18 May 1984, Volume 224, Number 4650 SCIENCE

Windows on a New Cosmology

George Lake

Summary. The standard Big Bang cosmology requires very special initial conditions: (i) an excess of matter over antimatter, (ii) delicate tuning of the expansion rate at an early time to produce the long-lived Universe we see today, and (iii) a conspiracy among parts of the Universe out of causal contact to produce the currently observed degree of homogeneity. New theories that unify the strong and electroweak interactions may remove the necessity of specifying these conditions as initial data by introducing an inherent matter-antimatter asymmetry in physical laws and changing the early dynamic history of the Universe.

At first sight, there seem to be remarkably few probes of the early Universe. The expansion of the Universe was first discovered by Edwin Hubble. Together with the 2.7-K blackbody radiation discovered by Penzias and Wilson (*1*), this expansion implies that the Universe was once very hot and dense (the Universe expands adiabatically, cooling in the process). There are other signatures of this hot, dense state of the Universe. Nucleosynthesis of the lighter elements (*2*) occurred when the Universe was at temperatures of 10^8 to 10^9 K (a thermal energy of 0.1 to 1 MeV; hereafter such energy units will often be used to characterize the temperature), producing the currently observed abundances of ^{2}H, ^{3}He, ^{4}He, and ^{7}Li.

The measurements of these elemental abundances are in such good accord with the calculations of primeval nucleosynthesis that one is encouraged to consider the state of the Universe at still higher temperatures and densities, ones that will never be available in the laboratory. Consideration of these conditions places cosmology in a new position with respect to fundamental physics. It can provide a unique laboratory for the understanding of new physical laws, perhaps in the same way that Galileo's observations of Jupiter led to the formulation of the laws of gravity or in the way that nuclear physics was stimulated by the mystery of energy generation in stars. A key to further discovery may well be the tension of an ill fit between particle physics and cosmology. At present they are sewn together in an awkward manner, similar to the patchwork-quilt nature of the fundamental forces before the electroweak unification (*3, 4*). I will make no attempt to hide the ugly stitches of the seam, as they are likely to be the source of the most lively research in the coming years.

The author is with the Physical Research Laboratory, AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

Old Global Symmetries and New Gauge Symmetries of Grand Unified Theories

The use of symmetries has a long history in the formulation of physical laws. In the *Phaedo*, Plato explains that objects fall toward the center of the earth as there is no other direction for them to go. Even in modern physical theory, all things are permitted unless specifically excluded by symmetry laws. [In this regard, it is interesting that, only 40 years ago, baryon number was invented as a symmetry to exclude proton decay (*5*). For the purpose of our discussion here, baryon number may be considered to be the number of protons and neutrons minus the number of their antiparticles.] Until recently, the types of symmetry considered were nearly all "global" symmetries: C or charge symmetry—the replacement of particle by antiparticle; P or parity—spatial inversion; T—time reversal, isospin, and strangeness. All these symmetries are now known to be only approximate (*3*). There are also "local" or "gauge" symmetries. In these there are continuous transformations which differ from place to place. One example is the phase changes that can be put into the electric field when accompanied by a corresponding transformation of the magnetic vector potential.

At present, all global symmetries are observed to be violated with the exception of baryon and lepton number. It is hoped that experiments on proton decay and neutrino mass will reveal that these symmetries are also violated, which would leave only gauge symmetries as the true symmetries of nature. If baryon number is not conserved, this opens the possibility of generating the matter content of the Universe, as will be discussed later.

A symmetry may be either violated (which is to say it is not really a symmetry) or "broken" (*6*). A macroscopic example of a broken symmetry is the absence of translational and rotational invariance in crystals despite the existence of these symmetries in the electromagnetic forces that determine the structure of the crystal. This occurs because the physical states of a system can have a lower energy by breaking the symmetries of the Hamiltonian. This can even occur in the vacuum state of a field theory. In this way the properties of the space between particles can become dynamically important to cosmology, as will be explained later.

Symmetry is also important in cosmology. In the next section, I use the assumptions of homogeneity and isotropy to construct a tractable description of universal expansion. In the following sections, I describe the exquisite tuning required for these symmetries and attempt to relate them to physics at very high energies, producing a model known as the inflationary Universe.

Dynamical Equations for the Expansion

To get a good handle on the conundrums of cosmology, we need to start by describing the dynamical history of the Universe. The gross features should be clear at the outset. The gravitational force of matter pulls everything in the Universe toward every other thing. It halts expansion and speeds contraction. A pressure affects the expansion in two ways: (i) its energy density has an equivalent mass density, which contributes to the gravitational forces, and (ii) it does work (PdV, where P is the pressure and V is the volume) as the Universe expands.

Gravity determines an expansion law, where the declaration or acceleration is determined by the density and pressure. The evolution of density is determined by the equation of state, which relates the density to the pressure (which is different for radiation, matter, and the vacuum) and the conservation of energy, which balances the books for the dilution of density by the expansion and the work done by the pressure. With these features in mind, we can easily construct a mathematical description of the expansion.

In order to describe the expansion of the Universe, we start with a Newtonian calculation of the expansion of a small spherical volume. The reason this suffices is that the matter outside the volume exerts no force (by spherical symmetry), and in this small volume the forces are small and the expansion velocities are nonrelativistic. As the expansion is universal, the laws derived for any small volume must apply to the Universe as a whole (7). There are some features that must be added to this approach. The equivalent "mass" of energy or radiation, E/c^2, is important in the early history of the Universe. Another problem is that over large enough distances, the Universal expansion approaches and even exceeds the speed of light, c. The full treatment of these effects requires general relativity; but the most important aspect is that, when the Newtonian calculation indicates that two regions are expanding faster than light, it really means that they are not in communication.

Consider a particle at a distance $L(t)$ from the center of the small spherical volume. First, we introduce a coordinate r, which is a "comoving coordinate" and does not change as the expansion occurs. A scale function $a(t)$ describes the expansion in that

$$L(t) = a(t)r \qquad (1)$$

The acceleration of the particle at $L(t)$ is given by the gravitational attraction of the matter inside a shell of radius $L(t)$, that is,

$$\ddot{L}(t) = -\frac{GM(\text{int})}{L(t)^2} = -\frac{4\pi}{3}G\rho\frac{L^3(t)}{L^2(t)} = -\frac{4\pi}{3}G\rho L(t) \qquad (2)$$

where G is Newton's gravitational constant. Here the mass interior to L, M(int), and the density, ρ, include the normal matter density as well as the "mass" of energy or radiation, E/c^2. If we divide by r, we find the dynamical equation for the scale factor:

$$\ddot{a}(t) = -\frac{4\pi}{3}G\rho a \qquad (3)$$

Pressure has been neglected here. When it is included, the correct equation reads

$$\ddot{a}(t) = -\frac{4\pi}{3}G(\rho + 3P)a \qquad (4)$$

where P is measured in units that are somewhat unfamiliar; this pressure is the familiar one divided by c^2. Normally one is used to only gradients in the pressure having significance. The pressure here is defined as $-(dE/dV)$ where E is the energy and V the volume. Its absolute value is important because of its equivalent mass density. This effect only arises in the full general relativistic treatment, which is why it has to be inserted here "by hand." We add to this the first law of thermodynamics (which is nearly the conservation of energy, except for the PdV work done in the expansion):

$$\frac{d\rho}{da} = -\frac{3(P + \rho)}{a} \qquad (5)$$

To complete the description of the evolution of the density and the scale factor, an equation of state relating the density and the pressure has to be specified. There are three cases normally considered: (i) "dust": $P = 0$; (ii) radiation: $P = \rho/3$; and (iii) vacuum energy: $P = -\rho$. We can now look at the evolution of the density and scale factor in these three cases.

"Dust," $P = 0$: The solution to Eq. 5 in this case is $\rho_{\text{dust}} = \rho_0(a_0/a)^3$, where the subscript zero denotes the current epoch. Integrating Eqs. 4 and 5, we find

$$\dot{a}^2 = \frac{8\pi}{3}G\rho_{\text{dust}}\,a^2 - \epsilon A^2 \qquad (6)$$

where ϵA^2 is an integration constant with $\epsilon = -1, 0, +1$ and $A > 0$. If we redefine $a \rightarrow Aa$ and $r \rightarrow A^{-1}r$, we have

$$\dot{a}^2 = \frac{8\pi}{3}G\rho_{\text{dust}}\,a^2 - \epsilon \qquad (7)$$

If $\epsilon = +1$, the scale factor expands to a maximum value and then the Universe recontracts. If $\epsilon = 0$ or -1, $a(t)$ can increase monotonically from zero to infinity. The parameter ϵ is called the sign of the spatial curvature. To show the physical significances of this number, we rearrange Eq. 7 and evaluate it at our current epoch to find

$$\frac{-\epsilon}{a_0^2} = \left(\frac{\dot{a}_0}{a_0}\right)^2 - \frac{8\pi}{3}G\rho_0 \equiv H_0^2 - \frac{8\pi}{3}G\rho_0 \qquad (8)$$

Here (ϵ/a_0^2) is the present spatial curvature and H_0 is the current expansion rate (the Hubble constant), given by $\dot{a}/a$ at the current epoch. The current value of the Hubble constant is 50 km sec^{-1} Mpc^{-1} (1 megaparsec is 3.2×10^6 light-years). This number is uncertain, with some techniques indicating a value nearly twice as large. The spatial curvature is then positive, negative, or zero depending on whether the density ρ_0 is greater than, equal to, or less than

$$\rho_{\text{crit}} = \frac{3H_0^2}{8\pi G} = 4.9 \times 10^{-30}\left(\frac{H_0}{50\text{ km sec}^{-1}\text{ Mpc}^{-1}}\right)^2 \text{ g cm}^{-3}$$

This shows that the Newtonian distinction of a system being bound, unbound, or marginally bound is equivalent to the general relativistic "sign of the spatial curvature." An unbound Universe ($\rho < \rho_{\text{crit}}$) is usually called "open" and expands forever. A bound Universe ($\rho > \rho_{\text{crit}}$) is "closed"; the force of gravity is sufficient to reverse the expansion and cause collapse. A marginally bound Universe ($\rho = \rho_{\text{crit}}$) is "flat," as the spatial curvature is zero.

The evolution of the scale factor in a matter-dominated Universe is shown in Fig. 1. This plot shows the infinite expansion of an open ($\epsilon = -1$) or flat ($\epsilon = 0$) Universe and the recollapse of a closed ($\epsilon = +1$) Universe.

"Radiation," $P = \rho/3$: In this case, Eq. 5 becomes $\rho_{\text{rad}} = \rho_0(a_0/a)^4$. Once again, integrating Eqs. 4 and 5 and transforming as before, we find

$$\dot{a}^2 + \epsilon = \frac{4\pi}{3}G\rho_{\text{rad}}\,a^2 \qquad (9)$$

I will say more on this case later.

"Vacuum energy," $P = -\rho$: Equation 5 now yields the particularly simple form; $\rho_{\text{vac}} = \rho_0$, a constant. As the Universe expands, the energy density of the vacuum remains unchanged. For $\epsilon = 0$, this case yields a qualitatively different equation:

$$\ddot{a} = \frac{8\pi G\rho_{\text{vac}}}{3}a \qquad (10)$$

This is the form of Einstein's equations with a cosmological constant, where $8\pi G\rho_{vac}$ is normally referred to as Λ. This form was first introduced to satisfy a "perfect cosmological principle": the Universe is the same at any time and place. The particular form chosen by Einstein was a combination of matter density and vacuum energy that gave $\rho = -3P$, making the scale factor constant in time. It was not treated as just another equation of state because of the unphysical feature of either negative pressure or density. After the discovery of the universal expansion by Hubble, the cosmological constant was later denounced by Einstein as the biggest blunder of his life. That Einstein would regard this as such a colossal blunder is easy to see. Since gravity is a Universal force, Newtonian physics and general relativity automatically lead to the picture of a dynamic Universe, except in the very special case of a purely homogeneous, isotropic Universe with $\rho = -3P$. Even in this idealized case, small perturbations will immediately set a Universe in motion. To come so close to a prediction of an expanding, dynamic Universe and yet abandon the idea because of preconceived belief must surely have seemed to Einstein a grave mistake. The backlash led to a veritable banishment of the cosmological constant in physical cosmology. The revival of this term is due to its modern use as the energy density of the vacuum (which can have a negative pressure) which arises in field theories.

When the vacuum energy density is large, the curvature, ϵ, is unimportant and the evolution of the scale factor is given by

$$a = a_0 e^{\left(\frac{8\pi}{3} G\rho_{vac}\right)^{1/2} t} \tag{11}$$

The scale factor evolves exponentially with time, t. This will be a critical feature later.

A Matter and Radiation Universe

The observed Universe expands at a rate of 50 km sec^{-1} Mpc^{-1}, contains a matter density equal to $\sim 10^{-30}$ g cm^{-3}, and has a radiation density equal to 10^{-33} g cm^{-3}. Matter now dominates radiation by a factor of approximately 10^3. The expansion of the Universe diminishes the matter density by $(\text{volume})^{-1}$ or $[a(t)]^{-3}$, while the radiation density evolves as $[a(t)]^{-4}$. This can be viewed in terms of the PdV work done by the radiation or, equivalently, the red-shifting of individual quanta by the expansion. As a consequence, when the scale factor was one-thousandth of its current value, the density of matter and radiation were equal. Before this time, the dynamics of the Universe was dominated by the radiation.

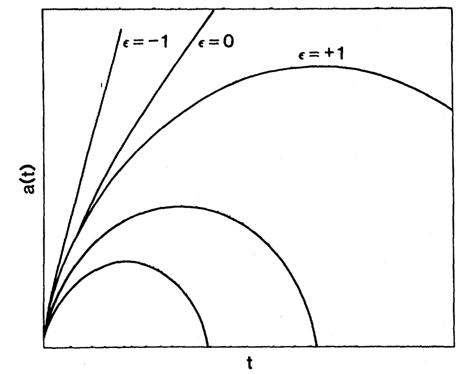

Fig. 1. The evolution of the scale factor $a(t)$ with time in a matter-dominated Universe. The $\epsilon = -1$ curves correspond to open, infinitely expanding Universes. A Universe with positive curvature ($\epsilon = +1$) is closed; the force of gravity is sufficient to halt the expansion and cause collapse. The $\epsilon = 0$ Universe is "flat"; small perturbations will cause it to behave as either the open or the closed case.

The Bang and Causal Structure

If we integrate the equations for our expanding Universe backward, we find a time when the scale factor is zero. As this is the natural starting point for the Universe, we call this $t = 0$. At $t = 0$, the scale factor in the radiation case is proportional to $t^{1/2}$ (at these early times, the curvature term is unimportant and the Universe is radiation-dominated). The expansion rate at $t = 0$ is singular, the Universe starts with a Bang.

In the case of a Universe comprised of radiation and matter, the singular expansion has an important consequence for the causal structure of the Universe. Although starting from an initially small volume, the region of the Universe in causal contact (defined as having the ability to have exchanged a signal at the speed of light) is initially zero and grows as the rate of expansion decreases and time elapses for signals to propagate. This leads to the concept of a particle horizon, that region of space that it is possible to have seen. Note that once seen, always seen.

The case of a "vacuum energy" is different in that the expansion rate is uniform, and, looking to the past, one sees a Universe that is forever shrinking, but the scale factor only approaches zero as time goes to $-\infty$. A forward evolution shows a fundamental difference in the causal structure. As the particles are carried apart by expansion, the recessional velocity between them increases. Eventually the inferred speed approaches and even exceeds the speed of light. At this point the particles drop out of communication. In this type of Universe, the volume of communication remains fixed and expansion carries particles out of this volume through an event horizon (just as particles falling into a black hole pass through a point of last communication). Whereas particles say hello and never farewell in a radiation-matter Universe with particle horizons, there are only farewells in a Universe with a vacuum energy density and event horizons.

Starting Up the Universe

What constitutes "initial data" in cosmology? (8). In the standard model of a radiation-dominated Universe, $t = 0$ is a singular point and as such it is a poor place to specify any parameters. The fundamental constants, Newton's gravitational constant, G; Planck's constant characterizing the quantum aspects of matter, $\hbar$; and the speed of light, c, can be used to define units of mass (or equivalently energy or temperature), time, and length and thus yield a starting point for a discussion of the early Universe. These are known as Planck units and are given by

$$M_{PL} = \left(\frac{\hbar c}{G}\right)^{1/2} = 2.2 \times 10^{-5}\text{g} = 1.2 \times 10^{19}\,\frac{\text{GeV}}{c^2}$$

$$t_{PL} = \left(\frac{\hbar G}{c^5}\right)^{1/2} = 5.3 \times 10^{-44} \text{ second}$$

$$\ell_{PL} = \left(\frac{\hbar G}{c^3}\right)^{1/2} = 1.6 \times 10^{-33} \text{ cm}$$

A particle of Planck energy is one whose Compton wavelength (the radius where its quantum behavior becomes evident, $\hbar/mc$) is the same as its Schwarzschild radius (the radius inside which it becomes a black hole, GM/c^2), a ghastly beast beyond the realm of known physical law. We will restrict our discussion to temperatures of less than 1 percent of the Planck temperature, that is, $T < 10^{17}$ GeV. This regime is interesting as the symmetry breaking which separates the strong and electroweak interactions is believed to occur at 10^{15} to 10^{16} GeV.

In the standard scenario, the Universe cools to a temperature of 10^{17} GeV just 10^{-39} second after the initial singularity. The Hubble constant has the dimensions of (velocity/distance) or (1/time) and de-

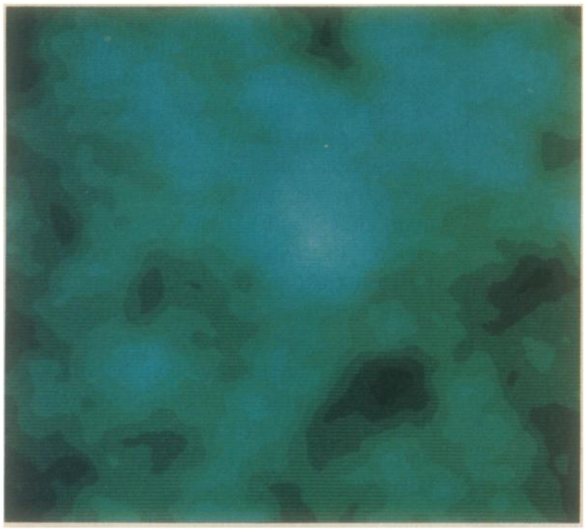

Fig. 2. The distribution of galaxies observed on a limiting exposure plate taken with the 4-meter telescope at Kitt Peak National Observatory. The cluster of galaxies clearly visible spans several million light-years in radius and is approximately 2×10^9 light-years away. [Courtesy of J. A. Tyson]

fines the current age of the Universe of roughly 10^{18} seconds. Comparing this age to the Planck time, we find that the Universe has clearly persisted for a very large number (10^{61}) of fundamental time units ($t_{PL} = 10^{-43}$ second). To do so requires an exquisite tuning of the density and the expansion rate at early times. If we rearrange Eq. 8 into

$$[H(t)]^2 = \frac{8\pi}{3} G\rho(t) - \frac{\epsilon}{a^2(t)}$$

we can consider the relative importance of the mass density and curvature term in determining the expansion rate. Currently ρ is poorly known, but the first term is in the range of a tenth to ten times the second. If we trace both in time $\rho \propto a^{-3}$ in a matter-dominated Universe (or a^{-4} for a radiation-dominated Universe). At a time $t \sim 1$ second, when the light elements were formed, the first term is larger than the second term by some 14 orders of magnitude. At a temperature of 10^{17} GeV ($t = 10^{-39}$ second), the ratio of the two numbers is 10^{50}. This is seen in Fig. 1 as the rapid divergence of the curves that are nearly identical close to the origin. What we find from this is that the initial data (expansion rate and density), specified at a time 10^{-39} second after the Bang, have to be tuned to an accuracy of one part in 10^{50}, even though ρ is so poorly known today. This tuning would have to be done to 1 part in 10^{14} even if we imagined starting the Universe just before the light elements were formed.

To describe this in more physical terms, we note that the basic time scale for gravitational instability is the free-fall time $[(4\pi/3)G\rho]^{-1/2}$. The expansion of the Universe can be described as a time-reversed free fall, so the age of the Universe is roughly the free-fall time. This implies that the time scale for deviations from a flat Universe to grow and the age of the Universe are the same at any time. Any irregularities at $t = 10^{-39}$ second would have rapidly become black holes or regions of space that expand so rapidly as to now appear empty and devoid of structure. Clearly this does not describe the Universe around us. This difficulty is known as the "flatness problem."

The flatness problem is compounded by what is known as the "horizon problem." In the standard model of a Universe with only matter and radiation, the volume in causal contact is constantly growing. The Universe we see today was 10^{83} separate horizon volumes when the temperature was 10^{17} GeV. Not only do we have to specify expansion parameters to 1 part in 10^{50}, it has to be done 10^{83} times in each of the tiny causally disconnected regions!

Are these really problems? Isn't the most simple specification of a Universe just the statement that it is described by a homogeneous and isotropic flat cosmology? This is a statement whose simplicity and elegance might be overpowering, if only it were true. The problem is that it is only approximately true (approximate simplicity and elegance are not powerful tools), as the Universe is known to exhibit structure on scales up to the largest explored by examining the distributions of galaxies. Figure 2 shows the distribution of galaxies seen on a photographic plate at the limit of the observing power of large telescopes. It is only on still larger scales, probed by fluctuations in the 2.7-K background, that the Universe appears structureless. There are clearly small fluctuations; the puzzle is why they are so very small, yet there.

Both the horizon and flatness problems may be solved by a model first proposed independently by Guth (*9*) and by Brout, Englert, Gunzig, and Spindel (*10*) and termed the inflationary Universe by Guth (*9*). In this model there is an epoch of exponential expansion derived by the energy of the vacuum. During this time an initially small volume is expanded by a large factor. In this scenario, two particles that are initially in causal contact (contained in one another's particle horizon) are carried through event horizons. At the end of the period of exponential expansion, the Universe reverts to one with particle horizons. As time passes, the two particles eventually are again contained in each other's particle horizons. This time they are not strangers, having met before the inflation. If at a temperature of 10^{17} GeV the scale factor of the Universe is increased by 28 orders of magnitude, what was one small horizon volume at the beginning becomes the entire volume of the Universe now seen by us.

Can this scheme be realized from the microphysics of grand unified theories? It is allowed by the theory and is vaguely plausible but certainly not required, and it has some "tuning" problems of its own. In order to understand this microphysics, it is useful to first describe Heisenberg's model of ferromagnetism as an example of the type of symmetry breaking that is essential to theories of unification.

Ferromagnets display a spontaneous magnetic moment, one that persists in the absence of applied fields. This orientational order is promoted by the magnetic and exchange forces between the magnetic moments produced by the electron spins. Although the laws of motion that govern these materials are isotropic, that is, display no preference for direction in space, it is energetically favorable for neighboring spins to be aligned (there is no preference for an absolute direction, only a relative one). At high temperatures, when the thermal energy is large as compared to the interaction energy, no such alignment is seen; the system displays the full symmetry of the Hamiltonian at high temperatures. Inside the ordered material there may be many ferromagnetic domains, wherein the spins are locally aligned but domain walls separate regions that have selected different directions of alignment. This occurs to minimize the value of the energy contained in the magnetic field, while allowing most neighboring spins to be aligned.

The vacuum in grand unified theories has many properties analogous to those of a ferromagnet. The spin orientation of electrons is replaced by an order parameter for the orientation of a set of "Higgs fields," ϕ, entities that lead to the symmetry breaking. Interactions lead to local alignments and a lowering of the total energy by the correlations. Domains form, but the defects between them are not walls but point magnetic monopoles. As causality constrains each domain to be no larger than a horizon volume, if the Universe doesn't inflate there would be as many monopoles as there are protons. Detection of a monopole would be a sensational discovery, as it would provide direct evidence that the Universe was once at a temperature of 10^{17} GeV and would also provide some information on the correlations in the Higgs fields and the amount of inflation. Completing the analogy, at high temperatures the symmetry is restored and the expectation value of the field reverts to zero.

The energy $V_T(\phi)$, as a function of $<\phi>$ is shown for several temperatures in Fig. 3. Similar diagrams abound in studies of solids. From the standpoint of describing the dynamics of the fields only relative energies are important in Fig. 3, but the absolute value is important cosmologically as it contributes to the mass-energy density. Where is one to place the zero point? Since we know that the cosmological constant is now small, we will assume that Fig. 3 should have been drawn such that our T-nearly-zero state is the one at zero energy. We have to change Fig. 3, moving all the curves up by the quantity β shown there. This implies that at an early time, when the temperature was large, the vacuum had an enormous energy density which might trigger the inflation.

This scheme is an exciting one but not without its share of problems. The first one is the cavalier setting of the zero point which transpired in the last paragraph. The change in the zero point from that shown in Fig. 3 relative to what we require from current observations is in the ratio of 10^{110}, a problem of "tuning" that is orders of magnitude larger than any inherent in a Universe containing only matter and radiation. Even when one considers only the symmetry breaking in the electroweak interactions, the tuning of the final vacuum energy density must be "done by hand" to 1 part in 10^{60}. There is a further problem in getting the exponential phase to last long enough to change the scale factor by 10^{28}. As the Universe expands, it "supercools" to very low temperatures. Further parameter tunings are required to keep the Universe from rolling off the high-temperature state too soon or having various regions change at different times, which would yield "bubbles" of

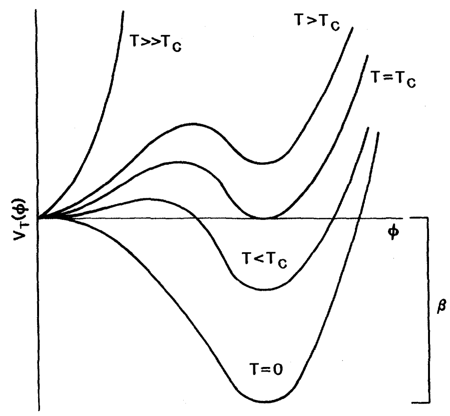

Fig. 3. The free energy density as a function of the order parameter in the Higgs field (ϕ) at various temperatures. The critical temperature, T_c, is the point where the two possible vacuum states are degenerate in energy; T_c is approximately 10^{15} GeV in current grand unified theories.

the standard model Universe inside an exponentially expanding vacuum.

The solution to some of these problems may be within sight. Gott (*11*) has proposed that, if "bubbles" form slowly, just one of them might contain the entire observable Universe. Linde (*12*) and Albrecht and Steinhardt (*13*) have shown that the Coleman-Weinberg (*14*) potential introduces a small potential barrier near the origin with a steep drop to a relative minimum at the origin of the $V_{T=0}(\phi)$ potential. This turns the obviously unstable state at $\phi = 0$, $T = 0$, shown in Fig. 3 into a metastable state and helps prolong the exponential expansion in what is known as the "new inflationary" model. The period of inflation is still too short and the amplitude of the fluctuations is too high (*15*), but this is progress in a difficult problem.

Another Problem: The Matter-Antimatter Asymmetry

There is a second problem inherent in starting up the Universe: where did the matter come from? In our Galaxy, the absence of antimatter in low-energy cosmic rays leads us to conclude that the Galaxy is all matter. At present we have no way to refute the hypothesis that the nearest Galaxy, the Andromeda nebula, is pure antimatter (*16*). To arrange for the degree of homogeneity that must have been present in the early Universe and yet separate matter and antimatter on scales of galaxies or clusters of galaxies seems an impossible task after nearly two decades of effort (*16*). Why then is there a pronounced matter-antimatter asymmetry in the Universe?

In order for an excess of matter to arise, charge-parity (CP) must be an imperfect symmetry (*17*). Somehow matter must be created in preference to antimatter. There is a second requirement that the process occur out of equilibrium. If the system were in equilibrium, CP violation would introduce a reaction channel between matter and antimatter and prevent any imbalance. Generally, it has been assumed that, at extremely high temperatures, gravitational interactions succeed in establishing an equilibrium where the initial baryon number (B, defined as the number of particles minus antiparticles) is zero. At a later time the expansion of the Universe breaks the equilibrium.

Before describing a scenario of net baryon generation, it is useful to enumerate the reasons that have convinced the physics community at large that baryon number is not conserved:

1) Gravity violates baryon number. A black hole made of particles is indistinguishable from one made of antiparticles, even if it evaporates by quantum processes (*18*).

2) Baryon nonconservation arises as a natural consequence of the standard Weinberg-Salam model of the electroweak interaction (*15*).

3) In unifying the strong and electroweak interactions, grand unified theories break the global symmetry of baryon number leaving only local gauge symmetries (*4*).

4) The Universe contains an excess of baryons over antibaryons.

Although items 1 and 2 have not been experimentally verified, they are such natural consequences of well-established theories that there is little doubt as to their existence. These first two violations of baryon number are far too small to account for the matter-antimatter asymmetry, but they serve as encouragement for constructing a grand-unified scheme that does.

The essential features of such a scheme are as follows. Consider the decay of a heavy particle, either a Higgs or gauge boson, X, and its antiparticle $\overline{X}$. The two particles have the following decay modes, branching ratios, and changes in baryon number:

$$X \rightarrow A + B \text{ at rate } r,\ \Delta B = B_1$$
$$X \rightarrow C + D \text{ at rate } 1 - r,\ \Delta B = B_2$$
$$\overline{X} \rightarrow \bar{A} + \bar{B} \text{ at rate } \bar{r},\ \Delta B = -B_1$$
$$\overline{X} \rightarrow \bar{C} + \bar{D} \text{ at rate } 1 - \bar{r},\ \Delta B = -B_2$$

When the Universe is at a temperature much greater than the mass of the X, these processes ensure that the initial baryon number is zero. When the expansion rate becomes more rapid than the decay rate, a baryon number excess is created (*3*). The resultant excess (calibrated to the number of photons, where all the energy from particle and antiparticle annihilations eventually winds up) is

$$\frac{n_b}{n_\gamma} = \begin{pmatrix} \text{fraction of all} \\ \text{particles that} \\ \text{are X's} \end{pmatrix} (r - \bar{r})\,(B_1 - B_2)$$

There are roughly 100 particles in the grand unified schemes, so the first number is $\sim 10^{-2}$. A reasonable value of $(B_1 - B_2)$ is 1. Since the observed value of n_b/n_γ is $\sim 10^{-9}$, the theory must place $(r - \bar{r})$ in the range of 10^{-7}. Numerous schemes have been devised to do so.

Baryosynthesis is quite different from nucleosynthesis in that laboratory determinations of the cross sections will not be rapidly forthcoming. There are other ways to corroborate the result. The first one is to find proton decay. This is currently a topic of intense experimental effort, with an upper limit of 10^{31} years to the decay into what is believed to be the principal mode (*19*). These results will help constrain the range of possible theories (the simplest model is already ruled out), but proton decay experiments will not directly yield the reaction rates needed in baryosynthesis.

The measurement of the neutron electric dipole moment is an experiment that has direct bearing on the problem of baryosynthesis. It is a cherished belief that the combination CPT (T is time) is an exact symmetry (there is currently no way to build a local field theory without it), so a violation of CP must be accompanied by a matching violation of T. The neutron electric dipole moment is a sensitive probe of T violation which complements the CP violation necessitated by baryosynthesis. The net baryon excess observed implies a lower bound on the neutron electric dipole moment of 3×10^{-28} e − cm, whereas the current limit is 1.6×10^{-24} e − cm. It will be a decade or two before the electric dipole moment experiments are in this range (*20*).

Looking Ahead

This survey has considered processes at energies of 10^{17} GeV, some 13 orders of magnitude above current accelerator experiments and only two orders of magnitude below the realm of quantum gravity. Many surprises are certain to lurk in unknown corridors. Where does one start to look? If the answers were not many and varied, we would be facing the end of a science; but of course they are. Figure 4 shows some of the tools that are

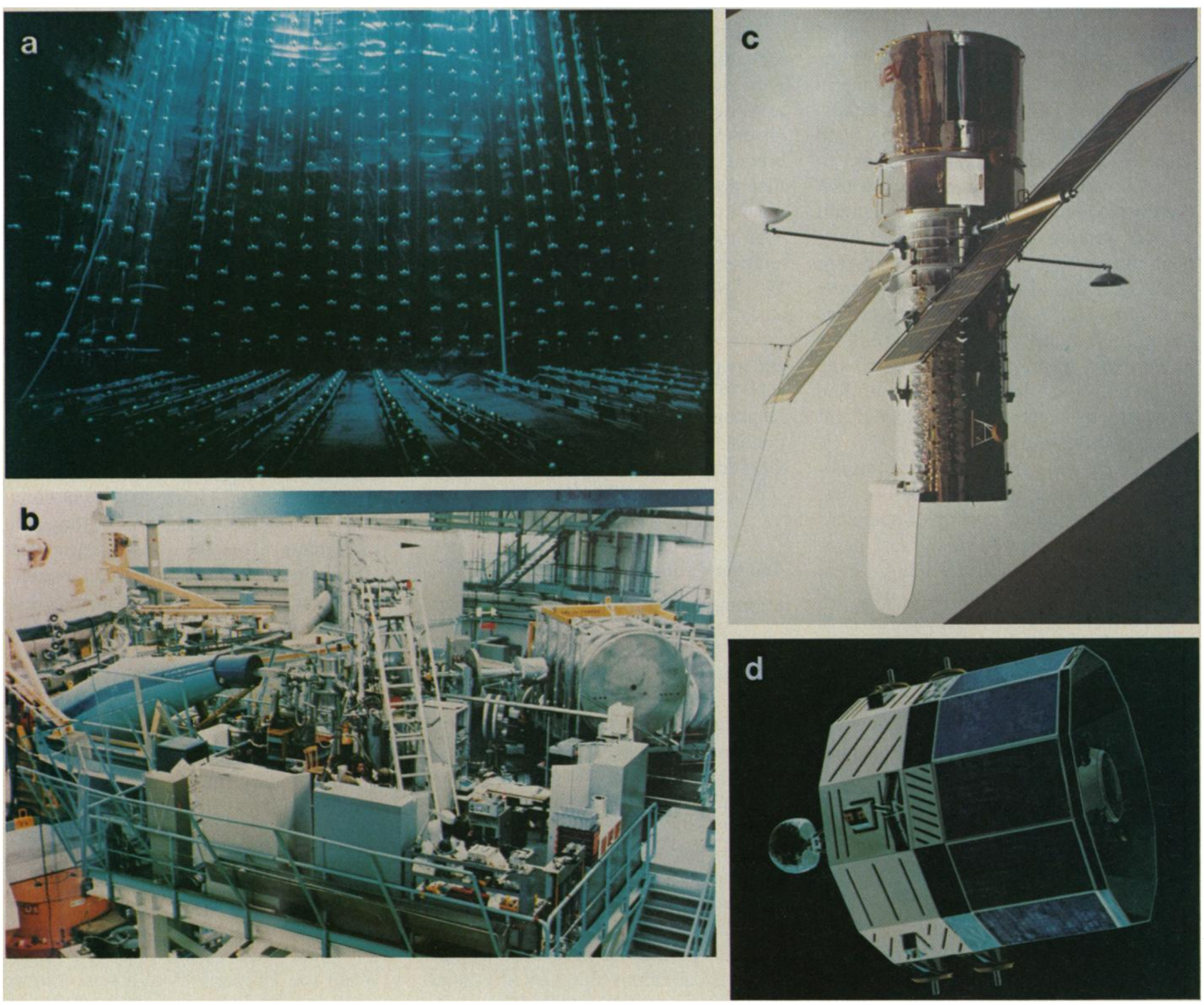

Fig. 4. Some of the windows on cosmology. (a) An underwater photograph of the Irvine-Michigan-Brookhaven detector "pool," which is part of the proton decay apparatus in the Morton Salt Mine outside of Cleveland [courtesy of K. Luttrell]. (b) The neutron electric dipole moment apparatus at Harvard [courtesy of N. Ramsey]. (c) A model of the Space Telescope. (d) An artist's rendering of the Cosmic Background Explorer satellite, which will probe the spectral characteristics and large-scale homogeneity of the 2.7-K background radiation.

windows on cosmological events and processes.

There is a clear experimental route toward elucidating the processes of baryosynthesis, as was described in the last section. In the next decade, experiments on proton decay and the neutron electric dipole moment (see Fig. 4, a and b) will either provide confirmation of the scheme sketched here or insights into new theories.

The horizon and flatness problems are certainly not yet solved. Examining the large-scale structure of the Universe by using large telescopes on the ground, the Space Telescope, and the Cosmic Background Explorer satellite (see Fig. 4, c and d) will help to elucidate the origin and evolution of the deviations from a homogeneous, flat cosmology. It will also shed light on the current value of the expansion rate, help to determine whether the Universe is open or closed, and place better limits on the current energy density of the vacuum.

An important problem in need of solution by theorists is the current tiny value of the vacuum energy density (the cosmological constant) relative to that expected in our current low-temperature vacuum state. The solution may involve a new, fundamental principle that is certain to have a broader impact than the resolution of the cosmological problems discussed here.

References and Notes

1. A. A. Penzias, *Rev. Mod. Phys.* **51**, 422 (1979).
2. P. J. E. Peebles, *Physical Cosmology* (Princeton Univ. Press, Princeton, N.J., 1971).
3. S. Weinberg, *Rev. Mod. Phys.* **52**, 515 (1980).
4. S. L. Glashow, *ibid.*, p. 539.
5. E. C. G. Stuckelberg, *Helv. Phys. Acta* **11**, 299 (1983); E. P. Wigner, *Proc. Am. Philos. Soc.* **93**, 521 (1949).
6. P. W. Anderson, *Science* **177**, 393 (1972).
7. C. Callan, R. H. Dicke, P. J. E. Peebles, *Am. J. Phys.* **33**, 105 (1965).
8. P. J. E. Peebles and R. H. Dicke, in *General Relativity: An Einstein Centenary Survey*, S. W. Hawking and W. Israel, Eds. (Cambridge Univ. Press, Cambridge, 1979), pp. 504–517.
9. A. H. Guth, *Phys. Rev. D* **23**, 347 (1981).
10. R. Brout, F. Englert, E. Gunzig, *Ann. Phys. (N.Y.)* **115**, 78 (1978); R. Brout, F. Englert, P. Spindel, *Phys. Rev. Lett.* **43**, 417 (1979).
11. J. R. Gott, *Nature (London)* **295**, 304 (1982).
12. A. D. Linde, *Phys. Lett. B* **108**, 389 (1982).
13. A. Albrecht and P. J. Steinhardt, *Phys. Rev. Lett.* **48**, 1220 (1982).
14. S. Coleman and E. J. Weinberg, *Phys. Rev. D* **7**, 1888 (1973).
15. G. 'tHooft, *Phys. Rev. Lett.* **37**, 8 (1976).
16. R. Omnès, *ibid.* **23**, 38 (1969); G. Steigman, *Annu. Rev. Astron. Astrophys.* **14**, 339 (1976).
17. A. D. Sakharov, *Pis'ma Zh. Eksp. Teor. Fiz.* **5**, 32 (1967).
18. C. W. Misner, K. S. Thorne, J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
19. R. Bionta *et al.*, *Phys. Rev. Lett.* **51**, 27 (1983).
20. J. Ellis, M. K. Gaillard, D. V. Nanopoulos, S. Rudaz, *Phys. Lett. B* **99**, 101 (1981).

Endocytosis: Relation to Capping and Cell Locomotion

Mark S. Bretscher

In this article, I draw together several aspects of the endocytic cycle initiated by coated pits in the cell's plasma membrane (*1–4*). I suggest that in motile cells this cycle causes a bulk flow of particular membrane components across the cell's surface. A natural consequence of this flow is that any large aggregate on the cell's surface will be swept to the rear of the cell. This movement of aggregates, which is known as capping, depends on the fluid nature of membranes. Finally, I suggest how the endocytic cycle may be part of the machinery a cell uses in locomotion, and what this implies for the difference between motile and nonmotile cells.

The Endocytic Cycle

The function of coated pits in adsorptive endocytosis was first deduced from studies on developing mosquito oocytes (*5*). In thin sections viewed by electron microscopy, coated pits appear as depressions or invaginations in the cell's surface. A coated pit is distinguished from other depressions by its characteristic, thick (about 200 Å) coat at its cytoplasmic surface. Coated pits on oocytes usually bear many yolk particles attached to their outer surfaces; in the same sections coated vesicles that bear yolk particles are seen inside the cell. Therefore, coated pits presumably bud into a cell to yield coated vesicles, and their function in oocytes is to bring yolk particles into the cell to form the yolk (*5*).

Coated pits have now been observed on the surface of almost all cells, except erythrocytes (*1, 6*). These pits serve to bring specific macromolecules into the cell, but which macromolecule is determined by the specific receptors present in the coated pit; and the specific receptors, in turn, depend on the cell type. Thus, oocytes have receptors for yolk proteins, which bring the yolk into those cells. Infant rat gut epithelial cells have immunoglobulin G receptors that bind antibodies from the mother's milk; the antibodies are internalized by coated pits and eventually they are transferred across the epithelium to the infant's blood circulation, a mechanism that provides the infant rat with passive immunity during its early life (*7*).

The properties of coated pits and their associated receptors have been studied most extensively in cells grown in culture. The best understood receptor is that for low density lipoprotein (LDL) (*8*). The LDL is a large particle originating in the liver that circulates in the plasma; it is taken up by many cells and degraded, and the cholesterol so liberated serves as the main source of cellular cholesterol. When fibroblasts in culture are starved of LDL, the number of LDL receptors is increased to about 10,000 per cell.

The distribution of these receptors on the cell surface has been determined by adding ferritin-conjugated LDL to human fibroblasts at 4°C (at this temperature the cells bind LDL, but do not undergo endocytosis). When thin sections of such labeled cells are examined, about two-thirds of the LDL-ferritin is found in coated pits; nevertheless, coated pits account for only about 2 percent of the cell's surface (*9*). If cells labeled with LDL-ferritin at 4°C are warmed to 37°C for a few minutes, much of the ferritin is found in coated vesicles or in smooth vesicles inside the cell. At later times (30 minutes at 37°C) the LDL is found in lysozomes. The LDL receptor is returned to the cell surface to be reutilized for many further cycles since the uptake of LDL is unaffected by the presence of inhibitors of protein synthesis over a period of several hours (*2*). Such experiments (*5, 8*) indicate that the

The author is a member of the staff of the Medical Research Council's Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, England.