density, and the chemical composition of the material along the line of sight to the source.

III. X-RAY ASTRONOMY TECHNIQUES

A. X-Ray Imaging Systems

Early images of the X-ray sky were produced using collimated X-ray detectors, where a metal tube (called a collimator) was placed in front of the X-ray detector to restrict the view direction to some fraction of a square degree. Resolving finer spatial detail at X-ray energies is best accomplished through the use of an imaging system to bring the X-rays to focus. Unlike visible-band photons, X-ray photons cannot be focused by refraction or by single scattering from a mirror at near normal incidence, since X-rays at near normal incidence will penetrate most materials. However, X-rays can be focused by scattering at near grazing incidence. Use of grazing incidence X-ray optics in astronomy was first proposed by Riccardo Giacconi and Bruno Rossi in 1960. The simplest X-ray telescope mirror configuration is a paraboloid of revolution, in which X-rays along the axis of the paraboloid can be brought to a focus by scattering at near-grazing incidence from the surface of the reflector. This type of X-ray mirror was used on the SAS-2 (*Copernicus*) satellite. Simple paraboloids of revolution can only bring on-axis rays to focus, however; off-axis sources suffer aberration that severely limits the field of view of the paraboloid mirror.

On-axis and off-axis X-rays can both be focused by multiple scattering at near grazing incidence. The most common mirror geometry used in X-ray astronomy is the "Wolter type I" configuration (designed by Hans Wolter in 1952), which consists of concentric pairs of paraboloid–hyperboloid cylindrical mirror shells, as shown in Fig. 1. The mirrors are typically composed of finely polished glass coated with a thin layer of dense material such as gold or iridium that has a sufficiently high X-ray reflectivity. In order to maximize the scattering area, the paraboloid–hyperboloid shell pairs are nested so as to provide a maximum scattering surface for the incident X-rays. X-rays entering the mirror parallel to the mirror axis scatter first

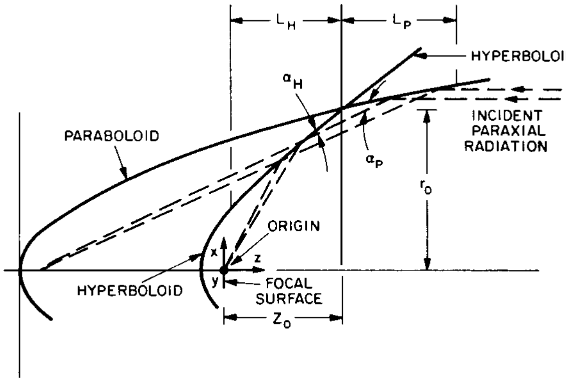


FIGURE 1 Focusing geometry of a Wolter type I grazing incidence X-ray telescope constructed by a paraboloid followed by a hyperboloid cylindrical mirror. In X-ray telescopes a number (2, 4, or hundreds) of such shells are "nested," or placed one inside the other, to improve the X-ray gathering power of the mirror.

off the paraboloid shells and then are brought to focus by scattering off the hyperboloid shells. The mirror collecting area, or "effective area," depends sensitively on photon energy and off-axis angle and is typically much less than 1 square meter in all cases (for example, the X-ray telescope on the *XMM-Newton* observatory, the largest X-ray telescope yet flown, has an effective area that varies from about 0.6 m^2 near 0.1 keV down to 0.1 square meter near 10 keV). The mirror focal length depends on photon energy due to the energy dependence of the scattering angle. In general the shape of the focal plane is curved, which means that, for example, images obtained by a plane parallel detector will show increasing distortion with increasing off-axis angle. Although the spatial resolving power of a telescope depends inversely on photon wavelength, the practical difficulty in bringing X-rays to focus means that the spatial resolution of X-ray telescopes is generally much worse than for an optical telescope of similar collecting area. The *Chandra* X-ray observatory contains the finest X-ray imaging telescope yet flown; the absolute limiting resolution of this telescope is about 0.5 seconds of arc. For comparison, the theoretical diffraction limit at 1 keV for *Chandra* is about 2.5×10^{-4} second of arc.

Wolter type I X-ray mirrors are typically very expensive to build since they need to be polished to very high tolerances, and are expensive to launch since they are generally heavy. One means of reducing costs is to approximate the Wolter type I reflecting surface through use of nested sets of thin, lightweight, conical aluminum foil shells. Although these thin-foil mirrors sacrifice some angular resolution (typically they can achieve spatial resolutions of a few minutes of arc), they can provide much greater throughput over a large energy range since hundreds of foils can be nested in a single mirror. Thin foil X-ray mirrors were used for the first time by the *Broad-band X-ray Telescope* (*BBXRT*) during the ASTRO-1 mission, and later by the *Advanced Satellite for Cosmology and Astrophysics* (*ASCA*) X-ray observatories.

B. X-Ray Detectors

Detectors used in X-ray astronomy include proportional counters, microchannel plates, and charge-coupled devices and other solid-state detectors. Proportional counters were the first type of X-ray detector used in astronomy and are the most common astronomical X-ray detector in use today. Proportional counters consist of an electrically neutral gas (usually argon or xenon) in a sealed chamber. As an X-ray enters the chamber, it can photoionize a gas atom, producing a photoelectron that can then be amplified and detected. Modern proportional counters can detect the position of the incident X-ray along with its time of arrival and energy, since the number of photoelectrons produced depends on the energy of the incoming X-ray. The relation between the energy of the incident photon and the amplitude of the electron pulse detected in the proportional counter is termed the gain of the detector. Examples of X-ray proportional counters include the Imaging Proportional Counter (IPC) flown on the *Einstein* X-ray observatory, the Position Sensitive Proportional Counter on the Röntgen Satellite X-ray observatory (*ROSAT*), and the Proportional Counter Array (PCA) on the Rossi X-Ray Timing Explorer (*RXTE*).

Use of charge-coupled devices (CCDs) is becoming increasingly common in X-ray astronomy. X-ray CCDs are similar to the more familiar optical CCDs, but have some important differences. In each case a photon interacts with a solid layer of material (usually silicon) over which an array of electrodes has been formed. The electrodes define potential wells that form the picture elements, or "pixels," of the image. Like proportional counters, CCDs can be used in photon counting mode to record the time of arrival, location, and energy of the incident X-rays. Because each pixel in a CCD must be read out by the detector electronics, X-ray CCDs are prone to "pileup," in which two or more photons arrive in a single pixel before it is read out and are counted as a single photon. In addition CCDs are prone to "charge transfer inefficiency" (CTI), in which the inferred energy of the photon may change during readout, producing a gain that varies spatially over the face of the detector. Examples of charge-coupled devices that have been used as X-ray detectors include the Solid-State Imaging Spectrograph used on *ASCA*, the Advanced CCD Imaging Spectrometer (ACIS) on the *Chandra X-Ray Observatory*, and the European Photon Imaging Camera (EPIC) on *XMM-Newton*.

Microchannel plates are composed of bundles of millions of extremely narrow lead-oxide glass tubes. The interior of each tube is coated with a photoelectrically sensitive material. An X-ray entering a tube collides with the interior of the tube wall and produces photoelectrons, which are then accelerated in an electric field, amplified as they move down the tube, and detected by an anode grid at the opposite end of the microchannel plate. Because each glass tube can be extremely thin (typically each tube is narrower than a human hair), microchannel plates offer the best spatial resolution of any type of X-ray detector. While providing extremely fine spatial resolution, the current generation of microchannel plate detectors is unable to provide any information about the energy of the incident X-ray. The first microchannel plate widely used as an X-ray detector in astronomy was the High Resolution Imager (HRI) on the *Einstein* observatory. A similar instrument was flown on *ROSAT*. The High Resolution Camera (HRC) on the *Chandra* X-ray observatory has the best spatial resolution of any microchannel plate detector yet flown.

C. X-Ray Spectrometers

By separating the radiation emitted by astrophysical objects into its component spectrum (the distribution of emitted intensity as a function of photon wavelength or energy), astronomers gain greater physical insight into the nature of the emission and the nature of the source of the emission. X-ray astronomical spectroscopy consists of measuring the spectrum of the X-rays emitted by a source in space. By studying the spectrum of the emitted X-rays, astronomers can determine important source parameters such as the temperature of the source, its composition, the magnetic field strength, and/or the relativistic electron population.

Crude spectroscopy can be achieved with proportional counters, since the size of the charge cloud induced by the incident X-ray is proportional to the energy of the X-ray. In practice the size of the charge cloud changes as it migrates to the anode in the counter body, so that the detected size of the cloud bears only a very approximate relation to the energy of the original photon. As a result proportional counters typically have very crude energy resolving power, $E/\Delta E \sim 1$–2 (where ΔE is the uncertainty in the derived photon energy E). As a result, the analysis of proportional counter spectra involves an assumption of an incident photon spectrum, which is convolved with an approximation of the energy smearing induced by the proportional counter (which can usually be statistically estimated to fairly high accuracy), and comparison of the “smeared” spectrum to the observed spectrum defined by the proportional counter. If the match is not satisfactory, the incident photon spectrum is modified and the process is repeated until a satisfactory match is achieved. In this sense, X-ray spectroscopy from proportional counter measurements is more akin to optical broad-filter photometry than it is to optical spectroscopy.

Higher X-ray energy resolution may be obtained by interposing an element in the light path to disperse the incident X-ray flux into its component spectrum. X-ray transmission gratings are typically made of a regular pattern of fine gold bands, and they operate in the same way as optical transmission gratings, by deflecting incident photons by an amount dependent on photon energy. The dispersed spectrum is then read out on a position sensitive detector (typically a microchannel plate or CCD). X-ray grating spectroscopy requires relatively large amounts of X-ray flux in order to generate a detectable signal in a reasonable amount of observing time. Gratings can obtain X-ray energy resolving powers of 1000.

Gratings work best with point sources, since extended sources produce complicated dispersed spectra that may be difficult to analyze. The spectra of extended sources (and faint point sources) can be obtained via nondispersive spectroscopy using solid-state detectors. Solid-state detectors called microcalorimeters can achieve resolving powers that equal or surpass the resolving power of transmission gratings. A microcalorimeter consists of a thermally sensitive absorber such as mercury telluride (HgTe) and a thermistor that changes its resistance when heated. An X-ray incident on the microcalorimeter results in a very slight temperature change (of the order of a few microkelvins) in the absorber; the rise in temperature of the absorber, and the resulting change in resistance of the thermistor, is proportional to the energy of the incident X-ray. Microcalorimeters need to be operated at extremely low temperatures (typically within a few thousandths of a degree of absolute zero) in order to measure the very small change in temperature produced by the incident X-rays. To reach temperatures this low, the device needs to be cooled using liquid helium as a refrigerant plus some type of active magnetic cooling system. Microcalorimeters that achieve energy resolving powers $E/\Delta E \sim 2000$ or higher have been constructed.

IV. A BRIEF HISTORY OF X-RAY ASTRONOMY

A. The Discovery of Cosmic X-Ray Sources

Development of reliable launch vehicles at the end of World War II ignited interest in the exploration of space, and especially in the hitherto unexplored regions of the electromagnetic spectrum that are blocked by the earth’s atmosphere. During this period there was an explosive increase in curiosity about the possible sources of emission and absorption of cosmic X-rays, especially at energies above 0.5 keV where absorption by the cold gas and dust in the galaxy is not a problem. Groups of scientists at the Naval Research Laboratory, American Science and Engineering (AS & E), the Massachusetts Institute of Technology (MIT), and the Smithsonian Astrophysical Laboratory (to name a few), and at other locations in the United States and abroad, began to identify mechanisms by which cosmic objects could emit and absorb X-rays, and to develop detectors and high-altitude and/or space-based platforms to observe this emission. X-ray astronomy currently is an international effort, with groups in the United States, the United Kingdom, Germany, Japan, Russia, France, the Netherlands, and other countries all making important contributions. In the following we present a brief discussion of some important milestones in X-ray astronomy, concentrating on technological developments and observational discoveries.

X-ray emission from the sun was first detected photographically in 1948 by a group from the Naval Research Lab in Washington, D.C., using as a detector a filtered photographic emulsion carried aloft on an Aerobee rocket. The sun, though, is a rather weak X-ray source (at least as cosmic X-ray sources go), and by extrapolation, it was judged that other cosmic X-ray sources would be difficult to observe with available detector technology if these sources were similarly faint but located at cosmic distances. It was already understood though that X-ray sources stronger than the sun could exist, and that objects such as supernova remnants, flare stars, and other magnetically peculiar stars might be sources of significant and detectable X-ray emission. The first celestial X-ray source outside the solar system was detected on 12 June 1962 by a group from AS & E led by Riccardo Giacconi using a set of geiger counter X-ray detectors on an Aerobee 150 sounding rocket. The primary purpose of this observation was to attempt to measure X-rays from the moon produced by the interaction of the solar wind with the lunar surface. During the course of this rocket flight, the detectors scanned a swath of sky and discovered a surprisingly strong X-ray source (soon dubbed "Sco X-1") in the constellation Scorpius, along with nonlocalized X-ray background emission of unknown origin. If Sco X-1 were a nearby star, its X-ray luminosity would be about 10–100 million times greater than the X-ray luminosity of the Sun, and indeed greater that the total solar luminosity over all wavelengths. This astonishing result pointed to the existence of a population of exceedingly bright X-ray sources in the sky. The next step was to find them and identify the physical mechanism giving rise to the emission.

B. The First Surveys

Spurred by the discovery of Sco X-1 and other strong sources, the next step was to try to spatially resolve and identify the sample of X-ray sources in the sky. The first successful all-sky survey of the X-ray sky was conducted by the first Small Astronomy Satellite (*SAS-1*, renamed *Uhuru* after launch). *Uhuru* was the first orbiting observatory dedicated to the study of X-ray emission. *Uhuru* was built by AS & E and launched from a floating platform off the coast of Kenya on 12 December 1970. *Uhuru* scanned the entire sky with two collimated proportional counters sensitive to X-rays in the 2–20 keV energy range. The collimators restricted the accuracy of source positions to only about 30 minutes of arc. This rather crude positional accuracy was sufficient to determine positions for all of the brightest X-ray sources in the sky, and to discover diffuse emission from clusters of galaxies. During its 2-year life it obtained positions for 339 X-ray sources in the 2–20 keV energy range and measured the temporal variability of the emission for many of these sources. Figure 2 shows a map of the *Uhuru* cosmic X-ray sources in galactic coordinates, clearly showing a population of X-ray sources along the plane of the galaxy along with a population of fainter sources off the galactic plane.

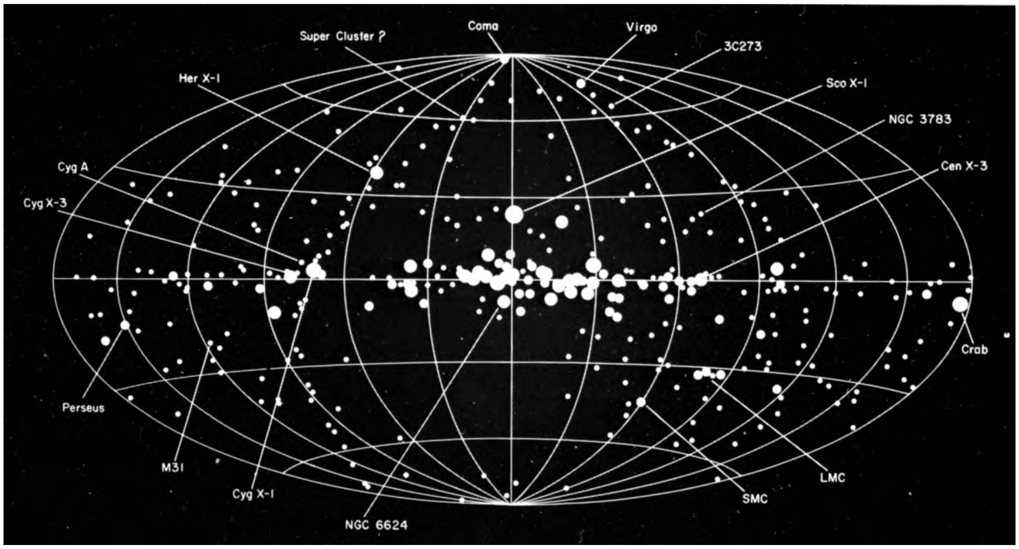


FIGURE 2 Map of *Uhuru* cosmic X-ray sources plotted in galactic coordinates and showing the relative intensity and source type for several major classes of source.

The richness of the X-ray sky uncovered by *Uhuru* ignited interest among astronomers to understand these sources and to discover new ones. In 1977 NASA launched the first in a series of large satellite observatories devoted to high-energy astronomy. The first High Energy Astrophysical Observatory, *HEAO-1*, was a designed to survey the entire sky every 6 months in the 0.1 keV–10 MeV energy range. The instruments on *HEAO-1* included sets of collimated proportional counters sensitive to low, medium, and high energy X-rays. *HEAO-1* mapped the position of bright X-ray sources to a precision of ∼30 seconds of arc (compared to ∼30 minutes of arc for *Uhuru*) with much higher sensitivity than *Uhuru*. *HEAO-1* produced a catalog of about 1000 X-ray sources and obtained identifications of hundreds of these sources because of the improvement in X-ray source position. *HEAO-1* produced a sensitive, brightness-limited catalog of sources off the plane of the galaxy and probed the distribution of extragalactic sources as a function of source brightness. In addition to carrying out the sky survey, *HEAO-1* was also operated in pointed mode, enabling sensitive studies of the time variability and energy spectrum of individual bright X-ray sources.

C. Imaging the X-Ray Sky

True X-ray images of the sun were first obtained by rocket flights in the 1960s. In the mid-1970s, a group at the Smithsonian Astrophysical Observatory led by Paul Gorenstein produced the first X-ray images of extra-solar objects using two orthogonal one-dimensional X-ray mirrors and a proportional counter detector. During rocket flights in March 1975 and December 1975 they used this instrument configuration to image the X-ray sky near the star Algol, and in June 1976 they imaged the extended X-ray emission around the Virgo cluster of galaxies. The first use of a Wolter nested grazing incidence telescope in X-ray astronomy was the Apollo Telescope Mount, which flew on *Skylab* in the early 1970's and obtained images of the solar corona. The first use of Wolter-type optics in extra-solar X-ray astronomy was by Saul Rappaport of MIT and collaborators who imaged X-ray emission from supernova remnants during the course of two rocket flights.

A major turning point in X-ray astronomy was the launch of the second High Energy Astrophysical Observatory (later christened the *Einstein* observatory) in November 1978. The *Einstein* observatory was an orbiting satellite observatory designed to provide the first direct images of large regions of the X-ray sky. Figure 3 shows a schematic drawing of the *Einstein* observatory. It carried the largest Wolter-type 1 X-ray mirror ever launched to that time, and was equipped with sensitive focal plane imaging detectors. *Einstein* carried on board two focal-plane cameras, the Imaging Proportional Counter and the High Resolution Imager microchannel plate detector. The

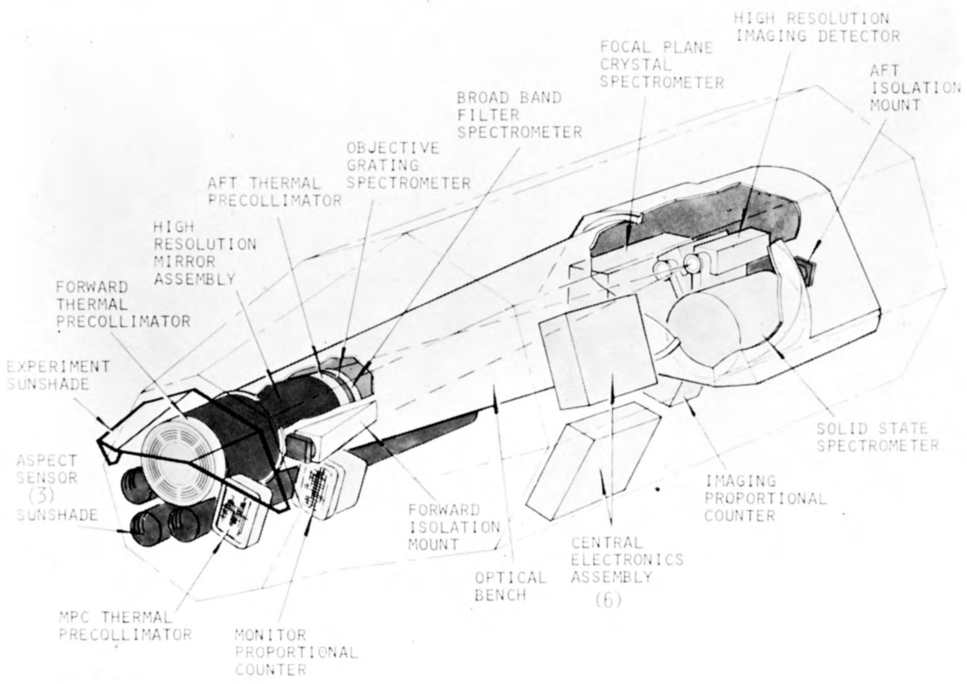

HIGH RESOLUTION IMAGING DETECTOR
FOCAL PLANE CRYSTAL SPECTROMETER
AFT ISOLATION MOUNT
BROAD BAND FILTER SPECTROMETER
OBJECTIVE GRATING SPECTROMETER
AFT THERMAL PRECOLLIMATOR
HIGH RESOLUTION MIRROR ASSEMBLY
FORWARD THERMAL PRECOLLIMATOR
EXPERIMENT SUNSHADE
ASPECT SENSOR (3) SUNSHADE
SOLID STATE SPECTROMETER
IMAGING PROPORTIONAL COUNTER
FORWARD ISOLATION MOUNT
CENTRAL ELECTRONICS ASSEMBLY (6)
OPTICAL BENCH
MPC THERMAL PRECOLLIMATOR
MONITOR PROPORTIONAL COUNTER

FIGURE 3 Overall layout of the *Einstein* X-Ray Observatory.

IPC provided spatial resolutions of a few minutes of arc in a 1° field along with crude sensitivity to photon energy. The HRI was insensitive to photon energy but could provide arcsecond-scale images over a 1° field. *Einstein* allowed astronomers for the first time to directly image the X-ray sky with a sensitivity two orders of magnitude greater than any previous X-ray observatory. *Einstein* also carried a focal plane spectrometer, the Solid State Spectrometer (SSS), for improved energy sensitivity in a 6-arcminute field. The highest energy resolution on *Einstein* could be obtained through use of the Objective Grating Spectrometer (OGS) or a Bragg crystal spectrometer (called the Focal Plane Crystal Spectrometer or FPCS), though dispersed spectra could be usefully obtained only for the brightest X-ray sources. In addition to the focal plane instrumentation, *Einstein* also carried a collimated proportional counter detector, the Monitor Proportional Counter (MPC). The *Einstein* mission lasted until April 1981. During this time *Einstein* produced more than 5000 images of the X-ray sky in the energy band 0.5–4.5 keV. *Einstein* revolutionized X-ray astronomy by showing the ubiquity of X-ray emission for almost all classes of astrophysical sources (from stars to galaxies), and allowed detailed study of important sources of X-ray emission such as supernovae, solar-type and non-solar-type stars, neutron star systems, and black holes, galaxies and galaxy clusters, as well as studies of the diffuse X-ray background.

The European X-ray Satellite observatory *EXOSAT* was launched in 1983, shortly after the end of the *Einstein* mission in 1981, and was operational for nearly 3 years. *EXOSAT* carried two Wolter type 1 X-ray grazing incidence mirrors for imaging of low-energy (0.05–2 keV) X-rays from space and a microchannel plate (the Channel Multiplier Array, or CMA) for high spatial resolution studies plus a proportional counter (the Position Sensitive Detector or PSD) for imaging studies at coarser spatial resolution but better spectral resolution. Unlike *Einstein*, which was placed in a circular orbit a few hundred kilometers above the earth, *EXOSAT* was placed in an elliptical orbit with apogee (maximum distance from earth) of 191,000 kilometers (about one-half the distance to the moon) and perigee (minimum distance from earth) of 350 kilometers. This orbit allowed efficient observations of X-ray sources, allowing up to 76 hours of uninterrupted observation out of every 90-hour orbit. Studies using *EXOSAT* showed the first evidence of "quasi-periodic oscillations" (QPOs) in the X-ray emission from low-mass X-ray binaries, discovered low-energy X-ray excesses in the emission from active galactic nuclei, and observed for the first time red- and blueshifted X-ray iron line emission from the low mass X-ray binary SS 433.

D. Expanding the Bounds of the X-Ray Universe

Recent advances in the development and deployment of X-ray satellite observatories have concentrated on probing the unexplored territory: direct X-ray imaging of large portions of the Universe, studies of extremely fast variations in X-ray brightness and spectral energy distributions, imaging of sources with extremely high spatial and spectral resolutions, and studies of the population of extremely faint X-ray sources.

1. *ROSAT*: Large-Scale Spectral Imaging

In 1990 the Röntgen Satellite (*ROSAT*) was launched. *ROSAT* was an international X-ray space observatory developed by Germany (who built the X-ray mirror, the spacecraft, and the Position Sensitive Proportional Counter detector for moderate spectral and spatial resolution studies), the United States (who provided a microchannel plate detector similar to the *Einstein* HRI for high spatial resolution studies, and who provided the launch vehicle), and the United Kingdom (who provided an extreme ultraviolet telescope coaligned with the X-ray telescope). The X-ray telescope on *ROSAT* had very fine imaging capability, achieving resolutions of about 1 minute of arc with the PSPC, and about 5 seconds of arc with the HRI, in the energy band 0.2–2.4 keV. The detectors also had extremely low internal backgrounds, which meant *ROSAT* was an ideal instrument to observe soft X-ray emission from low surface brightness objects. The observatory continued to make useful science observations up through the end of 1998, more than double the lifetime of other X-ray imaging observatories to that time. This longevity allowed *ROSAT* to detect more X-ray sources than any other X-ray observatory, and also allowed astronomers to undertake sensitive studies of X-ray variability on longer time scales than ever before. During the first 6 months after launch *ROSAT* conducted an "all-sky" survey in which the entire sky was imaged with the PSPC. After the all-sky survey phase, and for the next 8 years, *ROSAT* obtained deep pointed observations of particular regions of the sky containing X-ray targets of special astrophysical interest. *ROSAT* observations revolutionized almost all areas of X-ray astronomy in much the same way that *Einstein* did 12 years earlier. In addition to the many discoveries during the all-sky survey phase, in the pointed phase *ROSAT* made the first ever sensitive detection of sources in the so-called "Lockman Hole," a low column density region in the Milky Way, and showed that much of the hard X-ray background can be resolved into emission from active galactic nuclei. *ROSAT* also obtained the first X-ray image of a shadow in the soft X-ray background produced by an intervening molecular cloud and so provided the first firm lower limit on the location of a source of the soft X-ray background. *ROSAT* discovered a class of X-ray objects called “supersoft sources” and monitored the brightening of the X-ray emission from Supernova 1987A, the closest supernova explosion in more than 400 years. *ROSAT* made the first surprising detection of X-ray emission from comets and showed that the impact of a comet on Jupiter could produce observable X-ray emission. *ROSAT* also showed the ubiquity of X-ray variability for all classes of stellar X-ray sources. The *ROSAT* archive contains more than 11,000 pointed observations and a catalog of more than 90,000 X-ray sources, extending the number of known X-ray sources by about one order of magnitude.

2. *RXTE*: Fast Variability Studies

The Rossi X-Ray Timing Explorer (*RXTE*) was launched in 1996. *RXTE* was designed to provide extremely sensitive, high-time-resolution (to the microsecond level) studies of the variations of X-ray brightness and X-ray spectra from compact objects, active galaxies, and bright stars. A particular strength of *RXTE* is the capability to schedule coordinated observations with other observatories to allow observations of physical processes in astronomical objects over a wide range of wavelength regions. *RXTE*’s instrumentation includes the Proportional Counter Array (PCA), a set of six collimated proportional counters coaligned to point simultaneously at a specified area of the sky of roughly 1 square degree in size. *RXTE* also carries the High Energy X-ray Transient Experiment (HEXTE) designed to study source variability in the 15–250 keV range, and an All-Sky Monitor (ASM) designed to study variability of bright X-ray sources in the entire sky. One of the major discoveries of *RXTE* is the detection of the fastest periodic signals ever observed from an astronomical source. These “millisecond” pulsations (typically there are roughly 1000 X-ray flashes per second) are produced by some neutron star binary systems and are used to probe the condition of material in the extreme gravity very near the surface of the compact object. *RXTE* also showed that at least some of the hitherto mysterious objects known as soft gamma ray repeaters show X-ray pulsations as well and that these objects are probably types of compact X-ray binary systems with exceedingly strong magnetic fields. *RXTE* also produced the first detailed measure of the X-ray variability of Eta Carinae, a star that many astronomers think is the most massive star in the Milky Way; these observations showed a substantial, periodic eclipse of the X-ray emission every 5 years, probably produced by the presence of a companion star.

3. *BeppoSAX*: Solving the Gamma Ray Burst Riddle

BeppoSAX is a joint Italian–Dutch X-ray observatory that was launched in 1996; it consists of four narrow-field X-ray instruments (a set of medium-energy telescopes, a set of low-energy telescopes, a collimated gas scintillation proportional counter, and a “phoswich” detector system) plus two wide-field coded-mask cameras. *BeppoSAX* has made a major contribution to X-ray astronomy and high-energy astrophysics in general by detecting the first X-ray afterglow from a “gamma-ray burst” source. Gamma-ray bursts are brief flashes of gamma rays (a type of extremely high energy radiation) that occur nearly once per day. Though they were originally detected by space satellites in the 1970s, their origin remained a mystery because of the poor angular resolution of gamma-ray detectors. The error box of a typical gamma-ray burst position is generally many minutes of arc, and typically each error box contains dozens of candidate sources, making source identification difficult. However, in February 1997 a gamma-ray burst error box was imaged by the narrow-field X-ray detectors on *BeppoSAX*, and *BeppoSAX* discovered a bright, fading X-ray source near the gamma-ray burst position. This source was recognized as the X-ray “afterglow” of the gamma-ray burst, and thanks to the higher spatial resolution of the narrow-field X-ray detectors, an accurate position of the source could be obtained. Subsequent follow-up observations using this position by radio and optical telescopes associated this gamma-ray burst with a distant, star-forming galaxy. Because of this observation and others like it, it is now thought that some gamma-ray bursts signal incredibly powerful explosions produced by the destruction of very massive stars in distant, star-forming galaxies.

4. *Chandra*: The Finest X-Ray Camera

The *Chandra* X-ray observatory was launched in the summer of 1999. *Chandra*, named for the famed astrophysicist Subrahmanyan Chandrasekhar, is the third of the “Great Observatories” launched by NASA (the Hubble Space Telescope and the Compton Gamma Ray Observatory are the other two) and is designed to image the X-ray universe with the highest spatial and spectral resolution ever obtained. *Chandra* has the finest X-ray mirror ever made. Combined with *Chandra*’s CCD camera (the Advanced CCD Imaging Spectrometer) and microchannel plate detector (the High Resolution Camera), *Chandra* can study brightness and spectral variations on extremely small spatial scales (on the order of 0.5–1 second of arc). In addition, *Chandra* carries aboard three sets of transmission gratings that enable the study of X-ray spectra with high resolving power ($E/\Delta E \sim$ 100–2000) over a broad energy range

(0.1–2.0 keV with the Low Energy Transmission Grating, and 0.4–10.0 keV with the Medium and High Energy Transmission Gratings). Like *EXOSAT*, *Chandra* has been placed in a highly elliptical orbit; its apogee takes *Chandra* about one-third of the distance to the moon. Near apogee, *Chandra* is able to observe a large fraction of the X-ray sky for long intervals uninterrupted by earth eclipses, allowing for a much more efficient use of observing time. At the very start of its science life as of this writing, already among *Chandra*'s science achievements are the discovery of a previously unseen collapsed stellar core at the center of the Cas A supernova remnant, the identification of the physical processes by which comets emit X-rays, and sensitive spatial studies of X-ray jets from active galaxies and galactic neutron star systems.

5. *XMM-Newton*: Probing Extremely Faint Sources

The X-ray Multi-Mirror Mission was launched in December 1999 by the European Space Agency; after launch it was renamed in honor of Isaac Newton. *XMM-Newton* consists of three coaligned, high-throughput X-ray mirrors and three sets of focal plane detectors: a CCD camera (the European Photon Imaging Camera, or EPIC) behind one of the mirror modules, and reflection gratings (the Reflection Grating Spectrometers, or RGS) and cameras to read out the dispersed spectra behind the other two modules. *XMM-Newton* provides for simultaneous 0.1–15 keV X-ray imaging (on spatial scales of about 15 seconds of arc) with its CCD cameras and X-ray spectroscopy (with a resolving power between 200 and 800) in the energy range 0.3–2.5 keV via the reflection gratings. Because of the high throughput of the X-ray mirrors *XMM-Newton* can study faint sources in unprecedented detail. At the time of this writing *XMM-Newton* is just beginning to take scientific observations after its instrumental on-orbit checkout interval. Nevertheless, *XMM-Newton* has already discovered a cluster of galaxies hidden from previous X-ray observatories by the gas and dust in the plane of the Milky Way, obtained a high-resolution X-ray spectrum of the star HR 1099, and discovered hundreds of faint, unidentified sources.

V. PROPERTIES OF COSMIC X-RAY SOURCES

A. Sources of X-Ray Emission in Normal Galaxies

1. Normal Stars

Nearly all types of normal (i.e., noncollapsed) single stars generate X-ray emission at some level. Stellar X-ray emission is thought to be dominated by thermal emission arising from one or more of the following processes: magneto-hydrodynamic heating in the outer stellar atmosphere; frictional heating; and shock heating produced by instabilities in an outwardly flowing stellar wind. The sun is the brightest cosmic X-ray source, yet has a rather modest X-ray luminosity compared to other observed stars. The overall flux of X-rays produced by the sun is also small in comparison to the solar emission at other wavelengths. The solar X-ray flux at earth is about 1×10^{-8} W m^{-2} (in the wavelength range 1–8 $\times 10^{-8}$ cm), so that the solar X-ray luminosity is roughly 3×10^{26} erg s^{-1}. This is only one ten-millionth of the solar visible-band luminosity. The X-ray emission can vary by factors of 10,000 on short (minutes–hours) time scales. In the sun X-ray emission is produced in the corona, the outermost layer of the solar atmosphere. Figure 4 shows an image of the soft X-ray emission from the solar corona taken by the *Yohkoh* satellite. The temperature of the corona is about 2 million degrees and varies spatially by about a factor of 3. The corona is heated by the interaction of the stellar magnetic field and the ambient gas. The magnetic field is produced by the so-called dynamo mechanism, in which the rotation of the convective outer stellar envelope generates subsurface electric currents that produce a magnetic field

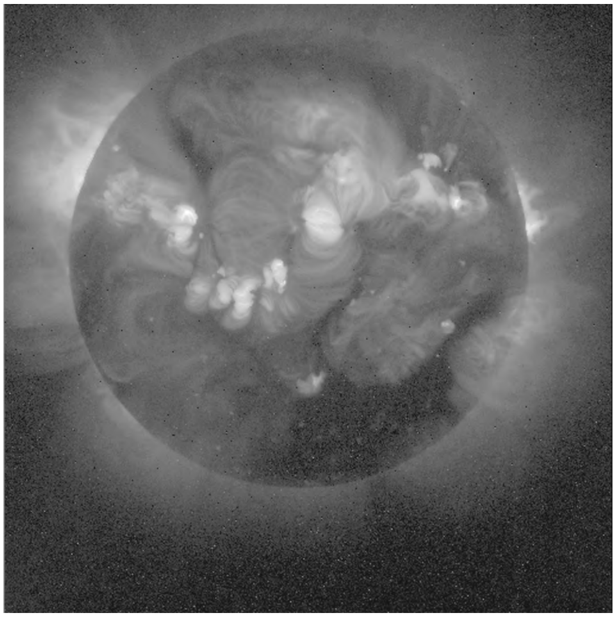


FIGURE 4 Image of the soft (0.25–4.0 keV) X-ray emission from the sun, from the the soft X-ray telescope on board the Japanese–U.S.–U.K. *Yohkoh* solar satellite observatory. X-ray emission from the sun is confined to the corona, the outermost layer of the sun's atmosphere. Thermal X-ray emission is produced by conversion of magnetic energy to heat, causing the outer solar atmosphere to reach temperatures of about 2 million degrees. [Credit: ISAS and the Yohkoh SXRT team.]

whose axis is roughly aligned with the rotational axis. The magnetic field threads the solar surface and the solar atmosphere. Heating of the corona is due to twisting of the field lines, which locally intensifies the field; magnetic pressure is converted to kinetic gas energy via disruption and reconnection of the magnetic field lines. Differential rotation in the outer solar envelope causes the strength of the magnetic field to vary with an 11-year cycle; this produces a cyclical variation in the numbers of sunspots on the solar surface, and a cyclical variation in solar X-ray emission. At times of sunspot maximum, the solar X-ray flux can increase by a factor of 100, and during solar flares, the solar X-ray emission can increase by a factor of 10,000 for brief periods. Detailed studies of the solar X-ray flux show that the spectrum is dominated by emission lines, indicating the predominance of thermal emission processes.

Solar-type stars (stars that have masses below about five times the sun's mass) are also known X-ray sources, though the nature of the emission in these stars is less well studied than in the solar case. Since these stars have convective outer envelopes and radiative cores like the sun, the emission mechanism is thought to be dominated by magnetic heating in the stellar coronae in a manner similar to the solar corona. Generally the spectrum of the emission has not been well studied, though modest resolution X-ray spectra show the presence of line emission, again suggesting the predominance of thermal emission processes. There is strong evidence that the X-ray luminosity is related to the stellar rotational velocity, indirect evidence of the importance of the stellar dynamo in generating X-rays in these stars. No clear periodicity or cyclical variability in the X-ray flux from solar-type stars has yet been identified, so it is unclear whether the type of activity cycle represented by the sunspot cycle is a general property of all solar-type stars. X-ray emission from solar-type stars is known to be variable, and rapid brief increases in the X-ray emission have often been seen; presumably these "flares" are similar to the better studied flaring activity seen in the sun.

Very massive stars (stars more massive than about 10 solar masses) have radiative envelopes and convective cores, which means that the stellar dynamo is not very effective in these stars. Naively, one expects that these stars would be weak X-ray sources, so it was somewhat of a surprise when the *Einstein* observatory conclusively detected X-ray emission from a sample of massive stars. It is thought that the emission from these stars is produced, not via magnetic heating, but via shock heating of a portion of the outwardly moving, unstable stellar atmosphere. Very massive stars produce a large flux of radiation at ultraviolet wavelengths and longer; matter at the stellar surface absorbs photon momentum from the UV radiation field and is driven outward as a massive stellar wind. The radiative driving mechanism is very unstable to velocity perturbations, and it is thought that such seed perturbations can steepen into strong shocks, converting wind kinetic energy to thermal gas energy; the relative velocities are such that temperatures should reach millions of degrees and produce observable X-rays. As yet, however, there is no predictive, quantitative physical model to describe the observed emission. Observationally, there seems to be a strong correlation of X-ray luminosity to the total amount of power radiated over all wavelengths, $L_x/L_{total} \sim 10^{-7}$. Since the total luminosities of these stars are generally hundreds of thousands to millions of times larger than that of the sun, massive stars generate about the same power in X-rays that the sun generates over all wavelength bands. Typically X-ray emission from massive stars shows little intrinsic absorption from the dense stellar winds, suggesting that at least some of the emitting region exists far from the photosphere. Generally the X-ray emission is not strongly variable, again suggesting that the emission is distributed over a relatively large spatial region.

2. Compact Objects and Accretion

Compact objects generally represent the remains of the stellar core after the rest of the star's atmosphere has been ejected in either a supernova explosion or by the formation of a planetary nebula. There are three classes of compact objects, representing increasingly higher maximum densities and gravitational fields. White dwarfs are objects in which approximately 1 solar mass of material is compressed to a volume of radius ~4000 km, with an average density of 10^6 g cm^{-3}. Neutron stars compress about 1 solar mass of material into a sphere of radius of only ~10 km and have an average density of 10^{14} g cm^{-3}. Black holes represent the ultimate crushing of matter by gravity, in which a singularity, or a region of infinite density, is produced after matter is compressed to greater than nuclear densities (densities in excess of 10^{14} g cm^{-3}). Though the singularity represents the greatest possible density achievable in nature, the actual average density of a black hole can be modest. The "size" of a black hole is given by the "Schwarzschild radius"

$$R_{sch} = 3(M/M_{\odot})\text{km} \tag{3}$$

where R_{sch} is the Schwarzschild radius, and $M/M_{\odot}$ the ratio of the mass of the black hole to the mass of the sun. Objects within 1 Schwarzschild radius of the singularity cannot communicate with more distant objects and may be considered "inside" the black hole. Thus the average density of a black hole (i.e., the amount of matter within the volume defined by the Schwarzschild radius) is $1.8 \times 10^{16}(M/M_{\odot})^{-2}$ g cm^{-3}. For a 1-solar-mass black hole, the average density inside the Schwarzschild radius is 1.8×10^{16} g cm^{-3}; however, for a supermassive, 100 million solar mass black hole, the average density inside the Schwarzschild radius is only1.8 g cm^{-3}.

Isolated neutron stars and white dwarfs are initially formed with surface temperatures representative of the temperature of the core of the progenitor star, $T \sim 10^7$–10^8 K, and as such should be sources of thermal blackbody X-ray emission, though the spectrum may be modified by the residual atmosphere of the compact object. Detection of this emission is difficult since the compact object cools quickly to temperatures too low to produce X-rays. The pointlike object at the center of the *Chandra* X-ray image of the supernova remnant Cas A shown in Fig. 5 may represent a young, hot isolated neutron star, though the emission appears too faint to be produced over the entire surface of a 10-km-radius neutron star and may rather represent emission from a small hot region on the neutron star surface.

Neutron stars may emit tightly collimated beams of radio radiation near the magnetic poles. If the rotational axis is not coaligned with the magnetic axis, these beams can be detected as pulses of radiation when the beam is directed towards earth. Such radio "pulsars" are often X-ray pulsars as well. Figure 6 shows a *Chandra* X-ray image of what is perhaps the most famous of this class of objects, the Crab Nebula, the remains of a massive star which exploded in our Galaxy in AD 1054. The Crab Nebula contains a spinning neutron star (which rotates 30 times per second) that powers a bright nebula surrounding the pulsar. The rotating neutron star is located at the center of the bright, disk-like rings of X-ray emission seen in the *Chandra* image. These X-ray rings are thought to represent nonthermal emission from high-energy particles flung outward over a distance of a lightyear from the neutron star. X-ray jets from the neutron star may be seen emanating from the pulsar perpendicular to the "rings." A subset of these X-ray pulsars, called anomalous X-ray pulsars (AXPs), has recently been identified. AXPs are thought to represent isolated neutron stars, with extremely high surface magnetic fields and extremely rapid spin rates of up to 1000 revolutions per second.

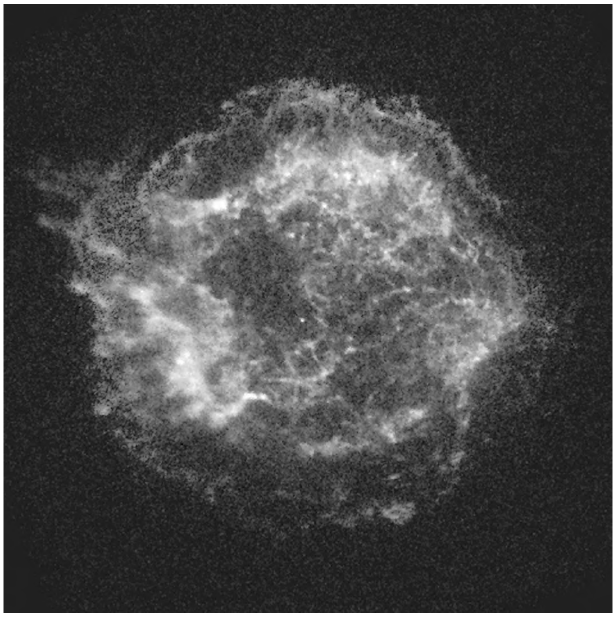


FIGURE 5 *Chandra* X-ray Observatory image of the Cas A supernova remnant. The point source at the center of the image is believed to be either a neutron star or black hole produced by the supernova explosion. [Credit: NASA/CXC/SAO.]

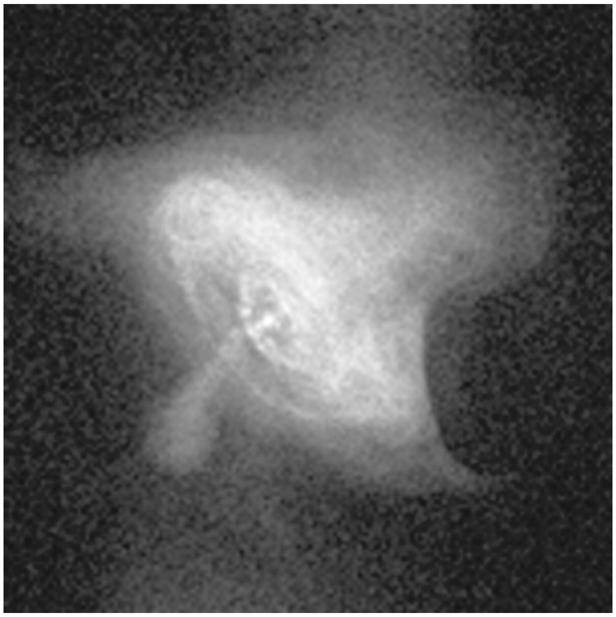


FIGURE 6 *Chandra* X-ray Observatory image of the Crab nebula. The source at the center of this image is the 33-ms Crab pulsar. The "rings" surrounding the pulsar are believed to be produced by nonthermal X-ray emission produced by populations of high-energy particles flung out from the pulsar to distances of more than a lightyear. X-ray "jets" can be seen perpendicular to the "rings." [Credit: NASA/CXC/SAO.]

Binary systems consisting of a "normal" star plus a compact companion are common sources of X-rays in the galaxy, and among the strongest galactic X-ray sources. In such systems the X-ray emission is generally produced by accretion of portions of the outer atmosphere of the "normal" star onto the surface of the companion. High-mass X-ray binaries (HMXBs) consist of a compact object plus a high mass ($M > 5\,\mathrm{M}_\odot$) companion star; since such high-mass stars generally possess strong stellar winds, X-rays can be produced by accretion of the stellar wind material onto the compact object. Low mass X-ray binaries (LMXBs) are systems in which the companion to the compact object is a low-mass star ($M < 1\,\mathrm{M}_\odot$).

In LMXBs, the transfer of material from the low-mass ("donor") star to the compact object usually occurs in the form of a thick stream from the donor to the compact object, and because of the angular momentum of the system, the transferred material does not fall directly onto the surface of the compact object, but instead forms an "accretion disk" around the compact object. Frictional heating can raise temperatures in the inner part of the disk to millions of degrees, so that the disks can radiate in X-rays. In addition, as material falls from the disk into the deep gravitational well of the compact object, kinetic energy is efficiently converted into heat, also producing X-ray emission. Accreted material on the surface of the compact object can reach sufficient densities and temperatures to start thermonuclear fusion of the accreted material, producing a flash of X-rays called an X-ray nova. Hot material can also be funneled toward the magnetic poles by the magnetic field to produce synchrotron X-rays. By some not-well-understood mechanism, material near the magnetic poles can become accelerated in a confined jet. The material in the jet can produce particle/antiparticle pairs that may produce very high energy X-rays and gamma rays. Such an object may be detectable as an X-ray pulsar if the beam is suitably aligned toward the earth. X-ray emission can also be produced by inverse comptonization of lower-energy photons as they interact with the accelerated charged particles near the magnetic poles. Binary systems in which one of the stars is a compact object may also show transient X-ray variability, i.e., flaring activity and/or fading of the emission, presumably due to the thermonuclear detonation of accreted material at the surface of the object. A now famous example of a single object which shows both regular (pulsing) variability and transient variability in X-rays is the so-called "Bursting Pulsar," GRO J1744-28. This system shows pulses of X-rays produced by a neutron star spinning with a period of 0.5 s in an 11.8-day orbit around a low-mass companion. Irregular bursts of X-ray emission from this star lasting from seconds to minutes have been observed by *RXTE*.

If the compact object is a black hole, emission should be dominated by the emission from the accretion disk, with no observable high-energy radiation from the black hole itself. Accretion onto black holes can result in the formation of collimated jets of material ejected from somewhere near the black hole that can produce observable X-ray emission. Low-mass X-ray binary systems containing black holes tend to show transient X-ray variability, while high-mass black hole X-ray binaries tend to show relatively constant X-ray emission. Of the 33 bright X-ray transients currently cataloged, 18 are thought to be LMXBs containing a black hole. On the other hand, the three best studied black hole systems that show presistent X-ray emission (LMC X-1, LMC X-3, and Cyg X-1) are all thought to have high-mass companion stars.

3. Supernova Remnants

A supernova is an extremely powerful explosion that accompanies the death of a massive star. During the explosion the stellar envelope and atmosphere of the star above the core are blasted outward. This ejected material moves at very high velocities (approaching 10% of the speed of light), and as it collides with the ambient gas and dust in the host galaxy, it deposits enormous amounts of energy into the galaxy, producing a roughly spherical shell of shock-heated, million-degree gas. Supernova remnants are generally strong, spatially resolved sources of thermal X-ray emission as exemplified by the *Chandra* Cas A image shown in Fig. 5. The X-ray emissivity of the remnant is expected to vary in the early stages as the remnant first expands into a low-density region carved out by the stellar wind of the precursor star prior to the explosion. Eventually the ejecta encounters denser material in the interstellar medium, producing a strong shock and substantial X-ray emission. With time the shocked material should cool and the X-ray emission from the remnant will fade. SN 1987 A, a supernova in the Large Magellanic Cloud (a companion galaxy to the Milky Way) that exploded in 1987, presents astronomers with a unique opportunity to study the evolution of the high-energy emission from a supernova explosion. The first X-ray observation of SN 1987A, by *ROSAT* during the all-sky survey in 1990, resulted in a rather large upper limit, though later that same year *BBXRT* put a much tighter constraint on the X-ray flux. Monitoring the source with *ROSAT* through the 1990s first detected the supernova as an X-ray source and measured the rise of its X-ray flux. An observation of SN 1987A with *Chandra* just 12 years after the supernova has already resolved a small ring of emission expanding outward from the explosion.

Study of the X-ray morphology of supernova remnants provides important clues to the poorly understood mechanism by which massive stars explode. Observation of the Cas A remnant with *Chandra* suggests the presence of an asymmetry in the remnant, possibly indicating that the explosion of the central star was nonspherical. Spatially resolved X-ray observations of supernova remnants allow astronomers to measure how the ejected material interacts with the local interstellar medium. An *ASCA* observation of the supernova remnant SN 1006 indicated two sites of nonthermal X-ray emission in the remnant, the first strong evidence that supernovae can accelerate particles to extremely high velocities and produce so-called "cosmic rays." Imaging proportional counter observations with *Einstein*, *ROSAT*, and *ASCA* and solid-state detector observations with *ASCA*, *Chandra*, and *XMM-Newton* allow astronomers to map out temperature and column densities in a number of supernova remnants such as the Cygnus Loop, the Puppis A remnant, and N132D, while high spatial resolution observations with the HRI instruments on *Einstein* and *ROSAT*, and the *Chandra* HRC, have allowed astronomers to study the filamentary structure of the hot shock interface in great detail.

Supernova explosions pollute their host galaxies with chemical elements heavier than hydrogen, produced by nuclear processing in the core of the star prior to the supernova, or created in the collision between the supernova ejecta and the interstellar medium. Spatially resolved X-ray spectroscopy is used to determine elemental abundances in the remnant so as to quantify the production of these heavy elements and to measure their modes of distribution into the galaxy. The *Chandra* observations of Cas A, for example, have revealed spatial variations in the abundances of iron, calcium, and silicon in the remnant, which may indicate the importance of mixing in the atmosphere of the star prior to the supernova explosion.

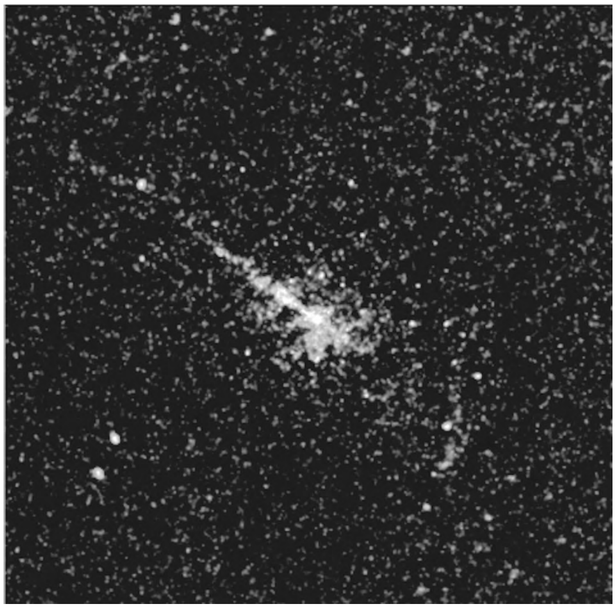


FIGURE 7 *Chandra* X-ray Observatory High Resolution Camera image of the radio galaxy Centaurus A (NGC 5128), at a distance of 10 million lightyears from earth, showing an X-ray jet emanating from the active nucleus of the galaxy. The active nucleus of the galaxy is also a source of X-rays and presumably houses a supermassive black hole. [Credit: NASA/CXC/SAO.]

B. Extragalactic X-Ray Sources

1. Active Galaxies

Active galaxies are those galaxies in which the emission is dominated not by the sum total of stars and gas, but by a localized region near the center of the galaxy. Such galaxies often show tightly collimated jets of emission stretching for millions of lightyears from the center of the galaxy, and are often violently variable on short (days or weeks) time scales. Figure 7 shows an X-ray image of the jet from the active galaxy Centaurus A obtained by the *Chandra* HRC. Because these galaxies are so bright, they can be used as probes of the very distant universe. Active galaxies are divided into two broad classes, quasars (quasi-stellar radio objects) and Seyfert galaxies. Current ideas (the so-called "unified model") suggest that all the phenomena associated with the various types of active galaxies are produced by a supermassive black hole (of a few million solar masses) surrounded by an accretion disk (which feeds the black hole) and an "obscuring torus" that can hide the black hole for certain viewing geometries; the various types of active galaxies are then simply the effect of viewing this system pole on, edge on, or at some intermediate angle. Active galaxies are strong X-ray sources, and the X-ray emission provides a means of probing the "central engine" of the active galaxy, referred to as the active galactic nucleus (AGN).

X-ray emission was first detected from the quasar 3C273 by early rocket and balloon flights. Three additional AGN were detected by *Uhuru*, and many more by *HEAO-1*, *Einstein*, *ROSAT*, and other modern X-ray observatories. Active galaxies typically show large X-ray luminosities ($L_x \approx 10^{43}$), emitting a substantial fraction of their total energy entirely in the X-ray band. The X-ray emission is variable, and typically X-ray variations occur on the shortest observed time scales, an indication that the X-ray emission arises from the smallest observable volume near the active nucleus. No AGN is known to show periodic variability (one putative periodic AGN, NGC 6814, was subsequently shown to be confused in early, low angular resolution observations with a nearby X-ray binary). The correlation between X-ray variability and variability at other wavelengths is not well known because of the difficulties of scheduling long-term near-simultaneous observations on X-ray and optical, ultraviolet and/or radio telescopes. The best near-simultaneous monitoring yet obtained is a study of the Seyfert galaxy NGC 7469 over about 1 month starting in June 1996 with *RXTE* in the X-ray band and the International Ultraviolet Explorer (*IUE*) in the ultraviolet. This monitoring effort showed that the X-ray variations were correlated with changes seen in the UV; the UV maxima seemed to precede the X-ray maxima by about 4 days, though the minima in each wavelength region were nearly simultaneous.

The X-ray emission from AGN shows a fairly complex spectral energy distribution, a combination of nonthermal and thermal continuum emission, along with narrow and broadened lines that are useful probes of universal expansion and the strong gravity environment in which the X-ray emitting accretion disk is embedded. An example of this is the active galaxy MCG-6-30-15, in which observations with the *ASCA* X-ray observatory of the iron line near 6.7 keV showed the line to be contorted into a "devil's horn" shape, which has been interpreted as a combination of the effects of motion of the inner edge of the accretion disk and gravitational effects just beyond the central black hole.

Absorption of the emitted X-rays also provides a useful probe of the environment in the galactic nucleus. Absorption edges due to oxygen at low X-ray energies and iron at higher X-ray energies have been detected for some active galaxies and seem correlated with the AGN subtype, generally in agreement with the "unified model." More recently X-ray absorption in discrete atomic spectral lines has been resolved by the *Chandra* transmission gratings in the active galaxy NGC 3783. The absorption lines in this system are systematically shifted toward negative velocities of about 400 km s^{-1}, suggestive of absorption in an outflowing medium.

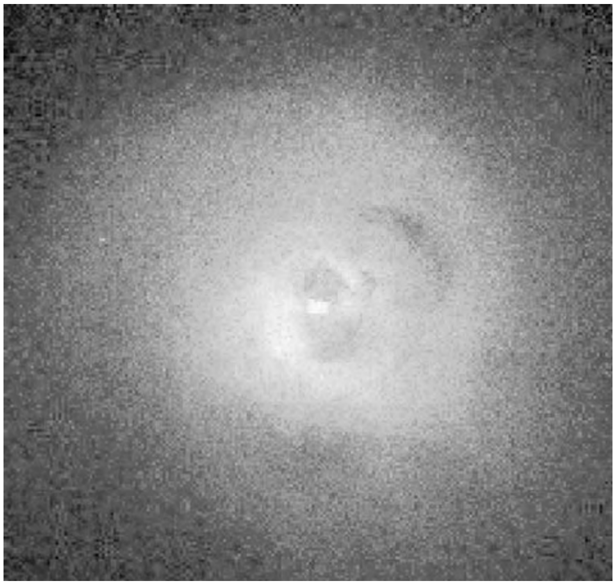


FIGURE 8 X-ray emission from the Perseus cluster of galaxies, as seen by the ACIS camera on the *Chandra* X-ray observatory. The image shows a region of about 640,000 by 640,000 lightyears. The most intense X-ray emission is associated with the largest member of the cluster, the supergiant galaxy Perseus A. Two dark cavities, each about half the size of the Milky Way, are thought to be buoyant magnetized bubbles of energetic particles produced by energy released from the vicinity of the black hole at the center of Perseus A. [Credit: NASA/IoA/A. Fabian *et al.*]

2. Galaxy Clusters

Clusters or gravitationally bound groups of galaxies often show extended X-ray emission in addition to the X-ray sources that may be associated with individual cluster members. Figure 8 shows the extended X-ray emission associated with the Perseus cluster of galaxies as seen by *Chandra*. Such extended X-ray emission is produced by optically thin thermal emission from intracluster gas at temperatures of millions of degrees. The heating mechanism is not well known, but presumably explosive motions of gas from supernovae in member galaxies, or heating due to the motion of galaxies through the intracluster medium, or residual heat from the gravitational collapse of the cluster may all play some role. Recent studies of galaxy clusters by *ROSAT* suggest that about one-half of all nearby clusters show evidence of extended X-ray emission.

The assumption that the hot intracluster gas is gravitationally bound to the galaxy provides the means for determining a lower limit to the cluster mass from the X-ray emission by equating the gravitational force per unit area on the X-ray gas to the thermal gas pressure. Cluster masses derived from X-ray measures are typically much larger than the masses estimated from the luminous matter in the cluster, suggesting that most of the matter in galaxy clusters is in the form of nonluminous, dark matter.

X-ray images of the extended cluster emission obtained by *Einstein* and *ROSAT* often show that the spatial distribution of the intracluster gas is centrally peaked. Such observations indicate a strong condensation of matter in the center of the extended emission, since the X-ray emissivity is proportional to the square of the gas density. These strong central peaks are usually interpreted in terms of a "cooling flow" in which the cooling of the dense gas at the center via X-ray emission produces a slow "infall" of outer material. As material moves into the center of the intracluster gas, inhomogeneities in the infalling matter should cause condensations of cool matter to fall out of the flow. Condensation rates of 10–100 solar masses per year are implied by the X-ray observations.

Since the intracluster medium will become polluted by heavy elements because of the explosion of massive stars in the member galaxies, the amount of heavy elements in the cluster gas is a clue to the efficiency of this process, and an indirect clue to the heating mechanism. Recent observations with *ROSAT* and *ASCA* suggest that many observed clusters have lower than solar iron abundance, suggesting that chemical pollution by supernovae is not so important. However, spatially resolved X-ray spectra of galaxy cluster emission by *ROSAT* show that the central emission regions tend to have higher iron abundances than the outer regions. This may suggest that mixing of polluted intracluster gas with primordial, low metal abundance gas is important, at least in the outer regions of the cluster.

C. The X-Ray Background

The X-ray band, along with the microwave band, are the only two regions of the electromagnetic spectrum in which the energy density of the emission from unresolved and/or diffuse sources is comparable to (or greater than) the emission from individual resolved sources. Whereas the microwave background is generally understood as the remnant thermal radiation produced by the Big Bang, the origins of the X-ray background are still somewhat of a puzzle, despite the fact that the X-ray background was the first cosmic background component to be discovered. The X-ray background seems to consist of two distinct components: hard emission ($E > 3$ keV) that is nearly isotropic, and soft emission that is spatially structured.

The isotropy of the hard component of the X-ray background suggests that it originates beyond the Milky Way. This background component is currently thought to be made up of unresolved X-ray emission from active galactic nuclei, an idea suggested by early results, especially with the *HEAO-1* observatory. Later, observations with the *Einstein* observatory began to resolve some of the hard X-ray background into point sources. More recently, deep X-ray images with *ROSAT* have shown that much if not all of the hard X-ray background can be resolved with sufficient angular resolution and sensitivity into point-source emission. Subsequent identifications of the point sources show that they are indeed primarily associated with active galaxies. The hard X-ray sky apparently shows the effects of “Olbers’ Paradox”, in which the sky glows since a source lies along the observer’s line of sight regardless of direction, and since the number density of sources increases with look-back time. Observations of the hard X-ray background thus provide useful information about the mass distribution in the Universe as a function of both space and time. Since mass accretion onto massive black holes is thought to power the X-ray emission from AGN, the measure of the hard X-ray background also provides a measure of the integrated mass accretion history onto massive black holes in the Universe.

Because of its sensitivity to low-energy X-rays, its low internal background, and its all-sky coverage, observations with the *ROSAT* observatory have recently played a dominant role in the interpretation of the soft X-ray background. At energies below 3 keV the X-ray background is highly anisotropic. Because absorption is much more important at soft X-ray energies, the soft component of the X-ray background traces not only the source of emission but the absorption of this emission by cold material in the galaxy. The importance of absorption is implied by an observed anticorrelation between the brightness of the diffuse X-ray background at $E \sim 1/4$ keV and the observed column density through the galaxy. This was best demonstrated by *ROSAT* images in which clear shadows of cold neutral hydrogen clouds could be seen in silhouette against the bright soft background. This anticorrelation of X-ray brightness and column density is not absolute, however, suggesting either that in some regions of the galaxy the hot and cold material may be extensively mixed, or that the sun is located in an asymmetric “local hot bubble” filled with X-ray emitting gas. More recent observations with *ROSAT* have shown the existence of soft background emission associated with the galactic halo and may suggest the existence of an extragalactic component to the observed soft background emission.

VI. FUTURE DIRECTIONS

As of this writing both the *Chandra* and *XMM-Newton* observatories are just beginning to make scientific observations. Even at these early stages the efficiency of discovery (i.e., the number of new results per observation) is extraordinary. Perhaps even more important than the unprecedented imaging these observatories can obtain are the high-resolution X-ray spectra they provide. High-resolution spectra of the type produced by the *Chandra* transmission gratings and by the *XMM-Newton* RGS will prove an increasingly important diagnostic of the physical conditions in the X-ray emitting region for all types of X-ray sources. Even higher resolution X-ray spectra will be obtained in the near future by X-ray calorimeters. Higher spatial resolution observations may be obtained by fabrication of larger X-ray mirrors, though without improvement in mirror construction technologies, creation of much larger mirrors than the *Chandra* or *XMM-Newton* mirrors may remain prohibitively expensive. The *Constellation-X* X-ray observatory, scheduled for launch in the first decade of this new millennium, will push the boundaries further by utilizing a system of small X-ray satellites that will work in unison to provide an increase in sensitivity of about two orders of magnitude over any previous (individual) X-ray observatory. X-ray interferometry, in which light from two or more X-ray telescopes is combined to give the effective spatial resolution of an X-ray mirror whose aperture size is equal to the separation of the X-ray telescopes, should provide extremely high spatial resolution if the significant technological challenges can be overcome. In theory interferometric techniques can allow astronomers to resolve almost arbitrarily small structures in X-ray sources (to image, for example, the accretion disk around a black hole). New results provided by *Chandra* and *XMM-Newton* along with the launch of new observatories guarantee that the upcoming decades will be an incredibly exciting time for X-ray astronomy.

SEE ALSO THE FOLLOWING ARTICLES

DARK MATTER IN THE UNIVERSE • GALACTIC STRUCTURE AND EVOLUTION • GAMMA-RAY ASTRONOMY • GRAVITATIONAL WAVE ASTRONOMY • INFRARED ASTRONOMY • NEUTRON STARS O PULSARS • QUASARS • SUPERNOVAE • ULTRAVIOLET SPACE ASTRONOMY

BIBLIOGRAPHY

Aschenbach, B., Hahn, H.-M., and Trümper, J. (1998). “The Invisible Sky: ROSAT and the Age of X-Ray Astronomy,” Copernicus, New York

Barcons, X., and Fabian, A. C. (1992). “The X-Ray Background,” Cambridge University Press, Cambridge, U.K.

Bradt, H. V. D., Ohashi, T., and Pounds, K. (1992). *Ann. Rev. Astron. Astrophys*. **30,** 391.

Charles, Philip A., and Seward, Frederick D. (1995). “Exploring the X-Ray Universe,” Cambridge University Press, New York.

Culhane, J. L., and Sanford, P. W. (1981). “X-Ray Astronomy,” Trinity Press, London.

Mushotzky, R. F., Done, C., and Pounds, K. A. (1993). *Ann. Rev. Astron. Astrophys*. **31,** 717.

Rosner, R., Golub, L., and Vaiana, G. S. (1985). *Ann. Rev. Astron. Astrophys*. **23,** 413.

Tanaka, Y., and Shibazaki, N. (1996). *Ann. Rev. Astron. Astrophys*. **34,** 607.

Tucker, W. H., and Giaconni, R. (1985). “The X-ray Universe,” Harvard University Press, Cambridge, MA.

Table of Contents