11
Inflation, Dark Matter and Dark Energy

Malcolm Longair and Chris Smeenk

11.1 Introduction

The story of how we arrived at the present picture of the structure and evolution of the Universe has concentrated largely upon observation, interpretation and the judicious application of theory through Chaps. 6 to 10. The developments in astrophysical and geometrical cosmology represent quite extraordinary progress in understanding the origins and evolution of our Universe and its contents. The contrast between the apparently insuperable problems of determining precise values of cosmological parameters up till the 1990s and the present era of *precision cosmology* in the first decades of the 21st century is startling. But these achievements also resulted in a significant change of perspective in that they involved the introduction of new aspects of physics into cosmology, largely as a result of the increased confidence in favour of the now-standard ΛCDM model. These in turn led to a better understanding of the energy budget of the Universe and a much clearer understanding of the early Universe.

We now need to pull all these strands together to address the major issues of cosmology as a science and pave the way for the considerations of Chaps. 12 and 13 which review potential future directions for contemporary cosmology.[1] The steps towards the realisation that dark matter and dark energy are essential components of the physical content of our universe will be briefly reviewed (Sect. 2). Then, the major physical problems which have to be addressed by observers and theorists are discussed (Sect. 3). In Sect. 4, a brief pedagogical interlude will help bring some of the issues into clearer focus. This leads to a critical discussion of the inflationary paradigm for the very early history of our Universe in Sect. 5 and subsequent sections.

11.2 Dark Matter and Dark Energy

11.2.1 Dark Matter

In the early days of astrophysical cosmology, there were many reasons why there should be dark matter in the Universe made out of familiar baryonic material – low mass stars, dead stars, interstellar and intergalactic gas, dust and so on. If the matter did not radiate in the optical waveband, it was invisible. The subsequent story breaks naturally into two parts – first

establishing the amount of dark matter present in the Universe and then determining whether or not it is baryonic. This endeavour was to require the full power of the information-gathering capacities of the new post-War astronomies, the advent of new technologies and associated astronomical facilities, and advances in interpretation and theory (Sect. 7.3). Among key astrophysical steps along the way were the following.

- Oort's pioneering determination of the mass density in the plane of the Galaxy from the velocity dispersion of stars perpendicular to the plane of the Galaxy (0.092 $M_\odot$ pc^{-3}) showed that gravitationally there was more mass present than the sum of the masses of all types of star in our vicinity (0.038 $M_\odot$ pc^{-3}) (Oort, 1932).
- Zwicky's remarkable pioneering demonstration of the enormous mass-to-light ratios of clusters of galaxies, as determined by application of the virial theorem to the velocity dispersion of galaxies in the Coma cluster, brought vividly to light just how much dark matter there had to be in these systems (Zwicky, 1933, 1937).[2]
- Once powerful long-slit optical spectroscopic facilities became available, the rotation curves of galaxies could be traced well beyond their central regions and flat rotation curves were observed by Vera Rubin and her colleagues (Rubin et al., 1980) (Fig. 11.1). At the same time, studies of spiral galaxies using the 21-cm line of neutral hydrogen determined the rotation curves to much greater distances from their centres than optical observations and showed that flat rotation curves are the norm, rather than the exception (Bosma, 1981).[3]
- Theoretical studies of the stability of the mass distributions in disk galaxies by Miller, Prendergrast and Hohl found that these were unstable. Ostriker and Peebles (1973) showed that the presence of dark matter haloes could stabilize disc galaxies.
- X-ray imaging of clusters of galaxies enabled the total mass distribution within the cluster gravitational potential to be determined and there was found to be much more mass present than could be attributed to galaxies, based on their average mass-to-luminosity ratios (Fabricant et al., 1980; Böhringer, 1994).
- The low mass-to-luminosity ratios for the visible parts of galaxies were consistent with the low baryonic mass density inferred from primordial nucleosynthesis (Sect. 6.7 and 10.7). While the parameters could be stretched to explain the dark matter in clusters by baryonic matter, it was at the verges of plausibility by the 1980s.
- Finally, the low limits to the fluctuations in the cosmic microwave background radiation forced cosmologists to take non-baryonic dark matter really seriously in the early 1980s in order to account for the formation of structure in the Universe by the present epoch, while depressing the predicted level of fluctuations in the Cosmic Microwave Background Radiation below the observational upper limits (Sect. 6.13).

Thus, from the early-1980s onwards, non-baryonic dark matter had to be taken seriously. From the point of view of the origin of cosmic structure, models were developed to study the astrophysical implications of different forms of dark matter candidates, for example, hot versus cold dark matter, top-down versus bottom-up approaches to structure formation and so on (Sect. 10.10). The observational and experimental challenges now shifted to developing more detailed models to understand the nature of the dark matter, either in terms of specific classes of astrophysical objects, or by following up clues from particle physics.

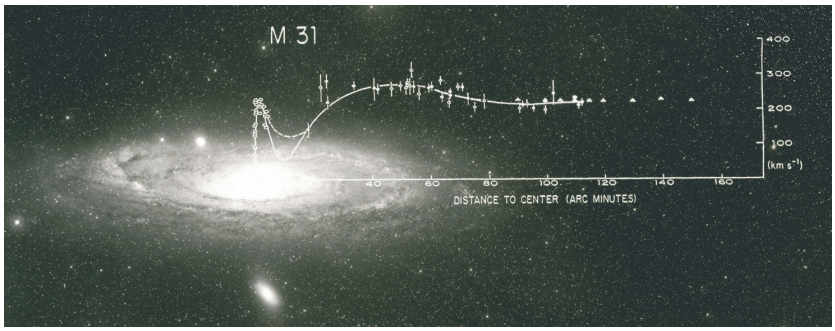


Fig. 11.1 The rotation curve for the nearby giant spiral galaxy M31, showing the flat rotation curve extending well beyond the optical image of the galaxy thanks to observations of the velocities of interstellar neutral hydrogen by Morton Roberts and his colleagues (Courtesy of the late Dr. Vera Rubin).

11.2.2 Constraining dark matter candidates

Baryonic Dark Matter. By *baryonic matter*, we mean ordinary matter composed of protons, neutrons and electrons and for convenience we include the black holes in this discussion. Certain forms of baryonic matter are very difficult to detect because they are very weak emitters of electromagnetic radiation. Important examples include stars with masses $M \leq 0.08 M_{\odot}$, in which the central temperatures are not hot enough to burn hydrogen into helium – they are known collectively as *brown dwarfs*. They have no internal energy source and so the source of their luminosity is the thermal energy with which they were endowed at birth. There could be a small contribution from deuterium burning, but even this is not possible for stars with masses $M \leq 0.01 M_{\odot}$. Brown dwarfs are normally classified as inert stars with masses in the range $0.08 \geq M \geq 0.01 M_{\odot}$. Below that mass, they are normally referred to as planets, $0.01 M_{\odot}$ corresponding to ten times the mass of Jupiter.

Until relatively recently, brown dwarfs were very difficult to detect. The situation changed dramatically with a number technical advances in optical and infrared astronomy. The 2MASS infrared sky survey, conducted at a wavelength of 2 μm, discovered many cool brown dwarfs. The NICMOS infrared camera on the Hubble Space Telescope (HST) discovered numerous brown dwarfs in nearby star clusters. The same techniques of high precision optical spectroscopy, which has been spectacularly successful in discovering extrasolar system planets, was also used to discover a number of brown dwarfs orbiting normal stars. Although the brown dwarfs are estimated to be about twice as common as stars with masses $M \geq 0.08 M_{\odot}$, they contribute very little to the mass density in baryonic matter as compared with normal stars because of their low masses. The consensus of opinion is that brown dwarfs could only make a very small contribution to the dark matter problem.

Black holes are potential candidates for the dark matter. The supermassive black holes in the nuclei of galaxies have masses which are typically only about 0.1% of the mass of the bulges of their host galaxies and so they contribute negligibly to the mass density of the Universe. There might, however, be an intergalactic population of massive black holes. Limits to their number density can be set in certain mass ranges from studies of the numbers of gravitationally-lensed galaxies observed in large samples of extragalactic radio sources. In

their VLA survey of a very large sample of extragalactic radio sources, designed specifically to search for gravitationally lensed structures, Hewitt and her colleagues set limits to the number density of massive black holes with masses in the range $10^{10} \leq M \leq 10^{12} M_{\odot}$. They found that the numbers found corresponded to $\Omega_{\rm BH} \ll 1$ (Hewitt et al., 1987). The same technique using very long baseline interferometry (VLBI) can be used to study the mass density of lower mass black holes by searching for the gravitationally lensed images on an angular scale of a milliarcsecond, corresponding to masses in the range $10^6 \leq M \leq 10^8 M_{\odot}$ (Kassiola et al., 1991). Wilkinson and his colleagues searched a sample of 300 compact radio sources for examples of multiple gravitationally lensed images but none were found. The upper limit to the cosmological mass density of intergalactic supermassive compact objects in the mass range $10^6 \leq M \leq 10^8 M_{\odot}$ corresponded to less than 1% of the critical cosmological density (Wilkinson et al., 2001).

Another possibility raised by Peter Mészáros (1975) was that the dark matter might consist of black holes of mass roughly 1 $M_{\odot}$. It cannot be excluded that the dark matter might consist of a very large population of very low mass black holes, but these would have to be produced by a rather special initial perturbation spectrum in the very early Universe before the epoch of nucleosynthesis. The fact that black holes of mass less than about 10^{12} kg evaporate by Hawking radiation on a cosmological timescale sets a firm lower limit to the possible masses of mini-black holes which could contribute to the dark matter at the present epoch (Hawking, 1975).

An impressive approach to setting limits to the contribution which discrete low mass objects, collectively known as MAssive Compact Halo Objects, or MACHOs, could make to the dark matter in the halo of our own Galaxy, has been the search for gravitational microlensing signatures of such objects as they pass in front of background stars. The MACHOs include low mass stars, white dwarfs, brown dwarfs, planets and black holes. These events are very rare and so very large numbers of background stars have to be monitored. The beauty of this technique is that it is sensitive to MACHOs with a very wide range of masses, from 10^{-7} to 100 $M_{\odot}$, and so the contributions of a very wide range of candidates for the dark matter can be constrained. In addition, the expected light curve of such gravitational lensing events has a characteristic form which is independent of wavelength. The time scale for the brightening is roughly the time it takes the MACHO to cross the Einstein radius of the dark deflector. Two large projects, the MACHO and the EROS projects, have made systematic surveys over a number of years to search for these events. The MACHO project, which ran from 1992 to 1999 used stars in the Magellanic Clouds and in the Galactic bulge as background stars and millions stars were monitored regularly (Alcock et al., 1993b). The first example of a microlensing event was discovered in October 1993 (Fig. 11.2), the mass of the invisible lensing object being estimated to lie in the range $0.03 < M < 0.5\ M_{\odot}$ (Alcock et al., 1993a).

By the end of the MACHO project, many lensing events had been observed, including over 100 in the direction towards the Galactic bulge, about three times more than expected. In addition, 13 definite and 4 possible events were observed in the direction of the Large Magellanic Cloud (Alcock et al., 2000). The numbers are significantly greater than the 2–4 detections expected from known types of star. The technique does not provide distances and masses for individual objects, but, interpreted as a Galactic halo population, the best statistical estimates suggest that the mean mass of these MACHOs is between $0.15 - 0.9\, M_{\odot}$. The statistics are consistent with MACHOs making up about 20% of the necessary halo mass, the

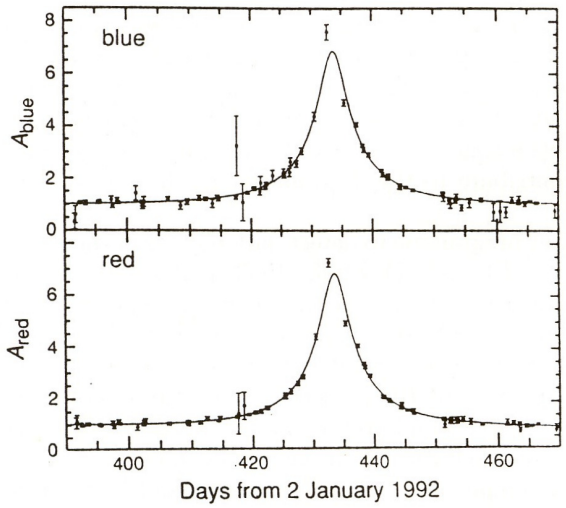


Fig. 11.2 The gravitational microlensing event recorded by the MACHO project in February and March 1993. The horizontal axis shows the date in days measured from day zero on 2 January 1992. The vertical axis shows the amplification of the brightness of the lensed star relative to the unlensed intensity in blue and red wavebands. The solid lines show the expected variations of brightness of a lensed star with time. The same characteristic light curve is observed in both wavebands, as expected for a gravitational microlensing event (Alcock et al., 1993b).

95% confidence limits being 8 – 50%. Somewhat fewer microlensing events were detected in the EROS project which found that less that 25% of the mass of the standard dark matter halo could consist of dark objects with masses in the range 2×10^{-7} to 1 $M_\odot$ at the 95% confidence level (Afonso et al., 2003). The most likely candidates for the detected MACHOs would appear to be white dwarfs, which would have to be produced in large numbers in the early evolution of the Galaxy, but other more exotic possibilities cannot be excluded. The consensus view is that MACHOs alone cannot account for all the dark matter in the halo of our Galaxy and so some form of non-baryonic matter must make up the difference.

As discussed in Sects. 6.7 and 10.7, a strong limit to the total amount of baryonic matter in the Universe is provided by considerations of primordial nucleosynthesis. A consequence of that success story is that the primordial abundances of the light elements, particularly of deuterium and helium-3, are sensitive tracers of the mean baryon density of the Universe. Steigman found a best estimate of the mean baryon density of the Universe of $\Omega_B h^2 = (0.0223 \pm 0.002)$ (Steigman, 2006). Adopting $h = 0.7$, the density parameter in baryonic matter is $\Omega_B = 0.0455$, compared with a mean density of matter in the Universe of $\Omega_0 \approx 0.25$ (see Sect. 10.7). Thus, ordinary baryonic matter is only about one tenth of the total mass density of the Universe, most of which must therefore be in some non-baryonic form.

11.2.3 Non-baryonic Dark Matter

The dark matter may consist of the types of particle predicted by theories of elementary particles but not yet been detected experimentally. Three of the most popular possibilities are described briefly in the following paragraphs.

Axions. The smallest mass candidates are the *axions* which were invented by particle theorists in order to 'save quantum chromodynamics from strong CP violation'. If these particles exist, they would have important astrophysical consequences (Kolb and Turner, 1990). If the axions were produced in thermal equilibrium, they would have unacceptably large masses, which would result in conflict with observations of the Sun and the supernova SN1987A. Specifically, if the mass of the axion were greater than 1 eV, the rate of loss of energy by the emission of axions would exceed the rate at which energy is generated by nuclear reactions in the Sun and so its centre would need to be hotter, resulting in a shorter age than is acceptable and greater emission of high energy neutrinos. There is, however, another non-equilibrium route by which the axions could be created in the early Universe. If they exist, they must have been created when the thermal temperature of the Universe was about 10^{12} K but they were out of equilibrium and never acquired thermal velocities – they remained 'cold'. Their rest mass energies are expected to lie in the range 10^{-2} to 10^{-5} eV. The role of such particles in cosmology and galaxy formation are discussed by Efstathiou and Kolb and Turner (Efstathiou, 1990; Kolb and Turner, 1990).

Neutrinos with finite rest mass. A second possibility is that the three known types of neutrino have finite rest masses. Laboratory tritium β-decay experiments have provided an upper limit to the rest mass of the electron antineutrino of $m_\nu \leq 2$ eV (Weinheimer, 2001). This measurement does not exclude the possibility that the two other types of neutrino, the μ and τ neutrinos, could have greater masses. However, the discovery of neutrino oscillations has provided a measurement of the mass difference between the μ and τ neutrinos of $\Delta m_\nu^2 \sim 3 \times 10^{-3}$ (Eguchi et al., 2003; Aliu et al., 2005). Thus, although their masses are not measured directly, they probably have masses of the order of 0.1 eV. This can be compared with the typical neutrino rest mass needed to attain the critical cosmological densitty of about 10–20 eV.[4]

WIMPs. A third possibility is that the dark matter is in some form of *Weakly Interacting Massive Particle*, or WIMP. This might be the gravitino, the supersymmetric partner of the graviton, or the photino, the supersymmetric partner of the photon, or some form of as yet unknown massive neutrino-like particle. In particular, the dark matter might be in the form of the lightest supersymmetric particle which is expected to be stable.

There must, however, be a suppression mechanism to avoid the problem that, if the WIMPs were as common as the photons and neutrinos, the masses cannot be greater than about 30 eV. The physics of this process is described by Kolb and Turner (1990).[5]. According to particle theorists, almost all theories of physics beyond the standard model involve the existence of new particles at the TeV scale because of the symmetries which have to be introduced to avoid proton decay and violations of the precision tests of the electro-weak theory. These considerations lead to the expectation of new particles at the weak energy scale.

An example of the type of experiment which could demonstrate the presence of new particles has been carried out at the LHCb and CMS experiments at CERN. The cross-section of the extremely rare decay of the B_s meson into two muons has been measured. The observed

branching fraction for this process compared with the predictions of the standard model provides a means of searching for physics beyond the standard model. The measurements were statistically compatible with standard model predictions and so allow stringent constraints to be placed on theories beyond the standard model. This experiment, involving the simplest of the routes to the detection of supersymmetric particles, gave a null result but this does not rule out the importance of this type of search for supersymmetric particles since there are other ways in which they could be involved in particle decays (CMS and LHCb Collaborations et al., 2015).

11.2.4 Astrophysical and Experimental Limits

Useful astrophysical limits can be set to the number densities of different types of neutrino-like particles in the outer regions of giant galaxies and in clusters of galaxies. The WIMPs and massive neutrinos are collisionless fermions and therefore there are constraints on the phase space density of these particles, which translate into a lower limit to their masses (Tremaine and Gunn, 1979).

Being fermions, neutrino-like particles are subject to the Pauli Exclusion Principle according to which there is a maximum number of particle states in phase space for a given momentum $p_{\max}$. It is a straightforwad calculation to show that the resulting lower bound to the mass of the neutrino is:

$$m_\nu \geq \frac{1.5}{(N_\nu \sigma_3 R^2_{\rm Mpc})^{1/4}} \text{ eV}\,, \tag{11.1}$$

where the velocity dispersion σ_3 is measured in units of 10^3 km s^{-1} and R is measured in Mpc.

In clusters of galaxies, typical values are $\sigma = 1000$ km s^{-1} and $R = 1$ Mpc. If there is only one neutrino species, $N_\nu = 1$, we find $m_\nu \geq 1.5$ eV. If there were six neutrino species, namely, electron, muon, tau neutrinos and their antiparticles, $N_\nu = 6$ and then $m_\nu \geq 0.9$ eV. For giant galaxies, for which $\sigma = 300$ km s^{-1} and $R = 10$ kpc, $m_\nu \geq 20$ eV if $N_\nu = 1$ and $m_\nu \geq 13$ eV if $N_\nu = 6$. For small galaxies, for which $\sigma = 100$ km s^{-1} and $R = 1$ kpc, the corresponding figures are $m_\nu \geq 80$ eV and $m_\nu \geq 50$ eV respectively. Thus, particles with rest masses $m_\nu \sim 1$ eV could bind clusters of galaxies but they could not bind the haloes of giant or small galaxies.

The search for evidence for different types of dark matter particles has developed into one of the major areas of *astroparticle physics*. An important class of experiments involves the search for weakly interacting particles with masses $m \geq 1$ GeV, which could make up the dark halo of our Galaxy. In order to form a bound dark halo about our Galaxy, the particles would have to have velocity dispersion $\langle v^2 \rangle^{1/2} \sim 230$ km s^{-1} and their total mass is known. Therefore, the number of WIMPs passing through a terrestrial laboratory each day is a straightforward calculation. When these massive particles interact with the sensitive volume of the detector, the collision results in the transfer of momentum to the nuclei of the atoms of the material of the detector and this recoil can be measured in various ways. The challenge is to detect the very small number of events expected because of the very small cross-section for the interaction of WIMPs with the nuclei of atoms. A typical estimate is that less than one WIMP per day would be detectable by 1 kilogram of detector material. There should be an annual modulation of the dark matter signal as a result of the Earth's motion through the Galactic halo population of dark matter particles.

A good example of the quality of the data now available is provided by the results of the Super Cryogenic Dark Matter Search (SuperCDMS) at the Soudan Laboratory. With an

exposure of 1690 kg days, only a single candidate event was observed, consistent with the expected background in the detector. The upper limit to the spin-independent WIMP-nucleon cross section is $(1.4\pm1.0)\times10^{44}\,\mathrm{cm}^2$ at 46 GeV c^{-2}. These results are the strongest limits to date for WIMP-germanium-nucleus interactions for masses greater than 12 GeV c^{-2} (SuperCDMS Collaboration et al., 2017).

Alternatives to dark matter have been proposed, including modifying Newtonian dynamics (MOND models). But the constraints and the undoubted successes of the standard picture based on standard general relativity sets a very high bar for alternatives to dark matter and will not be discussed further here.

11.2.5 The Dark Energy

The compelling evidence for the finite value of the cosmological constant Λ has been reviewed in Sects. 8.6.2 and 10.4 to 10.8. The evidence for an accelerating universe from the redshift-magnitude relation for Type 1A supernovae and, even more compellingly in the view of this author, from the many different aspects of analysing the properties of the power spectrum of the fluctuations in the Cosmic Microwave Background Radiation, is unambiguous. It is particularly impressive that, using the scalar power spectrum of the fluctuations, their polarisation power spectra and the large-scale power spectrum of the dark matter derived from the *Planck* observations, the six parameter family of the best-fit model can be derived without recourse to any other observations.[6] The independence of this result from all the other estimates is striking.

But there is more to it than that. The ΛCDM model solves the problem of creating the large-scale structure of the distribution of dark matter in a simple and elegant manner without the need to patch it up essentially arbitrarily with astrophysical phenomena, which is necessary in the other viable models.[7]

If dark matter is a hard problem, the dark energy is very, very hard. The contrast between the dark matter and the dark energy is striking. The estimates of the amount of dark matter depend on Newtonian gravity in domains in which we can have a great deal of confidence that it is the appropriate limit of general relativity. The dark matter is acted upon by gravity in the usual way, whereas the dark energy term in Einstein's equations does not depend upon the mass distribution, as can be seen from the expression for the variation of the scale factor a with cosmic time.

$$\ddot{a} = -\frac{4\pi G}{3}a\left(\varrho + \frac{3p}{c^2}\right) + \tfrac{1}{3}\Lambda a\ . \tag{11.2}$$

The Λ term provides a uniform background against which the evolution of the contents of the Universe unfold. It only makes its presence known on the largest scales a observable at the present epoch and becomes of decreasing importance at large redshifts.

There is also the issue of on which side of (11.2) the cosmological constant term should appear. The Einstein field equations are written by Matteo Realdi in his equation (3.3) above as follows:

$$G_{\mu\nu} - \tfrac{1}{2}\,g_{\mu\nu}\,G - \Lambda g_{\mu\nu} = -\kappa T_{\mu\nu}\ . \tag{11.3}$$

The left-hand side of this equation describes the geometry of space-time as described by $G_{\mu\nu}$ while the stress-energy tensor $T_{\mu\nu}$ appears on the right-hand side. Is the Λ-term part of the intrinsic geometry of the universe, in which case it should appear on the left-hand side, or is it a source term for the gravitational field in which case it should appear on the right-hand side

of (11.3). These are hard questions to answer observationally, but some aspects of them are feasible. For example, if the dark energy term were to change with cosmic epoch, that would imply that it is a physical field. Experiments such as the *Euclid* experiment of the European Space Agency and the *WFIRST* mission of NASA aim to tackle that very issue.

11.3 The Big Problems

The concordance model discussed in Sects. 8.6.2 and 10.4 is undoubtedly a remarkable triumph but, like all good theories, it raises as many problems as it solves. The picture is incomplete in the sense that, within the context of the standard Friedman world models, the initial conditions have to be put in by hand in order to create the Universe as we observe it today. How did these initial conditions arise? Let us review these basic problems.

11.3.1 The horizon problem

The horizon problem, clearly recognised by Dicke (1961) is the question 'Why is the Universe so isotropic?' At earlier cosmological epochs, the particle horizon $r \sim ct$ encompassed less and less mass and so the scale over which particles could be causally connected was smaller and smaller. We can illustrate this by working out how far light could have travelled along the last scattering surface at $z \sim 1000$ since the Big Bang. In matter-dominated models, this distance is $r = 3ct$, corresponding to an angle $\theta_{\rm H} \approx 2^\circ$ on the sky. Thus, regions of the sky separated by greater angular distances could not have been in causal communication. Why then is the Cosmic Microwave Background Radiation so isotropic? How did causally separated regions 'know' that they had to have the same temperature to better than one part in 10^5?

11.3.2 The flatness problem

Why is the Universe geometrically flat, $\Omega_\kappa = 1$? The flatness problem was also recognised by Dicke in 1961 and reiterated by Dicke and Peebles in 1979 for standard world models with $\Omega_\Lambda = 0$ (Dicke, 1961; Dicke and Peebles, 1979). In its original version, the problem arises from the fact that, according to the standard world models, if the Universe were set up with a value of the density parameter differing even slightly from the critical value $\Omega = 1$, it would diverge very rapidly from this value at later epochs. If the Universe has density parameter Ω_0 today, at redshift z, $\Omega(z)$ would have been given by

$$\left[1 - \frac{1}{\Omega(z)}\right] = f(z)\left[1 - \frac{1}{\Omega_0}\right], \tag{11.4}$$

where $f(z) = (1+z)^{-1}$ for the matter-dominated era and $f(z) \propto (1+z)^{-2}$ during the radiation dominated era. Thus, since $\Omega_0 \sim 1$ at the present epoch, it must have been extremely close to the critical value in the remote past. Alternatively, if $\Omega(z)$ had departed from $\Omega(z) = 1$ at a very large redshift, Ω_0 would be very far from $\Omega_0 = 1$ today. Thus, the only 'stable' value of Ω_0 is $\Omega_0 = 1$. There is nothing in the standard world models that would lead us to prefer any particular value of Ω_0. This is sometimes referred to as the *fine-tuning problem*.[8]

When Dicke described the horizon problem, the value of the overall density parameter was poorly known, but his argument was still compelling. Now we know that the value of the overall density parameter is $\Omega_\kappa = \Omega_{\rm D} + \Omega_{\rm B} + \Omega_\Lambda = 1.00 \pm 0.01$ (Planck Collaboration, 2016b) – there is no hiding place.

11.3.3 The baryon-asymmetry problem

The baryon-asymmetry problem arises from the fact that the photon-to-baryon ratio today is

$$\frac{N_\gamma}{N_{\rm B}} = \frac{4\times 10^7}{\Omega_{\rm B} h^2} = 1.6\times 10^9, \tag{11.5}$$

where $\Omega_{\rm B}$ is the density parameter in baryons and the values of $\Omega_{\rm B}$ and h have been taken from Table 10.6.2. If photons are neither created nor destroyed, this ratio is conserved as the Universe expands. At temperature $T \approx 10^{10}$ K, electron-positron pair production takes place from the photon field. At a correspondingly higher temperature, baryon-antibaryon pair production takes place with the result that there must have been a very small asymmetry in the baryon-antibaryon ratio in the very early Universe if we are to end up with the correct photon-to-baryon ratio at the present day. At these very early epochs, there must have been roughly 10^9+1 baryons for every 10^9 antibaryons to guarantee the observed ratio at the present epoch. If the Universe had been symmetric with respect to matter and antimatter, the photon-to-baryon ratio would now be about 10^{18}, in gross contradiction with the observed value (Zeldovich, 1965). Therefore, there must be some mechanism in the early Universe which results in a slight asymmetry between matter and antimatter. Fortunately, we know that spontaneous symmetry breaking results in a slight imbalance between various classes of mesons and so there is hope that this can be explained by 'standard' particle physics, but the precise mechanism has not been identified.

11.3.4 The Primordial Fluctuation Problem

What was the origin of the density fluctuations from which galaxies and large-scale structures formed? According to the analyses of Sect. 6.10.1, the amplitudes of the density perturbations when they came through the horizon had to be of finite amplitude, $\delta\varrho/\varrho \sim 10^{-4}$, on a very wide range of mass scales. Such density perturbations could not have originated as statistical fluctuations in the numbers of particles on, say, the scales of superclusters of galaxies. As discussed in the above Section, this problem led pioneers such as Lemaître, Tolman and Lifshitz to conclude that galaxies could not have formed by gravitational collapse. Others, such as Zeldovich, Peebles and their colleagues, pressed ahead and assumed that such fluctuations had their origin in the very early universe and followed up the consequences of that assumption. There must have been some physical mechanism which generated finite amplitude perturbations with power-spectrum close to $P(k) \propto k$ in the early Universe.

11.3.5 The Values of the Cosmological Parameters

The horizon and flatness problems, were recognised before compelling evidence was found for the finite value of the cosmological constant. The concordance values for the cosmological parameters create their own problems. The density parameters in the dark matter and the dark energy are of the same order of magnitude at the present epoch but the matter density evolves with redshift as $(1+z)^3$, while the dark energy density is unchanging with cosmic epoch. Why then do we live at an epoch when they have more or less the same values?

The tortuous history of the cosmological constant was recounted in Sects. 6.8.5 and 10.4. A key insight resulted from the introduction of Higgs fields into the theory of weak interactions (Higgs, 1964). The Higgs fields are *scalar* fields, which have negative pressure equations of

state, $p = -\varrho c^2$.[9] The theoretical value of ϱ_Λ can be estimated from quantum field theory and is found to be $\varrho_v = 10^{95}$ kg m^{-3}, about 10^{120} times greater than the value of ϱ_Λ at the present epoch, which corresponds to $\varrho_\Lambda \approx 10^{-27}$ kg m^{-3} (Carroll et al., 1992).[10] This is usually regarded as quite a problem.

As if these problems were not serious enough, they are compounded by the fact that the nature of the dark matter and the dark energy are unknown. Thus, one of the consequences of precision cosmology is the remarkable result that we do not understand the nature of about 95% of the material which drives the large scale dynamics of the Universe. The concordance values for the cosmological parameters listed in Sect. 10.6.2 really are extraordinary – many of our colleagues regard them as crazy. Rather than being causes for despair, however, these problems should be seen as the great challenges for the astrophysicists and cosmologists of the 21st century. It is not too far-fetched to see an analogy with Bohr's theory of the hydrogen atom, which was an uncomfortable mix of classical and primitive quantum ideas, but which was ultimately to lead to completely new and deep insights with the development of quantum mechanics (Longair, 2013).

11.3.6 The Way Ahead

In the standard Friedman models, the problems are solved by assuming that the Universe was endowed with appropriate initial conditions in its very early phases. To put it crudely, we get out at the end what we put in at the beginning. In a truly physical picture of our Universe, we should do better than this.

There are five possible approaches to solving these problems: (Longair, 1997).

- That is just how the Universe is – the initial conditions were set up that way.
- There are only certain classes of Universe in which 'intelligent' life could have evolved. The Universe has to have the appropriate initial conditions and the fundamental constants of nature should not be too different from their measured values or else there would be no chance of life forming as we know it. This approach involves the *Anthropic Cosmological Principle* according to which, in an extreme version, it is asserted that the Universe is as it is because we are here to observe it.
- The inflationary scenario for the early Universe. This topic is taken up in Sect. 11.5 and subsequent sections.
- Seek clues from particle physics and extrapolate that understanding beyond what has been confirmed by experiment to the earliest phases of the Universe.
- Something else we have not yet thought of. We can think of this in terms of what Donald Rumsfeld called the 'unknown unknowns – the ones we don't know we don't know'.[11] This would certainly involve new physical concepts.

Let us consider aspects of these approaches.

11.3.7 The Limits of Observation

Even the first, somewhat defeatist, approach might be the only way forward if it turned out to be just too difficult to disentangle convincingly the physics responsible for setting up the initial conditions from which our Universe evolved. In 1970, McCrea considered the fundamental limitations involved in asking questions about the very early Universe, his conclusion being that we can obtain less and less information the further back in time one asks questions about the

early Universe (McCrea, 1970). A modern version of this argument would be framed in terms of the limitations imposed by the existence of a last scattering surface for electromagnetic radiation at $z \approx 1000$ and those imposed on the accuracy of observations of the Cosmic Microwave Background Radiation and the large-scale structure of the Universe because of their cosmic variances.

In the case of the Cosmic Microwave Background Radiation, the observations made by the *Planck* experiment are already cosmic variance limited for multipoles $l \leq 2000$ – we will never be able to learn much more than we know already about the form of the scalar power-spectrum on these scales. In these studies, the search for new physics will depend upon the discovering discrepancies between the standard concordance model and future observations. The optimists would argue that the advances will come through extending our technological capabilities so that new classes of observation become cosmic variance limited. For example, the detection of primordial gravitational waves through their polarisation signature at small multipoles in the Cosmic Microwave Background Radiation, the nature of dark matter particles and the nature of the vacuum energy are the cutting edge of fundamental issues for astrophysical cosmology. These approaches will be accompanied by discoveries in particle physics with the coming generations of ultra-high energy particle experiments.

It is also salutary to recall that the range of particle energies which have been explored by the most powerful particle accelerators is about 200 GeV, corresponding to a cosmological epoch of about 1 microsecond from the Big Bang. This seems very modest compared with the Planck era which occurred at $t \sim 10^{-43}$ s. Is there really no new physics to be discovered between these epochs?

It is folly to attempt to predict what will be discovered over the coming years, but we might run out of luck. How would we then be able to check that the theoretical ideas proposed to account for the properties of the very early Universe are correct? Can we do better than bootstrapped self-consistency? The great achievement of modern observational and theoretical cosmology has been that we have made enormous strides in defining a convincing framework for astrophysical cosmology through precise observation and the basic problems identified above can now be addressed as areas of genuine scientific enquiry.

11.3.8 The Anthropic Cosmological Principle

There is certainly some truth in the fact that our ability to ask questions about the origin of the Universe says something about the sort of Universe we live in. The Cosmological Principle asserts that we do not live at any special location in the Universe, and yet we are certainly privileged in being able to make this statement at all. In this line of reasoning, there are only certain types of Universe in which life as we know it could have formed. For example, the stars must live long enough for there to be time for biological life to form and evolve into sentient beings. This line of reasoning is embodied in thei *Anthropic Cosmological Principle*, first expounded by Carter in 1974 (Carter, 1974) and dealt with *in extenso* in the books by Barrow and Tipler and by Gribbin and Rees (Barrow and Tipler, 1986; Gribben and Rees, 1989). Part of the problem stems from the fact that we have only one Universe to study – we cannot go out and investigate other Universes to see if they have evolved in the same way as ours. There are a number of versions of the Principle, some of them stronger than others. In extreme interpretations, it leads to statements such as the strong form of the Principle enunciated by Wheeler (1977),

Observers are necessary to bring the Universe into being.

It is a matter of taste how seriously one wishes to take this line of reasoning. To many cosmologists, it is not particularly appealing because it suggests that it will never be possible to find physical reasons for the initial conditions from which the Universe evolved, or for the values of the fundamental constants of nature. But some of these problems are really hard. Weinberg, for example, found it such a puzzle that the vacuum energy density Ω_Λ is so very much smaller than the values expected according to current theories of elementary particles, that he invoked anthropic reasoning to account for its smallness (Weinberg, 1989, 1997). Another manifestation of this type of reasoning is to invoke the range of possible initial conditions which might come out of the picture of chaotic or eternal inflation (Linde, 1983) and argue that, if there were at least 10^{120} of them, then we live in one of the few which has the right conditions for life to develop as we know it. We leave it to the reader how seriously these ideas should be taken, having first read Chaps. 12 and 13. Some of us prefer to regard the Anthropic Cosmological Principle as the very last resort if all other physical approaches fail.

11.4 A Pedagogical Interlude – Distances and Times in Cosmology

First, let us summarise the various times and distances used in the study of the early universe.[12] Some of the terminology used in the subsequent discussion may seem somewhat non-intuitive and so this short pedagogical interlude is intended to help the non-expert appreciate the importance of the physics which follows.

Comoving radial distance coordinate In order to define a self-consistent distance at a specific epoch t, we projected the proper distances along our past light cone to that reference epoch which we take to be the present epoch t_0. In terms of cosmic time and scale factor a, the comoving radial distance coordinate r is then defined to be

$$r = \int_t^{t_0} \frac{c\,\mathrm{d}t}{a} = \int_a^1 \frac{c\,\mathrm{d}a}{a\dot{a}} \,. \tag{11.6}$$

Proper radial distance coordinate The same problem arises in defining a proper distance at an earlier cosmological epoch. We *define* the proper radial distance r_{prop} to be the comoving radial distance coordinate projected back to the epoch t. From (11.5), we find

$$r_{\mathrm{prop}} = a \int_t^{t_0} \frac{c\,\mathrm{d}t}{a} = a \int_a^1 \frac{c\,\mathrm{d}a}{a\dot{a}} \,. \tag{11.7}$$

Particle horizon The particle horizon r_{H} is defined as the maximum proper distance over which there can be causal communication at the epoch t

$$r_{\mathrm{H}} = a \int_0^t \frac{c\,\mathrm{d}t}{a} = a \int_0^a \frac{c\,\mathrm{d}a}{a\dot{a}} \,. \tag{11.8}$$

Radius of the Hubble sphere or the Hubble radius The Hubble radius is the proper radial distance of causal contact *at a particular epoch*. It is the distance at which the velocity in the

velocity-distance relation at that epoch is equal to the speed of light. This Hubble sphere has proper radius

$$r_{\rm HS} = \frac{c}{H(z)} = \frac{ac}{\dot{a}} \,. \tag{11.9}$$

This is the maximum distance over which causal astrophysical phenomena can take place at the epoch t.

Event horizon The event horizon $r_{\rm E}$ is defined as the greatest proper radial distance an object can have if it is ever to be observable by an observer who observes the Universe at cosmic time t_1.

$$r_{\rm E} = a\int_{t_1}^{t_{\rm max}} \frac{c\,{\rm d}t}{a(t)} = a\int_{a_1}^{a_{\rm max}} \frac{c\,{\rm d}a}{a\dot{a}} \,. \tag{11.10}$$

The presence of an event horizon reflects the space-time structure of the universe in the infinite future.

Cosmic time Cosmic time t is defined to be time measured by a fundamental observer who reads time on a standard clock.

$$t = \int_0^t {\rm d}t = \int_0^a \frac{{\rm d}a}{\dot{a}} \,. \tag{11.11}$$

Conformal time The conformal time is found by projecting time intervals along the past light cone to the present epoch, using the cosmological time dilation relation. There are similarities to the definition of comoving radial distance coordinate:

$${\rm d}t_{\rm conf} = {\rm d}\tau = \frac{{\rm d}t}{a} \,. \tag{11.12}$$

Thus, according to the cosmological time dilation formula, the interval of conformal time is what would be measured by a fundamental observer observing distant events at the present epoch t_0. At any epoch, the conformal time has value

$$\tau = \int_0^t \frac{{\rm d}t}{a} = \int_0^a \frac{{\rm d}a}{a\dot{a}} \,. \tag{11.13}$$

The Past Light Cone

This topic requires a little care because of the way in which the standard models are set up in order to satisfy the requirements of isotropy and homogeneity. Because of these, Hubble's linear relation $v = H_0 r$ applies at the present epoch *to recessions speeds which exceed the speed of light*. Consider the proper distance between two fundamental observers at some epoch t

$$r_{\rm prop} = a(t)r \,, \tag{11.14}$$

where r is comoving radial distance. Differentiating with respect to cosmic time,

$$\frac{{\rm d}r_{\rm prop}}{{\rm d}t} = \dot{a}r + a\frac{{\rm d}r}{{\rm d}t} \,. \tag{11.15}$$

The first term on the right-hand side represents the motion of the substratum and, at the present epoch, becomes $H_0 r$. The second term on the right-hand side of (11.14) corresponds to the

velocity of peculiar motions in the local rest frame at r, since it corresponds to changes of the comoving radial distance coordinate. The element of proper radial distance is $a\,\mathrm{d}r$ and so, if we consider a light wave travelling along our past light cone towards the observer at the origin, we find

$$v_{\rm tot} = \dot{a}r - c\ . \tag{11.16}$$

This key result defines the propagation of light from the source to the observer in space-time diagrams for the expanding Universe.

We can now plot the trajectories of light rays from their source to the observer at t_0. The proper distance from the observer at $r = 0$ to the past light cone $r_{\rm PLC}$ is

$$r_{\rm PLC} = \int_0^t v_{\rm tot}\,\mathrm{d}t = \int_0^a \frac{v_{\rm tot}\,\mathrm{d}a}{\dot{a}}\ . \tag{11.17}$$

Notice that, initially the light rays from distant objects are propagating away from the observer – this is because the local isotropic cosmological rest frame is moving away from the observer at $r = 0$ at a speed greater than that of light. The light waves are propagated to the observer at the present epoch through local inertial frames which expand with progressively smaller velocities until they cross the *Hubble sphere* at which the recession velocity of the local frame of reference is the speed of light. Note that $r_{\rm HS}$ is a proper radial distance. From this epoch onwards, propagation is towards the observer until, as $t \to t_0$, the speed of propagation towards the observer is the speed of light.

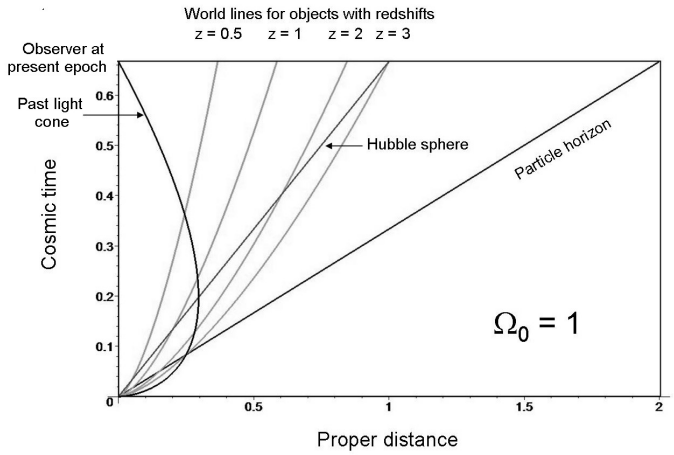
It is simplest to illustrate how the various scales change with time in specific examples of standard cosmological models. We consider first the critical world model and then our reference Λ model. It is convenient to present these space-time diagrams with time measured in units of H_0^{-1} and distance in units of c/H_0. The diagrams shown in Figs. 11.3 and 11.4 follow the attractive presentation by Davis and Lineweaver, but the time axis has been truncated at the present cosmological epoch (Davis and Lineweaver, 2004).

The Critical World Model $\Omega_0 = 1, \Omega_\Lambda = 0$. Two different versions of the space-time diagram for the critical world model are shown in Fig. 11.3a and b. The world lines of galaxies having redshifts 0.5, 1, 2 and 3 are shown. As expected, in Fig. 11.3a, the world lines of galaxies follow the relation $r \propto t^{2/3}$. When plotted against comoving radial distance coordinate in Figs. 11.3b, these become vertical lines. Using the conformal time coordinate, the Hubble sphere and particle horizon, as well as the past light cone, become straight lines. There is no event horizon in this model. The initial singularity is now stretched out to become the abscissa of Fig. 11.3*b*.

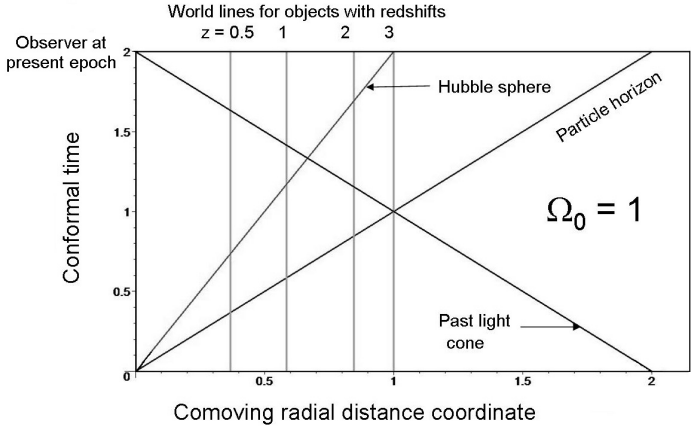
The Reference World Model $\Omega_0 = 0.3, \Omega_\Lambda = 0.7$. Taking $\Omega_0 = 0.3$ and $\Omega_\Lambda = 0.7$, the rate of change of the scale factor with cosmic time in units in which $c = 1$ and $H_0 = 1$ is

$$\dot{a} = \left[\frac{0.3}{a} + 0.7(a^2 - 1)\right]^{1/2}\ . \tag{11.18}$$

The diagrams shown in Fig. 11.4*a*, *b* have many of the same general features as Fig. 11.3*a*, *b*, but there are key differences, the most significant being associated with the dominance of the dark energy term at late epochs.



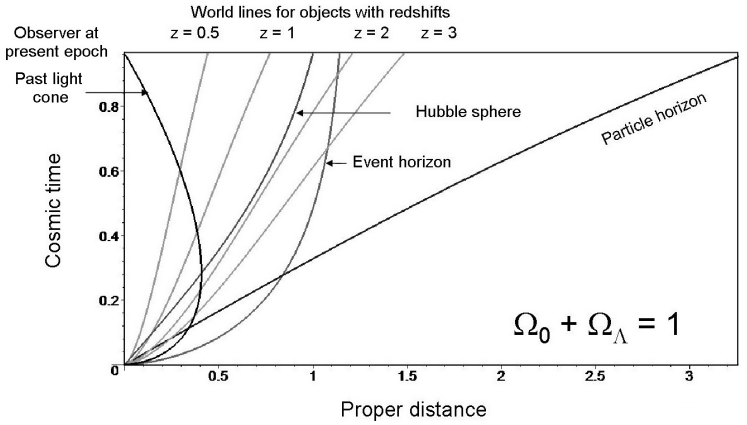
(*a*)



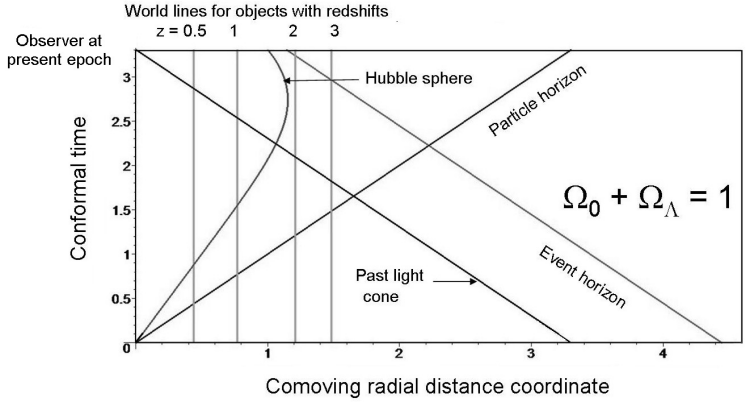
(*b*)

Fig. 11.3 Space-time diagrams for the critical cosmological model, $\Omega_0 = 1, \Omega_\Lambda = 0$. The times and distances are measured in units of H_0^{-1} and c/H_0 respectively. (*a*) This diagram is plotted in terms of cosmic time and proper distance. (*b*) The same space-time diagram plotted in terms of conformal time and comoving radial distance coordinate.

- Note that the cosmic time-scale is stretched out relative to the critical model.
- The world lines of galaxies begin to diverge at the present epoch as the repulsive effect of the dark energy dominates over the attractive force of gravity.
- The Hubble sphere begins to converge to a proper distance of 1.12 in units of c/H_0. The reason for this is that the expansion rate becomes exponential in the future while Hubble's constant tends to a constant value of $\Omega_\Lambda^{1/2}$.
- Unlike the critical model, there is an event horizon in the reference model. The reason is that, although the geometry is flat, the exponential expansion drives galaxies beyond distances at which there could be causal communication with an observer at epoch t. It



(*a*)



(*b*)

Fig. 11.4 Space-time diagrams for the reference cosmological model, $\Omega_0 = 0.3, \Omega_\Lambda = 0.7$. The times and distances are measured in units of H_0^{-1} and c/H_0 respectively (Davis and Lineweaver, 2004).

can be seen from Fig 11.4*a* that the event horizon tends towards the same asymptotic value of 1.12 in proper distance units as the Hubble sphere. To demonstrate this, we need to evaluate the integral

$$r_{\rm E} = a \int_a^\infty \frac{{\rm d}a}{\left[0.3a + 0.7(a^4 - a^2)\right]^{1/2}}. \tag{11.19}$$

For large values of a, terms other than that in a^4 under the square root in the denominator can be neglected and the integral becomes $1/0.7^{1/2} = 1.12$, as found above for the Hubble sphere. In Fig. 11.4*b*, the comoving distance coordinates of the Hubble sphere and the event horizon tend to zero as $t \to \infty$ because, for example, (11.8) has to be divided by a to convert it to a comoving distance and $a \to \infty$. This shrinking of the Hubble sphere is the origin of the statement that ultimately we will end up 'alone in the Universe'.

The papers by Davis and Lineweaver (2004) and by Ellis and Rothman (1993) repay close study. The remarkable Appendix B of the former paper indicates how even some of the most distinguished cosmologists and astrophysicists can lead the unwary newcomer to the subject astray.

11.5 The Inflationary Universe – Historical Background

The most important conceptual development for studies of the very early Universe can be dated to about 1980 and the proposal by Guth of the *inflationary model* for the very early Universe (Guth, 1981). Guth fully acknowledged that there had been earlier suggestions foreshadowing his proposal.[13] Zeldovich had noted in 1968 that there is a physical interpretation of the cosmological constant Λ in terms of the zero-point fluctuations in a vacuum (Zeldovich, 1968). Linde in 1974 and Bludman and Ruderman in 1977 had shown that the scalar Higgs fields of particle physics have similar properties to those which would result in a positive cosmological constant (Linde, 1974b; Bludman and Ruderman, 1977).[14] In 1975, Gurevich noted that an early initial vacuum-dominated phase would provide a 'cause of cosmological expansion', this solution having later to be joined onto the standard Friedman-Lemaître solutions. Starobinsky, a member of Zeldovich's group of astrophysicists/cosmologists, found a class of cosmological solutions which indeed did just that, starting with a de Sitter phase and ultimately ending up as Freidman-Lemaître models – he noted that the exponential de Sitter expansion could lead to a solution of the singularity problem by extrapolating the de Sitter solutions back to $t \rightarrow -\infty$. He also predicted that gravitational waves would be generated during the de Sitter phase at potentially measureable levels. Commenting on this work, Zeldovich also noted that the exponential expansion would eliminate the horizon problem.

Guth realised that, if there were an early phase of exponential expansion of the Universe, this could solve the horizon problem and drive the Universe towards a flat spatial geometry, thus solving the flatness problem, simultaneously. The great merit of Guth's insights was that they made the issues of the physics of the early Universe accessible to the community of cosmologists and spurred an explosion of interest in developing genuine physical theories of the very early Universe by particle theorists.

Suppose the scale factor, a, increased exponentially with time as $a \propto \mathrm{e}^{t/T}$. Such exponentially expanding models were found in some of the earliest solutions of the Friedman equations, in the guise of empty de Sitter models driven by what is now termed the vacuum energy density Ω_Λ (Lanczos, 1922). Consider a tiny region of the early Universe expanding under the influence of the exponential expansion. Particles within the region were initially very close together and in causal communication with each other. Before the inflationary expansion began, the region had physical scale less than the particle horizon, and so there was time for it to attain a uniform, homogeneous state. The region then expanded exponentially so that neighouring points in the substratum were driven to such large distances that they could no longer communicate by light signals – the causally-connected regions were swept beyond their particle horizons by the inflationary expansion. At the end of the inflationary epoch, the Universe transformed into the standard radiation-dominated Universe and the inflated region continued to expand as $a \propto t^{1/2}$.

Let us demonstrate to order of magnitude how the argument runs. The time-scale 10^{-34} s is taken to be the characteristic e-folding time for the exponential expansion. Over the interval from 10^{-34} s to 10^{-32} s, the radius of curvature of the Universe increased exponentially by

a factor of about $e^{100} \approx 10^{43}$. The horizon scale at the beginning of this period was only $r \approx ct \approx 3 \times 10^{-26}$ m and this was inflated to a dimension of 3×10^{17} m by the end of the inflationary era. This dimension then scaled as $t^{1/2}$, as in the standard radiation-dominated Universe so that the region would have expanded to a size of $\sim 3 \times 10^{42}$ m by the present day – this dimension far exceeds the present particle horizon $r \approx cT_0$ of the Universe, which is about 10^{26} m. Thus, our present Universe would have arisen from a tiny region in the very early Universe which was much smaller than the horizon scale at that time. This guaranteed that our present Universe would be isotropic on the large scale, resolving the horizon problem. At the end of the inflationary era, there was an enormous release of energy associated with the 'latent heat' of the phase transition and this reheated the Universe to a very high temperature indeed.[15]

The exponential expansion also had the effect of straightening out the geometry of the early Universe, however complicated it may have been to begin with. Suppose the tiny region of the early Universe had some complex geometry. The radius of curvature of the geometry $R_c(t)$ scales as $R_c(t) \propto a(t)$, and so it is inflated to dimensions vastly greater than the present size of the Universe, driving the geometry of the inflated region towards flat Euclidean geometry, $\Omega_\kappa = 0$, and consequently the Universe must have $\Omega_0 + \Omega_\Lambda = 1$. It is important that these two aspects of the case for the inflationary picture can be made independently of a detailed understanding of the physics of the inflation. There is also considerable freedom about the exact time when the inflationary expansion could have occurred, provided there are sufficient e-folding times to isotropise our observable Universe and flatten its geometry.

The problem with this realisation was that it predicted 'bubbles' of true vacuum embedded in the false vacuum, with the result that huge inhomogeneities were predicted. Another concern was that an excessive number of monopoles were created during the GUT phase transition. Kibble (1976) showed that, when this phase transition took place, topological defects are expected to be created, including point defects (monopoles), line defects (cosmic strings) and sheet defects (domain walls). Kibble also showed that one monopole is created for each correlation scale at that epoch. Since that scale cannot be greater than the particle horizon at the GUT phase transition, it is expected that huge numbers of monopoles are created. According to the simplest picture of the GUT phase transition, the mass density in these monopoles in the standard Big Bang picture would vastly exceed $\Omega_0 = 1$ at the present epoch (Kolb and Turner, 1990).

In Guth's original inflationary scenario, the exponential expansion was associated with the spontaneous symmetry breaking of Grand Unified Theories of elementary particles at very high energies through a first-order phase transition, only about 10^{-34} s after the Big Bang, commonly referred to as the GUT era. The Universe was initially in a symmetric state, referred to as a false vacuum state, at a very high temperature before the inflationary phase took place. As the temperature fell, spontaneous symmetry breaking took place through the process of barrier penetration from the false vacuum state and the Universe attained a lower energy state, the true vacuum. At the end of this period of exponential expansion, the phase transition took place, releasing a huge amount of energy.

The model was revised in 1982 by Linde and by Albrecht and Steinhardt who proposed instead that, rather than through the process of barrier penetration, the transition took place through a second-order phase transition which did not result in the formation of 'bubbles' and so excessive inhomogeneities (Linde, 1982, 1983; Albrecht and Steinhardt, 1982a). This

picture, often referred to as *new inflation*, also eliminated the monopole problem since the likelihood of even one being present in the observable Universe was very small.

11.6 New Inflation and the Nuffield Workshop

By the spring of 1982 several groups were at work fleshing out the details of the new inflationary scenario: Turner and Kolb at the University of Chicago and Fermilab, Steinhardt and Albrecht at the University of Pennsylvania, Guth at MIT, Linde and his collaborators in Moscow, Laurence Abbott at Brandeis, Hawking and others in Cambridge, and John Barrow in Sussex. With notable exceptions, such as Hawking and Barrow, nearly everyone in this research community came from a background in particle physics. They all met in Cambridge at a workshop sponsored by the Nuffield Foundation to hammer out the developing issues in the physics of the early Universe.[16]

Nearly half the lectures at the Nuffield workshop were devoted to inflation. One important focus of the conference was the calculation of density perturbations produced during an inflationary era. Steinhardt, Starobinsky, Hawking, Turner, Lukash and Guth had all realized that this was a "calculable problem", the answer being an estimate of the magnitude of the density perturbations, measured by the dimensionless density contrast $\Delta = \delta\rho/\rho$, produced during inflation. In this intense period of calculation and critical discussion, the particle physicists adopted Grand Unified Theories of particle physics as the basis for their calculations, particular attention being paid to Higgs fields which had just the right equation of state to drive inflation. Preliminary calculations of this magnitude disagreed by an astounding 12 orders of magnitude: Hawking found $\Delta \approx 10^{-4}$, whereas Steinhardt and Turner (1984) initially estimated a magnitude of 10^{-16}. After three weeks of effort, the various groups working on the problem had converged on an answer.

Mukhanov and Chibisov (1981) had argued that a de Sitter phase could generate perturbations by "stretching" the zero-point fluctuations of quantum fields to significant scales. This idea, carried out quite independently of Guth's work, would become the basis for the generation of seed perturbations in inflationary cosmology. Prior to the workshop, Hawking had circulated a preprint which argued that initial inhomogeneities in the scalar field ϕ would result in inflation beginning at slightly different times in different regions; the inhomogeneities reflect the different "departure times" of the scalar field. Hawking's preprint claimed that this resulted in a scale-invariant spectrum of adiabatic perturbations with $\Delta \approx 10^{-4}$, exactly what was needed in accounts of structure formation.

But others did not trust Hawking's method. At the heart of the debate was the "gauge problem", reflecting the fact that a "perturbed space-time" cannot be uniquely decomposed into a background space-time plus perturbations. Slicing the space-time along different surfaces of constant time leads to different magnitudes for the density perturbations. The perturbations "disappear," for example, by slicing along surfaces of constant density. In practice, almost all studies of structure formation used a particular choice of gauge, generally the synchronous gauge, but this leads to difficulties in interpreting perturbations with length scales greater than the Hubble radius. Length scales "blow up" during inflation since they scale as $R(t) \propto \mathrm{e}^{Ht}$, but the Hubble radius remains fixed since H is approximately constant during the slow roll phase of inflation.[17] For this reason it is especially tricky to calculate the evolution of physical perturbations using a gauge-dependent formalism.

Hawking and Guth pursued refinements of Hawking's approach during the Nuffield workshop, the centerpiece of these calculations being the "time delay" function characterizing the start of the scalar field's slow roll down the effective potential. This "time delay" function can be related to the two-point correlation function characterizing fluctuations in ϕ prior to inflation, and it is also related to the spectrum of density perturbations, since these are assumed to arise as a result of the differences in the time at which inflation ends (see Sect. 11.7.3).

Steinhardt and Turner then enlisted James Bardeen's assistance in developing a third approach; he had recently formulated a fully *gauge invariant formulation* for the study of density perturbations on all scales (Bardeen 1980). Using Bardeen's formalism, the three aimed to give a full account of the behavior of different modes of the field ϕ as these evolved through the inflationary phase and up to recombination. The physical origin of the spectrum was traced to the qualitative change in behavior as perturbation modes expand past the Hubble radius: they "freeze out" as they cross the horizon, and leave an imprint that depends on the details of the model under consideration. Despite the conflicting assumptions and other differences, the participants of the Nuffield workshop gave increasing credibility to these results because of the rough agreement between the three different approaches.

The key results were that inflation leads naturally to an almost Harrison-Zeldovich spectrum of density fluctuations and these have Gaussian phases (Bardeen et al. 1983). But reducing the magnitude of these perturbations to satisfy observational constraints required an unnatural choice of coupling constants. In particular, the self-coupling for the Higgs field apparently needed to be on the order of 10^{-8}, in contrast to the "natural" value which would be of the order of 1.

The Higgs model was not successful but it was clear how to develop a "newer inflation" model. Bardeen, Steinhardt and Turner suggested that the effective potential for a scalar field in a supersymmetric theory, rather than the Higgs field of a Grand Unified Theory, would have the appropriate properties to drive inflation. Finding a particular particle physics candidate for the scalar field driving inflation would provide an important independent line of evidence. The Nuffield workshop marked the start of this new approach, as the focus shifted to implementing inflation successfully, rather than starting with a candidate for the field driving inflation derived from particle physics. The introduction of an "inflaton" field, a scalar field custom-made to produce an inflationary stage, roughly a year later illustrates this methodological shift.

Following the demise of the minimal GUT models, there was an ongoing effort to implement inflation within new models provided by particle physics. Following the Nuffield workshop, inflation turned into a "paradigm without a theory," to borrow Turner's phrase, as cosmologists developed a wide variety of models bearing a loose family resemblance. The models share the basic idea that the early universe passed through an inflationary phase, but differ on the nature of the "inflaton" field (or fields) and the form of the effective potential $V(\phi)$. Keith Olive's review of the first decade of inflation ended by bemoaning the ongoing failure of any of these models to renew the strong connection with particle physics achieved in old and new inflation:

> A glaring problem, in my opinion, is our lack of being able to fully integrate inflation into a unification scheme or any scheme having to do with our fundamental understanding of particle physics and gravity. . . . An inflaton as an inflaton and nothing else can only be viewed as a toy, not a theory.[18]

Many different versions of the inflationary picture of the early Universe emerged, an amusing table of over 100 possibilities being presented by Shellard (Shellard, 2003) and

5-dimensional assisted inflation
anisotropic brane inflation
anomaly-induced inflation
assisted inflation
assisted chaotic inflation
boundary inflation
brane inflation
brane-assisted inflation
brane gas inflation
brane-antibrane inflation
braneworld inflation
Brans-Dicke chaotic inflation
Brans-Dicke inflation
bulky brane inflation
chaotic hybrid inflation
chaotic inflation
chaotic new inflation
D-brane inflation
D-term inflation
dilaton-driven inflation
dilaton-driven brane inflation
double inflation
double D-term inflation
dual inflation
dynamical inflation
dynamical SUSY inflation
eternal inflation
extended inflation
extended open inflation
extended warm inflation
extra dimensional inflation
F-term inflation
F-term hybrid inflation
false vacuum inflation
false vacuum chaotic inflation
fast-roll inflation
first order inflation
gauged inflation
generalised inflation
generalized assisted inflation
generalized slow-roll inflation
gravity driven inflation
Hagedorn inflation
higher-curvature inflation
hybrid inflation
hyperextended inflation
induced gravity inflation
induced gravity open inflation
intermediate inflation
inverted hybrid inflation
isocurvature inflation
K inflation
kinetic inflation
lambda inflation
large field inflation
late D-term inflation
late-time mild inflation
low-scale inflation
low-scale supergravity inflation
M-theory inflation
mass inflation
massive chaotic inflation
moduli inflation
multi-scalar inflation
multiple inflation
multiple-field slow-roll inflation
multiple-stage inflation
natural inflation
natural Chaotic inflation
natural double inflation
natural supergravity inflation
new inflation
next-to-minimal supersymmetric hybrid inflation
non-commutative inflation
non-slow-roll inflation
nonminimal chaotic inflation
old inflation
open hybrid inflation
open inflation
oscillating inflation
polynomial chaotic inflation
polynomial hybrid inflation
power-law inflation
pre-Big-Bang inflation
primary inflation
primordial inflation
quasi-open inflation
quintessential inflation
R-invariant topological inflation
rapid asymmetric inflation
running inflation
scalar-tensor gravity inflation
scalar-tensor stochastic inflation
Seiberg-Witten inflation
single-bubble open inflation
spinodal inflation
stable starobinsky-type inflation
steady-state eternal inflation
steep inflation
stochastic inflation
string-forming open inflation
successful D-term inflation
supergravity inflation
supernatural inflation
superstring inflation
supersymmetric hybrid inflation
supersymmetric inflation
supersymmetric topological inflation
supersymmetric new inflation
synergistic warm inflation
TeV-scale hybrid inflation

Fig. 11.5 Paul Shellard's table showing the proliferation of inflationary models from an archive search (Shellard, 2003).

shown in Fig. 11.5.

As a result, there is not a genuine physical theory of the inflationary Universe, but its basic concepts resolve some of the problems listed in Sect. 11.3. What it also does, and which gives it considerable appeal, is to suggest an origin for the spectrum of initial density perturbations as quantum fluctuations on the scale of the particle horizon.

11.7 The Origin of the Spectrum of Primordial Perturbations

As Andrew Liddle and David Lyth (2000) have written,

> Although introduced to resolve problems associated with the initial conditions needed for the Big Bang cosmology, inflation's lasting prominence is owed to a property discovered soon after its introduction: It provides a possible explanation for the initial inhomogeneities in the Universe that are believed to have led to all the structures we see, from the earliest objects formed to the clustering of galaxies to the observed irregularities in the microwave background.

The theory also makes predictions about the spectrum of primordial gravitational waves which are accessible to experimental validation.[19] The enormous impact of particle theorists taking these cosmological problems really seriously has enlarged, yet again, the domain of astrophysical cosmology. For the 'cosmologist in the street', the theory of inflation does not make for particularly easy reading, because the reader should be comfortable with many aspects of theoretical physics which lie outside the standard tools of the observational cosmologist – ladder operators, quantum field theory, zero point fluctuations in quantum fields, all of these applied within the context of curved space-times. Developing the theory of the quantum origin of density perturbations in detail cannot be carried out with modest effort. There is no question, however, that these remarkable developments are at the cutting edge of cosmological research and have the potential to reveal new physics.

Let us list some of the clues about the formulation of a successful theory.[20]

The equation of state. We know from analyses of the physical significance of the cosmological constant Λ that exponential growth of the scale factor is found if the dark energy has a negative pressure equation of state $p = -\varrho c^2$. More generally, exponential growth of the scale factor is found provided the strong energy condition is violated, that is, if $p < -\frac{1}{3}\varrho c^2$. To be effective in the very early Universe, the mass density of the scalar field has to be vastly greater than the value of Ω_Λ we measure today.

The duration of the inflationary phase. In the example of the inflationary expansion given in above, we arbitrarily assumed that 100 e-folding times would take place during the inflationary expansion. A more careful calculation shows that there must have been at least 60 e-folding times and these took place in the very early Universe, much earlier than those which have been explored experimentally by particle physics experiments. It is customary to assume that inflation began not long after the Planck era, but there is quite a bit of room for manoeuvre.

The shrinking Hubble sphere. There is a natural way of understanding how fluctuations can be generated from processes in the very early Universe. It is helpful to revisit the conformal diagrams for world models discussed in Sect. 11.4, in particular, Fig. 11.4b. Recall that these diagrams are exact in the sense that the comoving radial distance coordinate and conformal time are worked out for the reference model with $\Omega_0 = 0.3$ and $\Omega_\Lambda = 0.7$. The effect of using conformal coordinates is to stretch out time in the past and shrink it into the future. Notice that, because of the use of linear scales in the ordinate, the radiation-dominated phase of the standard Big Bang is scarcely visible.

In Fig. 11.6a, there are two additions to Fig. 11.4b. The redshift of 1000 is shown corresponding to the last scattering surface of the Cosmic Microwave Background Radiation. The intersection with our past light cone is shown and then a past light cone from the last scattering surface to the singularity at conformal time $\tau = 0$ is shown as a shaded triangle. This is another way of demonstrating the *horizon problem* – the region of causal contact is very small compared with moving an angle of 180° over the sky which would correspond to twice the distance between the origin and the comoving radial distance coordinate at 3.09.

Let us now add the inflationary era to Fig. 11.6*a*. It is useful to regard the end of the inflation era as the zero of time for the standard Big Bang and then to extend the diagram back to negative conformal times. In other words, we shift the zero of conformal time very slightly to, say, 10^{-32} s and then we can extend the light cones back through the entire inflationary era (Fig. 11.6*b*). This construction provides another way of understanding how the inflationary picture resolves the causality problem. The light cones have unit slope in the conformal diagram and so we draw light cones from the ends of the element of comoving radial distance at $\tau = 0$ from the last scattering surface. Projecting far enough back in time, the light cones from opposite directions on the sky overlap, represented by the dark grey shaded area in Fig. 11.6*b*. This is the region of causal contact in the very early Universe.

There is, however, an even better way of understanding what is going on. We distinguished between the Hubble sphere and the particle horizon in Sect. 11.4 – now this distinction becomes important. The particle horizon is defined as the maximum distance over which causal contact could have been made from the time of the singularity to a given epoch. In other words, it is not just what happened at a particular epoch which is important, but the history along the past light cone. Writing the exponential inflationary expansion of the scale factor as $a = a_0 \exp[H(t - t_\mathrm{i})]$, where a_0 is the scale factor when the inflationary expansion began at τ_i, $r_\mathrm{HS} = c/H$ and the

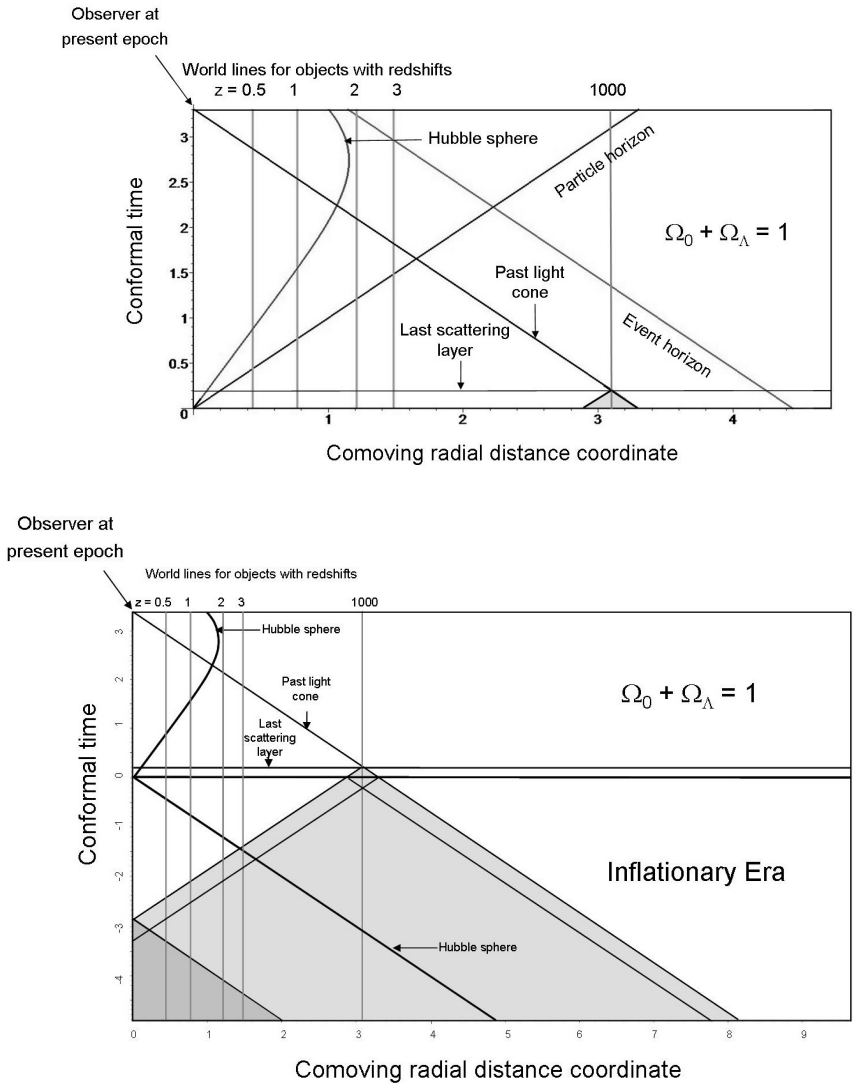


Fig. 11.6 (**a**) A repeat of conformal diagram Fig. 12.2c in which conformal time is plotted against comoving radial distance coordinate. Now, the last scattering surface at the epoch of recombination is shown as well as the past light cone from the point at which our past light cone intersects the last scattering surface. (**b**) An extended conformal diagram now showing the inflationary era. The time coordinate is set to zero at the end of the inflationary era and evolution of the Hubble sphere and the past light cone at recombination extrapolated back to the inflationary era.

comoving Hubble sphere has radius $r_{\rm HS}({\rm com}) = c/(Ha)$. Since H is a constant throughout most of the inflationary era, it follows that the comoving Hubble sphere *decreases* as the inflationary expansion proceeds.

We now need to join this evolution of the comoving Hubble sphere onto its behaviour after the end of inflation, that is, join it onto Fig. 11.6*a*. The expression for conformal time during the inflationary era is

$$\tau = \int \frac{{\rm d}a}{a\dot{a}}\,, \tag{11.20}$$

and so, integrating and using the expression for $r_{\rm HS}({\rm com})$, we find

$$\tau = {\rm constant} - \frac{r_{\rm HS}({\rm com})}{c}\,. \tag{11.21}$$

This solution for $r_{\rm HS}({\rm com})$ is joined on to the standard result at the end of the inflationary epoch, as illustrated in Fig. 11.6*b*. The complete evolution of the Hubble sphere is indicated by the heavy line labelled 'Hubble sphere' in that diagram.

Fig. 11.6*b* illustrates very beautifully how the inflationary paradigm solves the horizon problem. It will be noticed that the point at which the Hubble sphere crosses the comoving radial distance coordinate of the last scattering surface, exactly corresponds to the time when the past light cones from opposite directions on the sky touch at conformal time -3. This is not a coincidence – they are different ways of stating that opposite regions of the Cosmic Microwave Background were in causal contact at conformal time $t = -3$.

But we learn a lot more. Because any object preserves its comoving radial distance coordinate for all time, as represented by the vertical lines in Fig. 11.6, it can be seen that, in the early Universe, objects lie within the Hubble sphere, but during the inflationary expansion, they pass through it and remain outside it for the rest of the inflationary expansion. Only when the Universe transforms back into the standard Friedman model does the Hubble sphere begin to expand again and objects can then 're-enter the horizon'. Consider, for example, the region of the Universe out to redshift $z = 0.5$ which corresponds to one of the comoving coordinate lines in Fig. 11.6*b*. It remained within the Hubble sphere during the inflationary era until conformal time $\tau = -0.4$ after which it was outside the horizon. It then re-entered the Hubble sphere at conformal time $\tau = 0.8$. This behaviour occurs for all scales and masses of interest in understanding the origin of structure in the present Universe.

Since causal connection is no longer possible on scales greater than the Hubble sphere, it follows that objects 'freeze out' when they pass through the Hubble sphere during the inflationary era, but they come back in again and regain causal contact when they recross the Hubble sphere. This is one of the key ideas behind the idea that the perturbations from which galaxies formed were created in the early Universe, froze out on crossing the Hubble sphere and then grew again on re-entering it at conformal times $\tau > 0$.

Notice that, at the present epoch, we are entering a phase of evolution of the Universe when the comoving Hubble sphere about us has begun to shrink again. This can be seen in the upper part of Fig. 11.6*b* and is entirely due to the fact that the dark energy is now dominating the expansion and its dynamics are precisely those of another exponential expansion. In fact, the Hubble sphere tends asymptotically to the line labelled 'event horizon' in Fig. 11.6*a*.

11.7.1 Scalar Fields

As Baumann (2007) noted, there are three equivalent conditions necessary to produce an inflationary expansion in the early universe:

- The decreasing of the Hubble sphere during the early expansion of the Universe;
- An accelerated expansion;
- Violation of the strong energy condition, meaning, $p < -\varrho c^2/3$.

How can this be achieved physically? To quote Baumann's words, written before the discovery of the Higgs boson in 2012,:

Answer: scalar field with special dynamics! Although no fundamental scalar field has yet been detected in experiments, there are fortunately plenty of such fields in theories beyond the standard model of particle physics. In fact, in string theory for example there are numerous scalar fields (moduli), but it proves very challenging to find just one with the right characteristics to serve as an inflaton candidate.

The results of calculations of the properties of the scalar field $\phi(t)$, which is assumed to be homogeneous at a given epoch, are as follows. There are a kinetic energy $\dot{\phi}^2/2$ and a potential energy, or self-interaction energy, $V(\phi)$ associated with the field. Putting these through the machinery of field theory results in expressions for the density and pressure of the scalar field:

$$\varrho_\phi = \frac{1}{2}\dot{\phi}^2 + V(\phi) \qquad ; \qquad p_\phi = \frac{1}{2}\dot{\phi}^2 - V(\phi) \tag{11.22}$$

Clearly the scalar field can result in a negative pressure equation of state, provided the potential energy of the field is very much greater than its kinetic energy. In the limit in which the kinetic energy is neglected, we obtain the equation of state $p = -\varrho c^2$, where the c^2, which is set equal to one by professional field theorists, has been restored.

To find the time evolution of the scalar field, we combine (11.21) with the Einstein field equations with the results:

$$H^2 = \frac{1}{3}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right) \qquad ; \qquad \ddot{\phi} + 3H\dot{\phi} + V(\phi)_{,\phi} = 0\ . \tag{11.23}$$

where $V(\phi)_{,\phi}$ means the derivative of $V(\phi)$ with respect to ϕ. Thus, to obtain the inflationary expansion over many e-folding times, the kinetic energy term must be very small compared with the potential energy and the potential energy term must be very slowly varying with time. This is formalised by requiring the two *slow-roll parameters* $\epsilon(\phi)$ and $\eta(\phi)$ to be very small during the inflationary expansion.[21] These parameters set constraints upon the dependence of the potential energy function upon the field ϕ and are formally written:

$$\epsilon(\phi) \equiv \frac{1}{2}\left(\frac{V_{,\phi}}{V}\right)^2 \quad ; \quad \eta(\phi) \equiv \frac{V_{,\phi\phi}}{V} \quad \text{with} \quad \epsilon(\phi), |\eta(\phi)| \ll 1\ . \tag{11.24}$$

where $V(\phi)_{,\phi\phi}$ means the second derivative of $V(\phi)$ with respect to ϕ. Under these conditions, we obtain what we need for inflation, namely,

$$H^2 = \frac{1}{3}V(\phi) = \text{constant} \quad \text{and} \quad a(t) \propto \mathrm{e}^{Ht}\ . \tag{11.25}$$

At this stage, it may appear that we have not really made much progress since we have adjusted the theory of the scalar field to produce what we know we need. The bonus comes when

we consider fluctuations in the scalar field and their role in the formation of the spectrum of primordial perturbations.

11.7.2 The Quantised Harmonic Oscillator

The key result can be illustrated by the elementary quantum mechanics of a harmonic oscillator. The solutions of Schrödinger's equation for a harmonic potential have quantised energy levels and wave functions

$$E = \left(n + \tfrac{1}{2}\right)\hbar\omega \qquad ; \qquad \psi_n = H_n(\xi)\exp\left(-\tfrac{1}{2}\xi^2\right) , \tag{11.26}$$

where $H_n(\xi)$ is the Hermite polynomial of order n and $\xi = \sqrt{\beta}x$. For the simple harmonic oscillator, $\beta^2 = am/\hbar^2$, where a is the constant in the expression for the harmonic potential $V = \frac{1}{2}ax^2$ and m is the reduced mass of the oscillator.

We are interested in fluctuations about the zero-point energy, that is, the stationary state with $n = 0$. The zero-point energy and Hermite polynomial of order $n = 0$ are

$$E = \tfrac{1}{2}\hbar\omega \quad \text{and} \quad H_0(\xi) = \text{constant} . \tag{11.27}$$

The first expression is the well-known result that the oscillator has to have finite kinetic energy in the ground state. It is straightforward to work out the variance of the position coordinate x of the oscillator,[22]

$$\langle x^2 \rangle = \frac{\hbar}{2\omega m} . \tag{11.28}$$

These are the fluctuations which must necessarily accompany the zero-point energy of the vacuum fields. This elementary calculation sweeps an enormous number of technical issues under the carpet. Baumann's clear presentation of the proper calculation can be warmly recommended. It is reassuring that his final answer agrees exactly with the above results for the one-dimensional harmonic oscillator.

11.7.3 The Spectrum of Fluctuations in the Scalar Field

We need only one more equation – the expression for the evolution of the vacuum fluctuations in the inflationary expansion. The inflaton field is decomposed into a uniform homogeneous background and a perturbed component $\delta\phi$ which is the analogue of the deviation x of the zero point oscillations of the harmonic oscillator. Baumann outlines the derivation of this equation, warning of the numerous technical complexities which have to be dealt with. In Bertschinger's review of the physics of inflation, he deals with these issues and finds the following equation:

$$\delta\ddot{\phi}_k + 3\left(\frac{\dot{a}}{a}\right)\delta\dot{\phi}_k + (k_c^2 c_s^2 - 2\kappa)\,\delta\phi_k = 0 , \tag{11.29}$$

where κ is the curvature of space at the present epoch (Bertschinger, 1996). This has a familiar form which can be understood by comparing it with (6.8) for the evolution of density perturbations in the Friedman models

$$\frac{\mathrm{d}^2\Delta}{\mathrm{d}t^2} + 2\left(\frac{\dot{a}}{a}\right)\frac{\mathrm{d}\Delta}{\mathrm{d}t} = \Delta(4\pi G\varrho_0 - k^2 c_s^2) \tag{11.30}$$

where k is the proper wavenumber and c_s is the speed of sound.[23]

Since we are interested in flat space solutions, $\kappa = 0$. Furthermore, for matter with equation of state $p = -\varrho c^2$, the speed of sound is the speed of light, which according to Baumann's conventions is set equal to unity, and so we obtain an equation of the form (11.28). A big advantage of Baumann's proper derivation of (11.28) is that it can be applied on superhorizon scales as well as for those within the horizon, thanks to the use of Bardeen's gauge-invariant formulation of the perturbation equation.

We recognise that (11.28) is the equation of motion for a damped harmonic oscillator. If the 'damping term' $3H\delta\dot{\phi}_k$ is set equal to zero, we find harmonic oscillations, just as in the case of the Jeans' analysis of Sect. 11.3. On the other hand, for scales much greater than the radius of the Hubble sphere, $\lambda \gg c/H$, an order of magnitude calculation shows that the damping term dominates and the velocity $\delta\dot{\phi}_k$ tends exponentially to zero, corresponding to the 'freezing' of the fluctuations on superhorizon scales.

Both x and $\delta\phi_k$ have zero point fluctuations in the ground state. In the case of the harmonic oscillator, we found $\langle x^2 \rangle \propto \omega^{-1}$. In exactly the same way, we expect the fluctuations in $\delta\phi_k$ to be inversely proportional to the 'angular frequency' in (11.29), that is,

$$\langle (\delta\phi_k)^2 \rangle \propto \frac{1}{k/a} \propto \lambda \,, \tag{11.31}$$

where λ is the proper wavelength. Integrating over wavenumber, we find the important result

$$\langle (\delta\phi)^2 \rangle \propto H^2 \,. \tag{11.32}$$

At the end of the inflationary expansion, the scalar field is assumed to decay into the types of particles which dominate our Universe at the present epoch, releasing a vast amount of energy which reheats the contents of the Universe to a very high temperature. The final step in the calculation is to relate the fluctuations $\delta\phi$ to the density perturbations in the highly relativistic plasma in the post-inflation era. In the simplest picture, we can think of this transition as occurring abruptly between the era when $p = -\varrho c^2$ and the scale factor increases exponentially with time, as in the de Sitter metric, to that in which the standard relativistic equation of state $p = \frac{1}{3}\varrho c^2$ applies with associated variation of the inertial mass density with cosmic time $\varrho \propto H^2 \propto t^{-2}$ (see (9.7)). Guth and Pi used the time-delay formalism which enables the density perturbation to be related to the inflation parameters (Guth and Pi, 1982) (see Sect. 11.6.1). The end results is

$$\frac{\delta\varrho}{\varrho} \propto \frac{H_*^2}{\dot{\phi}_*} \,. \tag{11.33}$$

where H_* and ϕ_* are their values when the proper radius of the perturbation is equal to the Hubble radius.

This order of magnitude calculation illustrates how quantum fluctuations in the scalar field ϕ can result in density fluctuations in the matter which all have the more or less the same amplitude when they passed through the horizon in the very early Universe. They then remained frozen in until they re-entered the horizon very much later in the radiation-dominated era, as illustrated in Fig. 11.6*b*.

This schematic calculation is only intended to illustrate why the inflationary paradigm is taken so seriously by theorists. It results remarkably naturally in the Harrison-Zeldovich spectrum for the spectrum of primordial perturbations.

In the full theory, the values of the small parameters ϵ and η defined by (11.23) cannot be neglected and they have important consequences for the spectrum of the perturbations and the existence of primordial gravitational waves. Specifically, the spectral index of the perturbations on entering the horizon is predicted to be

$$n_{\rm S} - 1 = 2\eta - 6\epsilon \, . \tag{11.34}$$

Furthermore, tensor perturbations, corresponding to gravitational waves, are also expected to be excited during the inflationary era. Quantum fluctuations generate quadrupole perturbations and these result in a similar almost scale-invariant power spectrum of perturbations. Their spectral index is predicted to be

$$n_{\rm T} - 1 = -2\epsilon \, , \tag{11.35}$$

where scale-invariance corresponds to $n_{\rm T} = 1$. The tensor-to-scalar ratio is defined as

$$r = \frac{\mathit{\Delta}^2_{\rm T}}{\mathit{\Delta}^2_{\rm S}} = 16\epsilon \, , \tag{11.36}$$

where $\mathit{\Delta}^2_{\rm T}$ and $\mathit{\Delta}^2_{\rm S}$ are the power spectra of tensor and scalar perturbations respectively.

These results illustrate why the deviations of the spectral index of the observed perturbations from the value $n_{\rm S} = 1$ are so important. The fact that best fit value $n_{\rm S} = 0.961^{+0.018}_{-0.019}$ is slightly, but significantly, less than one suggests that there may well be a background of primordial gravitational waves. The detection of a background of gravitational waves is really a very great observational challenge, but they provide a remarkably direct link to processes which may have occurred during the inflationary epoch. To many cosmologists, this would be the 'smoking gun' which sets the seal on the inflationary model of the early Universe.

Whilst the above calculation is a considerable triumph for the inflationary scenario, we should remember that there is as yet no physical realisation of the scalar field. Although the scale-invariant spectrum is a remarkable prediction, the amplitude of the perturbation spectrum is model dependent. There are literally hundreds of possible inflationary models depending upon the particular choice of the inflationary potential. We should also not neglect the possibility that there are other sources of perturbations which could have resulted from various types of topological defect, such as cosmic strings, domain walls, textures and so on (Shellard, 2003). Granted all these caveats, the startling success of the inflationary model in accounting for the observed spectrum of fluctuations in the Cosmic Microwave Background Radiation has made it the model of choice for studies of the early Universe.

11.8 Topological Defects

Throughout the 1980s and 1990s the most important alternative account of the origins of structure was based on topological defects. These ideas were first studied in the 1970s prior to the introduction of the concepts of inflation, as a general feature of spontaneous symmetry-breaking phase transitions in the early universe. Several theorists took up the challenge of understanding whether defects formed in the early universe could produce the appropriate seeds for structure formation.[24]

Starting in the early 1970s the ideas of spontaneous symmetry breaking were applied to cosmology. Extrapolating the Friedman-Lemaître models back to very early epochs, the early

universe reaches arbitrarily high temperatures at early times. Kirzhnits (1972) suggested that symmetries in particle physics would be restored at sufficiently high temperatures, by analogy with symmetry restoration in condensed matter systems. Further calculations of symmetry restoration in the Standard Model of particle physics supported the idea that as the universe cooled it passed through a series of phase transitions that broke the symmetries between various interactions. Many symmetry breaking phase transitions in condensed matter systems lead to the formation of topological defects, such as vortices in liquid helium and so it is natural to expect that defects also arise in early universe phase transitions.

In a seminal paper, Kibble (1976) argued that topological defects would be produced as a result of the horizon structure of the early universe. Given that the correlation length of the order parameter is bounded by the horizon distance, the phase transition produces domains in which the order parameter takes on different values determined by random fluctuations. The implication is that there must be a "defect," namely a region of space in which the fields cannot reach the vacuum state, and instead remain trapped in a state of higher energy. The nature of these regions of higher energy is fixed by the structure of the manifold. In the case of a non-simply connected vacuum manifold, the phase transition leads to two-dimensional defects called "cosmic strings." There are several other possibilities. A phase transition breaking a discrete symmetry leads to regions in which the order parameter takes on discrete values separated by domain walls, which are three-dimensional surfaces in space-time. If the vacuum manifold has non-contractible two-spheres rather than circles, then the phase transition produces point-like defects, such as magnetic monopoles; for non-contractible three-spheres the corresponding zero-dimensional defects are called "textures," event-like defects that do not have a stable localized core.

Early studies showed that domain walls and some types of monopoles had disastrous consequences, conflicting with observational constraints by several orders of magnitude (see, for example, Zeldovich et al. 1975; Zeldovich and Khlopov 1978; Guth and Tye 1980). However, other types of defects – in particular, cosmic strings – were more plausible candidates for the seeds of structure formation. The defects are inherently stable regions of higher energy density, whose scale is set by the energy scale of the phase transition. The defects have an important impact on the dynamical evolution of the system following the phase transition, and in particular it is plausible that they provide seeds that are subsequently enhanced by gravitational instability, as described by linear perturbation theory. These theories passed an important initial test in that they lead to an approximately scale-invariant Harrison-Zeldovich spectrum of perturbations, compatible with the first generation of Cosmic Microwave Background Radiation observations and the general picture of structure formation described above. However, there are important general differences between the inflationary account and that provided by topological defects, and these were clarified by a substantial research effort throughout the 1980s and 1990s.

To determine whether topological defects suffice as the primary mechanism for producing seeds for structure formation, researchers had to tackle two challenging problems. The first was to describe the phase transition itself and determine the nature of the defects produced with sufficient quantitative detail to determine the consequences for the later stages of evolution. Second, one had to describe the subsequent evolution of the network of defects left over following the phase transition over a wide range of dynamical scales. Throughout the 1980s, for example, the general picture of how strings seeded galaxy formation changed considerably in light of numerical simulations establishing details regarding the size of typical closed loops

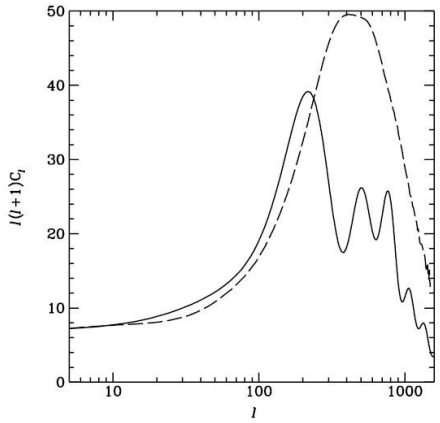


Fig. 11.7 This figure (from Albrecht et al. 1996) shows the predicted angular power spectrum of temperature fluctuations in the Cosmic Background Radiation from a particular model of cosmic strings (dashed line), and a generic inflationary model (solid line).

of strings and the behavior of open strings. These two problems are exacerbated by uncertainty regarding the relevant fundamental physics. The details of the phase transitions depend on specific features of the physics – specifically concerning proposed extensions of the Standard Model.

Despite these difficulties, by about 1997 there was a consensus regarding the generic consequences of structure formation through defects and the contrast with the consequences of inflation. Perturbations produced in defect theories "decohere", as first noted by Albrecht et al. (1996), in the sense that fluctuations at all wave-numbers are not in phase. This is a consequence of the non-linear evolution of the source term, which leads to mixing of perturbations across different modes. The perturbations are also non-Gaussian due to the correlations that this mixing produces between perturbations. Finally, defects generate scalar, vector, and tensor perturbations of roughly equal magnitude.

The most striking contrast with the inflationary theories is that inflation leads to phase coherence of the perturbations because the dynamics leads to synchronization of the Fourier modes with the consequent prediction of Doppler peaks. The position of the first peak also differs between the inflationary and topological defect models, with defect models generally predicting a primary peak at a larger multipole moment ($l \geq 300$) than inflation ($l \approx 200$). Observational results starting in the late 1990s and culminating in the WMAP results provided decisive support for inflation with respect to both of these features.

In addition to the physical contrast between the mechanisms for structure formation, there are important methodological contrasts between the two approaches. First, despite uncertainty regarding the detailed physics of the phase transitions, the account of structure formation via defects is sufficiently constrained by general theoretical principles to produce specific observational signatures. Physicists working on defects often highlighted this rigidity as a virtue

of the theory, characterizing it as "falsifiable" in a Popperian sense. Second, accounts based on topological defects do not address the problems related to initial conditions highlighted by Guth. Those who accepted Guth's approach to fine-tuning and initial conditions could still use defects, however. Inflation could still be invoked to solve the problems related to initial conditions (see, for example, Vilenkin and Shellard 2000), as long as inflation set the stage for subsequent phase transitions that would produce appropriate topological defects.

11.9 Baryogenesis

A key contribution of particle physics to studies of the early Universe concerns the baryon-asymmetry problem, a subject referred to as *baryogenesis*. In a prescient paper of 1967, Sakharov enunciated the three conditions necessary to account for the baryon-antibaryon asymmetry of the Universe (Sakharov, 1967). *Sakharov's rules* for the creation of non-zero baryon number from an initially baryon symmetric state are:

- *Baryon number* must be violated;
- C (charge conjugation) and CP (charge conjugation combined with parity) must be violated;
- The asymmetry must be created under *non-equilibrium conditions*.

The reasons for these rules can be readily appreciated from simple arguments (Kolb and Turner, 1990). Concerning the first rule, it is evident that, if the baryon-asymmetry developed from a symmetric high temperature state, baryon number must have been violated at some stage – otherwise, the baryon-asymmetry would have to be built into the model from the very beginning. The second rule is necessary in order to ensure that a net baryon number is created, even in the presence of interactions which violate baryon conservation. The third rule is necessary because baryons and antibaryons have the same mass and so, thermodynamically, they would have the same abundances in thermodynamic equilibrium, despite the violation of baryon number and C and CP invariance.

There is evidence that all three rules can be satisfied in the early Universe from a combination of theoretical ideas and experimental evidence from particle physics. Thus, baryon number violation is a generic feature of Grand Unified Theories which unify the strong and electroweak interactions – the same process is responsible for the predicted instability of the proton. C and CP violation have been observed in the decay of the neutral K^0 and $\bar{K}^0$ mesons. The K^0 meson should decay symmetrically into equal numbers of particles and antiparticles but, in fact, there is a slight preference for matter over antimatter, at the level of 10^{-3}, very much greater than the degree of asymmetry necessary for baryogenesis, $\sim 10^{-8}$. The need for departure from thermal equilibrium follows from the same type of reasoning which led to the primordial synthesis of the light elements. As in that case, so long as the time-scales of the interactions which maintained the various constituents in thermal equilibrium were less than the expansion time-scale, the number densities of particles and antiparticles of the same mass would be the same. In thermodynamic equilibrium, the number densities of different species did not depend upon the cross-sections for the interactions which maintain the equilibrium. It is only after decoupling, when non-equilibrium abundances were established, that the number densities depended upon the specific values of the cross-sections for the production of different species.

In a typical baryogenesis scenario, the asymmetry is associated with some very massive boson and its antiparticle, $X, \overline{X}$, which are involved in the unification of the strong and electroweak forces and which can decay into final states which have different baryon numbers. Kolb and Turner provided a clear description of the principles by which the observed baryon-asymmetry can be generated at about the epoch of grand unification or soon afterwards, when the very massive bosons can no longer be maintained in equilibrium (Kolb and Turner, 1990). Although the principles of the calculations are well defined, the details are not understood, partly because the energies at which they are likely to be important are not attainable in laboratory experiments, and partly because predicted effects, such as the decay of the proton, have not been observed. Thus, although there is no definitive evidence that this line of reasoning is secure, well-understood physical processes of the type necessary for the creation of the baryon-antibaryon asymmetry exist. The importance of these studies goes well beyond their immediate significance for astrophysical cosmology. As Kolb and Turner remark,

> . . . in the absence of direct evidence for proton decay, baryogenesis may provide the strongest, albeit indirect, evidence for some kind of unification of the quarks and the leptons.

11.10 The Planck Era

Enormous progress has been made in understanding the types of physical process necessary to resolve the basic problems of cosmology, but it is not clear how independent evidence for them can be found. The methodological problem with these ideas is that they are based upon extrapolations to energies vastly exceeding those which can be tested in terrestrial laboratories. Cosmology and particle physics come together in the early Universe and they boot-strap their way to a self-consistent solution. This may be the best that we can hope for but it would be preferable to have independent constraints upon the theories.

A representation of the evolution of the Universe from the Planck era to the present day is shown in Fig. 11.8. The *Planck era* is that time in the very remote past when the energy densities were so great that a quantum theory of gravity is needed. On dimensional grounds, this era must have occurred when the Universe was only about $t_{\rm Pl} \sim (hG/c^5)^{1/2} \sim 10^{-43}$ s old. Despite enormous efforts on the part of theorists, there is no quantum theory of gravity and so we can only speculate about the physics of these extraordinary eras.

Being drawn on a logarithmic scale, Fig. 11.8 encompasses the evolution of the whole of the Universe, from the Planck area at $t \sim 10^{-43}$ s to the present age of the Universe which is about 4×10^{17} s or 13.6×10^9 years old. Halfway up the diagram, from the time when the Universe was only about a millisecond old, to the present epoch, we can be confident that the Big Bang scenario is the most convincing framework for astrophysical cosmology.

At times earlier than about 1 millisecond, we quickly run out of known physics. This has not discouraged theorists from making bold extrapolations across the huge gap from 10^{-3} s to 10^{-43} s using current understanding of particle physics and concepts from string theories. Some impression of the types of thinking involved in these studies can be found in the ideas expounded in the excellent volume *The Future of Theoretical Physics*, celebrating the 60th birthday of Stephen Hawking (Gibbons et al., 2003). Maybe many of these ideas will turn out to be correct, but there must be some concern that some fundamentally new physics will emerge at higher and higher energies before we reach the GUT era at $t \sim 10^{-36}$ s and the Planck era at $t \sim 10^{-43}$ s. This is why the particle physics experiments being carried out at

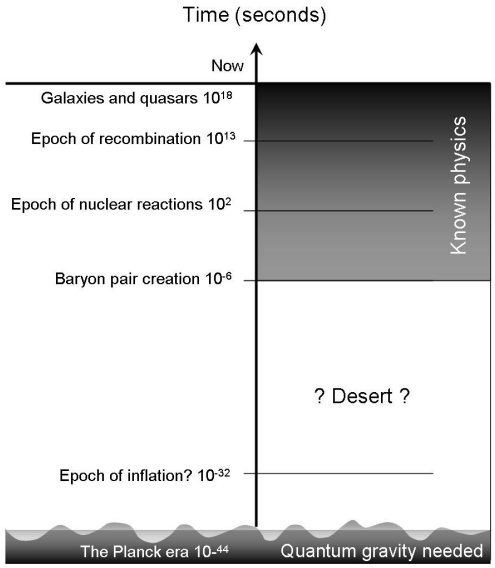


Fig. 11.8 A schematic diagram illustrating the evolution of the Universe from the Planck era to the present time. The shaded area on the right of the diagram indicates the regions of known physics.

the Large Hadron Collider at CERN are of such importance for astrophysics and cosmology. The discovery of the Higgs boson was a real triumph (Aad et al., 2012), providing essential support for our understanding of the standard model of particle physics. In addition, there is the possibility of discovering new types of particles, such as the lightest supersymmetric particle or new massive ultra-weakly interacting particles, as the accessible range of particle energies increases from about 100 GeV to 1 TeV. These experiments should provide clues to the nature of physics beyond the standard model of particle physics and will undoubtedly feed back into understanding of the physics of the early Universe.

It is certain that at some stage a quantum theory of gravity is needed which may help resolve the problems of singularities in the early Universe. The singularity theorems of Penrose and Hawking show that, according to classical theories of gravity under very general conditions, there is inevitably a physical singularity at the origin of the Big Bang, that is, as $t \to 0$, the energy density of the Universe tends to infinity. However, it is not clear that the actual Universe satisfies the various energy conditions required by the singularity theorems, particularly if the negative pressure equation of state $p = -\varrho c^2$ holds true in the very early Universe. All these considerations show that new physics is needed if we are to develop a convincing physical picture of the very early Universe.

Notes

[1]The contents of the book *Galaxy Formation* (Longair, 2008) by one of us (MSL) has been used extensively in preparing this chapter. It may be consulted for more of the technical details of the observations and the theoretical background.

[2]The method Zwicky used to estimate the total mass of the cluster had been derived by Eddington in 1916 to estimate the masses of star clusters. Eddington derived the *virial theorem* which relates the total internal kinetic energy T of the stars or galaxies in a cluster to the total gravitational potential energy, $|U|$, assuming the system to be in a state of statistical equilibrium under gravity (Eddington, 1916). The total mass of the cluster can be found from the virial theorem to be $M \approx 2R_{\rm cl}\langle v^2\rangle/G$. Zwicky measured the velocity dispersion $\langle v^2\rangle/$ of the galaxies in the Coma cluster and found that there was much more mass in the cluster than could be attributed to the visible masses of galaxies. In solar units of $M_\odot/L_\odot$, the ratio of mass-to-optical luminosity of a galaxy such as our own is about 3, whereas for the Coma cluster the ratio was found to be about 500. In other words, there must be about 100 times more dark, or hidden, matter as compared with visible matter in the cluster.

[3]The significance of these flat rotation curves can be appreciated from the following argument. For simplicity, assume that the distribution of mass in the galaxy is spherically symmetric, so that we can write the mass within radius r as $M(\leq r)$. According to Gauss's law for gravity, we can then find the radial acceleration at radius r by placing the mass within radius r, $M(\leq r)$, at the centre of the galaxy. Then, equating the centripetal acceleration at radius r to the gravitational acceleration, we find $M(\leq r) = v_{\rm rot}^2(r)\, r/G$. If the rotation curve of the spiral galaxy is flat, $v_{\rm rot}$ = constant, $M(\leq r) \propto r$ and so the mass within radius r increases linearly with distance from the centre. This contrasts dramatically with the distribution of light in the discs, bulges and haloes of spiral galaxies which decrease exponentially with increasing distance from the centre.

[4]See the discussion of Sect. 10.10.1.

[5]An outline of the physics involved is given in Sect. 13.3 of *Galaxy Formation* (2008).

[6]See Chap. 8.

[7]The reasons for this are illustrated in Sect. 11.4.2 of the book *Galaxy Formation* (2008). See also Fig. 14.10 of that text.

[8]Note that this is only one aspect of fine-tuning in order to ensure that there are observers capable of asking these questions in the Universe. See chap. 13.

[9]More exactly, the equation of state for a Higgs field takes this form if derivative terms are negligible and the effective potential is displaced from its true minima. The stress-energy tensor for the scalar field takes the form:

$$T_{ab} = \nabla_a\phi\nabla_b\phi - \frac{1}{2}g_{ab}\left(g^{cd}\nabla_c\nabla_d\phi - V(\phi)\right). \tag{11.37}$$

If the first two terms are negligible, we find $T_{ab} \approx \frac{1}{2}g_{ab}V(\phi)$, which is the equation of state discussed in the main text. See also Sect. 11.7.

[10]This calculation of the theoretical value of the cosmological constant was first carried out by Wolfgang Pauli in the 1930s, but he did not take the result seriously. See, for example, Rugh and Zinkernagel (2000).

[11]Donald Rumsfeld was President George W. Bush's United States Secretary of Defense and played a central role in the planning of the United States' response to the September 11 attacks, which included two wars, one in Afghanistan and one in Iraq.

[12]For a more detailed discussion of these topics, see Sect. 12.2 of *Galaxy Formation* (Longair, 2008).

[13]These earlier works, particularly the work of the Soviet theorists Sakharov, Zeldovich, Starobinsky and their colleagues are surveyed in detail by Smeenk (2005).

[14]A popular account of the history of the development of ideas about the inflation picture of the early Universe is contained in Guth's book *The Inflationary Universe: The Quest for a New Theory of Cosmic Origins* (Guth, 1997). The pedagogical review by Lineweaver can also be recommended. He adopts a somewhat sceptical attitude to the concept of inflation and our ability to test inflationary models through confrontation with observations (Lineweaver, 2005).

[15]In fact, the situation is somewhat more complex than this simple picture. Although the matter and radiation in the very early universe would have been homogenised on the small-scale, the matter and energy densities of everything other than the inflaton field are rapidly diluted during inflationary expansion – pre-existing matter and radiation are dynamically irrelevant after the first few e-folds. What explains the uniformity of different regions is instead the fact that the inflaton field decays and reheats the universe in the same fashion in different regions. Why then do the temperatures of the CMB in different patches of the sky agree? The inflaton field had a homogeneous state over some region, triggered an inflationary state that proceeded in the same way throughout this region, and then decayed into other types of matter and energy at the end of inflation. But, inflation does not guarantee that the outcome of inflation is a smooth universe by itself – it instead magnifies any non-uniformities that exist at much shorter length scales to cosmologically relevant scales (See the discussion of Sect. 11.7). The simple picture has to be combined with an assumption, which has generally agreed to be quite plausible, about the small-scale non-uniformities in a pre-inflationary patch.

[16]This section draws upon material contained in Smeenk (2005) and Smeenk (2018). See Vilenkin and Shellard

(2000) for a masterful overview of this line of research, which includes more detailed discussion of the historical development of the field and references to original papers.

[17] See Sect. 11.7.1

[18] Olive (1990), pp. 389

[19] There are now several recommendable books on this subject (Liddle and Lyth, 2000; Dodelson, 2003; Mukhanov, 2005).

[20] The pedagogical exposition by Baumann (2007) is a very helpful guide.

[21] See also footnote 9.

[22] See Longair (2008), Sec. 20.5.5.

[23] Equation (11.29) is often referred to as the Mukhanov-Sasaki equation for the evolution of linearised perturbations on sub- and super-horizon scales during the inflationary era.

[24] This section draws upon material contained in Smeenk (2005) and Smeenk (2018).