Vol. B13 (1982) ACTA PHYSICA POLONICA No 10-11

PHASE TRANSITIONS IN THE EARLY UNIVERSE*

By T. W. B. Kibble

Blackett Laboratory, Imperial College, London**

(Received March 24, 1982)

I briefly review the ideas of grand unification and the hot big bang which together suggest that the universe underwent a series of phase transitions in its early history. I then discuss the nature of these phase transitions and the structures produced at them — domain walls, which can normally be excluded, strings, which might form the basis of a theory of galaxy formation, and monopoles, for which the problem is to prevent an over-abundance.

PACS numbers: 98.80.Bp

1. Introduction

The most exciting developments in physics often occur where several very different disciplines come together. The subject of my lectures is an excellent example, for it involves cosmology and astrophysics, quantum field theory and particle physics, thermodynamics and statistical mechanics, and the physics of condensed matter.

There are two basic ingredients — the idea of grand unification and the hot big bang theory of the early universe. I shall review each of these briefly and then show how when combined they lead to the notion that the universe went through a series of phase transitions early in its history. Then I shall explain the various types of structures that can appear at the phase transitions — domain walls, strings and monopoles. Each of these poses some intriguing questions, which I shall discuss in turn. Domain walls can be excluded on the grounds that their gravitational effect would lead to unacceptably large anisotropy of the microwave background. This imposes interesting, and powerful, restrictions on the acceptable grand unified theories (or GUTs for short). Monopoles appear naturally in almost all GUTs, and are expected to be created in large numbers. The difficulty is to find a mechanism that will reduce their number to an acceptable level. Finally, strings, though rather more elusive, do appear in some GUTs. There has recently been a very interesting attempt to use them to construct a theory of galaxy formation which, though speculative, seems to be viable.

---

* Presented at the VI Autumn School of Theoretical Physics, Szczyrk, Poland, September 21–29, 1981, organized by the Silesian University, Katowice.

** Address: Blackett Laboratory, Prince Consort Road, London SW7 2BZ, England.

1.1. Grand unification

The unification of weak and electromagnetic interactions in an electroweak gauge theory based on the gauge group $SU(2) \times U(1)$ is now generally accepted as a fact. Certainly the Nobel Committee believes it! We also have in quantum chromodynamics (QCD) a very promising gauge theory of strong interactions, based on the colour group SU(3), which has at the very least had some remarkable qualitative successes. (For a review, see Marciano and Pagels 1978). It is thus quite natural to ask whether one can go on to combine the electroweak and strong interactions in a grand unified theory (GUT) based on a simple or semisimple group G that contains $SU(3) \times SU(2) \times U(1)$ as a subgroup. The smallest possible group is SU(5) (Georgi and Glashow 1974) though SO(10) is another popular choice (Fritzsch and Minkowski 1975, Georgi 1975, Chanowitz, Ellis and Gaillard 1977).

Non-Abelian gauge theories (unless there are too many other fields present) have the special feature of asymptotic freedom — their energy-dependent running coupling constants become smaller at higher energies, in a predictable fashion governed by the renormalization group. The SU(3) coupling decreases faster than that of SU(2) and even more so than that of U(1). Thus all three couplings can become equal at a grand unification mass which can be estimated to be around 10^{15} GeV. By the standards of particle physics this is enormous. Indeed it is only four orders of magnitude below the Planck mass

$$m_{\mathrm{P}} = G^{-1/2} = 1.2 \times 10^{19}\ \mathrm{GeV}$$

at which even gravity becomes strong. The unified value of the coupling at the unification mass is approximately

$$\alpha = g^2/4\pi \approx 0.025.$$

The unified theory is broken spontaneously in at least two stages, for example from SU(5) to $SU(3) \times SU(2) \times U(1)$ at a mass scale of 10^{15} GeV and then to $SU(3) \times U(1)$ at about 100 GeV. Between the two is the uninteresting "desert" though it is possible in more complicated theories to make the desert bloom with a whole succession of further stages of symmetry breaking.

2. The early universe

There are two crucial pieces of observational evidence for the hot big bang. The first is the cosmic red-shift, the observation that the spectral lines of distant galaxies are shifted to the red. If this is interpreted as a Doppler shift they are receding from us with velocity $v = Hr$, where r is distance and H is the Hubble constant, $H \sim 100\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$. Traced back in time, the galaxies must have been close together at a time in the past of order $H^{-1} \sim 10^{10}$ years ago.

The second observation is of the cosmic microwave background radiation discovered by Penzias and Wilson (1965) with approximately a blackbody spectrum corresponding to a temperature of 2.7 K. As the universe expanded, this radiation must have cooled

adiabatically, so at early times the universe was much hotter than it now is. The universe is now transparent but when it was much denser and hotter it was not. The radiation we see is the redshifted remnant of thermal radiation emitted when the temperature was around 4000 K. At even earlier times the temperature must have been higher still, hot enough to induce elementary particle reactions.

If the universe is assumed isotropic and homogeneous its geometry must be described by the Robertson-Walker metric

$$ds^2 = dt^2 - R(t)^2 d\sigma^2$$

where $d\sigma^2$ is the metric of a three-space of uniform curvature K. The rate of change of the scale factor $R(t)$ is given by Einstein's equation (see, for example Weinberg 1972)

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3}\varrho - \frac{K}{R^2} + \Lambda,$$

where ϱ is the density and Λ the cosmological constant which is observationally small and which I shall take to be zero for simplicity. This equation can be interpreted as expressing conservation of energy of a test particle on the surface of a comoving volume:

$$\tfrac{1}{2}\dot{R}^2 - \frac{G\varrho(\frac{4}{3}\pi R^3)}{R} = -\tfrac{1}{2}K.$$

Thus if $K > 0$, which means that the universe is closed, then the motion is bound: the universe will reach a maximum size and collapse back to a final singularity. On the other hand an open universe ($K \leqslant 0$) corresponds to unbound motion and will continue expanding forever. At early times, the value of K is unimportant, so we may for simplicity take $K = 0$, i.e. choose a Euclidean universe.

During at least part of its early history, when the temperature is large compared to all relevant masses, the matter in the universe may be treated as an ideal relativistic gas of massless particles. Its density is then

$$\varrho = (\pi^2/30)N_* T^4,$$

where N_* is the effective total number of particle species, counting the two helicity states of massless bosons separately (and with an extra factor of 7/8 for fermions). In the simplest SU(5) grand unified theory, $N_* = 160.75$. The expansion leads to an adiabatic cooling, with $T \propto R^{-1}$. Substituting in Einstein's equation we easily find that $R \propto t^{1/2}$. The relationship between temperature and time may be written

$$T^2 t = 0.3\, N_*^{-1/2} m_{\mathrm{P}}.$$

2.1. The initial state

At first sight it may seem bizarre to suppose that matter at a density far in excess of nuclear densities can be treated as an ideal gas. However, it is not really so strange. In the first place, because of asymptotic freedom, the matter is indeed weakly interacting at the relevant energies. Moreover, one can show that the mean free path is large compared

to the interparticle spacing. The number density of particles of each boson species is

$$n_s = \pi^{-1/2}\zeta(3)T^3,$$

while for fermions there is an extra factor 3/4. Thus there is approximately one particle of each species in a thermal volume of radius $1/T$. All particle cross-sections should be of order $\sigma \sim \alpha^2/T^2$, so the mean free path λ is given by

$$\lambda^{-1} = n\sigma \sim N_* T^3(\alpha^2/T^2) = N_*\alpha^2 T.$$

Since $N_*\alpha^2 \sim 1/15$, we see that λ is large compared to the thermal wavelength $1/T$ and *a fortiori* to the mean interparticle spacing $(N_*^{1/3}T)^{-1}$. This serves to validate the ideal-gas approximation.

Unfortunately there is another, less satisfactory comparison we can make. From the formulae above we see that the ratio of the mean free path to the expansion time is

$$\lambda/t \sim T/N_*^{1/2}\alpha^2 m_P,$$

which exceeds unity so long as T is greater than about 10^{17} GeV. In other words, there is no time for the system to reach equilibrium.

2.2. The homogeneity problem

This brings me to one of the oustanding puzzles about the early universe — the homogeneity problem or horizon problem (Guth 1981). Why should the universe be homogeneous out to large distances over which there can have been no prior causal contact? In particular, why should it be in thermal equilibrium when there has been no time to establish equilibrium? A randomly chosen initial state would be far more inhomogeneous. Though we talk of the "hot big bang", our universe was started apparently in a state of remarkably small entropy. (This circumstance seems to be crucial to our perception of an "arrow of time".)

The coordinate distance to the causal horizon is easily seen to be

$$\int_0^t \frac{dt'}{R(t')} \propto t^{1/2}$$

in the standard scenario. The existence of a finite causal horizon is inevitable unless this integral diverges, i.e. unless $R(t)$ varies as t or faster, which would mean that instead of $\varrho \propto T^4$ we would need $\varrho \propto T^\alpha$, $\alpha \leqslant 2$. (An intriguing alternative has been suggested by Zee (1979) in a theory of spontaneously generated gravity which yields $\varrho \propto T^4$ but $G \propto T^{-2}$.) If the integral did diverge, then all parts of the universe would be very briefly in causal contact. But even then, there would still be no time to establish equilibrium. There would also be no adequate mechanism for dissipating initial anisotropy.

It seems therefore that we must look elsewhere for a solution of the homogeneity problem. The most often canvassed solution is to invoke quantum gravity, which is undoubtedly crucial in the period before the Planck time. However until we have a proper theory of quantum gravity this remains pure speculation.

2.3. The flatness problem

There is another similar problem concerning the initial state which has been highlighted particularly by Guth (1981). Observationally, our universe is nearly flat. The condition for the curvature K to vanish is

$$\varrho = \varrho_c = \frac{3}{8\pi G}\left(\frac{\dot{R}}{R}\right)^2.$$

The observational limits on ϱ are

$$0.02 \lesssim \varrho/\varrho_c \lesssim 2.$$

If we follow the evolution back in time we find in the radiation-dominated era that while $\varrho \propto R^{-4}$ the difference $\varrho - \varrho_c$ varies much more slowly: $\varrho - \varrho_c \propto KR^{-2}$. Thus at the Planck time we must have had

$$|\varrho - \varrho_c|/\varrho_c \lesssim 10^{-58}.$$

This is an extraordinary degree of fine tuning. What is the explanation?

One speculative idea involves the so-called "anthropic principle". If ϱ had been much larger than ϱ_c, the universe would have collapsed back to a second singularity in a very short time; if it had been much smaller, the universe would have expanded exponentially and condensation of galaxies and stars would never have occurred. In neither case would we be here discussing the question. So on this principle, the universe is nearly flat because only in such a universe could we exist. Perhaps there are many other universes but only in a few lucky ones are there galaxies, stars, and people!

There is a very similar problem attached to the cosmological constant Λ. Observationally, it is tiny in natural units:

$$|\Lambda|/m_P^2 \lesssim 10^{-122}.$$

Again, however, if it were much larger we would not be here.

3. Spontaneously broken gauge theories

Next let me review some basic facts about spontaneous symmetry breaking in gauge theories. I shall consider a gauge theory described by the Lagrangian density

$$\mathscr{L} = \tfrac{1}{2}\,\mathrm{tr}\, F_{\mu\nu}F^{\mu\nu} + \tfrac{1}{2}\, D_\mu\phi \cdot D^\mu\phi - U(\phi)$$

$$+\bar{\psi} i\gamma^\mu D_\mu \psi - \bar{\psi}\Gamma\psi \cdot \phi + \text{gauge-fixing and Fadeev-Popov ghost terms.}$$

Here ϕ is a scalar Higgs field belonging to some representation (not necessarily irreducible) of the gauge group G, with real anti-symmetric generators T_a normalized so that

$$\mathrm{tr}\, T_a T_b = -\tfrac{1}{2}\,\delta_{ab}.$$

The covariant derivative of ϕ is

$$D_\mu\phi = \partial_\mu\phi - gA_\mu\phi, \qquad A_\mu = A_{a\mu}T_a,$$

where g is the gauge coupling constant. The gauge field is

$$F_{\mu\nu} = \partial_\nu A_\mu - \partial_\mu A_\nu + g[A_\mu, A_\nu].$$

$U(\phi)$ is a G-invariant polynomial of degree 4. This restriction is imposed to ensure renormalizability.

The spinor field ψ also belongs to some (generally different) representation, and its covariant derivative is defined similarly. The matrix Γ symbolizes the Yukawa interaction terms. However the spinor field does not play a very important role in the problems I shall discuss, and I shall ignore it. One must of course take account of it in any detailed calculation since, for example, it affects the renormalization group β function, making the coupling change more slowly than it would in a pure gauge theory.

As a convenient example that may serve to illustrate the discussion let us take the gauge group G to be SO(N) and ϕ to belong to the fundamental N-dimensional vector representation. The polynomial U is

$$U = \tfrac{1}{8} h^2(\phi \cdot \phi - \eta^2)^2$$

where η is a constant defining the scale of ϕ and h^2 is a coupling constant. I have chosen the coefficient of the quadratic term in U to be negative, so that U has a maximum at $\phi = 0$. Its minimum lies on the surface M which here is the $(N-1)$-sphere. Thus in the classical (tree) approximation, ϕ acquires a vacuum expectation value lying on M, and the symmetry is broken from SO(N) to SO($N-1$).

In the general case of which this is a particular example, if ϕ is a point on the surface M of minima of U and H is the corresponding isotropy subgroup of G, namely

$$H = \{g \in G : g\phi = \phi\},$$

then M may be identified with the quotient space G/H, whose elements are the left cosets of H in G. Each point on M defines one of the family of degenerate vacuum states. The symmetry group G is spontaneously broken, leaving H as the subgroup of unbroken symmetries.

The case $N = 2$ of the SO(N) model is just the Landau-Ginzburg model of superconductivity. In its relativistic context it is the Higgs model. Note that for any $N > 1$ there are as in superconductivity two characteristic length scales, the correlation length ξ and the penetration depth λ given by

$$\xi^{-1} = m_s = h\eta, \lambda^{-1} = m_v = g\eta.$$

Here m_s is the mass of the Higgs scalar, corresponding to the radial component of ϕ, while m_v is the mass of each of the $N-1$ vector bosons corresponding to broken generators of G. (Those corresponding to generators of H remain massless.)

3.1. Symmetry restoration at finite temperature

At finite temperature the equilibrium state is determined not by the potential $U(\phi)$ but by the *effective* potential $V(\phi)$, which is the minimum free energy density in states with a given expectation value of ϕ. It is calculated from the sum of all one-particle-irreducible

Feynman diagrams with zero-momentum external ϕ lines only (because of the assumed translational invariance of the vacuum). We must of course use the finite-temperature Feynman rules with a periodic imaginary-time variable of period $1/T$.

At large T the leading correction terms coming from the one-loop contribution to $V(\phi)$ yield

$$V(\phi) = U(\phi) - \frac{\pi^2}{90} N_* T^4 + \tfrac{1}{24} M_*^2(\phi) T^2 + O(T),$$

where as before N_* is the effective number of helicity states of "light" particles (those with masses $\ll T$), and M_*^2, which is ϕ dependent, is the sum of the squared masses of these helicity states (counting fermions with a factor $\frac{1}{2}$). For example in the simple SO(N) model I discussed earlier, with ϕ belonging to the vector representation, one finds $N_* = N^2$ and

$$M_*^2(\phi) = \tfrac{1}{2} Nh^2(\phi^2 - \eta^2) + h^2\phi^2 + 3(N-1)g^2\phi^2.$$

Thus we obtain

$$V(\phi) = \tfrac{1}{8} h^2(\phi^2)^2 + \tfrac{1}{2}(-\tfrac{1}{2} h^2\eta^2 + AT^2)\phi^2 + \phi\text{-independent terms}$$

where

$$A = 3(N-1)g^2 + \tfrac{1}{2}(N+2)h^2.$$

Clearly there is a transition temperature T_c given by

$$T_c^2 = h^2\eta^2/2A \sim \eta^2.$$

For $T > T_c$, the coefficient of ϕ^2 in V is positive, and so $\langle\phi\rangle = 0$. The system is then in a symmetric phase. When T falls below T_c, the coefficient of ϕ^2 becomes negative, and ϕ acquires an expectation value given by:

$$\langle\phi\rangle^2 = \eta^2(1 - T^2/T_c^2).$$

Thus the system is in an ordered phase, with $\langle\phi\rangle$ playing the role of the order parameter. Note that the correlation length ξ is given by

$$\xi = \xi_0(1 - T^2/T_c^2)^{-1/2}$$

where $\xi_0 = 1/h\eta$. It is of course infinite at $T = T_c$.

4. History of the phase transitions in the universe

If these ideas are broadly correct then we may conclude that the universe underwent at least three phase transitions during the course of its early history.

The first of these is the grand unification transition which occurs more or less at the temperature corresponding to the grand unification energy scale, 10^{15} GeV. This is only four orders of magnitude below the Planck mass. This transition happens when the universe is only 10^{-37} s old, at a redshift Z of 10^{28} (i.e. since then, the radius of a comoving volume has increased by a factor of 10^{28}). Before it, the universe was in a state with the full sym-

metry of the grand unified group, for example SU(5). After it, the symmetry is broken. If it breaks directly to $SU(3) \times SU(2) \times U(1)$ the next transition does not occur until the temperature has fallen to 100 GeV or so, after about 10^{-11}s (with $Z \sim 10^{15}$). This is the electroweak transition at which the $SU(2) \times U(1)$ part of the symmetry breaks to the U(1) of electromagnetism. In this case there is a huge uninteresting "desert" between 10^{15} and 10^2 GeV. However it is not hard to devise ways of making the desert bloom with multiple transitions if one wishes to do so.

The last phase transition is connected with quark confinement, the transition from a quark and gluon soup to a gas of identifiable hadrons. It presumably occurs at about 1 microsecond, when the temperature is a few hundred MeV. It is associated with the colour SU(3) group but is not a symmetry-breaking transformation. I shall not consider it further.

From that point on the history of the universe is rather less speculative and more conventional. As the temperature falls the various massive particle species annihilate and disappear, first the baryons, then the mesons and finally the leptons, leaving only the small excess demanded by nonzero average values of the baryon and lepton numbers. Once the temperature has fallen well below 1 MeV, when the universe is a few seconds old, it contains essentially only the stable or nearly stable particles — a plasma of electrons and protons together with photons and neutrinos. The remaining neutrons either decay or are incorporated in deuterons or helium nuclei. This plasma era lasts for about 10^4 to 10^5 years, until the decoupling time, at $T \sim 4000$ K, when the electrons and protons combine to form atomic hydrogen, and the radiation effectively decouples from the matter.

4.1. Structures formed at a phase transition

Let me now turn to the question of what happens when the universe goes through a phase transition (see also Kibble 1980). Above the transition the order parameter $\langle\phi\rangle$ vanishes. When T falls below T_c, $\langle\phi\rangle$ becomes nonzero and lies on the minimum surface M — but where on M?

The situation here is very similar to that of a ferromagnet cooled through its Curie point. It acquires a spontaneous magnetization, but the direction of magnetization is not entirely predictable and is determined in practice by stray external fields or random initial fluctuations. In the early universe too, the choice of one point rather than another on the minimum surface M is random. Moreover there is no reason why $\langle\phi\rangle$ should make the same choice everywhere. Indeed, it is hard to see how there could be any correlation between the choices in far separated regions, certainly when they are beyond the causal horizon.

In these circumstances there may be trapped defects, regions in which $\langle\phi\rangle$ is prevented from taking on a value on M by a topological obstruction. The various possibilities can be well illustrated by considering the SO(N) model for different values of N.

Let me take first the case $N = 1$, which is not a gauge theory at all, but simply a model of a single scalar field ϕ with a ϕ^4 interaction and a negative mass term. Here at low temperature there are two minima, close to the points $\phi = \pm\eta$. Thus we may expect a domain structure to form, with some parts of the universe choosing $\langle\phi\rangle = \eta$ and others $\langle\phi\rangle = -\eta$ Where the two meet there must be a domain wall. As we go through the wall from one

domain to the other, $\langle\phi\rangle$ must vary from η to $-\eta$, passing through 0 on the way. Thus the wall represents a local concentration of energy. No smooth change can eliminate it, except by shrinking one of the domains to the vanishing point.

Next, suppose that $N = 2$. Then M is a circle, and we can have string singularities, as in a superconductor or superfluid. As we go around the string, the angle specifying the point on M changes by 2π (or a multiple thereof). Once again, $\langle\phi\rangle$ must pass through zero somewhere inside such a loop, in the core of the string.

Finally, if $N = 3$ then M is a sphere and we can have point-like monopole singularities such as the famous monopole solution of 'tHooft (1974) and Polyakov (1974). Here $\langle\phi\rangle$ has the hedgehog configuration, pointing radially outwards all around the monopole. In the centre it must pass through zero.

4.2. Singularities and homotopy groups

It is easy to find general conditions for the existence of such singularities, in terms of the homotopy groups of the manifold M of degenerate vacuum states.

For example, the existence of string singularities is governed by the first homotopy group $\pi_1(M)$. The elements of this group are homotopy classes of closed curves starting and finishing at some designated point ϕ_0 in M. Two curves belong to the same class if one can be smoothly deformed into the other. The identity element of $\pi_1(M)$ consists of curves that are homotopically trivial, i.e. that can be smoothly shrunk to a point. Group multiplication is defined in an obvious way: if c_1 and c_2 are closed curves in M starting and finishing at ϕ_0, then $c_2 \cdot c_1$ is the closed curve obtained by following first c_1 and then c_2. For instance the homotopy classes of a circle S^1 are labelled by integers specifying the number of times the curve winds around the circle, so $\pi_1(S^1) = Z$. Similarly the homotopy classes of a torus are labelled by 2 integers: $\pi_1(S^1 \times S^1) = Z \times Z$.

What is required for the existence of strings is that the first homotopy group be non-trivial, $\pi_1(M) \neq 1$. (I prefer to use the notation 1 rather than 0 to symbolize a group of just one element.)

In exactly the same way, the condition for monopoles to exist is that the second homotopy group be non-trivial, $\pi_2(M) \neq 1$. The elements of this group are homotopy classes of closed surfaces in M, maps of the 2-sphere S^2 into M such that a designated point $x_0 \in S^2$ is mapped into $\phi_0 \in M$.

For domain walls what is relevant is $\pi_0(M)$ which in general is not a group but simply denotes the number of distinct connected components in M. (It may also be defined in terms of homotopy classes of maps of S^0 into M. Here S^0 is a set of just two points, but since one point is the base point x_0 which must always be mapped into the chosen base point ϕ_0 of M, what varies is the image of just one point. Two maps are homotopic if these images belong to the same connected component.) Domain walls are produced if and only if M is disconnected.

In our case, M may be identified with the quotient space G/H of the gauge group G by its unbroken subgroup H. This allows its homotopy groups to be computed in terms of those of G and H. In particular, let us suppose that G is not only a Lie group — which automatically ensures that $\pi^2(G) = 1$ — but also connected, so that $\pi_0(G) = 1$, and simply

connected, so that $\pi_1(G) = 1$. This last condition may be ensured by replacing the original gauge group by its simply-connected covering group, for example SO(3) by SU(2). When these conditions are satisfied there are standard theorems which show that

$$\pi_1(M) = \pi_0(H), \quad \pi_2(M) = \pi_1(H).$$

Thus the condition for the existence of strings is that H have disconnected pieces. Note that to apply this criterion to the SO(2) model, we have to replace this group by its simply connected covering group, i.e. to take $G = R$. Then $H = Z$, which is obviously disconnected. Indeed $\pi_1(M) = \pi_0(H) = Z$.

Similarly for monopoles to exist we require that H be non-simply-connected. In particular, any U(1) factor in H yields a factor Z in $\pi_2(M)$. Since the last stage of symmetry breaking leaves us with an unbroken U(1) symmetry, that of electromagnetism, we are guaranteed that at some transition, monopoles are produced.

Note that in both these cases the relevant homotopy groups may also contain finite factors such as Z_2.

4.3. Domain walls

Let me now discuss the three types of singularities in more detail, beginning with the case of domain walls.

Clearly a domain wall has a finite energy density. In equilibrium, its thickness is of the same order as the correlation length ξ at the relevant temperature. Thus its mass per unit area, which for these relativistic walls is the same thing as the surface tension σ, is in order of magnitude

$$\sigma \sim \xi \Delta f \sim h\eta^3(1-T^2/T_c^2)^{3/2},$$

where Δf is the height of the potential hump separating the two minima.

This is a huge number. It was pointed out by Zel'dovich, Kobzarev and Okun (1974) that even for domain walls formed at a transition at 100 GeV, the mass of a single wall stretched across the universe would vastly exceed that of all other matter and hence introduce an unacceptably large anisotropy in the microwave background radiation, as well as other gravitational effects. For GUT domain walls the situation would be much worse. Indeed it is clear that if such massive walls ever existed in the universe they must have disappeared a very long time ago.

Unfortunately this is not easy to arrange. Suppose we start with a more or less random distribution of the two types of domain. Then for energetic reasons we may expect that locally one type will grow at the expense of the other. Small isolated pockets of one will disappear, and the overall length scale of the domain structure will increase. However if in the initial configuration there was no overall bias as between the two types, then we cannot expect ever to eliminate either type throughout the whole universe. At the most we may expect the size of the typical domain to be comparable with the horizon distance, that is we may expect to see of the order of one domain wall across the visible universe.

The most likely application of this result is to rule out theories of spontaneously broken CP invariance. It is only possible to avoid the disastrous conflict with observation

by introducing some kind of bias. One can for instance have a nearly left-right symmetric theory, in which the approximate symmetry is broken spontaneously, but with a small intrinsic asymmetry that ultimately favours one type of domain over the other.

It has recently been suggested by Kuzmin, Shaposhnikov and Tkachev (1981) that the conflict may be avoided in a theory where the symmetry is broken and then restored at a lower temperature. But of course in such a case the spontaneous symmetry breaking can have nothing to do with the observed low-temperature CP violation, which must again be attributed to a small intrinsic asymmetry.

Another way out might be to arrange that the symmetry is never restored at any temperature, as can be done in suitably complicated models. (I shall have occasion to consider a similar suggestion later in connection with monopoles.) In a sense, however, this does not solve the problem, but merely removes it from the realm of dynamics to that of initial conditions. Even if the symmetry were broken right from the start, we could still ask why the universe chose one domain initially rather than the other, and indeed why the same choice was made everywhere.

It is possible that a resolution of the horizon problems which explains why far separated regions of the universe have the same temperature might also explain why they belong to the same domain, and hence resolve the problem of domain walls. If no such resolution is possible then it remains true that no theory involving purely spontaneous breaking of a discrete symmetry, without intrinsic asymmetry, is viable.

4.4. Strings

Next let me turn to the case of strings. The mass per unit length or tension, μ is given approximately by

$$\mu \sim \xi^2 \Delta f \sim \eta^2(1-T^2/T_c^2).$$

This number is still large. In conventional units the mass of a string produced at the grand unification transition is about 10 tons per fermi! However because strings are only one-dimensional rather than two, they would not if present in small numbers give an excessive contribution to the mass density of the universe.

Consider for example a single string stretched across the visible universe, of length $4t$. Its mass is $M_s = 4t\mu$. Let us compare this with the total mass of the visible universe, assuming that we are in the radiation-dominated era. The density is

$$\varrho \approx 0.03/Gt^2,$$

so the total mass is

$$M = (4\pi/3)\varrho(2t)^3 \sim \varrho t/G.$$

Thus the ratio of the mass of a string (well after the phase transition) to that of the universe is

$$M_s/M \sim 4G\mu \sim 4(\eta/m_P)^2 \sim 10^{-7}.$$

This is a sufficiently small number to cause no difficulties with the observational isotropy. It is interesting to note that it is independent of the epoch t.

It will be important for our subsequent discussions to estimate the typical length scale of the structures that are formed at the phase transition. Consider the situation when the temperature has fallen just below T_c. At this stage the central hump in the effective potential $V(\phi)$, of height Δf, is still quite small. Thus thermal fluctuations back and forth across the hump will be quite common. The temperature at which this ceases to be true is the Ginzburg temperature T_G (Ginzburg 1960). It is determined by the condition that the energy of a fluctuation back to $\phi = 0$ on a length scale of the order of the correlation length should be equal to T:

$$\xi^3 \Delta f \sim T.$$

For the case of weak interaction, T_G lies not far below the critical temperature T_c, in fact (Kibble 1976)

$$T_c - T_G \sim h^2 T_c.$$

The correlation length at this temperature is

$$\xi_G \sim 1/h^2 \eta.$$

Once T has fallen below T_G, fluctuations back to $\phi = 0$ are improbable except on a length scale too short to be of relevance. Thus topological features like strings may be regarded as frozen in. It follows that the initial length scale we may expect is of order ξ_G.

Their tension will tend to make strings straighten and shorten. Their subsequent evolution is quite a complex problem to which I shall return later, in the context of a recently proposed theory of galaxy formation.

5. *Monopoles*

Meanwhile, let me turn to the case of monopoles. As I pointed out, they are ubiquitous in the sense that any GUT, which starts with a simple or semisimple grand unification group G and ends up with an unbroken U(1) subgroup, is bound to produce monopoles at some transition. Moreover they are liable to be produced copiously, but since they carry a conserved charge can only disappear by annihilation which is usually a slow process. There is thus a serious difficulty in avoiding an overabundance of monopoles in the universe now.

Using the same simple argument as before, one would arrive at an estimate for the monopole mass of

$$\xi^3 \Delta f \sim (\eta/h)\,(1 - T^2/T_c^2)^{1/2}.$$

This is at best a lower limit, since it completely ignores the contribution to the energy of the gauge field. In fact a much better estimate of the zero temperature mass, which is a good approximation except for very small values of h/g is (Bogomol'nyi 1976)

$$m_M \simeq 4\pi\eta/g.$$

For monopoles produced at the grand unification transition, this is of order 10^{16} GeV

A simple estimate of the initial density of monopoles is obtained as before from the correlation length at the Ginzburg temperature. We expect the position of $\langle\phi\rangle$ on the surface M of minima of V to vary randomly from point to point, with a correlation length of order ξ_G. This gives

$$n_M \sim p\xi_G^{-3},$$

where p is a numbrical factor of order 0.1 (whose value can be predicted from the structure of the group G).

A convenient parameter to use in discussing the density of monopoles is the ratio $r = n_M/T^3$, because in the absence of annihilation r will be nearly constant during the adiabatic expansion of the universe. (It would be more accurate to use the monopole-to--entropy ratio, but the difference is not siginificant when discrepancies of many orders of magnitude are in question). The estimate made above gives

$$r \sim ph^6,$$

so a reasonable order-of-magnitude guess for the initial value of r might be 10^{-7}.

5.1. Annihilation of monopoles

The subsequent annihilation of monopoles has been discussed by Preskill (1979) (see also Zel'dovich and Khlopov (1979)). The monopoles and antimonopoles must diffuse towards each other and then become bound in Coulomb orbits before annihilating. As the universe expands this becomes an increasingly slow process. Thus Preskill concludes that if r initially exceeds 10^{-10}, it will be reduced to about that level but no further.

This level is far too high to be acceptable. Such heavy monopoles might have escaped direct detection, but would contribute to the overall mass density of the universe. Imposing the requirement that the total mass of the monopoles in the universe now cannot exceed the observational upper limit on the mass of all matter gives $r < 10^{-24}$. One might possibly argue that in the recent history of the universe monopoles have migrated preferentially to special places, such as centres of galaxies, where their annihilation rate is enhanced. Even so, there is another limit that comes from an earlier stage in the history of the universe. The amount of helium synthesis is very sensitive to the expansion rate of the universe and hence to its density at the time. Requiring that monopoles should not excessively increase this density gives the limit $r < 10^{-19}$.

Many authors have suggested ways of avoiding this paradox. One of the most straightforward was proposed by Bais and Rudaz (1980) who argued that the correct expression for the initial monopole density when the monopoles freeze out at the Ginzburg temperature should be the equilibrium density for particles of mass equal to the monopole mass at that temperature, i.e.

$$r = n_M/T_G^3 = [m_M(T_G)/2\pi T_G]^{3/2} \exp[-m_M(T_G)/T_G].$$

They calculate the value of the exponent and show that it has the form

$$m_M(T_G)/T_G = Ch/g = Cm_s/m_v,$$

where C is a model-dependent constant whose value is somewhat uncertain — mainly because of uncertainty in the precise specification of the Ginzburg temperature. However the factors in C that can be calculated are quite large, of order 30. If in addition, we assume that the Higgs coupling constant is rather large, say $h/g \gtrsim 2.5$, then we can reduce r to an acceptable level, 10^{-32} or less.

This explanation would favour a large value of the Higgs boson mass, $m_s/m_v \gtrsim 2.5$, which may be difficult to reconcile with GUT phenomenology. A more serious objection to this proposal has been raised by Linde (1980a) who pointed out that in the circumstances considered, the probability of finding close monopole-antimonopole pairs is much larger than for monopoles alone. In such a configuration there are large cancellations of the long-range gauge-field contributions. Thus there is no exponential suppression of monopole pairs and we should in fact expect a monopole density closer to the original estimate. The argument of Bais and Rudaz may be correct in a modified form, in which m_M is replaced by the mass of a monopole in a plasma of monopole-antimonopole pairs.

Linde (1980b) has also suggested an interesting alternative mechanism for eliminating excess monopoles. This is based on an analysis of the infrared behaviour of non-Abelian gauge theories. Linde suggests that in addition to the "electric" gauge-boson mass of order gT which is calculable in perturbation theory there must be a non-perturbative "magnetic" mass of order g^2T. Linde then argues that below about 10^3 GeV monopoles and antimonopoles would be connected by a relatively light string, or tube of magnetic flux, which will lead to rather rapid annihilation. It is not at all clear however whether this is the correct interpretation of magnetic mass. A different interpretation is suggested by recent lattice calculations (Billoire, Lazarides and Shafi 1981).

5.2. First-order transitions

A considerable number of authors have proposed ways of eliminating monopoles involving first-order phase transitions (Guth and Tye 1980, Einhorn, Stein and Toussaint 1980, Cook and Mahanthappa 1981, Billoire and Tamvakis 1981, Guth and Weinberg 1981, Einhorn and Sato 1981, Kennedy, Lazarides and Shafi 1981). So let me now turn to that subject.

For several reasons the transitions are in fact likely to be first-order rather than second-order as we have hitherto assumed.

First, there are one-loop radiative corrections to the effective potential at zero temperature of the form

$$+(m^2/8\pi^2)^2 \ln (m^2/\mu^2),$$

where μ^2 in some renormalization point. These corrections are significant in special cases — if the Higgs coupling is very weak, $h^2 \sim g^4$ (Linde 1976), or if there are no quadratic terms in $U(\phi)$ (Coleman and Weinberg 1973). The effect is to make the effective potential $V(\phi)$ have a minimum at $\phi = 0$ separated from the absolute minima on M by a hump, thus giving a first-order transition.

In many cases, there are cubic terms in $U(\phi)$ which will have a similar effect. For example, in SU(5) with a Higgs field in the adjoint representation, $U(\phi)$ would be expected

to contain a term proportional to tr (ϕ^3). Even if no cubic terms exist at $T = 0$ they will appear in the temperature-dependent corrections, which involve terms like

$$-(m^3/12\pi)T.$$

In all such cases, we may expect a degree of supercooling. At the critical temperature T_c where the minima at $\phi = 0$ and nonzero ϕ are degenerate nothing will actually happen. Instead, the universe will cool further and the transition will occur discontinuously and in an inhomogeneous way, by the spontaneous formation of bubbles of the new phase which then expand to meet and fill all of space. This may happen either due to thermal fluctuations or by quantum tunnelling.

Let us consider first thermal fluctuations. The energy of a small bubble of the new phase may be written in the thin-wall approximation as

$$E(r) = -(4\pi/3)r^3\Delta f + 4\pi r^2\sigma,$$

where Δf is the difference in free energy density between the symmetric and asymmetric minima, and σ is the surface tension of the wall separating the phases (related to the integral over the hump in V). If the bubble radius once reaches the critical value $r_c = 2\sigma/\Delta f$, it will continue to grow spontaneously. Thus the probability of bubble formation by thermal fluctuations is proportional to $\exp(-E_c/T)$ with $E_c = 16\pi\sigma^3/3\Delta f^2$. This probability peaks at a temperature not far below T_c. If the minimum value of E_c/T is reasonably small, the transition should occur rapidly. In that case we have a weakly first-order transition, not very different in character from a second-order one. In this case the initial monopole density would be of the same order of magnitude as the density of bubbles,

$$n_M \simeq p n_b,$$

which would again be much too large.

On the other hand, if E_c/T is never small, thermal fluctuations are unlikely to trigger the phase transition. In that event we may have extreme supercooling. That is the possibility I examine next.

5.3. Strongly first-order phase transitions

What is relevant now is the probability of bubble formation as a quantum tunnelling process. This probability is of the form Ae^{-B}, where B is the Euclidean action for the "bounce" solution (for details see Coleman 1977, Callam and Coleman 1977). It is possible to estimate the value of B in specific models, (see for example Cook and Mahanthappa 1981, Guth and Weinberg 1981, Einhorn and Sato 1981) but we shall not need to do so here.

If the transition is strongly first-order, and occurs by a rather improbable tunnelling event, then the size of bubbles produced will typically be large, so the monopole density should be small. Moreover when the transition eventually occurs at a temperature $T \ll T_c$, it releases large latent heat, generating entropy which further dilutes the monopoles by increasing the denominator of the monopole-to-entropy ratio. By the same token, of course, it reduces the baryon-to-entropy ratio. Thus one must arrange that the latent

heat released reheats the universe at least to a temperature equal to the Higgs mass m_s, to allow baryon number to be regenerated.

There are grave difficulties with this scenario, arising from the rapid expansion of the universe during supercooling. Once T has fallen well below T_c, the density ϱ is dominated by the vacuum energy

$$\varrho_v \sim h^2\eta^4,$$

which plays the role of a cosmological constant. In particular, the Einstein equation

$$(\dot{R}/R)^2 = 8\pi G\varrho_v/3$$

leads to exponential expansion, as in the de Sitter universe:

$$R \propto e^{Ht}, \quad H \sim h\eta^2/m_P.$$

If the maximum value of e^{-B} is rather small, then because those parts of the universe still in the old phase are expanding exponentially, the growing bubbles of new phase never catch up, and the transition is never complete. The universe would then be very inhomogeneous — impossibly so unless the visible part of it is entirely contained in one original bubble.

The exponential expansion has another effect too. It introduces event horizons. The usual particle horizon or causal horizon of the universe is the surface beyond which no causal contact could have occurred in the past. An event horizon is a surface beyond which no *future* contact can occur. The distance to the event horizon is $H^{-1} \sim m_P/h\eta^2$.

Hut and Klinkhamer (1981) have pointed out that the existence of event horizons implies mode-mixing of positive and negative frequencies for modes with wavelengths greater than H^{-1}, or frequencies $\omega < H$. They argue therefore that once T falls below H, the phase transition will be induced by mode mixing, so that further supercooling is halted. Similar conclusions have been reached on rather different grounds by Abbott (1981) and by Horibe and Hosoya (1981). For the grand unification transition, $H \sim 10^{11}$ GeV, so this argument seems to rule out supercooling by more than about four orders of magnitude — not enough to reduce the monopole density to an acceptable level.

5.4. Multiple transitions

Another possible escape from the problem of overabundant monopoles is to invoke a series of transitions.

One explanation of this type is due to Langacker and Pi (1980). It was noted by Weinberg (1974) that it is sometimes possible to have a phase transition in which the symmetry of the high-temperature phase is lower than that of the low-temperature one. An example of a substance that displays such behaviour is Rochelle salt (sodium potassium tartrate). Langacker and Pi constructed a model involving three Higgs doublets that exhibits the symmetry-breaking pattern

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

The U(1) factor in the symmetry group of the second phase means that monopoles are present, but they become unstable and disappear at the second transition. Of course, there

must be an unbroken U(1) factor in the symmetry group of the final phase. It appears at a third phase transition, which can be made to occur at a relatively very low temperature. Although some monopoles may be produced at that transition there need not be many. Their number may be suppressed by a large exponential factor because the temperature at the transition is far below the monopole mass. (Alternatively if the monopoles were light they would be relatively insignificant for cosmology.)

The general case of two successive transitions

$$G \to H \to H'$$

has been analyzed by Bais (1981). The question is, if monopoles are produced at the first transition, what happens to them at the second? There is in fact a wide range of possibilities. The H' phase may exhibit monopoles of either or both of two distinct kinds — light monopoles which carry only an H' charge and heavier monopoles that carry both H and H' charges.

A monopole in the H phase may do any one of several different things on passing through the second transition. It may remain essentially unaltered, or decay to a lighter H' monopole or to the vacuum. It may acquire a flux tube that joins it to another monopole and pulls them together, leading to annihilation either to the vacuum or to an H' monopole. Finally, there is an intriguing possibility pointed out by Steinhardt (1980) — it may become unstable to radial expansion. Effectively what happens in this last case is that the core of the monopole makes the transition to the H' phase and when it becomes energetically favourable to do so, expands to form the vacuum of the new phase. Where these expanding bubbles meet, regions of the old H phase may be trapped to form new monopoles.

The simplest case to analyze is when H' is a subgroup of H. In that case, what happens to a monopole can be described in terms of the homomorphism

$$\psi : \pi_1(H') \to \pi_1(H)$$

of the fundamental group of H' into that of H. If the H monopole corresponds to an element a of $\pi_1(H)$ belonging to the image im ψ, then it converts to an H' monopole with a magnetic charge corresponding to one of the elements in $\psi^{-1}a$, generally one of those with the smallest possible total H' charge. On the other hand if the H monopole corresponds to an element a that does *not* belong to imψ, then it must acquire a flux tube, a relatively light string that ties it to another monopole. In this event the two will be drawn together and will annihilate (Bais and Langacker 1981).

If H' is not a subgroup of H (nor H of H') then it is usually not possible to say of a particular monopole that it *must* transmute in a particular way. Rather, it has finite probabilities of making various different transitions. The details are certainly complex, and no very general analysis seems to be possible. It is clear however that if both H and H' phases can exhibit monopoles then it is extremely difficult to arrange that all the H monopoles disappear without creating H' monopoles in their place. The number may be reduced but not by many orders of magnitude.

One possible solution of the monopole problem might be to arrange that all H monopoles disappear and are replaced by lighter H' monopoles. If the second transition occurs at a sufficiently low temperature, say 10^5 GeV, and the number density is limited by some degree of supercooling, it might just be possible to achieve sufficient reduction. Against this, however, is the fact that the observational limits on monopole density are considerably more stringent for very light monopoles.

For strings there are similar options at a second transition. Most strings either survive unaltered or decay, though there is also in principle the possibility that they may become joined by relatively light walls. As with monopoles it is possible for strings to nucleate the transition.

6. String theory of galaxy formation

I now want to turn to the final topic of these lectures, the still rather speculative idea of using strings to explain the formation of galaxies.

Let me start by surveying briefly our present understanding of this problem (see Peebles 1980). Consider a uniform gas cloud of density σ. If a small scale density perturbation appears, it will tend to die away but there is a minimum scale beyond which such a perturbation will grow. This occurs when the gravitational attraction can overcome the gas pressure. The required minimum scale is the Jeans length

$$L_J = c_s/(G\varrho)^{1/2},$$

where c_s is the sound velocity. So long as the gas is ionised, $c_s = 1/\sqrt{3}$. Thus, during the radiation-dominated era, we have

$$L_J = c_s t = t/\sqrt{3},$$

so that the Jeans length is comparable with the horizon distance. However after the decoupling time, when the electrons and protons combine to form neutral hydrogen atoms and effectively decouple from the radiation, c_s falls rapidly to the value characteristic of atomic hydrogen, namely

$$c_s = (5T/3m_H)^{1/2}.$$

Then $L_J \ll t$. On scales larger than L_J, the density perturbation $\delta\varrho/\varrho$ can start to grow, essentially like $t^{2/3}$. Thus to achieve a density contrast of order unity at about 10^{16} s one requires $\delta\varrho/\varrho \sim 10^{-3}$ at the decoupling time ($t \sim 10^{12}$ s).

There are essentially two existing theories of galaxy formation. The first is based on *adiabatic* perturbations, for which

$$(\delta\varrho/\varrho)_{rad} \sim (4/3)\,(\delta\varrho/\varrho)_{matter}.$$

Before these come inside the horizon, i.e. for times t such that $t < \lambda$ (where λ is the wavelength of the perturbation at time t, a quantity that initially grows like $t^{1/2}$), the density contrast suitably defined grows like t. However during decoupling these perturbations are

heavily damped by photons diffusing out of the high-temperature regions on all scales up to the Silk mass,

$$M_{\text{Silk}} \geqslant 10^{12}\, M_{\odot}.$$

(Here $M_{\odot}$ is of course the solar mass.) Thus only the largest perturbations can survive. The resulting theory, largely due to Zel'dovich (1970), is often called the pancake theory because it envisages the formation of a layered structure that later breaks up into individual galaxies.

There are several difficulties with this scenario. The predicted large-scale inhomogeneity is close to the limit allowed by observation of the isotropy of the microwave background radiation. There is a problem with the generation of unwanted secondary isothermal perturbations at the decoupling time. Finally there is no satisfactory explanation of the required initial spectrum of density perturbations. Press (1979) has suggested one possible explanation, based in fact on a phase transition occuring very early in the history of the universe. However his theory depends on the assumption that Higgs fields are so special that their gravitational interaction differs from that of everything else. This is not easy to accept.

The alternative theory of galaxy formation (Peebles 1980) is based on isothermal perturbations, for which $(\delta\varrho/\varrho)_{\text{rad}} = 0$. In this case, the perturbation does not grow before coming inside the horizon — $\delta\varrho/\varrho$ is essentially constant for $t < \lambda$. These perturbations are not much damped at the decoupling time and on scales larger than the subsequent Jeans mass (about $10^5 M_{\odot}$) they can begin to grow, like $t^{2/3}$. Here the initial condensations are on the scale of stellar clusters which subsequently clump together to form galaxies. The major problem with this theory is that any mechanism that generates baryon number will destroy whatever isothermal perturbations are present initially. So the question is — what then generates $\delta\varrho/\varrho$? There seems to be no good answer to this question.

Since neither of these existing theories is very satisfactory, it is natural to look for an alternative. Strings offer one possible source of initial density perturbations. The question is — can they survive long enough to be relevant? If so, can they generate density perturbations of the right size?

The first suggestion that strings might be useful in this context was made by Zel'dovich (1980). However, a rather more attractive version of the theory has been put forward by Vilenkin (1981 b, c).

6.1. Early evolution of strings

To explain how this theory works I have to go back to the strings that are produced in some models and discuss how they evolve in time (Kibble 1976).

Initially they are produced in a rather random tangle. They are moving in a dense medium, so their motion is quite strongly damped. In fact, the damping force per unit length on a string moving with velocity v through a medium of density ϱ is of order $\xi_0 \varrho v$ (In later stages, this is no longer true, for reasons I will explain, but it is adequate for an initial discussion.) Thus the typical damping time is

$$t_{\text{d}} \sim \mu/\xi_0 \varrho \sim h\eta^3 t^2/m_{\text{P}}^2.$$

Initially t is roughly the time of the phase transition, i.e.

$$t \sim t_c \sim m_P/\eta^2.$$

This leads to $t_d \sim h/\eta$. This may be compared with the typical length scale L of the system of strings which as I showed earlier is of order

$$L \sim \xi_G \sim 1/h^2\eta.$$

It follows that $t_d \ll L \ll t$.

However, this situation does not last long, as I shall show. During the period of heavy damping the strings will tend to shorten and straighten. The total length of string will decrease through this process and the shrinkage and disappearance of closed loops. Moreover strings may sometimes cross and exchange partners, producing new kinks that straighten out in turn.

Consider a section of string with radius of curvature r. It will experience an initial inward acceleration of order $\mu/\mu r = 1/r$ but damping will quickly bring it to a limiting velocity of order t_d/r. Hence the typical time scale for straightening of kinks is r^2/t_d. We may expect therefore that the length scale of the structure will grow at a rate given by

$$\frac{1}{L}\frac{dL}{dt} \sim \frac{t_d}{L^2}.$$

Initially, this means that L grows like the square root of the elapsed time. However on a longer time scale, t_d is not constant because ϱ is falling like T^4 or t^{-2} so that as noted above t_d is proportional to t^2. This yields

$$t_d \propto t^2, \quad L \propto t^{3/2}.$$

Clearly t_d will eventually catch up with L. In fact this occurs according to this analysis when

$$t_d \sim L \sim t \sim m_P^2/h\eta^3 \sim 10^{-32}\text{s}.$$

A more careful analysis (Everett 1981) shows that this conclusion must be slightly modified. Once the temperature has fallen well below T_c, the typical wavelengths of particles are large compared to the thickness of a string. In these circumstances, the effective cross section presented by a string is not its thickness but the particle wavelength. Thus the damping is considerably increased. In fact, t_d now behaves not like t^2 but like $t^{3/2}$. Thus we have

$$t_d \propto t^{3/2}, \quad L \propto t^{5/4}.$$

However it remains true that t will eventually catch up with L, though at the somewhat later time

$$t_d \sim L \sim t \sim 10^{-28}\text{ s}.$$

This is still very early in the history of the universe.

6.2. Later stages of evolution

What happens subsequently is less clear. Since the motion of the strings is no longer heavily damped, they may acquire relativistic velocities. It is clear that the typical length scale of the configuration of strings cannot grow faster than the causal horizon distance, i.e. $L \lesssim t$. Let us for the moment assume (as does Vilenkin 1981 b) that in fact $L \sim t$. This means that at any epoch there is approximately one string across the visible universe.

This implies that the strings must be losing energy. I remarked earlier that a single string stretched across the visible universe at any time within the radiation-dominated era would contribute a fraction $4G\mu \sim 10^{-7}$ of the total mass. On the other hand the strings are constantly gaining energy from the expansion of the universe (because of work done against the string tension). So if it is true that $L \sim t$, there must be an energy loss mechanism that constantly transfers energy from the strings to the radiation.

Does such a mechanism exist? The answer is certainly yes, namely gravitational radiation, though whether this is the dominant mechanism is rather hard to assess. Vilenkin argues that the crucial step is the formation of closed loops. From time to time a string will self-intersect, forming a closed loop, typically with a radius of order t. Alternatively, larger closed loops may form by chance but will have no effect until they come within the horizon and can as it were see that they are closed.

Once formed, a closed loop cannot escape eventual disappearance, unless it should happen again to cross a longer string and become reattached. It will oscillate back and forth under its own tension, with a frequency $\omega \sim 1/l$ for a closed loop of roughly circular shape, of length $2\pi l$. One energy loss mechanism that certainly operates is gravitational radiation. (Exactly circular loops would not radiate, but are unlikely to form by chance.) The rate of energy loss is roughly

$$\dot{M} \sim -GM^2\omega^2 \sim -(2\pi\mu)^2 G,$$

since $M \sim 2\pi l\mu$. This leads to a lifetime τ given by

$$1/\tau \sim \dot{M}/M \sim 2\pi G\mu/l$$

or

$$\tau \sim 10^6\, l.$$

If this is the dominant energy loss mechanism, the loops live a long time. Those born at time t live to about $10^6\, t$ before finally annihilating into particles of various kinds. In total the many small closed loops that have not yet disappeared contribute much more to the total mass than the long strings with a length scale of order t. In fact the contribution of the loops gives

$$\varrho_{\text{loops}}/\varrho \sim 10^{-3}.$$

This is very encouraging because this is precisely the order of magnitude of density constrast needed to trigger galaxy formation.

Another good feature of this theory is that the scale of the loops is about right. A mass M enters the horizon at a time given roughly by

$$t \simeq 10^{-5}\,(M/M_\odot)\text{s}.$$

Thus for example a galactic mass $\sim 10^{12} M_{\odot}$ enters at about a year. Loops produced at that time with a galactic scale survive to about 10^6 years, that is well past the decoupling time.

An intriguing feature of the strings is their rather curious gravitational field (Vilenkin 1981a). The space-time around a string is in fact flat. A test particle in the neighbourhood would experience no gravitational force. However, they do have an effect. The space near a string is cone-shaped, as though a wedge of small angle had been removed and the two sides joined together. The missing angle is of order

$$\delta \simeq 8\pi G\mu \sim 10^{-6}.$$

Thus objects beyond a string might show double images. The string acts as a cylindrical gravitational lens. This might possibly explain the observation of double quasars with almost identical spectra.

6.3. Problems with the string theory

There are however difficulties with this attractive theory. One question that must be asked is whether any non-gravitational energy loss mechanisms exist. Though strings do not carry for instance electric charge they certainly do interact indirectly with the electromagnetic and other fields, and it is far from clear that accelerated strings do not generate higher-multipole electromagnetic radiation (but see Everett 1981).

Another possibility that has not been taken account of in our discussion so far is that an oscillating loop may self-intersect and break in two. Since smaller loops lose energy more rapidly, this will reduce the mean lifetime. It is hard to estimate the probability that a loop in an initially random configuration will subsequently intersect itself. However, it seems plausible to suppose that the probability of this occurring in any oscillation cycle is some fixed and non-negligible number, $2\pi p$, say. In that event a typical loop would fragment into two within a time of order l/p. The pieces would then oscillate faster and break again after a time $l/2p$. Clearly, the total lifetime would be reduced to about $2l/p$. Unless it could be shown that p were very small indeed, this would mean that the total contribution of loops to the mass density would be far too small to be of relevance to galaxy formation.

It would be possible to avoid this problem by adopting the original suggestion of Zel'dovich (1980), who hypothesized a transition much closer to the Planck mass, at about 10^{17} GeV. Strings formed at such a transition would be heavier and so even if the loops were rather short-lived might contribute enough mass to provide the requisite density contrast. It should be noted however that in that case the surviving loops would all be rather large. All galaxy-sized loops would have disappeared by the decoupling time. Thus in this scenario it would be clusters of galaxies that condense first. It is also important to ask whether realistic grand unified theories exist that exhibit strings. The relevant criterion, as I indicated earlier, is non-triviality of the first homotopy group $\pi_1(M)$, or equivalently, if G is taken to be simply connected of $\pi_0(H)$. The simplest GUT based on SU(5) does not produce strings, but it is not difficult to invent models that do so.

Possibly the simplest example that might be considered at all realistic is based on the

grand unification group SO(10). If the Higgs field is chosen in the 54-dimensional symmetric tensor representation, the potential may be arranged so that the symmetry breaking scheme is

$$SO(10) \to S[O(6) \times O(4)].$$

This subgroup is locally isomorphic to $SU(4) \times SU(2)_L \times SU(2)_R$, which of course breaks at a second transition to $SU(3) \times SU(2)_L \times U(1)$.

Note that the unbroken subgroup is not simply $SO(6) \times SO(4)$. In addition to rotations within the two subspaces of dimensions 6 and 4, it contains rotations through π in planes such as (67). It follows that

$$\pi_1(M) \simeq \pi_0(H) \simeq Z_2.$$

The model therefore produces "mod-2" strings — rather than strings carrying an additive quantum number.

This is not a very satisfactory model for other reasons (Lazarides, Magg and Shafi 1980). In particular it seems difficult to avoid a whole series of transitions in rapid succession with symmetries of the form $S[O(10-k) \times O(k)]$.

7. *Conclusions*

If we accept the idea of grand unification, which is certainly attractive, then we are forced to accept also the existence of phase transitions in the early universe. Various interesting structures can be formed at these transitions. Strings in particular may possibly provide a new theory of galaxy formation. It will be necessary first to support the arguments about energy loss from loops by more detailed dynamical calculations.

Possibly the most serious problem concerns the cosmic monopoles whose production is an almost inevitable feature of GUTs. Though there are several proposals for reducing their number to an acceptable level, none can at present be regarded as wholly satisfactory.

In conclusion I would like to thank the organizers for their hospitality at what has proved to be a very enjoyable and successful autumn school. I also acknowledge helpful comments from several of the participants at the school.

REFERENCES

Abbott, L. F., *Nucl. Phys.* **B185**, 233 (1981).
Bais, F.A., *Phys. Lett.* **98B**, 437 (1981).
Bais, F. A., Langacker, P., CERN preprint, TH.3142-CERN, 1981.
Bais, F. A., Rudaz, S., CERN preprint TH.2885-CERN, 1980.
Billoire, A., Lazarides, G., Shafi, Q., *Phys. Lett.* **103B**, 450 (1981).
Billoire, A., Tamvakis, K., CERN preprint, TH. 3019-CERN, 1981.
Bogomol'nyi, E. B., *Yad. Fiz.* **24**, 861 (1976) [*Sov. J. Nucl. Phys.* **24**, 449 (1976)].
Callan, C. G., Coleman, S., *Phys. Rev.* **D16**, 1762 (1977).
Coleman, S., *Phys. Rev.* **D15**, 2929 (1977).
Coleman, S., Weinberg, E., *Phys. Rev.* **D7**, 1888 (1973).

Cook, G. P., Mahanthappa, K. T., *Phys. Rev.* **D23**, 1321 (1981).
Einhorn, M. B., Sato, K., *Nucl. Phys.* **B180**, 385 (1981).
Einhorn, M. B. Stein, D. L., Toussaint, D., *Phys. Rev.* **D21**, 3295 (1980).
Everett, A. E., *Phys. Rev.* **D24**, 858 (1981).
Fritzsch, H., Minkowski, P., *Ann. Phys.* **93**, 193 (1975).
Georgi, H., *Particle and Fields*, Ed. C. A. Carlson (AJP), 1975.
Georgi, H., Glashow, S., *Phys. Rev. Lett.* **32**, 438 (1974).
Ginzburg, V. L., *Fiz. Tverd. Tela* **2**, 2031 (1960) [*Sov. Phys. Solid State* **2**, 1824 (1960)].
Guth, A. H., *Phys. Rev.* **D23**, 347 (1981).
Guth, A. H., Tye S. H., *Phys. Rev. Lett.* **44**, 631 (1980).
Guth, A. H., Weinberg, E., *Phys. Rev.* **D23**, 876 (1981).
Horibe, M., Hosoya, A., Osaka University preprint OU-HET. 41, 1981.
Hut, P., Klinkhamer, F. R., *Phys. Lett.* **104B**, 439 (1981).
Kennedy, A., Lazarides, G., Shafi, Q., *Phys. Lett.* **99B**, 38 (1981).
Kibble, T. W. B., *J. Phys. A* **9**, 1387 (1976).
Kibble, T. W. B., *Phys. Rep.* **67c**, 183 (1980).
Kuzmin, V. A., Shaposhnikov, M. E., Tkachev, I. I., *Phys. Lett.* **105B**, 167 (1981).
Langacker, P., Pi, S-Y., *Phys. Rev. Lett.* **45**, 1 (1980).
Lazarides, G., Magg, M., Shafi, Q., *Phys. Lett.* **97B**, 87 (1980).
Linde, A. D., *Zh. Eksp. Teor. Fiz. Pis'ma* **23**, 73 (1976) [*JETP Lett.* **23**, 64 (1976)].
Linde A. D., Lebedev Physical Institute preprint No: 125 (BI-TP 80/20), 1980a. Published in part in Linde (1980b).
Linde, A. D., *Phys. Lett.* **96B**, 293 (1980b).
Marciano, W., Pagels, H., *Phys. Rep.* **36C**, 138 (1978).
Peebles, P. J. A., *The Large Scale Structure of Space-Time*, Princeton U. P., 1980.
Penzias, A. A., Wilson, R. W., *Ap. J.* **142**, 419 (1965).
Polyakov, A. M., *Zh. Eksp. Teor. Fiz. Pis'ma* **20**, 430 (1974) [*JETP Lett.* **20**, 194 (1974)].
Preskill, J. P., *Phys. Rev. Lett.* **43**, 1365 (1979).
Press, W. H., *Phys. Scr.* **21**, 702 (1979).
Steinhardt, P. J., *Phys. Rev.* **D24**, 847 (1980).
'tHooft, G., *Nucl. Phys.* **B79**, 276 (1974).
Vilenkin, A., *Phys. Rev.* **D23**, 852 (1981a).
Vilenkin, A., *Phys. Rev. Lett.* **46**, 1169, 1496 (E) (1981b).
Vilenkin, A., *Phys. Rev.* **D24**, 2082 (1981c).
Weinberg, S., *Principles and Applications of the General Theory of Relativity*, Wiley, 1972.
Weinberg, S., *Phys. Rev.* **D9**, 3357 (1974).
Zee, A., *Phys. Rev. Lett.* **44**, 703 (1979).
Zel'dovich, Ya. B., *Astron. Astrophys.* **5**, 84 (1970).
Zel'dovich, Ya. B., *Mon. Not. Roy. Astron. Soc.* **192**, 663 (1980).
Zel'dovich, Ya. B., Khlopov, M. Y., *Phys. Lett.* **79B**, 239 (1976).
Zel'dovich, Ya. B. Kobzarev, I. Ya., Okun, L. B., *Zh. Eksp. Teor. Fiz.* **67**, 3 (1974) [*JETP* **40**, 1 (1974)].