N76-26800

Chapter 1

THEORIES OF THE ORIGIN AND NATURE OF THE UNIVERSE

A. G. W. CAMERON[1]

Harvard College Observatory, Cambridge, Massachusetts USA

COSMOLOGY

Radio and optical telescopes enable us to see only a very small fraction of the universe. All of our knowledge of physics and astronomy is limited to a relatively local part of the universe; it is necessary to extrapolate this knowledge to a much larger scale of time and distance. The cosmologist starts by assuming a broad, physical, "cosmologic principle" to govern his theory of behavior of the universe, which is subject to experimental and observational verification like any other law of physics. The cosmologist can, at present, only assume his broad principle and some framework of the physics of general relativity, work out the mathematical consequences of a model of the universe, and determine if the predicted phenomena are experimentally or observationally verifiable [36, 46, 56 or 58].

Most systems of cosmology are based upon a cosmologic principle, according to which, on a large-scale average, every point in space is indistinguishable from every other point in space, at a given cosmologic time.

Individual points in space are not observable as such, only objects in space. Consequently, the cosmologic principle can only be applied to the contents of space which can be observed. Each piece of matter in space has its own hierarchy of structure, and possesses its own intrinsic motion. Therefore, it is necessary to consider the average motion of many pieces of matter existing in a particular region of space as representing the general behavior of that region of space. The cosmologist must also ignore that in observable space, matter is collected into compact bodies while most of the intervening space is relatively empty. He assumes, in constructing his theory, that the mass of compact bodies in space is smeared out into a medium of average density, and he ignores special effects which may be caused by condensation of matter into compact bodies.

"Big Bang" theory. A consequence of general relativity, as formulated by Einstein, and applied to cosmology by Friedmann, is that only a universe containing no matter at all can be truly static. If matter is present, then the universe must be in either a state of expansion or contraction. Observations indicate that our observable universe is in a state of expansion. Distant matter in the universe, not gravitationally bound to our local system of galaxies, is receding from us. According to cosmologic principle, on any one of those distant pieces of matter, the same behavior would be observed from the point of view in space that we see in our own: all distant masses will be receding from the observer.

If distant masses in the universe are receding at present with a velocity proportional to their

[1] The work of A. A. Mikhailov (USSR) in the compilation of material for this chapter is gratefully acknowledged.

distance from us, then, going into the past, it would be found that local density of matter becomes progressively higher. If one goes far enough into the past, it would be discovered that matter in the universe progressively draws closer together, until at some finite time in the past, matter would appear to be gathered together at a single point. Such a picture is described as a "Big Bang" cosmology, matter exploding away from infinite density at a finite time in the past, which can be regarded as a sort of creation event. The relatively simple mathematical models of cosmology based upon this point of view are generally called Friedmann universes, on which most discussions of cosmologic processes are based.

Steady state cosmology. A number of British cosmologists were not satisfied with the "Big Bang" cosmologic principle. One of the corollaries of Einstein's general theory of relativity is that time has many dimensional attributes similar to those of space. Since space and time are joined in a four-dimensional geometry of space-time, it was therefore asked why the cosmologic principle should apply only to space, and not also to time. A broader, cosmologic principle, the "perfect cosmologic principle," was proposed which postulates that every observer in the universe, on a large-scale average, should see the same picture at every point in space and at every time.

The universe will always exhibit the same average behavior, according to this principle. Since the universe is observed to be expanding, and since the perfect cosmologic principle requires that the average density of matter in space should remain constant, then it is also necessary to postulate that additional matter must be created in space to replace loss through expansion. Since matter being lost is observed in the form of galaxies and clusters of galaxies, it is necessary to postulate that newly created matter must condense to form new systems of galaxies. This general picture is called steady state cosmology.

The continuous creation of matter, it must be emphasized, is a postulate of a physical process which has never been observed experimentally. Thus, the perfect cosmologic principle requires assuming new and untested laws of physics which need not be invoked in the cosmologies based only upon the ordinary cosmologic principle.

When Einstein first applied his theory of general relativity to cosmology, it was widely assumed that the universe was static. He realized that his ordinary theory of general relativity predicted either an expanding or contracting universe, experimented with the form of his equations, and found that by adding a term to his general equation, called the "cosmologic constant," it had the property of rendering the universe nearly static for very long periods. After it was discovered that the universe appears to be expanding, Einstein dropped his experiments with the cosmologic constant, although other cosmologists have continued such experiments, producing various types of cosmologic models based upon this constant.

The expansion of the universe was discovered by Hubble, resulting from his work on measuring the spectra of galaxies. The fainter the galaxy, he found, the greater the shift toward the red of the characteristic spectral lines in the light of that galaxy. The red shift of the spectral lines is interpreted as a Doppler shift, indicating that distant galaxies are receding from us. A distant galaxy is receding from us at a rate which appears to be proportional to its distance, a relationship known as the "Hubble law." If motions of distant galaxies are extrapolated backwards in time, it appears that the "Big Bang" must have occurred between 1 and 2×10^{10} years ago.

The red shifts of galaxies were essentially the only concrete observational data available to cosmology until the mid-1960s.

Background radiation. In 1965, Penzias and Wilson discovered microwave background radiation, and found that radio waves, at a wavelength of 7.5 cm, are constantly impinging on the Earth with an equal intensity from all directions in space. Other investigators have corroborated this throughout the observable radio wavelength region, and observations based upon interstellar molecules have confirmed that the radiation has approximately a blackbody shape down to a wavelength of about 1 mm at 2.7° K. The simplest interpretation of this background radiation is that it is a relic of a hot early stage in the universe, when matter in the universe was closely compacted and much higher in temperature than at present. When density of matter and temperature are high enough, then matter becomes ionized and fully in equilibrium with blackbody radiation. When temperature falls low enough so that matter, primarily composed of hydrogen, can recombine to neutral form, interaction between radiation and matter becomes very small, and radiation is free to expand into space with the universe. This expansion dilutes radiation, and photons suffer a red shift, so that the shape of the spectrum remains that of a blackbody while temperature is progressively decreased.

The idea of a very hot initial medium was proposed by Gamow [17] who predicted a temperature of 6° K instead of the actual 2.7° K for the present epoch. Doroshkevich and Novikov [11] pointed out that relic radiation would exceed the combined radiation of radio sources and stars in the 1–10 cm wavelength region. Dicke et al [10], beginning an experimental search for relic radiation, learned of the findings of Penzias and Wilson, and were immediately able to explain the results.

That isotropic background radiation is necessarily thermal in origin or that it has a Planck shape is not agreed upon by all astronomers. Suggestions have been made that radiation could arise from great numbers of unknown emitters isotropically distributed in space. Most suggestions involve very complex models of these unknown emitters and numerous arbitrary assumptions. An alternative model, complex though it may be, is necessary if the true cosmologic model should be steady state, for example, which never undergoes a high-density, high-temperature phase.

Within the context of a Friedmann model of the expanding universe, if density of matter presently in the universe is less than a critical value of a little less than 10^{-29} g/cm^3, the universe will always expand, and is described as an open universe. On the other hand, if density is greater than the critical value, expansion of the universe will coast to a halt, and the universe will start to contract again toward infinite density. This cosmologic model is called a closed universe. Critical density depends upon the precise value of the constant of proportionality in the Hubble law.

Many variations of Einstein's general theory of relativity have been suggested, all of which produce slight variations in their associated cosmologies. Only one of these is mentioned here because it has received a fair amount of attention in recent years: the scalar-tensor theory of general relativity. Einstein's theory of relativity is characterized by a tensor gravitational field, and added to this is a scalar gravitational field. This theory can predict some differences in the behavior of a cosmologic model compared to those in a Friedmann universe, including a very rapid expansion in the early history of the cosmologic model, and a variation with time of the gravitational "constant" in the Newtonian force law of mass attraction between bodies. The directly measurable consequences of addition of a scalar to a tensor field are very small, and it has not yet been possible to make an experimental choice between the two theories [36].

PHYSICS OF THE EARLY UNIVERSE

Considerable interaction between high energy particle physics and cosmology during the last few years has been stimulated by the discovery of microwave background radiation, which implies that the universe was once highly dense, very hot, and originated from a big bang. This interaction raises the very important issue of whether the universe is symmetric in matter and antimatter, or unsymmetric with an excess of one over the other. This difference can have an important influence on the behavior of the very early universe; some aspects of the difference may persist until the present; allowing a possible test of the degree of symmetry of the universe. In discussing initially the behavior of the unsymmetric universe, it is assumed that all galaxies visible in space are composed of ordinary matter.

Unsymmetric Universe–Open Model

The story begins when the universe was approximately 10^{-43} s old. The characteristic length associated with the universe, known as the "Hubble radius," is $r=ct$, the age of the universe multiplied by the velocity of light, a distance of only 3×10^{-33} cm. This is much smaller than the characteristic radius of any of the elementary particles with which we deal in the universe today, called the "Compton radius," which is inversely proportional to the mass of the particle. Ordinary neutrons and protons have radii around 10^{-13} cm so it might be said that the universe is not yet old enough to contain ordinary neutrons and protons at 10^{-43} s [19].

If any matter exists in the universe at this time, it must consist either of a highly excited baryonic state of the neutron or proton, or of some completely unknown form of matter. In any case, particles in the universe at 10^{-43} s can be expected to have masses of the order of 10^{-5} g in order to fit into the Hubble radius. Such particles are entirely hypothetical, lying far beyond the scope of the existing particle theory and its extrapolations.

The universe at this time is best described as chaotic. The Hubble radius has a value comparable to expected quantum fluctuations in the structure of the universe, if general relativity is to be unified with quantum mechanics in some way. If the age of the universe is multiplied by the energy content, including rest mass, of the volume contained within a Hubble radius, the result is a number of the order of Planck's constant, h. This is the cosmologic expression of the Heisenberg uncertainty principle. It does not make any physical sense to inquire what happened in the universe earlier than 10^{-43} s, because such times are not meaningful due to the energy uncertainty.

As the universe grows older, the Hubble radius grows and encompasses many types of particle. These particles can interact and achieve a thermodynamic equilibrium, and the mass of particles which can exist in the universe becomes progressively lower as the Hubble radius increases.

The character of physics at this expansion epoch is rendered uncertain by the unknown form of the mass excitation spectrum of the baryon. Hagedorn has suggested that the number of baryon states per unit mass interval rises exponentially, which leads to an expectation that the universe will have a finite maximum temperature that could have been attained in early times, of the order of 10^{12} °K. This limiting temperature is asymptotically attained as the rest mass of the baryons present goes to infinity. In the Hagedorn version of early cosmology, the early universe would have this limiting temperature, and as it expands, the characteristic masses of particles present would progressively decrease [18]. Significant variations in the ratio of baryons to antibaryons may exist in this picture. Only when the universe has expanded enough so that most of the particles expected to be present are neutrons and protons, would the temperature decrease appreciably below 10^{12} °K.

However, if the number of mass states of the baryon should increase with mass at a rate slower than exponential, then there is no upper limit on the temperature of matter, and this can become arbitrarily high. In such a case, the earliest universe would contain both matter and antimatter in profusion, but with a slight excess of matter over antimatter, at least in that region of the universe which will form our own galaxy. In this version of cosmology, matter and antimatter progressively annihilate, until, when temperature falls below 10^{12} °K, only baryons which are in excess of antibaryons remain unannihilated, and the local part of the universe then contains only matter in the form of baryons (see under next section, **Symmetric Universe – Open Model**).

At temperatures in the universe between 10^{11} and 10^{12} °K, it is not possible for baryonic pairs to exist together, but pi and mu mesons can exist in profusion. As expansion occurs and the universe cools, the pi mesons disappear, followed by the mu mesons which also disappear. This leaves matter consisting of some neutrons and protons, electron-positron pairs with a small excess of electrons, muon neutrinos and antineutrinos, and electron neutrinos and antineutrinos, as well as photons. After mu mesons have disappeared, mu meson neutrinos and antineutrinos no longer interact with the rest of the particles. When temperature falls below 10^{10} °K, electron neutrinos and antineutrinos no longer interact with ordinary matter, and shortly afterwards, electron-positron pairs disappear through annihilation [33, 56, 58]. Both electron and muon neutrinos and antineutrinos contribute energy density and pressure which help to drive the expansion of the universe, but it is only in this indirect way that these particles can have further influence upon physical events in the universe [36].

During the time that electron-positron pairs were present in the expanding universe, they were interacting with neutrons and protons present, causing interconversion of one into the other. This sets up an equilibrium between neutrons and protons, in which protons are somewhat more abundant than neutrons, since their mass is slightly less than that of neutrons. About seven times as many protons as neutrons are expected at the time that electron-positron pairs disappear through annihilation [36].

When temperature falls to 10^9 °K, it becomes possible for complex nuclei to exist. The first nucleus formed, deuterium, results from combination of a neutron with a proton with the emission of a photon. At the same time, absorption of photons from the radiation field by the resulting deuterium nuclei will result in their photodisintegration back into neutrons and protons. At first, photodisintegration greatly predominates, and very little deuterium can be present. However, as the temperature falls, equilibrium shifts toward deuterium in the medium.

Deuterium can in turn capture neutrons to form tritium. This nucleus is radioactive with a half-life of 12 years, which is very long compared to the expansion time in this discussion, when the expansion age of the universe is just a few minutes. Thermonuclear reactions then take place between deuterium and tritium, forming helium. At the expected matter densities for this stage in the expansion of the universe, these reactions go nearly to full completion, converting almost all of the neutrons in the mixture into helium. Since there were more protons present than neutrons, excess protons are left over following completion of nuclear reactions. About one-fourth of the matter becomes helium under these conditions. Very little deuterium and tritium is left when universal expansion has reached an age of half an hour [36].

The amount of helium formed in this way is comparable to that which appears to exist in all stars, both very old and quite new. The amount of helium produced by cosmologic nucleosynthesis varies from about 28% in a closed universe, to about 24% in an open universe. A variety of observational and theoretical evidence indicates that this is an approximately universal abundance of helium observed in stars everywhere in space, but it is not yet known whether the abundance of helium in these stars is sufficiently close to a single value to indicate that the uniform process of cosmologic nucleosynthesis has been responsible.

The amount of deuterium and ^{3}He produced in a closed universe model is negligible, but in an open universe it is quite possible to produce amounts of deuterium and ^{3}He comparable to those inferred in the primitive solar nebula from which the solar system evolved. If no alternative method should be found for producing deuterium and ^{3}He, this would constitute strong evidence in favor of this unsymmetrical model of the universe, and would identify our universe as an open model [39, 53]. However, this claim cannot yet be made because not all alternative methods of producing deuterium and ^{3}He have yet been ruled out.

As the universe continues to expand, it is composed of hydrogen, helium, electrons, and photons, with various types of neutrinos and antineutrinos arising from interaction with other particles. Photons continue to interact strongly with matter, and because most of the pressure is in photons, matter is unable to fragment into self-gravitating bodies, such as galaxies or clusters of galaxies at this stage in the expansion of the universe.

When the expansion of the universe is nearly 1 million years old, the temperature of matter and radiation will have fallen close to 3000° K. Already, helium will have recombined to the neutral form, and, at this temperature, hydrogen also recombines to form neutral hydrogen atoms. This makes a tremendous difference in the interaction between matter and radiation. As long as hydrogen was ionized, radiation photons could move only relatively short distances before undergoing Compton scattering on the free electrons. However, after recombination, only Rayleigh scattering is possible with neutral hydrogen atoms, and the mean free path of photons becomes larger than the Hubble radius of the universe. Hence, photons can expand indefinitely into space after recombination, and these constitute the isotropic background microwave radiation with a present temperature of 2.7° K in this type of cosmologic model. It is possible for gravitational instabilities in matter to grow only after decoupling from radiation.

Symmetric Universe—Open Model

What is necessary in a symmetric universe? There is, at present, as much matter as antimatter among visible galaxies in the universe. It can be assumed that imbalance between matter and antimatter arose from fluctuations in the relative numbers of baryons and antibaryons at different places in the early expanding universe. This type of assumption has an ad hoc character which does not lead to a natural explanation for anything. An attempt to find processes which can lead to separation of matter and antimatter on at least the galactic scale, starting from an intimate microscopic mixture of baryons and antibaryons, has been carried out in recent years by Omnes. The following theoretical developments are generally ascribed to him.

First, new developments have been reported by Parker at the Aspen Workshop, June 1972, on The Physics of the Early Universe. Returning to a universal age of 10^{-43} s and universal chaos, chaos is defined as different parts of the universe, merging with Hubble radii of 3×10^{-33} cm which introduce fluctuating gravitational potentials containing tremendous energy. Parker investigated the consequences of rapid nonisotropic expansion of the universe under these circumstances, and found that baryon-antibaryon creation can occur in which strong chaotic gradients in the gravitational potential are largely converted into rest mass and kinetic energy of baryon-antibaryon pairs. Furthermore, pair creation occurs over distances slightly greater than the Hubble radius, thus introducing a possible mechanism which can lead to large-scale homogeneity in expansion of the universe. This mechanism predicts production of equal numbers of baryons and antibaryons with any microscopic volume.

Zel'dovich and Starobinsky pointed out the importance of pair creation in an anisotropically expanding universe. The influence of created pairs on expansion tends to make expansion isotropic, which is a step toward an explanation of the Friedmann expansion law.

At temperatures greater than about 3×10^{12} °K (Omnes' basic hypothesis), there exists a phase separation between baryons and antibaryons [29]. The existence of these phases is very controversial and according to Omnes, in a statistical sense, baryons and antibaryons cannot approach each other too closely and retain their identity, whereas baryons can approach indefinitely closely to baryons, and antibaryons to antibaryons. This leads to an expected separation of baryons and antibaryons into two condensed phases, with separate blobs of baryons and antibaryons growing to dimensions of around 3×10^{-4} cm by the time the temperature has fallen to about 3×10^{12} °K, each patch of material containing about 10 kg of material.

After the temperature has fallen below 3×10^{12} °K, the thermodynamic condition which favored separation of the baryon from the antibaryon phase would no longer be the lowest thermodynamic energy state. Therefore, Omnes expects baryons and antibaryons in the separated patches to start diffusing together, leading to mutual annihilation. This mutual annihilation will continue until temperature has fallen to around 3×10^{8} °K. At this time, most of the original matter and antimatter will have disappeared in mutual annihilation, feeding energy into the photon field, and reducing the number of baryons or antibaryons to about 10^{-8}/photon.

At this stage, separate patches of matter and antimatter still exist. These patches continue to annihilate baryons and antibaryons along the borders between patches, but as temperature in the universe continues to fall, a new process sets in which Omnes calls the coalescence stage. Annihilation causes relative motions among patches of matter and antimatter, with a matter patch being repelled from an antimatter patch due to annihilation along the common boundary, but with mergers occurring when a matter patch meets other matter patches, or antimatter patches meet other antimatter patches.

These conditions have prevented the formation of helium or other light elements when the temperature was approximately 10^{9} °K. Under these conditions, there is no longer an electron-positron pair plasma present everywhere in space, hence there is no longer interconversion of neutrons to protons at that temperature. Because neutrons, which have remained after disappearance of electron-positron pairs, can diffuse faster than protons, there has been preferential annihilation of neutrons and antineutrons during the annihilation stage. Thus, in this type of symmetric universe there is no formation of large quantities of primordial helium (or antihelium), and some other type of event would be required to form large quantities of helium in the pregalactic stage of evolution of the universe.

As the universe expands, material in separate patches of matter and antimatter during the coalescence phase continues to grow rapidly. This is partially due to rather rapid motions of matter and radiation taking place within the universe at this time, because the speed of sound is not much less than the speed of light. However, when the characteristic radiation temperature falls to 3000° K, so that hydrogen recombines to the neutral form, the speed of sound in the decoupled matter decreases drastically, and the motion of such matter in the universe is greatly decreased. In effect, this brings the coalescence stage to an end and practically ends the mutual annihilation of matter and antimatter on the boundaries between their separate patches. Omnes has estimated that the patches contain an amount of matter at least as great as that of a typical galaxy at this stage. It is not out of question that the patches should contain enough matter to form clusters of galaxies. The discussion in the section that follows, THE PREGALACTIC ERA, would be more consistent with the Omnes cosmology if the separate patches had grown to the size of clusters of galaxies.

In summary, symmetric and unsymmetric types of cosmology lead to very similar results throughout the universe, with symmetric cosmology predicting that clusters of galaxies may tend to be either all of matter, or all of antimatter, and the unsymmetric cosmology predicting that visible galaxies are all composed of matter. There is no practical way to distinguish between these pictures by direct observation. However, the symmetric cosmology predicts that helium will not be formed in large amounts by cosmologic nucleosynthesis, and the unsymmetric cosmology not only predicts such a large cosmologic nucleosynthesis of helium, but also possibly that of deuterium and ^{3}He as well, if we live in an open cosmologic model.

THE PREGALACTIC ERA

The problem of gravitational instability within a large self-gravitating medium was first considered many years ago by Sir James Jeans. There is a critical length, called the Jeans length, such that disturbances of wavelengths longer than the Jeans length are unstable against gravitational contraction within the medium, whereas disturbances of wavelengths smaller than the Jeans length will be damped out and will propagate as giant sound waves.

Peebles and Dicke [37] postulated that following hydrogen recombination in the universe, matter may contain density perturbations over a wide range of scales, and that the most frequent of the perturbations which can become gravitationally unstable will have wavelengths close to the Jeans length. However, there is no theoretical basis at present to estimate the expected magnitude of such density fluctuations within the universe at a particular time. If the density fluctuation is very small, then it takes a long time for it to be felt within the expanding universe, and for the matter associated with the fluctuation to fragment and condense away from its neighboring matter. Thus, fragmentation time within the expanding universe may be very long.

A comprehensive study of perturbations superimposed on the isotropic homogeneous Friedmann solution (carried out by Lifshitz [24]) included simultaneous perturbations of radiation and matter density, rotational motion, and gravitational waves. Doroshkevich, Novikov, and Zel'dovich [12] added the perturbation of matter density on the unperturbed radiation (photon) background. This type of perturbation remains hidden until recombination occurs.

At the time of the hydrogen atom recombination, both matter and radiation are expanding with the universe. Following recombination, radiation continues to expand uniformly with the universe. Matter has a large outward expansive velocity, and density perturbations can only grow comparatively slowly. Thus, any given blob of matter must continue to expand for a long time before its self-gravitational forces are able to prevail and bring expansion to a halt, to be followed by subsequent contraction toward a denser state.

The smaller the density fluctuations within matter following recombination, the greater the subsequent expansion of a blob of matter before it reaches its minimum density. For example, if density fluctuation following recombination should amount to about 1%, then the subsequent expansion and recollapse of matter will take about 3 or 4 × 10^9 years. Rapid formation of galaxies within the universe is only possible if density fluctuations should be much larger than 1%. Perhaps such density fluctuations are more likely to occur in a violently stirred universe such as that of Omnes than in a rather uniform unsymmetric universe.

Primitive Globular Clusters

Following recombination, the Jeans length in matter encompasses nearly 10^6 solar masses of material. On this basis, a theory was formulated [37] of the formation of primitive globular clusters, noting that globular clusters have about 10^5 to 10^6 solar masses of stars. However, this value of 10^6 solar masses may not have much significance, since it is merely the minimum mass which is unstable against gravitational collapse. All density fluctuations representing larger masses of material are also unstable against collapse. Such density fluctuations may include large galactic masses or galactic material clusters. Such large-scale density fluctuations presumably are of smaller scale than smaller density fluctuations near the Jeans length, hence it is more likely that smaller density fluctuations will grow faster within the expanding universe. Thus, the emerging picture is one in which gravitational instability can occur on a very wide range of scales, with individual masses of the order of 10^6 solar masses condensing, and these in turn falling toward mutual gravitating centers, which in turn fall toward other centers on a larger scale, and so on.

However, as matter continues to expand toward its minimum density configuration, the Jeans length includes less matter. Matter is now decoupled from radiation, so that it can become extremely cold near the position of its maximum expansion. Temperature in matter may fall below 1° K. Under these circumstances, the Jeans length may enclose only 1000 solar masses. Thus, it is possible that the most frequent fluctuations which develop within the expanding universe have masses more characteristic of a few thousand solar masses.

As matter falls back from its position of maximum expansion, the rise in temperature due to adiabatic compression eventually becomes limited due to cooling processes within the gas. The most important of the cooling processes results from hydrogen molecule formation due to a few electrons left over from the recombination stage in the expansion of the universe. These electrons can be captured by hydrogen atoms, forming negative hydrogen ions which readily capture other hydrogen atoms, forming hydrogen molecules, and again releasing the electron. Thus, the electron acts as a catalyst in the process as long as it does not encounter a positive hydrogen ion and becomes captured, forming a neutral hydrogen atom. When temperature is high enough, of the order of 250° K, to produce significant excitation of the lower excited states of the hydrogen molecule, infrared radiation from de-excitation of the molecules can lead to cooling within the contracting gas. As gas contraction accelerates, formation of hydrogen molecules is limited through gradual elimination of all remaining electrons, and the increasingly rapid gas compression gradually raises the temperature until thermal decomposition of hydrogen molecules takes place. At this time, when the density has risen to about 10^{10} hydrogen atoms/cm^3, the Jeans length in the compressing material encloses only about 60 solar masses. Presumably, the gas can thus fragment into stars having masses in the range 10^2–10^3 solar masses [54].

Pregalactic Stars

There is a natural expectation that the immediate result of expansion and contraction of matter in the early universe is to form a set of pregalactic stars, all quite massive. These stars may be composed of pure hydrogen, if the symmetric universe is a correct model, or they may have about one-fourth their mass as helium. In either case, they do not behave like ordinary stars in the universe today. Massive stars composed of only hydrogen and helium have been studied by Ezer and Cameron [14]. These stars become much more compact than ordinary stars of the same mass, owing to very high internal temperatures needed to convert hydrogen into helium without initially such catalyst nuclei as those of carbon, nitrogen, and oxygen. This means that the radii of stars are abnormally small, and their surface temperatures are approximately 10^5 °K. These stars will emit a tremendous flood of ultraviolet radiation, which will certainly ionize any residual hydrogen and helium in the universe which have not been formed into stars at this stage.

Stars composed of pure hydrogen, corresponding to the symmetric universe, have been studied little so far. It is possible that the dynamic collapse which forms these stars may lead to such high central temperatures that there is sudden conversion of hydrogen to neutrons by electron capture, followed by massive helium formation and explosion of stars back into space. If stars can form stable objects, they should behave much the same as the hydrogen-helium stars described above. The major question is whether such stars can form so much helium and eject it into space, so as to contaminate the remaining gas in space to the extent of one-fourth by mass of helium, in order to account for the essentially universal hydrogen-to-helium ratio in stars throughout space, which seems to require a pregalactic helium formation.

Stars in this massive range may eventually explode in supernova explosions or collapse to form black holes. It is tempting to argue that both events will occur. This primordial generation of stars is presumably concentrated in space into masses at least of the order of clusters of galaxies. The visible amount of matter in galaxies in clusters of galaxies is inadequate, by approximately 1 order of magnitude, to provide for the gravitational binding of such clusters of galaxies, keeping them together against dispersion in space during the history of the universe. This mass discrepancy could be accounted for if a majority of the primordial generation of stars were to collapse to form black hole remnants, distributed more or less uniformly throughout the volume occupied by a cluster of galaxies [50].

Evidence, on the other hand, indicates that supernova explosions could also occur among the primordial generation of stars. It seems difficult to account for the observed heavy element content of stars in the galactic halo, unless the heavy metal content of these stars was also formed in the pregalactic era. If supernova explosions occur among some of the hydrogen-helium or pure hydrogen stars formed in the primordial stellar generation, some conversions into heavier elements are likely to take place (to be described in a subsequent section). These heavy elements will spread in the gas which will subsequently condense into galaxies [50].

In any event, it appears that unconsolidated gas, which did not succeed in becoming incorporated in the primordial generation of stars, together with gas ejected from stars in supernova explosions or as a result of stellar variability and stellar winds, is subject to a great deal of heating. Intense ultraviolet output from the primordial generation of stars is likely to heat gas at relatively low densities to a temperature of the order of 10^5 °K. In addition, supernova explosions occurring in this gas may provide heating to somewhat higher temperatures, which has been stressed [12]. This will probably lead unconsolidated gas to expand throughout the volume occupied by a cluster of galaxies, forming a common gas envelope throughout all of this volume. Into this common gas envelope will be injected in a random manner the products of stellar nucleosynthesis, including heavy metals, and possibly large amounts of helium. The heavy metals, in particular, should greatly enhance the cooling efficiency of the gas, leading to local cool spots within the gas. Since local cooling within the gas leads to a reduction in local pressure, such cool spots will undergo density increases, thus forming density fluctuations about which gravitational instabilities can grow within the gas. This is a probable mode of nuclei formation of galaxies within the common gas envelope filling a cluster of galaxies. After a galactic nucleus is formed, matter is likely to continue falling toward such galactic nuclei, allowing the galaxies to grow in size. There is considerable indication that such mass infall is still continuing within our galaxy today [30].

How newly formed galaxies acquire their angular momentum has been the subject of debate. It was suggested [34] that the galactic angular momentum is acquired as a result of tidal distortion during the fragmentation process in the expansion stage of the universe. This theory has been quantitatively questioned by Oort [30], but further defended by Peebles [35]. Oort, in turn, preferred a picture in which the expanding universe contains highly turbulent matter, and presumed that the galaxies have acquired their angular momentum as a result of vortices within the turbulent matter. The turbulent theory is developed also by Ozernoy et al [31]. The point of view that initial perturbations are characterized by acoustical waves in a photon-ionized gas mixture is advocated by Doroshkevich, Sunyaev, and Zel'dovich [13, 58].

THE EVOLUTION OF GALAXIES

Whatever the mechanism which initiates the collapse of galactic masses of gas to form galaxies, quite a bit about the nature of this collapse can be inferred from morphologic forms exhibited by the resulting galaxies. Galaxies show a tremendous variety of forms, and likely these arise largely from subtle differences in the angular momentum distribution of the initially collapsing gas clouds.

Elliptical galaxies. The nearly spherically symmetric distribution in space of stars in elliptical galaxies can be maintained only if the stars have nearly radial orbits; such stars fall close to the center of their galaxy and then recede outward to large distances from the center. Such galaxies have very little total angular momentum. As the total angular momentum of a galaxy increases, the degree of flattening of the distribution of stars increases. When the system of stars becomes rather flattened, gas and dust start to appear in the central plane of the galaxy, and as the degree of flattening becomes even greater, large amounts of gas and dust appear. Evidently stars in such galaxies have orbits which are more nearly circular, and certainly stars formed from gas and dust lying in the central plane of the galaxy will have almost circular orbits.

Systems with a great deal of gas and dust usually exhibit spiral arms; these are rendered more prominent because spiral arms are the locations of current star formation in such galaxies, and the hot, massive blue stars form effective markers for these arms. Spiral arms exhibit a wide range of curvature and numbers of turns about the center of the galaxy. In some of these galaxies, central stars exhibit an approximately spherical distribution, but in others they form an elongated bar. Central bars of barred spiral galaxies are thought to arise from details of angular momentum distribution of infalling gas; if this gas, upon forming a disk, rotates with approximately constant angular velocity, it is unstable against deformation into a bar. Some galaxies have no strong concentration whatever of stars toward the center; these systems, which contain a large ratio of gas to stars, are called irregular galaxies.

Spiral galaxies. Knowledge of the features of spiral galaxies is the most complete of any type of galaxy, since our own galaxy is apparently a typical spiral.

The time required for primordial galactic gas to collapse from the volume of space presently occupied by the galactic halo to form the disk is about 2×10^8 years. This estimate is necessarily rough, since the initial volume occupied by gas at the start of collapse may have been considerably greater than that now occupied by stars in the approximately spherical galactic halo. It is natural to suppose that stars in the halo were formed by condensation from gas during the collapse stage. The motions of these halo stars have large radial components.

The abundances of heavier elements in halo stars are less than in the Sun, typically by a factor of about 3, although in a few extreme cases, by factors of several hundred. It is rather unlikely that the collapsing gas had time to form an initial population of stars which could evolve and form heavy elements by nucleosynthesis, eject these elements back into the collapsing gas in supernova explosions, to be followed by formation of a second generation of stars incorporating these heavy elements. Thus, the existence of heavy elements in these stars provides an argument for a stage of pregalactic stellar nucleosynthesis.

After gas has collapsed into the form of a disk, subsequent star formation will be of stars in approximately circular orbits in the plane of the disk. When such a star has been formed, its orbit may suffer gravitational perturbations as it approaches other stars, star clusters, or interstellar gas clouds. As a result of such perturbations, the star may swing to increasingly greater distances away from the central plane of the galaxy as it grows older. Older stars within the disk population in our galaxy have a distribution in distance away from the central plane of the galaxy corresponding to a thickness of about 400 parsecs. Stars formed in more recent galactic history, on the other hand, are confined to a narrower thickness of about 200 parsecs.

Density Wave Theory

The most quantitative and promising modern theory of spiral arms appears to be the density wave theory [25, 26, 41, 42]. According to this theory, stars in the galactic disk can undergo collective gravitational clustering for significant periods, and the gravitational cluster, or density wave, progresses through the stellar distribution in the disk of the galaxy at about half the rate of rotation of the stars. For example, at the solar distance from the center of the galaxy, a star such as the Sun can be expected to spend about 10^8 years traveling in the interarm region from one spiral arm to the next, and then an additional 10^8 years within the next spiral arm which it encounters. At the end of this total interval, it has made approximately 1 revolution around the center of the galaxy, according to a simple two-armed model of a spiral galaxy. The actual model for our own galaxy may be slightly more complicated.

In order to understand how star formation currently takes place within our galaxy, it is necessary to consider first properties of the interstellar medium [47] which consists of gas and dust between the stars. The dust is composed of small solid particles with dimensions slightly less than 1 μm, intimately mixed with gas in the interstellar medium. The dust is responsible for extensive absorption of starlight in certain directions within our galaxy, showing up as dark patches on photographs of portions of the Milky Way.

Interstellar gas resembles the Sun in composition; about ¾ of the mass is hydrogen, about ¼ is helium, and only about 1.5% is in the form of heavier elements, mostly condensed into interstellar grains.

Hydrogen Ionization

When electromagnetic radiation, having photon energies greater than 13.6 electron volts (eV), is incident on interstellar hydrogen atoms, that radiation is capable of ionizing hydrogen. Consider some hydrogen located close to a very hot, massive, highly luminous star. Such a star, because of its high surface temperature, will emit most of its radiation far into the ultraviolet portion of the spectrum. Most of its emitted photons may have energies $\leq$ 13.6 eV. Hence, hydrogen close to the star will be ionized very quickly by the enormous ultraviolet flux, incident upon it.

Now consider a continuous hydrogen gas cloud surrounding this star which extends a considerable distance into the interstellar medium. Ultraviolet radiation from the star will start ionizing nearby hydrogen. As hydrogen becomes ionized, it no longer absorbs the outpouring radiation from the star, so that ultraviolet rays from the star can reach farther and farther into space, ionizing hydrogen on the way. If hydrogen were to stay ionized, this process could go on indefinitely, and it would be expected that the star could ionize hydrogen gas to indefinite distances.

Hydrogen, however, will not stay ionized indefinitely. Internal recombination will take place continuously, leading again to neutral hydrogen. Such a neutral atom will not remain in this state for very long before becoming ionized again by ultraviolet light. However, the greater the distance at which the star succeeds in ionizing hydrogen surrounding it, the greater the region in which hydrogen recombination will occur simultaneously, with consequent weakening of the ultraviolet radiation which proceeds to further distances as neutral hydrogen is again ionized.

A natural limit is created in this way for the distance to which starlight can ionize hydrogen. This limit is reached when the total number of neutral hydrogen atoms which are being ionized in every second becomes equal to the total number of hydrogen atoms which are recombining to the neutral state in every second. In this way, a hot star can create a sphere of ionized hydrogen surrounding it, with the result that there is a remarkably sharp transition from ionized hydrogen to neutral hydrogen at the surface of the sphere.

Neutral Hydrogen Regions

Absorption of ultraviolet photons by neutral hydrogen results in heating of the resulting ionized hydrogen. Most of the absorbed photons have an energy somewhat greater than 13.6 eV. Hence, in the absorption process, 13.6 eV are used to split up the hydrogen atom, and the remaining energy appears as kinetic energy of the proton and electron which are emitted. This excess energy is rapidly shared with other particles, leading to high temperature of the ionized gas.

Ionized gas can cool in various ways. The typical process involves collisions between a proton and an electron in which a photon can be emitted but the proton and electron remain ionized, with the electron departing from the proton having less kinetic energy than when it approached. The photon which is emitted is generally free to escape from the region in which it is created, thus removing some of the heat energy from that region.

Energy input into an ionized hydrogen region is limited by the number of photons from the hot star which neutral hydrogen atoms in the region can absorb. On the other hand, the rate of radiation from collisions between protons and electrons increases rapidly with temperature, hence a balance will be achieved at a temperature where the rate of energy input from absorption of ultraviolet photons is balanced by the rate of energy radiation from collisions between protons and electrons. The temperature which gives this equality between absorption and radiation of energy from ionized hydrogen gas is ca 10 000° K.

Thus, in a region in the interstellar medium in which hydrogen gas is neutral, the ultraviolet spectrum of interstellar starlight has been cut off above 13.6 eV, the ionization potential of the hydrogen atom. Photons more energetic than this limit have been absorbed in ionizing hydrogen in the ionized region.

Some of the elements with ionization potentials below 13.6 eV, including all the heavier elements, will be ionized in a neutral hydrogen region. The main elements which will not be ionized, in addition to hydrogen, are helium, nitrogen, oxygen, and neon. About three-fourths of the ions in the neutral hydrogen region are of carbon.

Ionization processes within neutral hydrogen regions also provide heat input to these regions, although of a considerably smaller magnitude than in the ionized regions. A variety of cooling processes are also present in neutral hydrogen regions, involving collisions among electrons, ions, neutral atoms, molecules, and interstellar grains. These collisions result in cooling when they raise one of the particles into a more highly excited energy state, from which deexcitation by radiation can occur, with radiation escaping from the neutral hydrogen region. The balance between heating and cooling of gas typically occurs at a temperature of about 100° K, somewhat lower in denser neutral hydrogen regions [21].

Pressure Equilibrium

Pressure in gas is proportional to density, also to temperature. In considering adjacent regions of neutral and ionized hydrogen, if each of these regions were to have about the same number density of particles, pressure in the ionized region would be about 100 times that in the neutral region, and gases would then be set into violent dynamic motion due to pressure imbalance. Pressure equality could be achieved if density in the ionized hydrogen region were 100 times less than density in the neutral hydrogen region, in inverse proportion to the temperatures in these two regions. It appears that the interstellar medium always seeks to achieve such a pressure equilibrium throughout, so that in general, ionized hydrogen regions are of much lower density than neutral hydrogen regions. However, newly formed stars and supernova explosions provide continual variation in local rates of heat input into the interstellar medium, so that the readjustment process of the configuration, in order to approach pressure equality, is ongoing, leading to significant dynamic motions of gas within the interstellar medium.

Comparable amounts of ionized and neutral hydrogen appear to be in the interstellar medium. Since the density of ionized gas is much less than that of neutral gas, the neutral gas will obviously occupy only a small fraction of the volume in the interstellar medium. The neutral hydrogen clumping which thus occurs leads to the description of the interstellar medium as consisting of neutral hydrogen clouds embedded within an ionized region.

The foregoing description is oversimplified. Other sources of ionization exist in the interstellar medium, such as cosmic rays and soft x-rays, which modify the details of the description given, but not its most essential features.

In a typical interstellar cloud, self-gravitational forces which would tend to make the cloud collapse are generally very much weaker than thermal pressure forces which attempt to expand the cloud. Expansion is prevented by a lower density medium surrounding the cloud, which has the same pressure as that in the cloud; the pressure throughout the interstellar medium is a consequence of the gas being held within the general gravitational potential field of the galaxy. Thus, under normal conditions, there is no tendency for collapse of any part of the interstellar medium, and hence no reason for star formation to take place.

Spiral Arm Shock

Consider the sequence of events which occurs when gas in the interstellar medium flows into a spiral arm [41, 42]. Since the spiral arm consists of a concentration of stars forming a gravitational potential well, gas undergoes shock deceleration when it flows into the arm, becoming hotter and denser. The interstellar magnetic field, whose lines of force are tied to interstellar gas via the ions within the gas, and which must move with the gas, is also compressed and strengthened in passing through the spiral arm shock. Instabilities will probably then arise in the magnetic field configuration, with large parts of the field becoming buoyant and rising out of the galactic plane, allowing gas attached to the lines of force to collect into pockets, or clouds, near the galactic plane by sliding along the lines of force, which it is free to do. This appears to be the general mechanism in which interstellar clouds can be formed as gas flows into a spiral arm.

After a cloud has formed, there is gradual diminution of heavier elements and ions within the cloud, many of which are mainly responsible for cooling processes within the cloud. These atoms and ions are likely to stay attached to interstellar grain surfaces when they strike those surfaces in the normal course of their thermally agitated motion within the gas. Thus, the newly formed cloud may start out with relatively low temperature, but as time goes on, temperaturє in the cloud will rise as its cooling agents decrease in number.

As temperature in the cloud rises, density correspondingly decreases, and probably when gas emerges from the spiral arm after residing there about 10^8 years, the clouds expand and disappear into a medium which again has become rather homogeneous in density. There is no evidence that interstellar gas clouds exist in the region between spiral arms.

New Star Formation

The optimum time for causing collapse of an interstellar cloud to form stars appears to occur at its formation, since some clouds may be formed with somewhat greater density than others, so that self-gravitational forces become stronger. However, some clouds may suffer dynamic compression at a later stage, being subjected to surface heating if there is nearby formation of a number of hot, massive stars, or nearby explosion of a supernova.

A typical interstellar cloud may have a mass a few hundred or few thousand times that of the Sun; the more massive clouds appear more likely to undergo collapse to form stars. A repeated process of fragmentation of the cloud into separate condensation centers evidently takes place during collapse of the cloud, so that the ultimate product of the collapse is formation of a great number of new stars. If only a portion of gas is formed into new stars, and the remaining gas is quickly expelled from the region of star formation through ultraviolet ionization when new hot stars are formed, then the newly formed stars are not gravitationally bound together, but are free to expand into space. Numerous examples of such expanding associations, containing stars which are massive and very luminous, have been observed in the galaxy.

It is likely that a certain minimum density of gas in the interstellar medium is needed in order that star formation can follow the flow of gas into the spiral arm shock [38]. It might be expected that this minimum density is closely approached through depletion of gas by star formation soon after collapse of the gas to form the galactic disk. In an isolated galaxy, the star formation process could occur subsequently only from gas returned to the interstellar medium as a result of stellar evolution. While this process is observed to occur, it is doubtful that the amount of gas thus returned to the interstellar medium is nearly enough to account for the number of new stars currently observed to be formed in the galaxy.

A number of high-velocity clouds observed in the galactic halo have been interpreted as gas infalling to the galaxy for the first time from intergalactic space [30]. If there is a continuous infall of this sort of gas, which would not be entirely unexpected according to the description in the preceding section, then this gas would collect in the interstellar medium and provide the excess density needed to maintain a steady state of star formation when gas passes into the spiral arms [38].

In an elliptical galaxy, only the first stages of this type of galactic evolution can occur. Infall of gas to form the galaxy presumably results in direct formation of stars from the collapsing gas. Because of the small amount of angular momentum in the system, residual gas, which does not form stars during the collapse process, is likely to collect near the center of the galaxy, forming stars in the vicinity of the center. Henceforth, introduction of new gas into such a system will result in collection of that gas near the center, with additional star formation there, and so on. However, the formation rate of new stars in such a galaxy is likely to be rather small unless there is a fairly substantial continuing input of gas from the intergalactic medium, which there appears to be in our own galaxy.

No existing theory describes the interaction between gas and stars in irregular galaxies. The geometry of such systems is too irregular and complicated for application of the density wave theory, but it is evident from the presence of many hot young stars in irregular galaxies that star formation is a vigorous and continuing process.

STELLAR EVOLUTION AND NUCLEOSYNTHESIS

A star is a large sphere of hot gas held together under its own gravitational forces (for details of stellar structure, see [9]). The structure of the star does not vary appreciably with time. It can therefore be concluded that the interiors of stars are in hydrostatic equilibrium; there is an exact balance between the force of gravity, directed downwards, and expansive thermal pressure of gas, opposing the pull of gravity. At any given point in the star interior, the gas pressure must be high enough to support the weight of overlying layers. Moving inward from that point, there is more matter in the overlying layer, hence, higher pressure is needed to sustain this increased weight.

Pressure increases from the surface of the star toward the center, and in general, density and temperature of the gas also increase from the surface toward the center. Typically, in the central regions of ordinary stars, such as the Sun, temperature will be a few million degrees Kelvin. Under such conditions, electrons are stripped from their atomic orbits, and constituents of the gas become fully ionized. Hence, the gas continues to act like a perfect gas even at densities more than 100 times that of water, which is characteristic of the center of the Sun, because atomic nuclei are very much smaller than atoms in size, hence, the particles have plenty of room to move about between collisions.

A star generates energy in the interior and radiates this energy electromagnetically from the surface, under ordinary circumstances. In order for energy to reach the surface from the deep interior where it is generated, it is necessary that temperature continually decrease from a high value at the center to a low value at the surface. In principle, this energy flow can occur in the deep interior of a star in three ways: conduction, radiation, and convection. Conduction is very inefficient in the interior of the star and usually contributes negligibly to the flow of energy. If the temperature gradient in the interior of the star is not very strong, energy transport is mainly by radiative transfer. In this process, radiation is emitted at any given point in the interior, but since it is emitted proportionally to the 4th power of temperature, more radiation is emitted in a region of higher temperature rather than lower temperature. Thus, photons diffuse in the interior of a star from higher to lower temperature regions.

If temperature gradients become sufficiently high, convection is possible. This is mass transport of heat. A blob of gas in the interior starts rising because of buoyancy forces, and after traveling some distance, it will break up and individual parts will mix into the surroundings. Gas moving vertically upwards in the interior will, under these circumstances, be hotter than the surrounding gas into which it mixes, thus, energy can be efficiently transported by the mass motion of the gas.

Theoretical determination of the structure of a star requires not only a precise balance of pressure at every point, a condition of hydrostatic equilibrium, but also determination of how energy flows in the interior. In successful models of stars, the various differential equations describing these processes are solved simultaneously.

The energy emitted from a star over billions of years is derived from nuclear reactions in its deep interior. These are fusion reactions, in which two relatively light nuclei collide and fuse to make a heavier one, releasing energy as a result. Since thermonuclear reaction rates are extremely sensitive to temperature, energy released from these reactions is strongly concentrated toward the center of the star, where temperature is highest.

In a typical star like the Sun, three-fourths of the mass is composed of the lightest element, hydrogen; since there is only one unit of charge on each nucleus, two protons will have the least coulomb repulsive forces between them of any possible combination of nuclei. Consequently, hydrogen nuclei will fuse in thermonuclear reactions at the lowest temperatures, since the least amount of relative kinetic energy is needed to bring two protons sufficiently close together. It also happens that the most efficient energy generation processes are those which involve hydrogen nuclei.

Four protons can fuse to make one helium nucleus in a variety of ways. Some of these processes commence with the proton-proton reaction, which makes deuterium. Deuterium in turn can capture another proton, making ^{3}He. The ^{3}He nuclei may combine with one another, making ^{4}He with the release of two protons. Alternatively, a ^{3}He nucleus can combine with a ^{4}He nucleus to make an isotope of beryllium, ^{7}Be, and after several more nuclear reactions, an additional proton is captured making a ^{4}He nucleus and releasing the original ^{4}He nucleus which acted as a catalyst for these reactions.

Still other nuclear reactions involve the combination of hydrogen with isotopes of carbon, nitrogen, and oxygen. These reactions form a linked triplet of nuclear reaction cycles in which carbon, nitrogen, and oxygen nuclei are successively transformed into one another. Thus, they act as catalysts in the process, resulting in four hydrogen nuclei combined to form one ^{4}He nucleus.

Main Sequence Stars

Stars converting hydrogen into helium at the centers are called main sequence stars. Since the bulk of available nuclear energy release from nuclear transformations occurs in the hydrogen-to-helium burning stage, the stars spend most of their active lifetimes in this stage. The Sun is expected to take about 10^{10} years to convert its central hydrogen into helium; it is about halfway through this process. When approximately the central 10% of a star like the Sun has fully burned its hydrogen into helium, then it is no longer stable for long periods, but starts to change its structure relatively quickly.

After all hydrogen has been converted into helium in the core of such a star, energy generation ceases there. The energy flow from the center of the star to the surface must continue, therefore energy which had been generated by nuclear reactions at the center must instead be generated by contraction of the central regions, releasing gravitational potential energy. However, this contraction does not last very long, since shrinkage of the star's core leads to a temperature rise of the material not only at the center, but also in the rest of the star's inner parts. Soon this temperature rise is enough to ignite the hydrogen burning thermonuclear reactions in the region surrounding the core which still contains hydrogen. These reactions then form a shell source of energy generation, which provides the energy flow toward the surface needed to maintain the star once again in a state of relatively long-term equilibrium, although the structure is changing at a rate considerably higher than previously during the main sequence hydrogen burning. Since the structure again becomes relatively stabilized, the higher temperature at the center of the core is no longer needed to maintain energy generation, hence, energy flow away from the center to regions of lower temperature tends to make the inner core, now completely helium, have much the same temperature throughout.

Red giant stars. As time passes, the hydrogen burning shell source converts a considerable amount of hydrogen into helium in the star's interior, then this added helium becomes part of the helium core of the star. Meanwhile, the hydrogen burning shell source moves outward in the mass of the star. The core continues to grow in mass and continues to contract slowly, remaining approximately isothermal. However, the outer parts of the star behave differently. Stellar evolution calculations show that, with a hydrogen-burning shell source, and corresponding discontinuity in composition between the helium core and largely hydrogen envelope, as this core contracts, the outer envelope of the star must expand. At the same time, because energy generation lies closer to the surface relative to mass distribution in the star, the star's luminosity increases. Meanwhile, the outer envelope of the star expands. Hence, the surface area of the star increases, and for a given luminosity, the temperature needed to radiate energy output of the star decreases. The surface of the star thus decreases in temperature, and the now relatively giant stars tend to appear red. Hence, they are called red giants.

Reliable calculations of stellar evolution extending beyond the red giant stage are relatively few. However, these calculations do indicate several features of stellar evolution. When a star becomes a red giant, enough mass has been added to the core by the hydrogen-burning shell so that the core starts to contract relatively rapidly. There is no longer time for energy to flow out of the core at a rate sufficient to maintain an approximate isothermal condition; hence, the central core temperature rises considerably above the temperature in the hydrogen-burning shell source. When this temperature reaches a value in the range $1-2\times10^8$ °K, helium-burning thermonuclear reactions commence. In these reactions, three helium nuclei can fuse together to make a nucleus of ^{12}C, and then a further helium nucleus captured by the ^{12}C nucleus to make ^{16}O.

Horizontal branch stars. In a star of relatively low mass, the onset of helium-burning causes quite a large expansion of the core. While the preceding contraction of the core caused expansion of the envelope, expansion of the core now causes contraction of the envelope, and the surface of the star increases in temperature. The star, then, is no longer a red giant, but the tendency is to call it a "horizontal branch star," from its position in a diagram in which luminosity of the star is plotted as a function of surface temperature (Hertzsprung-Russell diagram). In more massive stars, basically the same happens, but contraction of the outer envelope is much less pronounced.

Stellar evolution calculations that are available suggest that helium-burning will finish in the core, and the core will again start to contract. This will ignite helium-burning thermonuclear reactions in the material surrounding the carbon and oxygen core, so that the star in this stage may possess both hydrogen and helium-burning shells. These shells will again progress outward in the mass of the star, adding helium to the region below the hydrogen-burning shell, and adding carbon and oxygen to the central core below the helium-burning shell. At this stage, if the star is not more massive than the upper limit of stability of a white dwarf star, about 1.4 solar masses, it appears likely that the helium-burning shell source will die out after a certain time and the hydrogen-burning shell source will approach the surface quite closely.

White dwarf stars. Under these circumstances, it appears probable that the great bulk of the star at this stage in its evolution will form a white dwarf star. In this star, the electrons form a degenerate gas, thus contributing sufficient pressure to maintain the star against further gravitational contraction. The outer layer of hydrogen, which may still exist when the star is in this late stage of evolution, is probably expelled by one means or another. Possible mechanisms include stellar wind, bulk ejection of a shell of material leaving behind the nucleus of a planetary nebula, or nova explosions in which the hydrogen shell is ejected explosively. Details of the evolution which may lead to one of these possibilities are not yet understood. Since a considerable amount of mass may have been lost from the star during the red giant phase of its evolution, probably by a stellar wind process, the masses of stars which may end as white dwarfs can initially have been considerably larger than one solar mass, perhaps as much as five solar masses.

Further Evolutionary History of a Star

Among theoretical astrophysicists, there is considerable dispute about details of the ultimate evolutionary history of a star too massive to form a stable white dwarf. The general course of its evolution must lead the core of the star toward higher temperatures and higher densities. The two main pathways which may be followed are either complete collapse of the star (implosion) or nuclear explosion of the star [49].

Supernova explosions. Stars with initial masses of 4–8 solar masses were thought, until recently, to be destroyed in supernova explosions. In this sequence, as the carbon and oxygen core grows in size, and the hydrogen and helium-burning shells progress outward in the mass of the star, the core gradually contracts faster and faster, and temperature starts to rise rapidly. When temperature exceeds 10^9 °K, there is ignition of carbon thermonuclear reactions with two carbon nuclei reacting with each other to form various heavier nuclei. Because the core contains a degenerate electron gas, the equation of state is very dissimilar to that in a perfect gas, rather, there is an almost unique relation between pressure and density, with very little dependence of pressure on temperature. Thus, rising temperature in the core has very little effect on its structure. However, since thermonuclear reactions burn at rates that depend upon a high power of temperature, the ignition of carbon thermonuclear reactions will lead to a rapidly increasing rate of energy generation, hence a thermal runaway at the center of the star. This runaway can progress until carbon thermonuclear reactions have exhausted the carbon, and oxygen thermonuclear reactions have begun. In these reactions, two oxygen nuclei react to form heavier nuclei. All these reactions can be expected to be completed before significant changes in the structure have occurred, but meanwhile, the temperature will have risen so high that the center of the core is no longer degenerate.

The sequence of reactions would be sufficient to commence a detonation wave which would progress outward in the core, exploding the carbon and oxygen nuclei there, and probably allowing thermonuclear reactions to burn all the way to the vicinity of iron. Here, the nuclei have the greatest possible nuclear binding energy per nucleon, where their abundances are determined by principles of nuclear statistical equilibrium. If this process could continue, then all of the energy generated by thermonuclear explosion in the core would form a gigantic shock wave which would race through the outer layers of the star, ejecting them into space. This would result in a supernova explosion.

One apparent difficulty with this scheme, however, has been that it should lead to the complete explosion of the star, leaving no stellar remnant behind. While statistics of supernovae in our galaxy and others indicate that stars with masses as low as four solar masses are undoubtedly required to participate in the supernova process, the statistics of pulsars, which are believed to be neutron star remnants of supernova explosions, indicate that such neutron stars should be formed in the majority of supernova explosions in this mass range.

Resolution of this puzzle may have been provided by Paczynski [32]. He noted that when ignition of carbon thermonuclear reactions starts in the core of the star, convection commences at the center, in order to carry away some of the heat generated by carbon reactions. At this point the URCA convection process starts which is a powerful cooling mechanism. At high densities in the core, degenerate electrons may have a Fermi energy of a few MeV. These electron energies will be high enough to initiate electron capture on some of the heavier nuclei present in the core. When such a nucleus is transported by convection to higher densities, electron capture will take place, with emission of a neutrino, and when the same nucleus is subsequently transported toward lower densities, it can emit an electron, transforming back to the original nucleus, with emission of an antineutrino. Neutrinos and antineutrinos emerging from the core of the star can thus carry away tremendous amounts of energy, preventing the thermonuclear ignition from running away to explosive conditions, but nevertheless permitting continued burning of carbon. Under these circumstances, carbon- and oxygen-burning may take place in the core, followed by silicon-burning and formation of the iron equilibrium peak, if, in the meantime, the core has not initiated extensive electron capture, leading to collapse toward nuclear densities in the core, and the formation of a neutron star remnant. It is possible that the pulsar emission mechanism, which is still not understood, but which apparently derives its energy from the rotational energy of a neutron star, may provide enough prompt energy generation in the center to blow off the outer envelope of the star and provide the supernova explosion.

Black hole formation. Among more massive stars, carbon and oxygen-burning is expected to take place at lower densities, before the core of the star undergoes electron degeneracy. When the core of such a star eventually undergoes collapse to form a neutron star remnant, the mass of overlying layers may be too great to be ejected by radiation emitted by the newly formed pulsar. As the mass continues to accrete onto this neutron star remnant, the neutron star may be overwhelmed by general relativistic effects, and collapse to form a general relativistically collapsed object, called a black hole. In such cases, a supernova explosion would not be seen.

Among still more massive stars, perhaps about 50 solar masses or more, a different effect is expected. Such stars can achieve very high temperatures, about 10^9 °K, through large parts of their mass, even though density in most of these parts of the mass is relatively low. Under these circumstances, electron-positron pairs come into equilibrium with the radiation field of photons, having been created from energy in the photons, and this renders the star unstable against collapse [16]. However, the region in which pairs form contains tremendous amounts of carbon and oxygen, and collapse will progress only until thermonuclear detonation occurs in this carbon and oxygen. The result should be a thermonuclear explosion in which all massive outer layers of the star are ejected into space. If a stellar remnant would be formed under these circumstances is not clear, but in any case, a supernova explosion can be expected.

Nucleosynthesis

During the last decade, a great deal of progress has been made in understanding how nuclear reactions taking place during stellar evolution are related to the abundances of elements in nature [49]. This understanding has, to a great extent, resulted from construction of a good abundance distribution curve for elements in solar system material, which has relied in part upon spectroscopic determinations of the abundances of elements in the Sun, but these are rarely better than a factor of 2 for any given element.

Some of the more striking regularities of nuclear abundances are not brought out by using the fairly crude data available for solar abundances. Abundances determined in terrestrial materials are of little value, because the Earth has been extensively differentiated chemically, and it is very difficult to determine the total abundance of any given Earth element. On the other hand, certain types of meteorites have been an extremely valuable source of abundance information, and it appears that among nonvolatile elements, meteorites of the fragile stony variety, known as Type I carbonaceous chondrites, appear to exist with essentially nonvolatile constituents in the original proportions in which they existed in the interstellar medium.

Elements in Solar System Material

Hydrogen constitutes about three-fourths of the mass of solar system material, and helium is the second most abundant element with about one-fourth of the mass. The bulk of helium may have been created by cosmologic nucleosynthesis (described above), or possibly in the first generation of stars formed prior to formation of the galaxies, following the original collapse of matter in the universe after it decoupled from radiation. The remaining elements add up to only about 1.6% of the mass: among these, the third most abundant element, oxygen, has only 10^{-3} as many atoms as hydrogen in the solar system.

The next elements, lithium, beryllium, and boron, have extremely small abundances, whereas carbon, nitrogen, and oxygen are in much greater abundance. Helium-burning in stars jumps directly from helium to carbon and oxygen, and the intervening elements are destroyed in stellar interiors. It can thus be understood how this striking feature of element abundances is directly related to features of nuclear reactions in stellar interiors. Nitrogen is a product of coupled cycles involving carbon, nitrogen, and oxygen isotopes, in which hydrogen is converted into helium in main sequence stars.

Current calculations in the theory of nucleosynthesis indicate that elements in the neon through nickel range are made by explosive carbon, oxygen, and silicon-burning processes in supernova explosions. This range of elements has an intermediate abundance in nature, but, of course, it contains the most important elements present in a planet such as Earth. Of the nonvolatile elements in this range, the most abundant are magnesium, silicon, and iron, which have comparable abundances. Such elements as sodium, aluminum, and calcium are considerably less abundant. Sulphur is half as abundant as silicon, but because it is more volatile than the other elements named, its abundance varies widely in the solid materials of the solar system. Of the various supernova environments previously described, the one most consistent with production of these elements involves massive stars, of 50 solar masses and more, which undergo supernova explosions as a result of the electron-positron pair instability process.

Formation of elements. Current studies of nucleosynthesis have also shown how elements beyond nickel can be formed by a variety of secondary reactions during the process of stellar evolution. Some elements are made by a process of neutron capture taking place on a slow time scale, in which neutrons produced by thermonuclear reactions on relatively minor constituents of the medium during helium-burning are captured on the nuclei of the iron peak which happen to be present in the medium. This leads these nuclei, by a succession of neutron captures and beta decays, toward the heavy element region. This process can form elements only as heavy as lead and bismuth, and it forms only a few isotopes at most of each element.

Another process which has produced heavy elements is neutron capture taking place on a rapid time scale. Current calculations indicate that the most likely astrophysical environment in which this will occur is composed of material in the outer part of a stellar core which has imploded to nuclear densities. This material is compressed and largely converted to neutrons by capture of electrons on protons, before being ejected, by a process not yet understood. It has been shown that in the expansion of material of this sort, charged particle reactions build a variety of nuclei of medium mass number, which then can capture the remaining neutrons in which they are imbedded to form much heavier nuclei. This process is responsible for producing all the elements beyond bismuth which exist in nature, including all of the heavy radioactive nuclei present in nature.

The remaining isotopes of heavy elements which are not formed by neutron capture on some time scale, are likely to be formed by a process of proton capture taking place on a rapid time scale in supernova explosions. Such processes can occur in the outer regions of an exploding star when the supernova shock wave sweeps through a region still containing the original hydrogen. Heavy elements incorporated in this material at the time the star was formed may then capture protons and form neutron-deficient nuclei.

At present, the nuclear physical aspects of nucleosynthesis in stars are generally understood quantitatively better than the astrophysical aspects. Observations of the element abundances in stars show that very few stars are depleted in elements heavier than helium, relative to the Sun, by factors of 100 or more. Most stars in space appear to have abundances of these elements within a factor of 3 of that which is found in the Sun. From these results, together with some attempts to trace the history of nucleosynthesis in our galaxy, it has been concluded that the rate of star formation was probably very rapid in the early history of our galaxy or even in the pregalactic period, so that the formation of heavier elements occurred quite promptly on the galactic time scale [50]. As a consequence, most of the stars likely formed in space after that very early period of rapid star formation, contained sufficient heavy elements to be able to possess planetary systems containing planets having earthlike conditions accompanying them.

FORMATION OF THE SOLAR SYSTEM

The problem of the origin of the solar system has occupied scientists for more than three centuries, since the time of Descartes. For most of this time, the number of facts about the solar system which scientific theories attempted to explain were pitifully few: the regularity of planetary orbits, the alignment of angular momentum vectors within the solar system, and the slow rotation of the Sun. It now appears that the last of these features may have nothing to do with the origin of the solar system; an initially rapidly rotating Sun can easily have been slowed down to its present rate of rotation as a result of magnetic interaction of the Sun with the outflowing solar wind. It is not surprising that such a small number of boundary conditions could give rise to a very large number of widely different theories to explain them. During the last two decades, the whole problem of the origin of the solar system has been greatly transformed through acquisition of new boundary conditions of a physical and chemical character, resulting from meteorite research, from space probe analyses of distant bodies within the solar system, and from manned exploration of the Moon. This has required theories of the origin of the solar system to become more sophisticated, although several varieties are still current.

Nearly all the cosmogonical theories of the solar system developed within the last three centuries can be classified as either monistic or dualistic. In a monistic theory, the development of the Sun and planets occurs within a closed system, in which there is no interaction with any external system. A dualistic theory is one in which an external system, usually another star, is involved in the cosmogonical process. It is possible for a theory to start with dualistic features, and to end with essentially monistic features. Despite such possible ambiguities, the classification is useful.

The first monistic theory was proposed by Descartes, who published a vortex gaseous theory on formation of the solar system in 1634. The first dualistic theory appeared in 1745 when Buffon suggested that a comet made a grazing collision with the Sun, tearing from it sufficient material to form the planets. In those days, comets were thought to have masses comparable to the Sun, so this was the first of the invading star dualistic theories. Since these two theories were advanced, and until three decades ago, opinion has swung back and forth between proponents of monistic and dualistic theories. The review by Ter Haar and Cameron [48] relates this historical development.

Any form of dualistic theory has, at present, been almost universally discounted by astronomers. There are many difficulties associated with attempts to explain formation of the planets by material withdrawn from the Sun. These include various hydrodynamic difficulties: the amount of material drawn from the Sun would be very hot, hence would tend to expand and disperse in space, rather than undergo chemical condensation to form the planets. If matter is drawn from the Sun at the time of close passage of another star, by far the overwhelming bulk of this material would gain so little angular momentum that it would fall back into the Sun after the other star had departed.

The modern theory of solar evolution indicates that the Sun would have destroyed any initial deuterium in its constitution by thermonuclear reactions at an early stage in its evolution, at the time it was fully convective and hence fully mixed. Thus, the deuterium present in the oceans of Earth could not have come from the Sun at any time following its very early deuterium-burning stage, leading to supposition that the remaining material in the Earth was not derived from the Sun either.

Most current theories of the formation of the solar system are monistic, involving some assumptions about the properties of a primitive solar nebula from which the planets and possibly the Sun were believed to have been formed.

The major exception to this general picture among currently advocated theories is perhaps that of Arrhenius and Alfvén [3], who believe that nonequilibrium plasma processes were dominant in formation of the planets and their satellites. They postulate that interstellar gases become enhanced in density and partially ionized as they approach the Sun, so that the resulting plasma flows through the solar system, where some of it may be trapped by magnetic fields, and where various solid grains can condense from the plasma and accumulate into larger bodies via gravitational self-focusing processes forming "jet streams." According to this theory, the Sun was formed as an isolated body by an unspecified mechanism, and at the time the planets were formed, the Sun had a magnetic field which was very much stronger than at present, and no solar wind was in operation which would prevent convergence of the interstellar plasma upon the immediate neighborhood of the Sun. This may be regarded as a dualistic theory, although the external system involved is gas in the interstellar medium, rather than an invading stellar body, and no material is required to be torn from the Sun to form the planets.

Solar Nebula Theories

Modern theories of the primitive solar nebula may be divided into two subclasses: the minimum solar nebula and the massive solar nebula. The minimum solar nebula contains about 0.01 solar masses of material, containing just enough condensible matter to form the present planets, with a large amount of excess volatiles which must be removed following planetary formation. The massive solar nebula may contain about 2 solar masses of material, and is distinguished from the minimum solar nebula through not having the Sun initially present at its center; it is a pure rotating disk of gas from which the Sun is required to form.

Minimum solar mass theories. Most solar nebula theories discussed within the last three decades have been minimum solar mass theories. The earliest involved discussions by von Weizsacker, Kuiper, and Ter Haar [48], stressing the role of turbulence in physical processes within the primitive solar nebula. In fact, although turbulence is likely to be important within the solar nebula, it probably does not have the overwhelmingly important role ascribed to it in these theories. The authors assumed that turbulence must be present if the Reynolds number of the system should be very large, which is practically an inescapable situation in any large cosmic system of gas. However, they did not provide specific mechanisms, calculated in any detail, by which energy could be fed into large eddies in the turbulence, thus maintaining turbulence against natural dissipation processes. Nevertheless, these theories were very important historically, since they marked the first introduction of turbulence as a process which must be taken into account in the theory of the origin of the solar system. They also showed that gas dynamic processes can lead to a remarkably rapid dissipation of the solar nebula, the time scale for the dissipation being only some 10^2 or 10^3 years.

Meanwhile, monistic theories were being proposed in Russia. Schmidt [45] proposed a theory of the origin of the solar system in which the Sun was assumed to capture a swarm of smaller particles into a surrounding disk, suggesting that these small particles accreted to form planets. He later modified this to include gas together with the dust in the surrounding nebula. These ideas were further developed and modified by Levin [23] and Safronov [44], who attempted to calculate details of the formation of planets from the disk of dust and gas and to consider dissipation processes taking place within it.

The term *minimum solar nebula* describes a quantitative process carried out by Hoyle [20] in a theory developed a decade ago, and later elaborated by Fowler, Greenstein, and Hoyle [15]. It was noted by Hoyle that the icy and rocky components of solar matter constitute about 0.015 of the mass, and that the rocky constituents alone make up about 0.003 of the mass. Thus, the minimum amount of mass required to have been present in the solar nebula constitutes about 300 times the masses of the inner planets, which are rocky in composition, together with the masses of Jupiter and Saturn, which he assumed to be essentially solar in composition, with 60 or 70 times the masses of Uranus and Neptune, which he assumed to be composed mainly of ice and rock. Such minimum solar nebula theory requires that suitably condensed material be collected into planets with nearly 100% efficiency.

In Hoyle's theory, a fragment of a collapsing interstellar cloud was assumed to lose a great deal of angular momentum as a result of torque applied by an interstellar magnetic field interacting with the interstellar surroundings of the cloud. Material was then assumed to collapse to a radius somewhat less than that occupied by the orbit of Mercury. Part of the interstellar magnetic field was assumed to be compressed with collapsing material into the proto-Sun; when the collapse had been completed, this magnetic field was assumed to exert a torque on the equatorial regions of the proto-Sun, spinning them off in the form of a nebula, and increasing the radius of this nebula to some tens of astronomical units (AU).

In the elaboration of this theory, Fowler, Greenstein, and Hoyle suggest that chemical condensations will occur within this nebula, and chemically condensed material will form bodies typically tens of meters in diameter, which would be left behind as the magnetic field accelerated the gas toward larger radial distances. Following this, it was assumed that a great deal of magnetic energy would be dissipated in the proto-Sun, and that part of the dissipation would result in acceleration of charged particles to energies of several hundred million electron volts. These energetic particles were then supposed to bombard condensed bodies left behind in the inner part of the solar system, and to produce by nuclear spallation reactions the lighter elements which are not produced in the normal course of nuclear reactions in stellar interiors: lithium, beryllium, and boron. Deuterium in meteorites and in the Earth was assumed to be formed as a result of direct spallation, also by capture of spallation-produced neutrons by hydrogen in ice in the bodies.

The particle bombardment aspects of the theory of Fowler, Greenstein, and Hoyle have fallen into general disfavor because of predicting variations in certain isotopic ratios of some elements in meteorites which have been diligently searched for but not found. The theory had made very specific predictions concerning the presence of such variations, which stimulated a great deal of research on meteorites to search for these variations. However, other aspects of the discussion of formation of the solar nebula in this and other minimum solar nebula theories tend not to be predicted in any detail, thus, these aspects of the theories tend not be easy to test, to accept, or to reject.

Massive solar nebula theories. The principal advocate of a massive solar nebula has been Cameron [5, 7, 8]. This theory attempts to link formation of the solar system to the processes by which stars appear to be formed in space within the galaxy, which has been discussed previously. Random turbulence in the collapsing interstellar gas cloud is considered to have contributed so much total angular momentum to that fragment of the cloud which will form the solar system that gas is forced to form a flattened disk, without any stellar body at the center. The amount of material needed in this massive primitive solar nebula is considerably greater than the mass of the Sun for two reasons:

> newly formed stars are observed to be rapidly losing mass into space in what is called their T Tauri stages, and not all mass within the primitive solar nebula can be dissipated into the Sun;
>
> some must remain at large distances to take up the angular momentum lost by material dissipating to form the Sun, and this gas will later be ejected when the T Tauri phase mass loss of the Sun begins.

Cameron and Pine [8] have calculated detailed numerical models of the massive primitive solar nebula, with thermodynamic conditions linked to the history of the interstellar cloud collapse stage. They found that certain parts of their models were unstable against thermally driven convection, whereas other parts tended to be in radiative equilibrium. They postulated that these models would be dissipated to form the Sun in a few thousand years, due to internal angular momentum transport associated with circulation currents within the disk.

These recurring short dissipation times constitute a challenge to all theories of the primitive solar nebula. It is very much easier to accumulate solids into larger bodies in the presence of a gas than in a vacuum. Solid bodies collide under vacuum conditions today in the asteroid belt, resulting in shattering of the bodies into smaller pieces, rather than their assemblage into a larger body. The presence of a gas assures that different size bodies will move with different relative velocities, thus promoting collisions between them, and under most circumstances preventing the collisions from being rapid enough to cause shattering.

Cameron [7] has investigated general accumulation processes in the context of his massive primitive solar nebula theory. He finds the key to the rapid assembly of planetary bodies in the presence of gas is the start of accumulation processes during the collapse stage of the interstellar cloud which will form the primitive solar nebula. In the presence of turbulence in such gas, interstellar grains acquire significant relative velocities, and if it is assumed that they stick together upon collision, they will accumulate into bodies several centimeters in radius by the time the primitive solar nebula is formed. Such bodies can accumulate fairly rapidly near the central plane of the gaseous primitive solar nebula, where pressure gradients in the gas will induce relative velocities among the condensed solids, leading to their rapid accumulation into much larger bodies.

It has been estimated by Cameron that bodies of planetary size can be built up in times of the order of a few thousand years throughout much of the primitive solar nebula, from regions of relatively small radial distance where terrestrial-type planets would form, to much greater distances where the planets formed would be largely icy in composition. Perhaps the greatest uncertainty in these calculations is associated with the assumption that bodies stick upon contact; far too little is known about the character of the surfaces of chemically condensed material in the primitive solar nebula to evaluate the sticking probability upon collision at different velocities.

Theories of formation of the solar system, as they undergo further development, will undoubtedly benefit from important new developments in the study of meteorites from which physical conditions in the primitive solar nebula can be inferred. This is the development of cosmothermometers and cosmobarometers by Anders and colleagues [1]. When meteorites accumulate at some temperature in the primitive solar nebula, many elements will be completely condensed from gas and be quantitatively incorporated into meteorites, other elements will remain mostly in the gas and only traces of them will appear in the meteorites, but some elements will be partially condensed, and their abundances may be highly variable in the meteorites. These last elements constitute a form of cosmothermometer; a meteorite containing a larger amount of such an element will have been formed at a lower temperature than one which contains little of the element, and a precise thermodynamic analysis of the spread in abundances allows estimates to be made of these temperatures.

These techniques tend to indicate that ordinary chondritic meteorites accumulated in the primitive solar nebula at 400°–500°K, whereas carbonaceous chondrites have tended to accumulate closer to 300° K. Relative oxygen isotope ratios also act as a cosmothermometer. The precise mineral phases in which certain elements are condensed can act as crude cosmobarometers, as can the partial pressure of absorbed gases, such as argon in meteorites. These clues tend to indicate that ordinary chondritic meteorites accumulated at total pressures in the primitive solar nebula of the order of 10^{-5} atm. These thermodynamic conditions exist simultaneously at a suitable radial distance in the massive primitive solar nebula models of Cameron; however, minimum solar nebula models tend to have much lower pressures than are indicated by these cosmobarometers, although such models can be somewhat arbitrary since it is usually assumed that temperature distribution in such models arises from solar heating, with usually some partial dust-shielding.

THE EARTH-MOON SYSTEM

Problems associated with the formation of the Moon have, for the most part, been discussed independently of the general cosmogonies of the solar system outlined above. The Moon is a rather anomalous body within the solar system with a much greater mass relative to the mass of its primary planet than any of the other satellites in the solar system. It has a remarkably low density of only 3.34 g/cm^3, on the average, which is less than the density of meteorites, and very much less than the mean density of the inner terrestrial planets. These unusual characteristics have suggested to many that the formation of the Moon involved some rather unusual event within the history of the solar system.

Origin of the Moon

Four general theories have been advanced to explain the origin of the Moon [22, 28]: fission theories, atmospheric condensation theories, twin planet theories in which the Moon is assumed to be assembled in orbit about Earth, and capture theories in which the Moon is assumed to be formed elsewhere in the solar system and captured by Earth. All these theories have their modern advocates. Recent investigations of the Moon and of lunar materials have provided a large number of boundary conditions which must be satisfied by such theories, but apparently they have not yet allowed the unique selection of one class of theory.

Fission. In a fission theory, the Moon is postulated to be placed in orbit about the Earth as the result of some kind of disruption of the Earth. George Darwin was the first to suggest such a theory in the last years of the 19th century. He assumed that the initial Earth would be rotating with a period of about 4 hours, which he believed to be about twice the resonant period of Earth, so that huge tides would be raised on Earth, leading to separation of one of the tidal bulges to become the Moon. This theory was rejected by H. Jeffreys in 1930 on the grounds that the tidal dissipation would be too large foı the separation to occur. More recently, in modified forms of the tidal theories suggested, the Earth is assumed to be rotating even faster in its initial state, close to rotational instability, which is assumed to occur when the formation of the iron core of the Earth takes place, reducing the moment of inertia and increasing the rotation rate above that required for disruption. If the Earth were rotating fast enough for this, its initial angular momentum would be much greater than that presently possessed by the Earth-Moon system, and a major process for removal of angular momentum is required.

Atmospheric condensation theories, or precipitation theories, as Ringwood [40] calls them, are motivated by the desire to form the Moon with little content of internal iron. These theories rely upon the assumption that the Earth formed very quickly in space, so that a great deal of the gravitational energy of accumulation was retained by the growing Earth, and its outer layers would be so hot that silicates would be contained in gaseous form, together with their decomposition products. If this system collides with a major planetesimal, which sets it spinning rapidly, the outer part of the atmosphere can be shed into orbital motion, from which the silicate materials required to form the Moon are then precipitated [5]. The Moon is then assumed to form from this silicate debris in orbit about Earth.

Twin planet. Many have suggested that the Moon was formed near Earth as an independent body at the time the solar system was formed. The major problem faced by this kind of hypothesis is to account for the low density of the Moon, and no simple mechanism has been suggested by which the Moon could form with a very much smaller density than Earth if the materials from which the two bodies were assembled were similar. The Moon would have to be formed rather close to the Earth; some calculations backwards in time of the orbital motions of the Moon tend to indicate that the Moon was close to Earth much later than the time at which Earth formed, but these calculations generally assume a constant tidal phase lag, and the uncertainty in the actual phase lag introduces corresponding uncertainties into the time scale.

Various modifications of a simple twin planet formation have been suggested. For example, MacDonald [27] suggested two possible theories: (1) a small terrestrial satellite was formed in orbit about the Earth, which acted as a target which collided with a larger incoming body, thus allowing capture of the larger body; and (2) many smaller bodies are assumed to accumulate in orbit about the Earth, very similar to satellites of the giant planets, with the innermost one being more massive and receding from Earth due to tidal drag, gobbling up the others in the course of its recession.

Capture. Many different lunar capture hypotheses have both a geochemical and a dynamical aspect. In a geochemical version of a capture theory, possible conditions are discussed whereby the low mean density of the Moon can be produced, but the tendency is not to consider the details of capture dynamics. A dynamic theory tends not to be concerned with geochemical details, but to deal with the mechanism by which the independent motion of the Moon can be dissipated, leading to capture, and of the subsequent dynamic history of the lunar orbit.

Dynamic theories have recently been discussed in some detail by Kaula [22]. In their original form, as suggested by Gerstenkorn, capture theories required that the Moon initially approach the Earth in a retrograde orbit, with the subsequent elliptical orbit flipping over the poles of the Earth to become a prograde orbit. Various difficulties have been encountered with these versions of the theory, resulting from neglecting important dynamic details, and the most recent formulation of a dynamic theory of prograde capture is due to Singer. Kaula concludes that dynamic capture theories are improbable but not impossible.

All of these classes of theory must attempt to cope with the fundamental importance of the low mean density of the Moon. One of the major aspects of both fission and atmospheric condensation theories is to provide a single interactive system within which chemical differentiation can occur prior to the separation of Earth and the Moon into two distinct bodies. This essential interactive feature is missing from theories in which the Moon is assumed to be independently assembled in orbit about Earth from similar material, and this constitutes a prime objection against such theories. There is a little more freedom in capture theories, provided conditions can be found elsewhere in the solar nebula in which a body of lunar composition might be assembled.

Other theories. Several years ago it was believed that the Sun contained a much smaller abundance of iron relative to silicon than meteorites and terrestrial planets. This led Urey [52] to a theory of formation of the Moon, assuming that the Moon was formed from the condensible fraction of solar material, thereby being relatively low in iron and having a small mean density. It was supposed to be but one of a large number of primary condensation objects within the solar system. A great majority of these were assumed to collide with one another, producing fractionation of silicates and iron, leading to a concentration of iron in the surviving planets. The Moon was then supposed to be left over from this process, a surviving primary object, which was captured by the Earth. However, errors have been discovered recently in the oscillator strengths of the iron lines used to determine the solar abundance of iron; it now appears that the iron-to-silicon ratio in the Sun is essentially the same as in meteorites and terrestrial planets.

Analyses of lunar samples recently have indicated that the upper portion of the Moon is abnormally rich in aluminum, calcium, and titanium. Since oxides and silicates of these metals are the first major refractory substances to condense out of gas at high temperature, this has led P. W. Gast to propose that when the Moon was assembled, the outer layers were made from such very high temperature condensates. This point of view has recently been extended by Anderson [2], who has suggested that the entire Moon resulted from a complete chemical fractionation of such high temperature condensates. If this is correct, it must place the formation of the Moon in a part of the primitive solar nebula where the temperature is much higher than that in which the main part of Earth was formed. Cameron [6] has suggested that this locates the formation site of the Moon inside the orbit of Mercury, so that perturbations of the initial lunar orbit by Mercury can cause the Moon to acquire a highly elliptical orbit from which it can be captured by Earth, and at the same time, the surprisingly high orbital eccentricity of Mercury can be produced.

Earth scientists have turned up a great deal of evidence bearing on the physical and chemical history of the Earth. Nevertheless, major controversies remain as a result of their investigations, which render knowledge of the earliest history of Earth very unreliable. The oldest rocks, determined from a decay of radioactivities in the Earth, are only some 4×10^9 years old. On the other hand, a great deal of information indicates that the solar system is 4.6×10^9 years old. The first several hundred million years of Earth's history remain an enigma as far as direct geological investigation is concerned.

Formation of the Earth

A major controversy is whether Earth was formed in a very cold or very hot state. In the early years of this century, geologists favored a picture of an entirely molten Earth in its earliest stages. No doubt this point of view was influenced by dualistic theories for the origin of the solar system, in which the Earth was pictured as hot condensation from a filament of hot gases torn from the Sun. However, some two decades ago, this prevailing point of view was challenged by Urey [51], who pointed out that certain of the more volatile elements present in large quantities in the Earth could not be present if Earth were to form from condensates of such a gas filament at high temperature. Urey concluded that the condensed material which assembled to form the Earth was cold, not more than a few hundred degrees Celsius. However, it is sufficient that small condensed bodies which accumulate upon Earth be rather cool, since the accumulation process may produce a very hot body, but the more volatile elements can be retained by this body as soon as it has an appreciable mass.

The initial temperature in the Earth's interior is a strong function of the time interval required for accumulation of the Earth. If a body is pictured where growth is steady by accumulation of relatively small particles onto a much larger nucleus, then the particles will release gravitational potential energy when their infall is stopped by contact with the surface of the growing body. The bulk of released gravitational potential energy would then be radiated away from the surface into surrounding space. However, the faster the rate of accumulation, the higher the required surface temperature would be which would radiate the bulk of energy into space. This radiating surface temperature becomes a measure of the Earth's interior temperature.

If the Earth forms on a time scale characteristic of gaseous dissipation of the primitive solar nebula, of the order of 10^3 years, then interior terrestrial temperatures of the order of 5000 to 10 000° K would be produced [7]. At these high temperatures, the bulk of solid materials could only exist in gaseous form, and would form a hot extended atmosphere of the Earth. This is the basis for atmospheric condensation theories of the Moon's origin. However, this hot atmosphere would soon lose the bulk of its heat by radiation into space, and likely it would have condensed into liquid rock form in a few thousand years.

Related to these uncertainties in the thermal history of Earth are problems of the origin of the atmosphere and oceans of Earth. A great deal of geologic evidence was assembled by Rubey in 1951 [43] to show that the oceans had been outgassed from the interior of Earth. He concluded that the outgassing was a continuing process, the outgassing of water still occurring at present. Subsequently, however, no evidence has accumulated that any completely primitive water is still being outgassed from the interior of the Earth, with present evidence suggesting that the bulk of water now being outgassed has been recirculated from the surface through the interior of the Earth.

At about the same time, Brown [4] pointed out that the abundances of oxygen and nitrogen in Earth's atmosphere were several orders of magnitude greater than the abundances of rare gases. Thus, this implied that the atmosphere had originated largely in a secondary fashion by outgassing from the interior of Earth, since only chemically combined elements could be brought into the Earth in large amounts by small cold bodies which participated in the accumulation process. Present evidence indicates that both oceans and atmospheres of the Earth have been formed largely as a result of outgassing from the interior. The abundances of rare gases do not exhibit the solar abundance pattern, but rather a highly fractionated pattern which is characteristic of the abundances of rare gases absorbed in meteorites. Even the small amounts of rare gases present in Earth's atmosphere probably were also brought into the Earth by small-sized bodies.

Origin of Life

These considerations have important bearing upon problems of the origin of life on Earth. The present indication is that any primordial atmosphere that the Earth had, resulting from accumulation from the primitive solar nebula, was probably swept away by the very intensive T Tauri stage of the solar wind. Subsequently, the present atmosphere and oceans of the Earth were outgassed from the interior. Many biochemical investigations relating to the origin of life have assumed that the primitive atmosphere of Earth should be basically derived from solar composition, so that gases would be composed essentially of hydrogen, methane, ammonia, and water vapor. However, a secondary atmosphere will have much less hydrogen. Nitrogen may be outgassed partly in the form of ammonia, and carbon in the form of methane or other gaseous organic compounds, but the bulk of carbon is likely to be outgassed as carbon dioxide. Hence, such biochemical investigations of the early history of life on Earth might better proceed from the assumption that the composition of gases was mainly water vapor and carbon dioxide, with somewhat more minor constituents of excess hydrogen in the form of ammonia and methane.

The origin of life on Earth is thus seen arising as a natural consequence of a long series of physical and chemical processes taking place in association with the evolution of the universe, many details of which are still being unraveled. The seeming inevitability of this long sequence of events has led to the prevalent view that planetary systems are extremely widespread throughout our galaxy and the universe, that planets suitable for life exist in a large fraction of these planetary systems, and that the general geochemical conditions associated with such planets will be rather similar to those on Earth, so that the development of life and possibly of intelligent beings should be an exceedingly widespread phenomenon.

REFERENCES

1. Anders, E. Physico-chemical processes in the solar nebula as inferred from meteorites. *In*, Reeves, H., Ed. *On the Origin of the Solar System*, pp. 179–201. Paris, Cent. Nat. de Rech. Sci., 1972.
2. Anderson, D. L. Structure and composition of terrestrial planets. *In*, Cameron, A. G. W., Ed. *Cosmochemistry.* Boston, Reidel, 1973.
3. Arrhenius, G., and H. Alfvén. Fractionation and condensation in space. *Earth Planet. Sci. Lett.* 10:253–267, 1971.
4. Brown, H. Rare gases and the formation of the Earth's atmosphere. *In*, Kuiper, G. P., Ed. *The Atmospheres of*

the Earth and Planets, pp. 258–266. Chicago, Univ. Chicago Press, 1952.

5. Cameron, A. G. W. Formation of the Earth-Moon system. *EOS, Trans. Am. Geophys. Union* 51:628–633, 1970.
6. Cameron, A. G. W. The orbital eccentricity of Mercury and the origin of the Moon. *Nature* 240(5379):299–300, 1972.
7. Cameron, A. G. W. Accumulation processes in the primitive solar nebula. *Icarus* 18(3):407–450, 1973.
8. Cameron, A. G. W., and M. R. Pine. Numerical models of the primitive solar nebula. *Icarus* 18(3):377–406, 1973.
9. Cox, J. P., and R. T. Giuli. *Principles of Stellar Structure*, Vol. I, II. New York, Gordon and Breach, 1968.
10. Dicke, R. H., P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson. Cosmic black-body radiation. *Astrophys. J.* 142(7):414–419, 1965.
11. Doroshkevich, A. G., and I. D. Novikov. Average density of radiation in the metagalaxy and some problems of relativistic cosmology. *Dokl. Akad. Nauk SSSR* 154(4):809–811, 1964.
12. Doroshkevich, A. G., Ya. B. Zel'dovich, and I. D. Novikov. The origin of galaxies in an expanding universe. *Astron. Zh.* 44(2):295–303, 1967.
13. Doroshkevich, A. G., R. A. Sunyaev, and Ya. B. Zel'dovich. The formation of galaxies in Friedmannian universes. *In*, Longair, M., Ed. *International Astronomical Union Symposium* 63:200, 1973.
14. Ezer, D., and A. G. W. Cameron. The evolution of hydrogen-helium stars. *Astrophys. Space Sci.* 14:399–421, 1971.
15. Fowler, W. A., J. L. Greenstein, and F. Hoyle. Nucleosynthesis during the early history of the solar system. *Geophys. J.* 6:148–220, 1962.
16. Fraley, G. S. Supernovae explosions induced by pair-production instability. *Astrophys. Space Sci.* 2:96–114, 1968.
17. Gamow, G. On the temperature of the hot universe. *Kgl. Dan. Viedens. Selsk. Mat. Fys. Medd.* 27(10), 1953.
18. Hagedorn, R. Thermodynamics of strong interactions at high energy and its consequences for astrophysics. *Astron. Astrophys.* 5:184–205, 1970.
19. Harrison, E. R. Particle barriers in cosmology. *Comments Astrophys. Space Phys.* 4:187–191, 1972.
20. Hoyle, F. On the origin of the solar nebula. *Quart. J. Roy. Astron. Soc.* 1:28–55, 1960.
21. Kaplan, S. A., and S. B. Pikel'ner. *Mezhzvyozdnaya Sreda.* Moscow, Izv. Akad. Nauk SSSR, Ser. Fiz., 1963. (Transl: *Interstellar Environment*). Cambridge, Mass., Harvard Univ. Press, 1970.
22. Kaula, W. M. Dynamical aspects of lunar origin. *Revs. Geophys. Space Phys.* 9:217–238, 1971.
23. Levin, B. Yu. Structure of the earth and planets and meteoritic hypothesis of their origin. *Priroda* 10:3–14, 1949.
24. Lifshitz, E. M. On gravitation stability of the expanding world. *J. Exp. Theor. Phys.* 16:587, 1946.
25. Lin, C. C., and F. H. Shu. On the spiral structure of disk galaxies. *Astrophys. J.* 140:646–655, 1964.
26. Lin, C. C., and F. H. Shu. On the spiral structure of disk galaxies. III. Comparison with observations. *Astrophys. J.* 155:721–746, 1969.
27. MacDonald, G. J. F. Tidal friction. *Revs. Geophys.* 2(8):467–541, 1964.
28. Marsden, B. G., and A. G. W. Cameron, Eds. *The Earth-Moon System.* New York, Plenum, 1966.
29. Omnes, R. On the origin of matter and galaxies. *Astron. Astrophys.* 10:228–245, 1971.
30. Oort, J. H. The formation of galaxies and the origin of the high-velocity hydrogen. *Astron. Astrophys.* 7:381–404, 1970.
31. Ozernoy, L. M., and A. D. Chernin. The fragmentation of matter in a turbulent metagalactic environment. *Astronom. Zh.* 44(6):1131–1138, 1967. (Transl: *Sov. Astron.*) 11(3):907–913, 1968.
32. Paczynski, B. Carbon ignition in degenerate stellar cores. *Astrophys. Lett.* 11:53–55, 1972.
33. Peebles, P. J. E. Primeval helium abundance and the primeval fireball. *Phys. Rev. Lett.* 16:410, 1966.
34. Peebles, P. J. E. Origin of the angular momentum of galaxies. *Astrophys. J.* 155:393–401, 1969.
35. Peebles, P. J. E. Rotation of galaxies and the gravitational instability picture. *Astron. Astrophys.* 11(3):377–386, 1971.
36. Peebles, P. J. E. *Physical Cosmology.* Princeton, N.J., Princeton Univ. Press, 1972.
37. Peebles, P. J. E., and R. H. Dicke. Origin of the globular star clusters. *Astrophys. J.* 154(12):891–908, 1968.
38. Quirk, W. J. On the gas content of galaxies. *Astrophys. J.* 176:L9–L14, 1972.
39. Reeves, H., J. Audouze, W. A. Fowler, and D. N. Schramm. On the origin of light elements. *Astrophys. J.* 179:909, 1973.
40. Ringwood, A. E. Origin of the Moon. (Clarke Memorial Lecture) *Earth Planet. Sci. Lett.* 8:131–140, 1970.
41. Roberts, W. W. Large-scale shock formation in spiral galaxies and its implications on star formation. *Astrophys. J.* 158:123–143, 1969.
42. Roberts, W. W., Jr., and C. Yuan. Applications of the density wave theory to the spiral structure of the Milky Way system. III. Magnetic field: large-scale hydromagnetic shock formation. *Astrophys. J.* 161:877–902, 1970.
43. Rubey, W. W. Geologic history of sea water. An attempt to state the problem. *Bull. Geol. Soc. Am.* 62:1111–1148, 1951.
44. Safronov, V. S. Dimensions of the largest bodies falling on planets in the process of their formation. *Astron. Zh.* 42(6):1270–1276, 1965.
45. Schmidt, O. Y. The astronomical age of the earth. *Dokl. Akad. Nauk SSSR* 46:293–295, 1945.
46. Sciama, D. W. *Modern Cosmology.* New York, Cambridge Univ. Press, 1971.
47. Spitzer, L., Jr. *Diffuse Matter in Space.* New York, Wiley, 1968.
48. Ter Haar, D., and A. G. W. Cameron. Historical re-

view of theories of the origin of the solar system. *In*, Jastrow, R., and A. G. W. Cameron, Eds. *Origin of the Solar System*, pp. 4–37. New York, Academic, 1963.

49. TRURAN, J. W. Theories of nucleosynthesis. *In*, Cameron, A. G. W., Ed. *Cosmochemistry*. Boston, Reidel, 1973.

50. TRURAN, J. W., and A. G. W. CAMERON. Evolutionary models of nucleosynthesis in the galaxy. *Astrophys. Space Sci.* 14:179–222, 1971.

51. UREY, H. C. *The planets, Their Origin and Development.* New Haven, Yale Univ. Press, 1952.

52. UREY, H. C. Primary and secondary objects. *J. Geophys. Res.* 64:1721–1737, 1959.

53. WAGONER, R. V., W. A. FOWLER, and F. HOYLE. On the synthesis of elements at very high temperatures. *Astrophys. J.*, Pt. 1, 148(4):3–49, 1967.

54. YONEYAMA, T. On the fragmentation of a contracting hydrogen cloud in an expanding universe. *Pub. Astron. Soc. Japan* 24(1):87–98, 1972.

55. WHEELER, J. A. *Magic Without Magic*. New York. W. H. Freeman, 1972.

56. ZEL'DOVICH, Ya. B., and I. D. NOVIKOV. *Relativistic Astrophysics*. Moscow, Nauka, 1967.

57. ZEL'DOVICH, Ya. B., and A. A. STAROBINSKIY. The birth of particles and vacuum polarization in an anisotropic gravitation field. *Zh. Eksp. Teor. Fiz.* 61(6):2161–2175, 1971. (Transl: *J. Exp. Theor. Phys.*) 39(6):1159–1166, 1972.

58. ZEL'DOVICH, Ya. B., and I. D. NOVIKOV. *Relativistic Astrophysics.*, Vol. 2. Chicago, Univ. Chicago Press, 1975. (In press)