Chapter 19: **Cosmology**

- general relativity and the curvature of space; black holes
- elementary particles and high-energy physics
- expansion of the universe
- timeline of the history of the universe
- the future; string theory; a few unanswered questions
- sample questions

General Relativity and the curvature of space; black holes

Recall that the distinction between special and general relativity (see, e.g., the introductory chapter) is between uniform motion and accelerated motion. Special relativity is associated with the equivalence between energy and rest mass, the famous $E = mc^2$. General relativity is associated with the equivalence between gravitational mass and inertial mass. For example, the effect of being accelerated at 9.8 m/s^2 in a spacecraft well away from any gravitational influences is equivalent to experiencing the 9.8 m/s^2 acceleration of gravity at the surface of the Earth. In relativity, both special and general, space and time are woven together into one unified fabric of spacetime. A rule for measuring the distances between two points in spacetime is called a metric, a "Pythagorean theorem" for four dimensions. In general relativity that fabric of spacetime is not necessarily flat. The presence of energy density warps spacetime. Warped spacetime tells particles how to move, following a path called a geodesic. Because spacetime can be warped and because of this coupling between the warping and the geodesics, the equations of general relativity are not always easy to work with. They are what are called coupled differential equations and involve 4-dimensional tensors. The latter are matrices, in this case containing expressions for density and pressure and for the terms in the metric.

General relativity explicitly permits space to be curved, meaning the geometry could be different from the flat Euclidean geometry most of us were taught in school. Stick to two dimensions for a moment, where it is easier to visualize what's going on. Positively curved space is like the surface of a sphere. Imagine two sailors, starting south from the equator, each following a line of longitude. At the equator those look like parallel lines. Unlike on a flat surface, however, those lines do not remain equidistant; our travelers meet at the south pole. Or, consider a triangle. This only requires one traveler: start out heading west along the equator from, say, 80 degrees west longitude, just off Ecuador; at, say, 155-ish degrees west, take a 90 degree turn to the north and follow a line of longitude to the north pole, passing Hawaii on the way; at the north pole, take another 90 degree turn and follow the line of longitude at 80 degrees west to return to the equator at your starting point. We just completed a triangle containing *three* 90-degree angles. That isn't possible on a flat surface. Negatively curved two-dimensional surfaces look more like a saddle; starting from the narrowest point of a saddle, "parallel" line diverge, and the sum of the internal angles in a triangle is *less* than 180 degrees. The figures below illustrate a triangle on a positively curved surface and two parallel lines diverging on a negatively curved surface. Remember, these are *two*-dimensional surfaces. Good luck trying to visualize how this works in *four* dimensions!

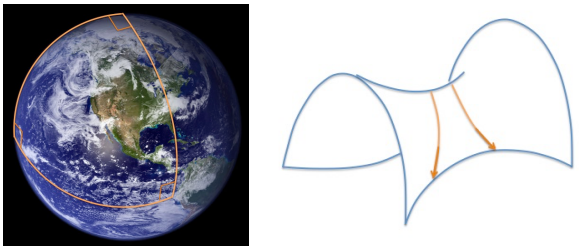


Figure 19.1: a) positively (left) and b) negatively curved (center) two-dimensional surface

http://eoimages.gsfc.nasa.gov/images/imagerecords/57000/57723/globe_west_540.jpg

In introducing the expansion of the universe we considered the view from another galaxy and the argument that anyone anywhere looking in any direction would see the same thing. This is a fairly important principle in cosmology, one that is going to make Einstein's equations of general relativity much more tractable. In the context of the solar system, the principle that we do not occupy a special, privileged location is called the Copernican Principle; in the context of the universe as a whole it is called the Cosmological Principle. There are a couple of caveats: First, we *do* occupy a special location in *time*. The existence of the Cosmic Microwave Background tells us that the universe used to be hotter and denser than it is now. It has not been the same at all times in the past nor will it be the same in the future as it is now. Second, scale matters. One cubic meter of rock is obviously not the same as one cubic meter at the center of the Sun or as one cubic meter in the empty space between the planets. One cubic parsec containing a solar system is different from one cubic parsec of interstellar space; one cubic Mpc containing several galaxies is different from one cubic Mpc in one of the voids between superclusters. But a cube 100 Mpc on a side is pretty much like any other cubic 100 Mpc box, meaning that although there is quite a bit of variety within that box, its average energy density is about the same as the average energy density of any other similar-sized chunk of universe sampled at the same time. There are two terms that distinguish two parts of this concept, namely *homogenous*, meaning the same in any place, and *isotropic*, meaning the same in any direction.

Math note: In a universe that is isotropic and homogeneous, path lengths may be described by the Robertson-Walker metric:

$$ds^2 = -c^2dt^2 + a(t)^2[dr^2 + S_k(r)^2 d\Omega^2],$$

where *a(t)* is the scale factor (= 1 at the present time), *k* represents the curvature and is either +1, 0, or –1, and

$$d\Omega^2 = d\theta^2 + \sin^2\theta d\varphi^2$$

$$S_k(r) = R\sin(r/R)_{k=+1};\ = r_{k=0};\ = R\sinh(r/R)_{k=-1}$$

where θ, φ are the polar and azimuthal angular coordinates and the function S_k provides the appropriate multiplier for the angular portion of the metric for the three possible cases of positive, flat, and negative curvature.

As we attempt to determine what the universe is like at large distances we need to keep in mind the fact that we are seeing the universe the way it was at some time in the past and that space might not be flat. Keep in mind also the fact that space locally, e.g., around a black hole, may be curved quite differently than the overall space that makes up the universe on large scales. The universe as a whole could be flat even though light passing a massive cluster of galaxies will follow a curved path due to the fact that the fabric of spacetime near the galaxies is warped by the mass in the cluster. Let's look at two examples, first, how the distribution of distant structures varies with global curvature, and second, gravitational lensing, that warping of spacetime, near a cluster of galaxies.

Global curvature. Suppose that we could observe structures in the distant universe whose linear size we knew. It isn't necessarily a given that this is possible; we know, for instance, that early galaxies were not quite like the galaxies today. Perhaps, though, galaxies of a given mass would be about the same size. Another possibility has to do with the anisotropies in the Cosmic Microwave Background radiation. The early universe wasn't totally smooth; if there hadn't been clumps where the mass density was higher than average, galaxies would not have formed. The relic information that a region of space was more or less dense is preserved as slight temperature variations in the CMB (more on this below). *If* we had reason to believe that we could predict theoretically how large those density enhancements were when the universe was ~380,000 years old (the time we are sampling when we observe the CMB) then we would have something of known size here as well. An object whose linear size we would know is referred to as a standard ruler.

How would our observations of the *angular* size of these structures of known *linear* size differ depending on the global curvature of the universe? Here is a figure to illustrate what would happen. Imagine that we are looking out into space and counting the number of galaxies or clumps in the CMB, whatever our known linear structure is, in a given angle on the sky. The blue triangle represents a flat universe. There are 5 plus signs along the far side of the blue triangle. In a negatively curved universe, the reddish lines that represent the sides of our triangle will diverge as we get farther away. We count more than 5 plus signs in the same angular chunk of sky. Conversely,

in a positively curved universe, the greenish lines representing the sides of our triangle will converge and we will see fewer objects in a given angle on the sky.

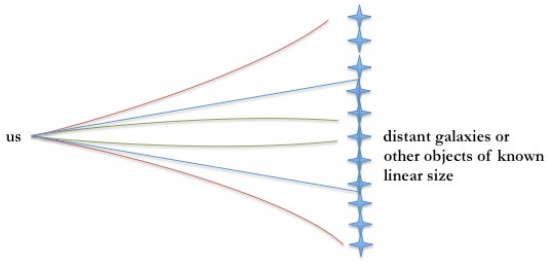


Figure 19.2:
Variation of angular size for objects of a given linear size in different geometries

Of course the extent of the difference is exaggerated quite a lot in this sketch!

Gravitational lensing is the term for for the fact that light passing a massive object will follow the local curvature of spacetime, making it appear to us as though the light's path had been bent. It's called lensing because it is similar to what happens when light is refracted by a glass lens. One of the earliest opportunities to confirm the predictions of general relativity came about during the total solar eclipse of May 1919. Astronomer Sir Arthur Eddington photographed the eclipse from the eastern Atlantic island of Principe. Stars are visible quite near the limb of the Sun during a solar eclipse because the Moon blocks the light from the Sun's photosphere (near, not exactly at, the limb of the Sun, because of the brightness of the corona). The Sun's mass warps space. Positions of known stars appear to be offset away from the limb of the Sun because the path of the starlight passing the Sun follows a curved path. Here, greatly exaggerated, is what happens:

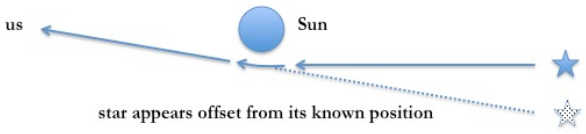


Figure 19.3:
Gravitational lensing by the Sun

Of interest for cosmology, foreground galaxies or clusters of galaxies will act as gravitational lenses for more distant galaxies. If the alignment of the foreground and background objects is almost perfect, the background galaxy will appear as a ring (called an Einstein Ring) around the foreground object. The following is a montage of such rings observed with the HST:

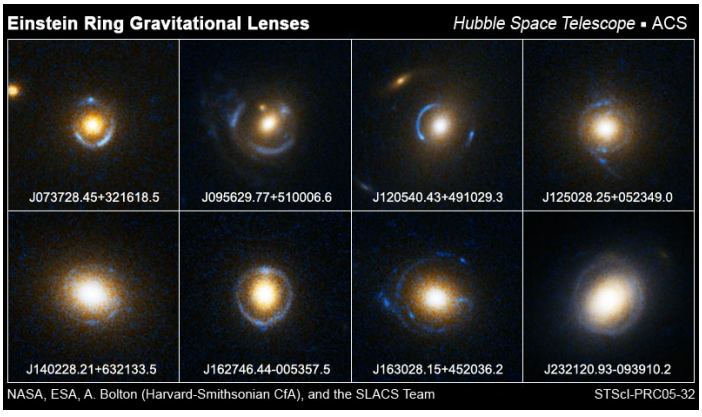
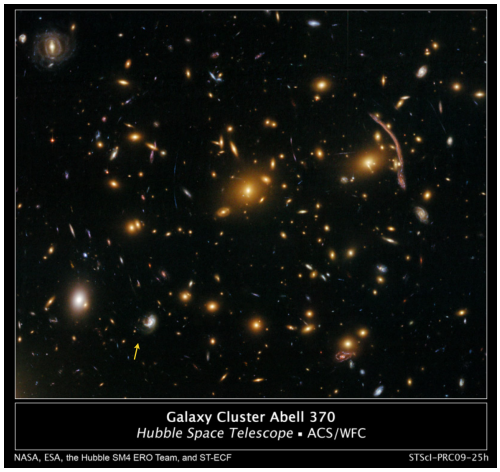


Figure 19.4: Einstein rings;
each image is 8 arcseconds wide.
http://hubblesite.org/newscenter/archive/releases/2005/32/image/a/

A team of astronomers actively looked for lensing by foreground giant elliptical galaxies using the Sloan Digital Sky Survey to identify candidate objects and the HST's Advanced Camera for Surveys (ACS) to capture deep enough images to confirm the lens. The foreground galaxies are themselves 2 to 4 billion light years away, with redshifts

ranging from z = 0.08 (lower right galaxy) to 0.32 (upper left). The background objects have redshifts ranging from z = 0.47 to 0.79.

More extended foreground objects such as galaxy clusters, possibly lensing galaxy clusters or background galaxies at a range of distances, create more complicated patterns. Sometimes the lensed objects are more easily identified than others. The following HST image is of a galaxy cluster called Abell 370, z = 0.375. In the central panel is an expanded view of a more distant spiral galaxy nicknamed "the Dragon". The background objects are not all at the same distance and most don't show up so clearly. At the other extreme of identifiability, one of the first galaxies to be found with a $z > 6$ is hiding just off one of the foreground galaxies, indicated by the arrow in the HST image. It's called HCM-6A and it's not visible in this image. The right-hand image shows that this background source emits in the hydrogen Lyman-α line. That's emitted at 121.6 nm; here, redshifted by 6.56, Ly-α shows up in the near IR. This image is taken with an 11.8-nm-wide filter centered at 915.2 nm using the Keck 10-m telescopes.



http://hubblesite.org/newscenter/archive/releases/2009/25/image/ap

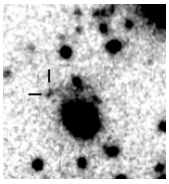


Figure 19.5 c): HMC-6A; E.M. Hu et al., 2002, *Ap.J.* **568**, pp. L75-L79

Figure 19.5 a): Gravitational lens Abell 370; b): detail

Example: check this redshift:

$z = \Delta\lambda/\lambda$. Here, $\Delta\lambda = 6.56 \cdot 121.6$ nm = 797.7 nm

121.6 nm + 797.7 nm = 919.3, which is in the filter bandpass centered at 915.2 nm.

One of the benefits of lensing, and magnifying, distant galaxies is that we might be able to catch galaxies in the act of forming from smaller clumps. The following images are of a cluster called Abell 2218; the colorful distended "tadpoles" are the background objects. In the inset image blow-up are two images of one tiny piece of the background galaxies; this little red dot appears to be an object with only about a million stars. The foreground cluster is about 600 Mpc away; this background object is ~4 Gpc away (i.e., ~13 billion light years). The fact that images of many foreground clusters include a number of lensed objects is what allows us to determine the distribution of mass, especially the dark matter, in the foreground cluster.

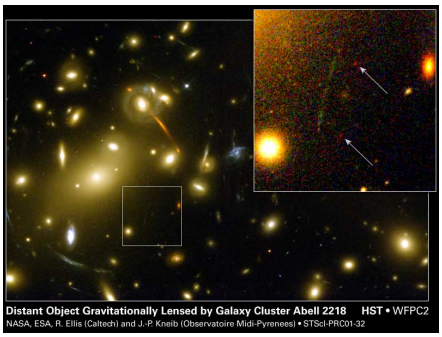


Figure 19.6: Abell 2218; observations by HST and the 10-m Keck telescopes in Hawaii.

http://hubblesite.org/newscenter/archive/releases/2001/32/image/a/

In 1993 R.A. Hulse and J.H. Taylor, Jr., won the Nobel Prize in Physics for their study of PSR 1913+16, a binary system of two neutron stars, one of which is a pulsar, first described by Hulse and Taylor in 1975. Neutron stars are remnants of some types of supernovae. Pulsars are rapidly rotating neutron stars with strong magnetic fields that emit electromagnetic radiation beamed along the axis of the magnetic field. If that beam sweeps past us periodically, like the beam from a lighthouse, we will receive periodic pulses in synch with the rotation period of the pulsar. In this particular case the pulsar rotation period is 59 milliseconds. This pair of neutron stars, each with a mass of about 1.4 solar masses, is in a close orbit, with a period of 7.75 *hours*. Because of the pulsar we can observe the fact that the orbit is decaying: the pulses arrive a bit ahead and then a bit behind as the pulsar's orbit takes it toward us and then away and over time the orbit has sped up. The system is losing energy precisely as predicted by general relativity.

Just as accelerated electric charges radiate electromagnetic radiation, accelerated masses should emit gravitational radiation. In particular, two massive objects, such as neutron stars or black holes, in a tight orbit should emit gravitational energy, causing the fabric of spacetime itself to stretch and oscillate as the wave passes.
The gravitational waves themselves were first detected in the fall of 2015, almost exactly 100 years after Einstein published his theory of general relativity. As described in the introductory chapter, in 2015 LIGO, the Laser Interferometer Gravitational wave Observatory, detected the merger of two 30-ish solar mass black holes (and, in 2017, the merger of two neutron stars). The energy lost by the black holes in the few milliseconds before they merged was equivalent to about 3 solar masses. That energy caused the fabric of spacetime to ripple enough to be detected in the miniscule stretching of the arms of the LIGO interferometer. This first detection was of black holes merging about 1.3 billion years ago, but it is possible both that we will soon be able to detect gravitational wave events from the more distant past and that gravitational waves in the early universe might have left their imprint on the light we see today as the Cosmic Microwave Background.

It's not too hard to visualize what should happen. Consider a round region of space and imagine what will happen to it as a gravitational wave moves through:

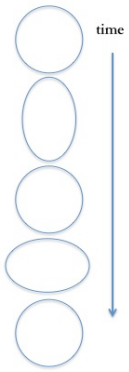


Figure 19.7: As a gravitational wave passes, moving into the plane of the page, space will oscillate, expanding in one direction / contracting in the perpendicular direction and then the opposite.

Because two polarizations are possible for gravitational waves, there's also a wave oscillating 45 degrees offset from this one. The wave amplitudes are often denoted h_+ and h_x to indicate the oscillations in the *x*-*y* directions and the 45° offset directions.

Black holes take the curvature of space to the extreme. Everywhere within the event horizon, or the Schwarzschild radius, of a black hole the escape speed is greater than the speed of light. Since escape speed is given by $v_{esc} = \sqrt{2GM/r}$, the radius at which the escape speed = *c* is

$$R_S = 2GM/c^2 .$$

For an object with the mass of the Sun, that's about 3 km. For the 4.3 million solar mass black hole at the center of the Milky Way, it's 4.3 million times larger, which is still less than 1/10 AU.

For an external observer the event horizon seems to be a place where time stops. Imagine an observer, safely far away from a black hole, but near enough to watch a probe falling toward the event horizon. Imagine further that the probe has an internal clock and a strobe light that flashes outward every second. The external

observer measures the pulses of light less and less frequently as the probe falls toward the black hole. Relative to the outside world, the probe's clock is ticking slow due to the gravitational field it experiences. The wavelength of the light also lengthens; it is gravitationally redshifted, losing energy as it climbs out of the gravitational potential well of the black hole. The last pulse sent by the probe as it crosses the event horizon hangs there forever, the light simultaneously moving outward at the speed of light and falling inward toward the black hole at the same speed. The probe, were it sentient, wouldn't notice anything odd as it crossed the event horizon. (Maybe. . .there are some calculations that suggest that the event horizon would be incredibly hot, like a "ring of fire", vaporizing anything that got close.) Its fate would now be sealed, though, constrained to move forward toward the singularity at the center of the black hole.

Singularity is the term for a non-rotating massive object of zero volume and infinite density. If we could imagine taking the mass of the Sun and crunching it down to smaller than its Schwarzschild radius, we know of no force that could then stop it from collapsing to a point. In the case of a rotating black hole, the singularity is a tiny ring. Thinking again about massive objects warping the fabric of spacetime, the black hole is the extreme case, where the warp has become a bottomless hole.

Physics note: there are a number of ways of representing what is happening to space near and inside a black hole. One is to use light cones. Here space is represented along the horizontal axis and time on the vertical.

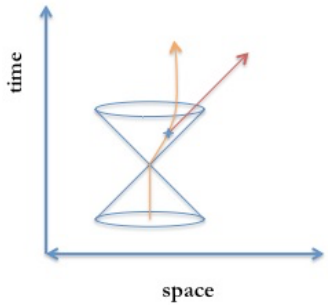


Figure 19.8: Light cone.

Here-and-now is at the intersection of the two cones. The yellow-ish line represents the world line of an object. It was stationary; in the future it moves to the right and emits a photon, indicated by the reddish arrow. Photons, moving at the speed of light, are shown by 45° lines in plots such as this. Events in the past light cone could influence what happens here-and-now.

Events outside the past light cone could not influence us. To do so would require travelling faster than the speed of light. In normal spacetime, objects can move in space but are constrained to move forward in time. Near a black hole, that changes. The light cones tilt.

At the event horizon the 45° line representing the path of a photon has tilted to be vertical, to represent the idea that a photon emitted "outward" at the event horizon would seem to hang there forever, its outward speed balanced by the rate at which it falls into the black hole. Inside the event horizon the light cone has tilted over on its side to indicate that the world line of our object now takes it inexorably forward in space toward the singularity (although at some level you could imagine it now being free to move both forward and backward in time, whatever that might mean!). And, to reiterate, this is only one way of representing space and time coordinates near a black hole.

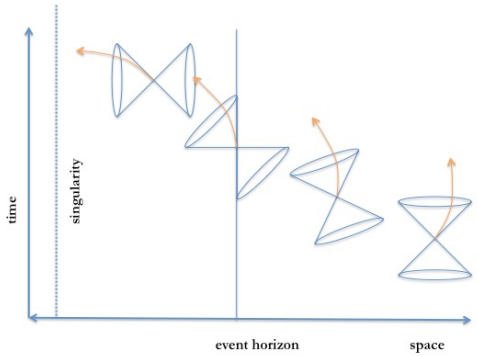


Figure 19.9: Light cones near a black hole.

Elementary particles and high-energy physics

The two dominant contributions to 20th-century physics are General Relativity, which is very successful for describing the large and massive, and Quantum Mechanics, which is very successful at describing the tiny and energetic. In the preceding section we considered some of the consequences of general relativity. Forces and elementary particles are described in the introductory chapter but let's review them here.

Fermions are particles that obey the Pauli Exclusion Principle, meaning that you can't put two of the same type in the same place at the same time doing the same thing. Bosons, such as photons, have no such restriction. There are four fundamental forces: gravity, the electromagnetic force, the strong nuclear force (which holds nucleons together), and the weak nuclear force (which governs radioactivity). Forces act on particles having the relevant charge; e.g., the electromagnetic force acts on particles that have electric charge. The charge for the strong nuclear force is called "color", in part because rather than two + / – charges there are three. Red, green, and blue would combine to produce black or white (subtractive or additive colors), which is neutral. For the weak force the charge is called "flavor". Forces are conveyed by a set of bosons; e.g., bosons called "gluons" carry the strong force, while photons carry the electromagnetic force.

Quarks are fermions with mass, color, flavor, and fractional electric charge. The two lowest mass are the up (u) quark (+2/3 electric charge) and the down (d) quark (−1/3 electric charge). A proton is composed of two ups and a down (uud) while a neutron is one up and two downs (udd). There are four higher-mass quarks, strange (s) and charm (c) and, higher mass still, bottom (b) and top (t) (called beauty and truth by some). Three-quark particles are called baryons and since protons and neutrons are ~2,000 times more massive than electrons, normal matter is often called "baryonic" matter even though it includes electrons as well as protons and neutrons. Speaking of electrons, electrons and two heavier cousins, the mu and tau, are leptons. Leptons don't have color and don't respond to the strong nuclear force. Leptons also include three neutrinos, very low-mass electrically neutral particles.

Particles have anti-particles, which differ in having opposite charge. The positron is like an electron except that it has a positive electric charge. Some particles, such as photons, are their own anti-particles. The universe is almost (but only almost) always symmetric and conservative, meaning, e.g., that if a nuclear reaction produces a charged particle it will also produce an anti-particle of opposite charge, or if a reaction creates a lepton it must also produce an anti-lepton. When enough energy is available, for instance in particle accelerators, energy can become particle – anti-particle pairs (recall that $E = mc^2$). If the particle and its anti-particle meet, they will annihilate back into gamma rays. Production of virtual particles can happen in our lower-temperature everyday world provided the particles don't live too long. Here we appeal to one formulation of the Heisenberg Uncertainty Principle, namely

$$\Delta E \cdot \Delta t \geq \hbar / 2;$$

there is a limit to how precisely the product of uncertainty in energy times uncertainty in time can be known. It is possible to "borrow" the amount of energy needed to create a pair of particles from the vacuum of spacetime but only for a short time.

The very early universe was very hot and energetic, meaning that energy became particles became energy again. If it weren't for that "only almost" always symmetric, we wouldn't be here: the universe cooled and became less energetic, eventually cool enough that there was no longer enough energy to produce particle – anti-particle pairs. If matter and anti-matter had been created in equal numbers, the last-created pairs of particles would have annihilated and then nothing would have remained except photons. The early universe must have been asymmetric by one part in a few billion to have left us with the particles that make up stars, galaxies, planets, us.

Physics note: the Casimir Effect demonstrates experimentally that pairs of virtual particles are being created all the time. To simplify just a bit, imagine two parallel uncharged metal plates, very close together (say, less than a micron), inside a larger evacuated container. The energy of the vacuum of spacetime is not zero. Virtual particles are effectively fluctuations in the vacuum and, thinking of them as waves, there will be fewer created

between the metal plates than outside because only the shorter possible wavelengths will fit in the gap between the plates. Whether we consider them as waves or as particles, there will be more interactions with the outer sides of the metal plates than the inner sides. In this set-up the plates will move toward each other (although different geometries can result in repulsion rather than attraction).

A related concept is the assertion that black holes could evaporate. In the 1970s Jacob Bekenstein considered the entropy of black holes. Entropy is sometimes considered a measure of disorder. More technically, it is a measure of the number of available states that a system can occupy. As an analogy, consider a parking lot with most spaces filled and only a few, say three, unoccupied. There are thus only three ways you could add your car to this system. If most of the cars were moving, though, driving (carefully, we trust!) around in the parking lot, you could join in with your car in quite a few more places. This version of the system has higher kinetic energy (higher "temperature") and more possible ways of arranging the particles. It has higher entropy. We could extend this analogy by considering what happens to the number of options for your car if there were fewer total cars or overall more space in the parking lot; in both cases, there would be more ways to add your car to the lot. One of the axioms of thermodynamics states that in any thermodynamic process the entropy of the systems involved doesn't decrease. In an everyday example, this says you expect that when you add ice cubes to a drink that the ice cubes will melt and the liquid will cool off, rather than the reverse, where the drink would get hotter and ice would spontaneously form into cubes. If you run a video of this, you can immediately tell which way is forward in time and which way is reverse.

Back to the black holes. Anything that falls into a black hole carries its entropy with it and the entropy of the black hole should increase. Objects with increasing entropy have a temperature and objects with temperature radiate. Therefore black holes should radiate. But how? Suppose that a particle – anti-particle pair happened to be created near the event horizon of a black hole. Suppose further that one of the pair fell across the event horizon before it had a chance to annihilate with its partner. The partner particle will now last, will become a measureable part of the rest of the universe. We must now account for the energy needed to produce it. The most readily available source of energy near an event horizon is the warp in the fabric of spacetime. Reducing the warping is akin to reducing the mass of the black hole. Stephen Hawking followed Bekenstein's work on entropy with theoretical work on the problem of producing the particles and this process is now generally referred to as Hawking Radiation.

Expansion of the Universe

Let's elaborate a bit on the observations and terminology for describing the expansion of the universe. Recall that the redshift $z = \Delta\lambda / \lambda$. Galaxies beyond the Local Group are receding from us and the farther away they are the faster they are receding. This distance – recession velocity relationship is often called the Hubble relation (or Hubble's Law). It follows from the assumption that the Robertson-Walker metric is an appropriate rule for determining the separation between points, i.e., from the assumption that the universe is isotropic and homogeneous. For small z this will be linear and thus

$$cz = H_0 d,$$

where H_0 is the slope of the Hubble relation, ~ 67 km/s / Mpc. At larger distances, where we are more obviously looking back in time, the slope is likely to be different. There is no requirement that the expansion rate should have been the same earlier in the history of the universe. More generally, H, with no subscript, is called the Hubble parameter; H_0 with the subscript refers to the slope of the Hubble relation at the present time.

All lengths are expanding at the same rate at any given time, meaning that the increase in wavelength obtained from z can immediately tell us about the difference in the size of the universe at the time the light of redshift z was emitted. Let a be the relative linear scale factor for the universe. Today $a_0 = 1$. Substitute a for λ in z:

$$z = \frac{\lambda_{obs} - \lambda_{emitted}}{\lambda_{emitted}} = \frac{\lambda_{obs}}{\lambda_{emitted}} - 1 \rightarrow$$

$$1 + z = \frac{a_0}{a(t)}$$

where *a(t)* means the scale factor at time *t*.

Math note: for those of you who've had some calculus, note that $H = \dot{a}/a$.

Also, likewise for the mathematically inclined, meet the Friedmann equation, that describes the expansion rate in an isotropic and homogeneous universe:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^3}\varepsilon - \frac{kc^2}{R_0^2 a(t)^2}.$$

The curvature is represented by *k*, introduced above, and how extreme the curvature is is characterized by R_0, the radius of curvature at the present time. ε represents the sum of the energy densities of all the components that make up the universe (radiation, matter, dark energy). It is also probable that the expansion rate is changing, and we can write an acceleration equation as

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\varepsilon + 3P),$$

where *P* represents pressure. Pressure can usually be related to the energy density; e.g., for radiation, $P = 1/3\ \varepsilon$, while for non-relativistic, not-very-dense matter, $P \sim 0$. Dark energy (see below) is most often treated as having a negative pressure, $P = -\varepsilon$.

Example:

We observe a galaxy with a redshift $z = 2$; how large was the universe when the light we are now seeing was emitted from that galaxy?

$a(t) = a_0 / (1 + z) = 1/3$ its current linear size.

Yet another math note: In a universe that is expanding and where the expansion rate changes with time, *distance* becomes a bit of a complicated concept. We can define the *proper distance* between two points as the pathlength between those points at a fixed scale factor. That's not an observable distance – we can't see a galaxy as it is now, we see it as it was then, when the light we are now seeing was emitted and when the scale factor was different than it is now. The proper distance can be calculated, *if* we know the expansion history of the universe, by the following integral:

$$d_p(t_0) = \int_{t_e}^{t_0} \frac{dt}{a(t)},$$

where t_e is the time when the light was emitted and t_0 is the time now. Because the rate of expansion has changed over time, there is no one neat function for *a(t)*, the scale factor as a function of the age of the universe. (Or for the inverse, namely the age of the universe at a given scale factor.) The proper distance is what is given by cz/H_0, which gives reasonable values of distance for $z < 0.2$. Beyond that, there is a very useful web site that will do the integration for you: http://www.astro.ucla.edu/~wright/CosmoCalc.html.

(The following expression is a polynomial fit to the calculated proper distance that is decent to $z \sim 1000$:

$$\log d = -4 \cdot 10^{-4}\, x^6 + 2 \cdot 10^{-5}\, x^5 + 0.0176\, x^4 - 0.0218\, x^3 - 0.2246\, x^2 + 0.702\, x + 3.5204,$$

where $x = \log z$.)

You might expect that we'd get the same distances when we measure the angular size an object of known linear size subtends on the sky (smaller = farther) or when we measure the flux received from an of known luminosity (dimmer = farther), but again, for large *z* the expansion of the universe intrudes. In a flat universe, the proper distance is related to the angular diameter and luminosity distances thusly:

$$d_p(t_0) = d_A(1+z) = {}^{d_L}\!/_{(1+z)}.$$

Think for a moment about the angular size on the sky for an object of known linear size: nearby, as you would expect, more distant objects look smaller on the sky, but objects from which the light was emitted in the very distant past were significantly closer to us when the light we are now seeing was emitted, meaning that their angular size on our sky is relatively large (even though they are faint).

Back to scale factor and expansion. Example:

Suppose that the numbers of photons and of baryons have been constant since some very early time. How has the energy density in the photons and in the baryons changed as the universe has expanded?

If the number of matter particles is constant, then the number density of baryons has decreased as the volume of the universe has expanded. The energy in a baryon is mostly due to its mass, via $E = m \cdot c^2$. Therefore as the universe gets older the energy density in the baryons decreases as a^3.

The energy in the photons is different. For light, $E = h \cdot c / \lambda$. As the universe expands, the wavelength increases. The photon number density goes down with increasing volume just as the baryon number density does, but the energy decreases by another factor of a because of the stretching of the wavelengths. Photon energy density decreases as a^4.

A result that follows from this example is that although the early universe was very hot and its energy was dominated by photons as it expanded and cooled it shifted into an era when its energy was dominated by matter. When the universe was hot, more than a few thousand K, it was opaque. The reason is that most of the atoms (almost entirely H and He) were ionized; free electrons are very efficient at scattering photons, meaning that light didn't get very far before scattered. Trying to see into the early universe would be like trying to see into a star. At about 380,000 years, the universe had cooled to about 3,000 K, cool enough that neutral atoms could form and the universe could become transparent. The universe was and is a black body, with an energy spectrum distributed according to the Planck function.

Example: What is the peak wavelength for the intensity of a black body of 3,000 K?

The peak wavelength is given by

$$\lambda_{\text{peak}} = \frac{2.898 \cdot 10^{-3}\,\text{m} \cdot \text{K}}{T\ \text{K}} \rightarrow \frac{2.898 \cdot 10^{-3}\,\text{m} \cdot \text{K}}{3000\ \text{K}} = 9.7 \cdot 10^{-7}\,\text{m} = 970\ \text{nm}.$$

Today that radiation has been redshifted into the microwave. In the early 1960s Arno Penzias and Robert Wilson were working at Bell Labs, in New Jersey, on a sensitive microwave antenna / receiver system. They realized that they had a low-level, isotropic, unidentified, source of static. The static remained even after one likely source, the droppings left by some resident pigeons, was removed. They discussed the issue with Robert Dicke, just down the road at Princeton University, who realized that their static was in fact a cosmic microwave background. Further observations revealed that the static was indicative of a blackbody universe, now registering a chilly 2.725 K. Penzias and Wilson shared the 1978 Nobel Prize for their discovery. The discovery of the CMB provided strong support for the Big Bang model for the universe, i.e., the assertion that the universe is expanding from an initial condition in which everything was tightly packed together and very hot.

Example: What is the peak wavelength for a blackbody of 2.725 K? How much smaller was the universe when the light of the CMB was emitted? What is the value z for the CMB?

$$\lambda_{\text{peak}} = \frac{2.898 \cdot 10^{-3}\,\text{m} \cdot \text{K}}{2.725\ \text{K}} = 0.001\ \text{m} = 1\ \text{mm}.$$

All lengths are stretched as the universe expands. The peak wavelength from the CMB is ~1100 times longer now than it was when it was emitted. That means that the scale factor, a, must have been ~1100 times smaller when the light of the CMB was emitted:

$$1+z=\frac{1}{a}\rightarrow z_{CMB}\approx\frac{1}{1/1100}=1100.$$

Whether z = 1100 or 1101 doesn't matter much!

Observations of distant Type Ia supernovae, e.g., around z = 1, suggest that the expansion of the universe is *accelerating*. If you've been watching the dates, you may have noticed that Einstein formulated his equations of General Relativity before Hubble, Slipher, and other astronomers had demonstrated that galaxies were receding. Einstein expected the universe to be stable. His equations suggest otherwise, though. Gravity should cause a universe with matter, photons, and possibly curvature, to contract. Einstein introduced a "cosmological constant", to which he assigned the symbol Λ, to counteract gravity (although if it's *constant* the universe would still be unstable and prone to collapse). Once the Hubble relation became apparent, Λ looked like a bad idea. Still, an expanding universe with matter, photons, and curvature should decelerate.

In the late 1990s two teams of astronomers published results of studies of Type Ia supernovae suggesting that the expansion of the universe is, actually, unexpectedly, accelerating. The standard model for Type Ia supernovae says that what is exploding is a white dwarf, a degenerate stellar remnant, pushed just over the Chandrasekhar stability limit of ~1.4 solar masses by mass from a companion star. Observations of Supernova 2012cg support this model: when this supernova exploded, in a galaxy 50 million light years away, the flash heated its companion star, making the companion temporarily brighter and visible. If this model is correct for all Type Ia supernovae, possibly with some allowances for compositional differences in the stars involved, then all these explosions are going to be quite similar, reaching similar peak (and very bright) absolute magnitudes. In that case, we have a "standard candle", an object of known luminosity, and thus an object whose distance can be determined. That's a big if, and there are observations today that suggest that some Type Ia supernovae could be due to collisions of two degenerate objects, likely two white dwarfs, rather than one white dwarf, or to the explosion of a single white dwarf caused by a shock from the ignition of an accreted helium layer. More work needs to be done to establish a way to separate possible sub-types and determine the absolute magnitude of any individual Type Ia supernova.

The following sketch shows several possible models for the scale factor, a, and redshift, z, vs time. "Now" has a_0 = 1, z = 0. The black line shows a steady state universe, eternally expanding, having no beginning; the existence of the CMB suggests that our universe is not like this model. The red line shows a universe with a beginning but expansion always decelerating, eventually to collapse (in a "Big Crunch"); observations suggest that there is not enough mass in our universe to halt the expansion and that our universe is not like this model. The blue curve, next to the red line, represents a universe that decelerates but not fast enough to halt the expansion and collapse; this would have been the preferred model prior to the mid-1990s. The green line begins by decelerating but in the not too distant past the expansion switches over to acceleration. The green oval shows the locus of observations of Type Ia supernovae with relatively high values of z, ~0.5 – 1. These supernovae are consistently fainter, and thus more distant, than had been expected for their z. If the universe had been expanding faster in the past when the light left those supernovae, they wouldn't have had to be as far away to show a given z. A universe expanding *less* rapidly in the past, means a given z is indicative of more distant objects, which is what we see. The conclusion is that the universe is now expanding more rapidly than it was in the past, i.e., that the expansion is accelerating. (For a description of the observations, see, e.g., Perlmutter, April 2003, *Physics Today*, pp. 53-64.) Saul Perlmutter, Brian Schmidt, and Adam Reiss shared the 2011 Nobel Prize in Physics for this discovery.

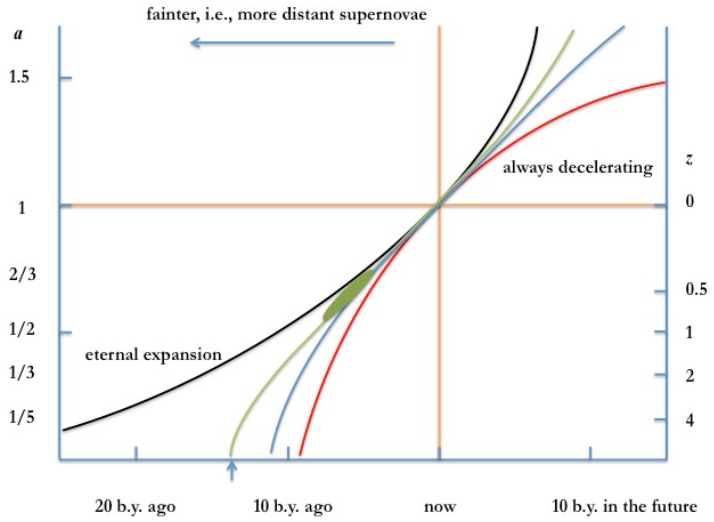


Figure 19.10:
Type Ia SNe (green oval) and various expansion histories.

We don't know why the expansion is accelerating. We can, of course, name whatever it is that's causing this effect: we call it Dark Energy. Observations are consistent with dark energy having a constant energy density and a negative pressure, filling all spacetime and not decreasing with changing scale factor *a*. Einstein's Λ has come back onstage as the symbol for this form of energy. Observations aren't adequate to rule out a dark energy that varies with time (or even with location), though, something that's often called "quintessence"; in 2024-25 there have been some observations that are consistent with a dark energy that decreases with time.

Today the universe is ~13.8 billion years old (indicated by the arrow on the time line above) and the energy density in the photons has dwindled to an almost insignificant fraction of the content of the universe. Our best current estimates, based on 2013 data from the ESA Planck spacecraft, are that the energy density in the universe today is 68.3% dark energy, 26.8% dark matter, and 4.9% baryonic (ordinary) matter and that the curvature is flat (or very very close to flat). The observations of distant Type Ia supernovae support the need for dark energy. Observations of the mass in clusters of galaxies provide a principle piece of evidence for the overall amount of matter (dark + baryonic) in the universe. Cluster mass comes from several observations, including 1) observations of gravitational lensing, as described in the previous section; 2) motions of galaxies within clusters, which, courtesy of the virial theorem, can be related to the cluster's gravitational potential; and 3) from the x-ray emission of intracluster gas, heated as material falls together to make the cluster. A crucial additional piece of evidence about the make up of the universe comes from the angular size of the anisotropies in the CMB.

There will have been primordial density fluctuations in the very earliest moments of the universe. Quantum mechanics pretty much forbids the universe from being absolutely smooth. There are theoretical predictions for the scale of those density perturbations. Along with everything else in the universe, the density perturbations will grow in size and will leave their mark on the CMB. Spacecraft such as the Wilkinson Microwave Anisotropy Probe and the European Planck spacecraft measured the temperature of the CMB around the entire sky with angular resolutions of ~13 arcmin and ~5 arcmin, respectively. The density perturbations result in temperature differences on the order of a few 10^{-5} K. The angular size of these anisotropic pieces (~1.2°) is in accord with what we would expect for a flat universe, given the scale of the initial quantum fluctuations and the growth of perturbations in the expanding early universe. The following all-sky map from the Planck data shows what the anisotropies look like. The plane of the Milky Way runs across the center of the image, which has been corrected for known sources of microwave emission in our galaxy. The image has also been corrected for a dipole (Doppler shift) term that arises because of our motion (around the Sun, around the galaxy, toward Virgo) with respect to the CMB.

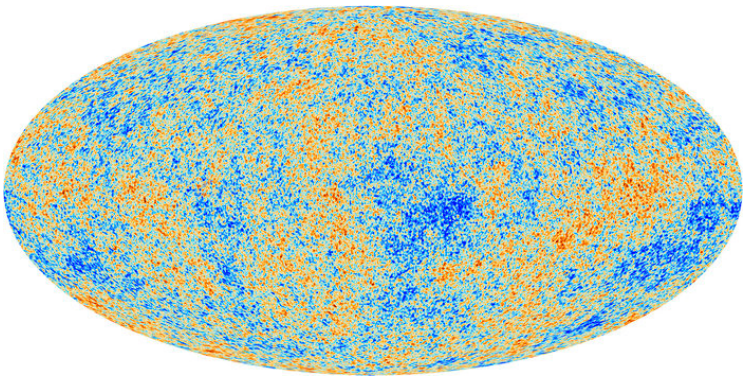


Figure 19.11: Anisotropies in the CMB; credit: ESA and the Planck Collaboration

http://www.esa.int/spaceinimages/Images/2013/03/Planck_CMB

The dipole term doesn't mean that there is a preferred direction for the CMB. Imagine driving into falling rain, and noticing that more raindrops hit your front windshield than hit the rear window. You could say, oh, I'm heading toward the center of the storm. But if you turn at the next corner, i.e., change your direction, there will still be more raindrops hitting your front windshield. The dipole term depends on our motion; it doesn't imply that there is a preferred orientation to the universe. Once we correct for the dipole term, on a large scale the universe looks pretty isotropic. Note, though, that it is still an active research question whether this is really the case; e.g., some researchers are asking whether the distribution of distant quasars might not actually agree with the CMB dipole. The standard model of cosmology still says that the universe is isotropic and homogeneous, but that doesn't mean it isn't legitimate to seek out observations that might challenge that model.

An interesting question to consider is whether the higher density patches are going to look warmer or cooler in the CMB observations. On one hand, if you compress a gas it gets hotter, suggesting that the denser regions would be warmer. On the other hand, light leaving denser regions will experience a larger gravitational redshift than light leaving less dense regions, which would suggest that the denser regions would appear cooler. On the other other hand, a warmer denser region might have stayed ionized, and thus opaque, longer, meaning that the universe had expanded noticeably more by the time the CMB from dense regions was emitted. On balance, it appears that the gravitational redshift piece (called the *Sachs-Wolfe Effect*) wins out on larger scales and the pressure waves (denser = hotter) win out on smaller scales (~ 1°). Note that interpreting the temperature variations isn't always straightforward. As an example, there's a region that's been dubbed the "Cold Spot" because it's a relatively large patch of sky (~ 5°) that on average is ~ 70 μK colder than the typical CMB temperature. That's big enough and cold enough to be statistically significantly unlikely. We aren't sure yet what it means.

The overall curvature of the universe depends on the total energy density, the total amount of light and matter and dark energy that the universe contains. There will be a critical density that is just exactly the right amount of energy to make the universe flat. We can express that critical energy density as

$$\varepsilon_{\text{crit},0} = \frac{3c^2}{8\pi G} H_0^2 .$$

Given our current best estimates for H_0, $\varepsilon_{\text{crit},0}$ ~5000 MeV/m^3, or $8 \cdot 10^{-10}$ J/m^3. If we want this in mass units, we can divide the latter figure by c^2, giving a critical mass density of ~$9 \cdot 10^{-27}$ kg/m^3. That's ~11 hydrogen atoms in 2 cubic meters of space. That is pretty low. The "empty" space of the interstellar medium is more than 100 times denser than this. Remember, though, that to get a representative box of universe we need a region with a size of ~100 Mpc, not 1 meter. When you average in the nearly empty voids between clusters of galaxies, matter (dark + baryonic) provides only about 1/3 the critical density.

One way to express the relative contributions of the various components of the universe is as a fraction of the critical energy density. Define

$$\Omega_{\text{component}} \equiv \frac{\varepsilon_{\text{component}}}{\varepsilon_{\text{critical}}}.$$

Using this notation, we can say that $\Omega_{\text{matter, 0}} \sim 0.32$. Today, the energy density in radiation –meaning adding up CMB photons, starlight, synchrotron emission from AGNs, relativistic neutrinos, etc. – is very low: $\Omega_{\text{radiation, 0}} \sim 8.4 \cdot 10^{-5}$. But, observations suggest that the universe is flat and contains dark energy. Current estimates are that $\Omega_{\text{dark energy, 0}} \sim 0.68$.

Math note: the Friedmann equation may be rewritten in terms of the quantities H and Ω:

$$\frac{H^2}{H_0^2} = \frac{\Omega_{\text{rad, 0}}}{a^4} + \frac{\Omega_{\text{matter, 0}}}{a^3} + \Omega_{\Lambda, 0} + \frac{1 - \Omega_0}{a^2}.$$

Here is a graphical way to represent the *possible* models for the current universe, overlain with the data from Type Ia SNe, the anisotropies in the CMB, and the matter content (dark + baryonic) from observations of galaxy clusters. This is based on a plot in Perlmutter, April 2003, *Physics Today*, pp. 53-64 (or posted at http://supernova.lbl.gov/PDFs/PhysicsTodayArticle.pdf). The possible ranges of values permitted for the densities are indicated by the ovals. The data ranges overlap at ~68.3% dark energy, 31.7% matter, and flat.

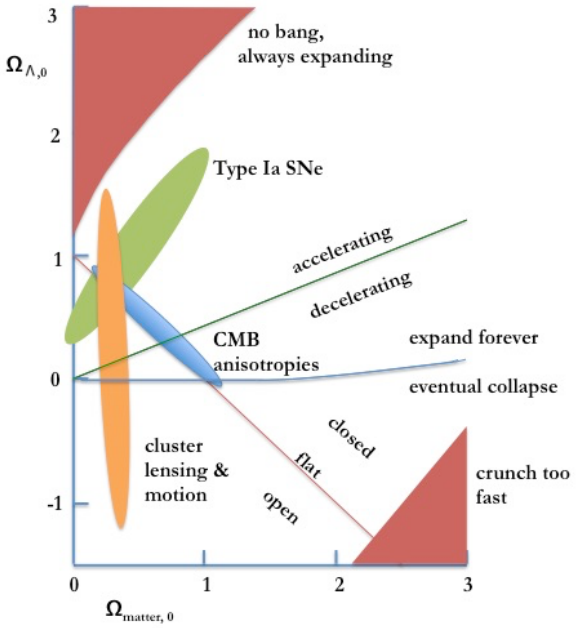


Figure 19.12:
CMB, SNe, and mass data today overlain on various possible values for component energy densities.

The lower right corner of this plot is excluded as a possible description of our universe not simply because we don't observe nearly enough matter but also because of the observed ages of stars. A universe with too much matter would start to expand but rapidly reach a maximum size and contract again in a Big Crunch. Our universe is too old to be like this. The upper left corner is excluded because it implies such a high rate of outward expansion that the universe could never have been in the tiny initial state of the Big Bang.

Remember, though, that the *data* on this plot are for the *current* universe; we could have added a third axis, with $\Omega_{radiation,0}$, but that value is very low today. If we drew this plot for an earlier time, we couldn't ignore radiation.

Here is a schematic showing the relative contents of the universe at four different times:

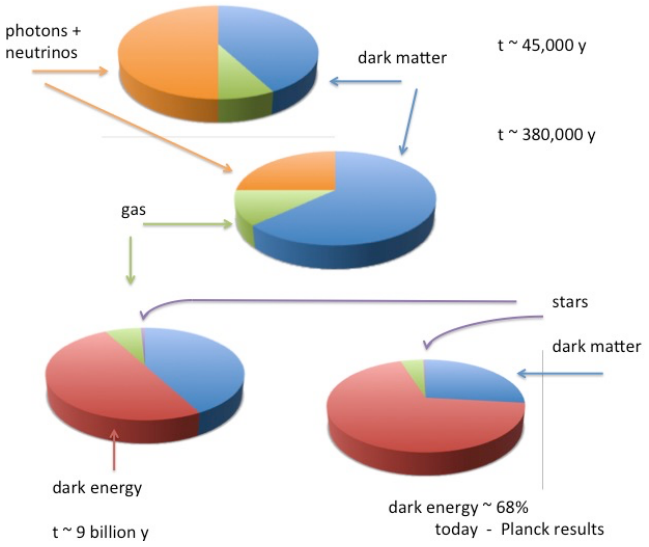


Figure 19.13:
Evolution of the relative contributions to the energy density of the universe with time.

Matter and radiation were both equally important when the universe was ~45,000 years old. By the time it was 380,000 years old (the time to which we are seeing in observations of the CMB) the universe was dominated by matter (gas + dark – no stars yet). About the time the solar system was forming, the energy density in the matter had fallen to the point where the ever-present dark energy was equally important. Today, the universe is dominated by the dark energy, although not yet overwhelmingly so.

Another way to show this change from radiation-dominated to matter-dominated to dark energy-dominated is to plot these three energy densities as a function of time, as in the following figure. The energy density in the radiation starts highest but falls off fastest (as $1/a^4$, recall, because the photon number density falls and the energy falls due to redshift), giving way to matter at ~45,000 years. The energy density in the dark energy is just there, constant (we think), starting to dominate over matter at an age of ~9 billion years.

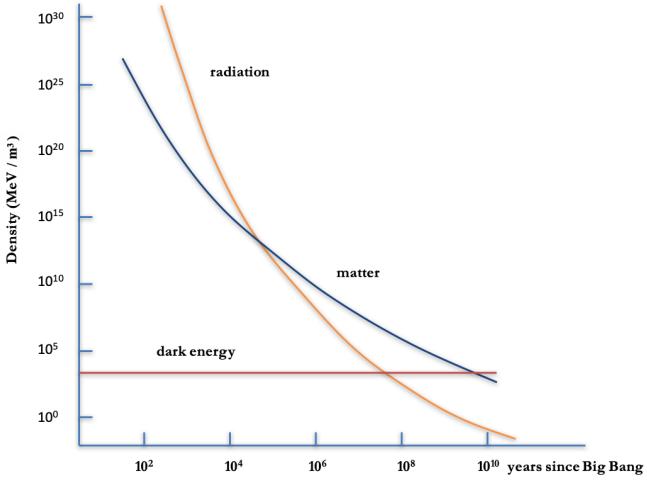


Figure 19.14: Absolute energy densities as a function of time.

History of the Universe

The preceding two plots lay out a very rough "history" of the universe, but there are quite a few additional pieces of observational or theoretical evidence that we can employ to get a more detailed history.

The Planck time. We do not, yet, have an adequate theory to permit us to understand what the universe might have been like at $t = 0$, or whether it is even possible to make sense of the beginning. A naïve expectation might be that we could simply play the expansion of the universe backward until we reached an infinitely hot, infinitely dense singularity at $t = 0$. But current theory fails us at ~ $5.4 \cdot 10^{-44}$ seconds and ~$1.6 \cdot 10^{-35}$ m, units called the Planck time and Planck length, respectively. How do these numbers arise? Recall that particles (e.g., electrons) can be shown to have wavelengths. Particles could also be described as having a Schwarzschild (black hole) radius. In fundamental units, these are equal to

$$\lambda_{particle} = \hbar / mc \text{ and}$$
$$\lambda_{blackhole} = Gm / c^2 ,$$

respectively, where "*h*-bar" is Planck's constant $h / 2\pi$. Setting these two quantities equal permits us to solve for a Planck mass (~$2.2 \cdot 10^{-8}$ kg), which yields a Planck length. The light travel time across a Planck length is the Planck time. In fundamental units the Planck time and length are given by

$$t_{Planck} = \sqrt{G\hbar / c^5} \text{ and}$$
$$\ell_{Planck} = \sqrt{G\hbar / c^3} ,$$

respectively, and, plugging in values, we get ~ $5.4 \cdot 10^{-44}$ seconds and ~$1.6 \cdot 10^{-35}$ m. We could, additionally, calculate the associated energy and temperature, using the fact that $E = mc^2$ and $E = kT$, where k is Boltzmann's constant. The Planck energy is ~$1.2 \cdot 10^{22}$ MeV, the Planck temperature ~$1.4 \cdot 10^{32}$ K. These units are small and hot, but not a singularity. Appealing once again to the Uncertainty Principle, the Planck units may be the smallest time and distance about which it makes sense to say anything physically realistic. This is the scale beyond which current physics cannot go.

It is possible that at or before this time all four of today's fundamental forces were unified as one fundamental force that acted without discrimination on everything (even if it's hard to imagine what *everything* might mean on a scale that is smaller than the smallest particle!).

Grand unification. Current theory can't explain how all four forces might look alike, but it can make at least partial sense of a slightly later time, when the strong – weak – electromagnetic forces were unified. Today, on length scales larger than the size of a nucleon, these three forces have distinctly different strengths. If we set the strength of the electromagnetic interaction = 1, then the strong force would be about 60 times stronger and the weak force would be about 10^{-4} as strong (and gravity is a miniscule 10^{-41}). At smaller scales and higher energies the difference is not so great. At the energies and length scales of the universe between the Planck time and ~ 10^{-36} seconds, it's possible that there was no difference. It's possible that different sorts of force-carrying bosons existed, capable of interacting equally well with both quarks and leptons. The Large Hadron Collider, which smashes beams of particles together at incredibly high energies, was by mid-2015 achieving collision energies of ~13 TeV. To replicate the conditions of the universe at grand unification scales would take ~ 10^{15} GeV. . . we're not going to replicate those conditions any time soon. Terminology: A Grand Unified Theory would explain the unification of the strong, weak, and electromagnetic interactions into one electronuclear interaction; a Theory of Everything would include gravity.

Inflation. Two interesting observations and one non-observation have influenced the development of models of what happens next, i.e., at an age of ~ 10^{-36} seconds. Let's look at horizons, flatness, and magnetic monopoles.

First, the horizon problem. The CMB is remarkably isotropic. Imagine a tiny speck of matter just hanging out in the universe when it was 380,000 years old. That's the time to which we are seeing when we observe the CMB. Our tiny speck could have been influenced by other nearby specks, perhaps regions that were a bit hotter or a bit denser, and our speck's properties might have adjusted accordingly, just as cold water coming into contact with warmer water would reach an equilibrium in-between temperature. We can say that these two regions are in "causal contact", in the sense that properties in one can cause changes in the properties of the other. Our speck could not possibly have been influenced by any parts of the early universe that were too far away for light to have reached it in the 380,000 years the universe had existed. That's the problem: We see the same temperature in the CMB on opposite sides of the sky. Places that we would have thought could not have been in causal contact when the universe was only 380,000 years old must somehow have been at some point before that because it would be incredibly unlikely for them to have the same properties otherwise.

Second, the flatness problem. The overall geometry of the universe is awfully close to flat. In terms of the density parameter, the quantity $|1-\Omega|$ is very small. The problem here is that if $|1-\Omega|$ isn't exactly zero, it should have grown with time. Planck spacecraft observations imply that the universe deviates from flat by less than 0.01 today. To be that small today, $|1-\Omega|$ must have differed from zero by less than 1 part in 10^{60} at the Planck time.

Third, the lack of magnetic monopoles. Magnets have north and south poles, and if we split one in two we don't get separate pieces of north and south, but simply two smaller magnets each with a north and a south pole. Why? Could there ever have been magnetic monopoles? It makes sense to posit a Grand Unified Theory, in other words, that when the universe was very small and hot interactions could be described by a single theory, rather than by separate strong, weak, and electromagnetic interactions. As the universe expanded and cooled the forces separated. The strong force should have separated from the still-unified electroweak force at an age of ~ 10^{-36} seconds. This sort of split is known as "spontaneous symmetry breaking". The idea is that when the universe was hotter interactions between quarks and leptons were symmetric; afterwards, they weren't. Imagine trying to make a baseball bat stand on end. It's likely to fall over. But suppose several people are reaching out, touching it lightly, pushing it back towards vertical before it has a chance to fall; they could probably keep the baseball bat close to vertical for quite a while. Eventually the people get tired and quit playing this game and the bat finally falls. Which direction does it fall? That's not likely to be something you could predict ahead of time, because while it is more or less vertical all directions are equally likely. There is symmetry among the directions. Once the bat falls, the symmetry has been broken. As another, slightly more relevant, example, consider a car windshield on a freezing day. Eventually ice crystals start to form, in several different places, on the windshield. A phase transition is taking place from water in tiny droplets to water as ice. The ice crystals spread, perhaps sending out ice "ferns" in one preferred direction, right up to the point where one bumps into another crystal. At the boundary between two crystals there is a discontinuity. When the early universe cooled to the point where the strong force split off from the electroweak it was like a phase transition, like the water going from a symmetric mist of tiny droplets to flat crystals, and that phase transition should have resulted in discontinuities, which in this case should manifest themselves as magnetic monopoles, with zero dimension, and one-dimensional cosmic strings (*not* string theory strings). Which we don't see. Perhaps there are no magnetic monopoles; on the other hand, if the Grand Unified Theories that predict them are correct, then we need to explain why we don't observe them.

All three of these problems are addressed by a proposal called inflation. Suppose that prior to ~ 10^{-36} seconds the universe was very tiny, and symmetric but unstable, like the baseball bat standing on end. The fabric of spacetime had a relatively high energy density, like the extra energy the baseball bat gets from several people pushing it lightly all the time. The technical phrase is that the state of the universe was a *false vacuum*. If the energy density of the universe was dominated by this extra high vacuum energy, a period of inflation happens, during which the scale factor would have increased exponentially, i.e.,

$$a(t) \propto e^{H(t)t}.$$

The universe expands amazingly fast until ~ 10^{-32} seconds, the temperature drops, the symmetry is broken (the baseball bat falls over) and the energy density settles into a lower level, called the *true vacuum*, which characterizes the dark energy today. How does this help with the horizon, flatness, and monopole problems? Before inflation the universe could have been very tiny. There is not one definitive model, so the following estimates for the size then are very uncertain. In particular, the energy destined to become our current visible part of the universe could have been contained in a sphere approximately a Planck length, give or take a few factors of 10, in diameter, and, importantly, at least several orders of magnitude *smaller* than the size of the visible horizon at that time. This means that all the pieces of the universe we see today could easily have been in causal contact before the period of inflation. During inflation, the scale of the universe increases by *at least* e^{60}; in some models it's over e^{100}. The scale factor, a, increases to approximately $10^{-30} \cdot a_0$. The size of what is today's visible universe becomes roughly the size of a grain of sand, about 1 mm or a bit less. (If you are doing the math on that, note that even though the universe is 13.8 billion years old, that doesn't mean that our visible horizon is 13.8 billion light years away, or that the visible universe is 27.6 billion light years across. . .we've been expanding during those years, and the diameter of the visible universe today is about 92 billion light years.) Blowing up the universe this fast drives it toward flat, just the way blowing up a balloon smoothes it out. It also, perhaps too conveniently, pushes any nearby "topological defects", the magnetic monopoles and cosmic strings, so far away that we don't expect to be able to detect any of them today.

When the false vacuum decays, a lot of energy must be released, which will produce a lot of photons and also lots of quarks and leptons and their associated anti-particles. (Remember that $E = mc^2$ says that if we have the energy we can turn it into particles.) Anti-particles are almost exactly like particles, just with opposite charge. That "almost" is critically important. *If* the universe were truly symmetric then we would expect there to have been equal numbers of particles and anti-particles produced. And that means, we wouldn't be here. If the particles and anti-particles are produced in pairs, those pairs will eventually annihilate back into energy, leaving nothing behind but photons. The fact that there is matter in the universe today says that it can't have been totally symmetric back then. Comparing the density of CMB photons (which way outnumber all the photons that all the stars have ever produced) with the density of matter particles, the universe must have been asymmetric by about a part in $10^{9\text{-}10}$. There are a few billion photons for every matter particle. Where this asymmetry arises is one of the challenges for particle physicists today.

Electroweak epoch. This describes the time period between the end of inflation, at about 10^{-32} seconds, and the point where the weak and electromagnetic forces separate, at about 10^{-12} seconds. It's hot and energetic, a plasma of quarks and gluons and bosons such as the Higgs (with a mass of 125 GeV/c^2), that helps explain why some particles have mass, and the W^+, W^- (80.4 GeV/c^2) and Z^0 (91.2 GeV/c^2) bosons that mediate the weak force. The Higgs is a boson but it is not a force-carrying boson; force-carrying bosons have spin = 1 whereas the Higgs boson has spin = 0. By about 10^{-12} seconds, the universe has cooled enough that there will no longer be enough energy to create these massive bosons. The remaining ones will rapidly decay – the Higgs has a lifetime of ~10^{-22} seconds and the weak interaction bosons have even shorter lifetimes, ~10^{-25} seconds. When there was one unified force the bosons were identical. Then the weak force bosons acquired mass and became different from the photons; once the weak bosons disappear, the effective range of the weak force shrinks and the weak and electromagnetic forces are separate.

Why some particles, in particular the three weak interaction bosons, have mass while others such as the photon do not has long been an intriguing question for particle physicists. In the 1960s several physicists, including Peter Higgs (d. 2024), proposed a mechanism whereby the symmetry that predicted that all four of these electroweak bosons should be massless could be broken and the $W^{+/-}$ and Z^0 bosons could acquire rest mass, leaving the photon massless. Later it became clear that the same mechanism could account for the rest mass of other elementary particles such as electrons and quarks as well. Note that the masses of composite particles such as protons and

neutrons are dominated by the kinetic energies of their internal particles rather than by the rest mass of their constituent quarks. The Higgs field is predicted to be everywhere and particles that are prone to do so acquire inertia by interacting with it. The Higgs is a scalar field, rather than a vector field, so it's just there, it doesn't have a direction associated with it at any given point. It is as if interacting particles acquired a third dimension: the photon has two polarizations, both transverse to the direction it is moving; the weak force bosons interact with the Higgs field and acquire another dimension (a third "degree of freedom"), an energy that slows them down. The more strongly a particle interacts with the Higgs field, the more that interaction results in the particle's resistance to acceleration. The more the particle resists acceleration, the more inertial mass we say it has.

The classic analogy is that of a famous person trying to cross a crowded room. In 1993 David Miller, of University College, London, used the example of the recent Prime Minister Margaret Thatcher to explain the Higgs and why Great Britain should support work on the CERN (the European Organization for Nuclear Research) particle collider in hopes of finding the Higgs boson. Imagine that Thatcher enters a room full of the party faithful, evenly distributed (the "field"). Those nearest her cluster around, eager for a chance to talk to her, creating a knot of people around her. Thatcher can move through the room, leaving people behind her while new people are accreted to the front of the knot, but she moves more slowly through the room than an unknown person could. It would also be harder to slow or turn Thatcher and the knot of people around her. The fact that she interacts with the "field" of party members means that she acquires an inertia.

The Higgs boson is the particle manifestation, the "excitation" of the field, that permits us to determine that the field really is present. To carry the political analogy a bit further, Miller imagined that instead of Thatcher entering the room full of party workers, someone at the door mentions a political rumor. The rumor can make its way across the room, as people lean in to hear what's up. The knot of people clustering around as they listen and pass along the rumor is like a "particle"; it has "mass", just like the knot of people who clustered around the Prime Minister had mass.

Detecting the boson would be evidence of the existence of the field. CERN took the lead, converting and expanding an existing particle accelerator facility into the Large Hadron Collider, the world's most powerful particle collider. The LHC occupies an underground tunnel, 27-km in circumference, under the region where France borders Switzerland. Protons are accelerated to nearly the speed of light in tubes that are held at a very high vacuum to ensure that the particles only collide at the specific points where they are supposed to. Several huge detectors record the evidence from the collisions. One objective for the LHC was to detect or rule out the existence of the Higgs boson. Because the Higgs so rapidly decays into other particles, the detection involved consistently observing the debris from the collisions and comparing that with what was expected from the decay of the Higgs boson, with more than one instrument, determining the mass of the particle that must have created the debris and ascertaining that it was not some other sort of particle. Unlike looking for a needle in haystack, where you'd know unambiguously once you'd found it, this has been described more as looking for hay in a haystack. . .an excess of pieces of hay all of the same mass and not explained by any other known particle. At a meeting on July 4, 2012, the discovery of a new particle with a mass of ~125 GeV/c^2 was announced. Additional reports, in spring 2013, confirmed that this new particle behaves the way a Higgs boson is expected to behave. In 2013, Peter Higgs and François Englert received the Nobel Prize in Physics for their role in this work. It remains to be seen whether there is more than one Higgs field, whether the earlier separation of the strong force from the electroweak force is similar to the splitting of the electromagnetic and weak forces, and why the Higgs boson has the mass that it does.

Quark epoch. At an age of 10^{-12} seconds it's still too hot (> 10^{12} K) for particles to combine. Today, quarks are always found in combinations, either a quark – anti-quark pair (a meson) or a three-quark baryon; collectively mesons and baryons are known as hadrons. When the universe was less than ~10^{-6} seconds old collisions were energetic enough to knock apart any hadrons that formed, so matter would have been a quark-gluon plasma. When there is enough energy the higher-mass quarks will also be present, not just the *u* and *d* quarks that make up protons and neutrons. There will be a small amount of time here when it's possible for high-mass quarks to form and combine briefly into high-mass hadrons, e.g., baryons containing strange or charmed or even bottom quarks. None of them are stable and all would rapidly decay. Top quarks are so massive that by the time it was cool enough for a

hadron to form it was too cool for top quarks to be created, so there weren't likely to have been too many hadrons containing top quarks. On the lepton side, energies were still high enough to produce the higher-mass τ and μ particles and anti-particles, but it was soon going to be too cool. Muons are ~200 times the mass of an electron, and once the temperature drops to ~10^{12} K the average pair of photons will no longer be able to turn into a muon – anti-muon pair.

Hadron epoch. By ~10^{-6} seconds / ~10^{12} K, quarks could combine into hadrons without instantly being broken apart. Protons have a mass of 938.3 and neutrons 939.6 MeV/c^2, respectively. The universe is now rapidly cooling below the temperature at which proton – anti-proton or neutron – anti-neutron pairs can be created and most particles are going to annihilate by the time the universe is ~ 1 second old, leaving only that slight excess of matter particles mentioned above. High-mass particles tend to decay into lower-mass particles. The proton (*uud*) is the lowest-mass baryon and appears to be stable, at least on scales much longer than the age of the universe. People have looked: for example, Super Kamiokande is a large (50 kiloton) tank of water in Japan equipped with detectors used, for instance, to detect neutrinos; as such it is also a ready supply of steadily observable protons, none of which have been observed to decay.

Physics note: If protons were to decay, one primary route would be

$$p^+ \rightarrow e^+ + \pi^0;\ \pi^0 \rightarrow 2\gamma,$$

where the π^0 represents a neutral pion. Pions are mesons (2-quark particles) containing up and down quarks and anti-quarks. There are four possibilities:

$$\pi^+ = u\overline{d};\ \pi^- = d\overline{u};\ \pi^0 = u\overline{u} \text{ or } d\overline{d}.$$

Observational limits put the proton lifetime at over 10^{33} years. Free neutrons (*udd*), on the other hand, are not stable and will decay with a mean lifetime τ = ~881 seconds. This means that after a time *t*, $e^{-t/\tau}$ of the original quantity will remain. Mean lifetime and half-life are related this way:

$$t_{1/2} = \tau \ln 2;$$

for the neutron, this means that the half-life is ~611 seconds.

When the universe is less than a second old, neutrons are for all intents and purposes stable, but the fact that they decay is going to start to matter quite soon. When they decay, neutrons decay into protons by beta (i.e., electron) decay:

$n \rightarrow p^+ + e^- + \overline{\nu}_e$. The bar over the neutrino signifies that it is an anti-particle.

Lepton epoch. By an age of ~1 second, T ~ $1.5 \cdot 10^{10}$ K, the exotic high-mass hadrons had decayed and the majority of the rest of the hadrons had annihilated with their anti-particle counterparts. The remaining baryons would be protons and neutrons, with a few more protons than neutrons, in part because the mass difference means that neutrons stopped getting created slightly earlier than protons and in part because a few of the neutrons will have decayed. The ratio still isn't very uneven, roughly $n_{p^+} / n_n \sim 10 : 6$. High-mass leptons are no longer being produced, but it's still just barely hot enough to produce electron – positron pairs. The universe does seem to conserve charge, meaning that the excess of electrons over positrons is the same as the excess of baryons over anti-baryons and the universe is electrically neutral.

Example: how hot does it have to be to produce an electron – positron pair?

The relevant equations are $E = kT$ and $E = mc^2$:

$$T \geq \frac{2m_e c^2}{k} = \frac{2 \cdot (9.1 \cdot 10^{-31}\text{kg}) \cdot (3 \cdot 10^8\,\text{m/s})^2}{1.38 \cdot 10^{-23}\,\text{J/K}} = 1.2 \cdot 10^{10}\,\text{K}.$$

The temperature in the early universe is falling as space expands, but there are a few bumps in that general downward trend. Think about particle creation and annihilation. When it's hot enough to produce particle – anti-particle pairs the energy in the particles and the energy in the photons is in equilibrium. When it cools to the point that we stop producing new particle – anti-particle pairs, and the remaining ones annihilate, that annihilation energy goes back into being photons. The photon energy is still in equilibrium with the remaining types of particles. For example, when the last electron – positron pairs annihilate, the energy from the annihilation is shared among the photons and the kinetic energy of the remaining electrons, protons, and neutrons. That boosts the temperature over what it would have been if it were simply falling uniformly.

At the point that the universe is a few seconds old, as the electrons and positrons are annihilating, there are still plenty of neutrinos around. They aren't going to share in the energy from the e^- / e^+ annihilation, though. Neutrinos have a very low probability of interacting with other particles (we say they have a low *cross-section*) because, being low mass, neutral leptons, and they pretty much only interact by weak interactions. By now the weak interaction is confined to very short range. At some point it was inevitable that the density of the universe would fall to the point where the time between scatterings, i.e., the time between neutrinos running into other particles and having a chance to interact, is going to exceed the age of the universe. By the time the universe is ~1 second old, neutrinos have effectively *decoupled* from the photons and other remaining particles. There should be a cosmic neutrino background with a predicted temperature of ~1.9 K. The CMB dates from the time when the photons decouple from matter particles; because the neutrinos decouple sooner, they have a head start on cooling, and we expect the neutrino background temperature to be lower than the background photon temperature. We don't, yet, have the technological capability to detect such low-energy neutrinos, so we don't yet have observational evidence to demonstrate the existence of the cosmic neutrino background.

The lepton era extends to an age of a few hundred seconds, plenty of time for more neutrons to decay, down to a ratio of roughly $n_{p^+} / n_n \sim 10 : 2$. It's just about cool enough now for protons and neutrons to combine to make deuterium, without having such energetic collisions with photons and other particles that the nucleus would immediately break apart.

Primordial nucleosynthesis. In some ways the early universe nucleosynthesis is easier than later stellar nucleosynthesis, because there are still free neutrons. In stellar nucleosynthesis the first step involves two protons overcoming their mutual Coulomb repulsion and slamming together, one of which has to turn into a neutron to make a stable nucleus of deuterium. That's a weak interaction, converting an up quark into a down quark, meaning fusion is less likely. On the other hand, unlike in the early universe, the cores of stars are not rapidly expanding and cooling, so collisions, and fusion, in stars will keep happening.

We can write the process of making primordial deuterium (D) like this:

$$p^+ + n \rightarrow \mathrm{D} + \gamma.$$

This process starts when the universe is only a few seconds old, but at that time it's hot enough for the reaction to run the opposite direction. Photodisintegration of a deuteron takes a photon with an energy greater than 2.23 MeV. By the time the universe is about 3 minutes old, it's cool enough that deuterons will survive. And the deuterons could fuse in various ways to make helium nuclei. We start nucleosynthesis with a ratio of 2 neutrons for every 10 protons. If all the neutrons combine into nuclei before there's a chance for any more of them to decay, then we expect that for every 12 nucleons, 4 will become a helium nucleus and the remaining 8 will be left-over protons. We thus predict, as a rough first estimate, that the mass fraction of primordial helium in the universe shouldn't be more than 4/12 or 1/3. Now, that was assuming no further neutron decay during the period of nucleosynthesis *and* that there were no other possible fusion reactions taking place at this time, both of which assumptions we should question. Let's look a bit more carefully.

Recall that the mean lifetime for the free neutron is ~ 881 seconds, which is not negligible compared with the ~200 seconds age of the universe when it's cool enough for the deuterium we are producing to hang around and not suffer photodisintegration. If we had a ratio of 1 neutron to 7 protons, rather than 1 to 5, then for every 16 nucleons (2 neutrons and 14 protons) we'd expect 4 to be bound into a helium nucleus 12 to be left over as free protons, meaning a helium mass fraction of 0.25 (4 nucleons / 16 nucleons).

Two deuterons could fuse directly to make a ^{4}He but there are other pathways that are a bit more energetically favorable. For instance:

$D + D \rightarrow {}^3H + p^+$ $\quad\quad$ $D + n \rightarrow {}^3H + \gamma$

$D + D \rightarrow {}^3He + n$ $\quad\quad$ $^3H + p^+ \rightarrow {}^4He + \gamma$

$^3H + D \rightarrow {}^4He + n$ $\quad\quad$ $D + p^+ \rightarrow {}^3He + \gamma$

$^3He + D \rightarrow {}^4He + p^+$ $\quad\quad$ $^3He + n \rightarrow {}^4He + \gamma$

And finally, as the temperatures fall,

$D + D \rightarrow {}^4He + \gamma$.

All of these reactions could go the opposite direction, given enough energy. None of them involve the weak interaction, and the universe is cooling, so we expect most of the tritium (^{3}H) and light helium that are produced will get processed into ^{4}He. Tritium is unstable and any that doesn't get processed into ^{4}He will ultimately (after ~18 years) decay to ^{3}He. When the universe is only a few minutes old, 18 years is effectively forever, so tritium can be considered stable for purposes of counting where the protons and neutrons get allocated in the primordial nucleosynthesis. Today, trying to observe primordial abundances of various isotopes, we'd assume any remaining tritium became light helium.

Could we go further, and produce heavier nuclei? A few. One problem is that the expansion and the falling temperature make collisions less frequent and less energetic and making it ever harder to overcome the Coulomb repulsion between the positively charged nuclei. Another significant roadblock is the fact that there are no stable nuclei with masses of 5 or 8 nucleons. The easiest reaction would have seemed to be adding either a proton or a neutron to a ^{4}He, but that won't work because ^{5}He and ^{5}Li are unstable and promptly decay – both have mean lifetimes on the order of 10^{-21} seconds. Similarly, we can't just fuse two ^{4}He nuclei together. That would produce ^{8}Be, with a mean lifetime of ~10^{-16} seconds. There are two stable isotopes of lithium and one isotope of beryllium that will get created in small quantities:

$^4He + D \rightarrow {}^6Li + \gamma$

$^4He + {}^3H \rightarrow {}^7Li + \gamma$

$^4He + {}^3He \rightarrow {}^7Be + \gamma$

There are still free electrons, and, with a mean lifetime of ~10 weeks, the ^{7}Be is likely to capture and electron and turn into ^{7}Li:

$^7Be + e^- \rightarrow {}^7Li + \nu_e$.

Some ^{7}Li will fuse with a proton and produce yet more ^{4}He:

$^7Li + p^+ \rightarrow 2\ {}^4He$.

We could represent these reactions graphically. In the following image, the number of protons (Z) is on the horizontal axis and the number of neutrons (n) is on the vertical axis. The various isotopes are indicated in boxes. The arrows show the reactions, with the parentheses listing what else goes into the reaction and what else the reaction produces. For example, the proton, in the lowest box, fuses with a neutron, emits a photon (γ), and leaves behind a deuteron, or, much more succinctly: p (n,γ) d. Remember that if it's hot enough, any of these reactions can run the other direction.

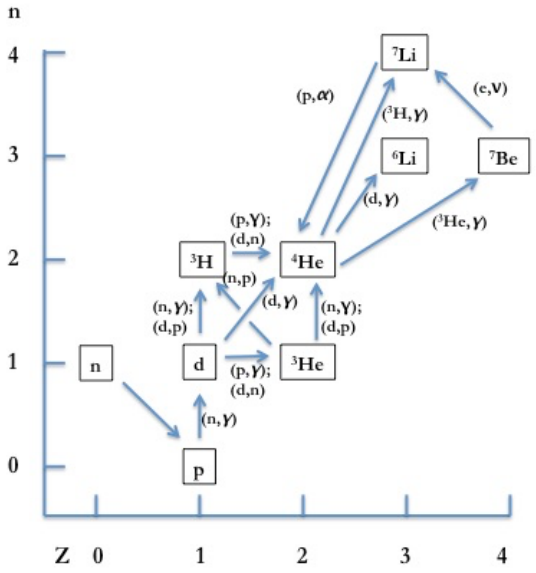


Figure 19.15:
Light-element nucleosynthesis

For all but the lowest mass stars, stellar fusion of hydrogen into helium is going to be followed, eventually, by the fusion of helium into carbon. In Big Bang nucleosynthesis, the expansion and the falling temperatures are going to make that almost impossible. On the whole, we expect that by the time the universe is a bit more than 10 minutes old, nucleosynthesis is over. The primordial nucleosynthesis should leave us with a universe composed primarily of ^{1}H and ^{4}He with lesser amounts of ^{2}H, ^{3}He, ^{6}Li and ^{7}Li. How far the nucleosynthesis proceeds depends critically on the density of the early universe. If, for instance, the ratio of baryons to photons had been higher (than the ~1 in a billion) we would expect nucleosynthesis to have gotten started a tad earlier, meaning we might expect to have had time for more of the intermediate products, the ^{2}H and ^{3}He, to be processed into ^{4}He, meaning a universe with a slightly, hopefully measurably, lower mass fraction of ^{3}He and ^{2}H. The expected fraction of ^{7}Li is complicated. On one hand, we produce it from fusion of ^{4}He and ^{3}H, so if there were more baryons and fusion lasts longer and more tritium goes into ^{4}He, the amount of ^{7}Li would decrease. On the other hand, some of the ^{7}Li is produced from ^{7}Be by electron capture, and that amount increases with increasing baryon density. We think that the primordial mass fraction of ^{4}He is ~0.24. But numerous paths lead to ^{4}He, making it not as sensitive an indicator as we would like of the primordial baryon density. It would be helpful if we could measure, accurately, the less abundant light isotopes, e.g., ^{2}H and ^{7}Li. Measuring primordial abundances is not easy, though, because those light isotopes quite readily get processed into something else in stellar fusion reactions.

The deuterium abundance would be most helpful because we don't expect any other reactions to produce more of it in the later universe; it should only decrease with time. Material in the local interstellar medium has already been processed somewhat through stars, so from local observations we could only get a lower limit on the primordial ratio of ^{2}H : ^{1}H, or D/H ratio. Observations of the ISM suggest that the nearby ratio is about $1.6 \cdot 10^{-5}$, so we expect the primordial D/H ratio to be larger than that. Ideally we'd like to look as far away, meaning as far back in time, as possible. Quasars at large z are helpful, because their light passes through surrounding clouds of neutral gas that will have been minimally affected by stellar nucleosynthesis. The fact that the gas is neutral matters, because it means that we could expect to see absorption lines from electron excitations. Deuterium absorption lines

are at slightly different wavelengths than ordinary hydrogen absorption lines because the mass of the nucleus has a bit of an effect on the electron energy levels. For example, the Lyman-α transition (n = 1 to n = 2) is at 121.57 nm in ordinary hydrogen and at 121.54 nm in deuterium. That's not much of a difference, but if we observe absorption lines due to a gas cloud at a large z, the redshift will make it a bit easier to tell the lines apart. Recent observations of the spectra of quasars at z ~3, for example, yield a primordial D/H ratio of ~$2.5 \cdot 10^{-5}$.

The primordial ^{7}Li abundance is a bit more problematic. Granting that in stars ^{7}Li will be destroyed in fusion reactions quite readily, it seems as though there is too little ^{7}Li in the oldest (Population II) stars in the halo of our galaxy. The primordial ^{7}Li abundance can be modeled, based on all the other light element abundances and the expected nucleosynthesis reactions. The abundance of ^{7}Li in old stars in the Milky Way may be low by a factor of ~3 compared with the expected primordial values, and that's more of a decrease than would have been expected. It is not yet clear whether the resolution of this question lies in a better understanding of the Big Bang production of ^{7}Li or the stellar destruction of ^{7}Li or a bit of both.

The following plot shows the sensitivity of the abundances of the light elements to the baryon-to-photon ratio.

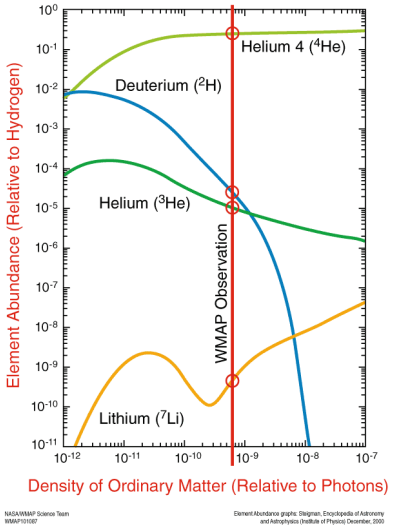


Figure 19.16: Light-element sensitivity to baryon:photon ratio

http://map.gsfc.nasa.gov/universe/bb_tests_ele.html

The observations of the abundances of the light elements provide a crucial piece of evidence about the condition in the universe in those first few minutes.

Recombination and decoupling of photons. The next Big Events in the history of the universe are often called recombination and photon decoupling and occur by the time the universe is ~380,000 years old and the temperature has fallen to ~3,000 K.

The era of nucleosynthesis ended once the temperature and density had fallen too far for the fusion reactions. As we've seen, at this point the universe was ~75% hydrogen (by mass), ~25% helium, and tiny amounts of the light elements. Even though it was no longer hot enough for fusion, it was still hot enough for most of the atoms to be spending most of their time ionized and that means there were lots of free electrons. Free electrons are very efficient at scattering photons so the universe at this time is effectively opaque; in other terms, we could say that the mean free path for a photon is short compared to the size of the universe or, equivalently, that the time between scatterings is short compared to the age of the universe. In the universe around us, we observe hydrogen atoms being ionized, usually followed by the protons recombining with the electrons to form a neutral atom once again. By analogy we call the process of atoms becoming neutral in the early universe recombination, although it might make a bit more sense to call this initial process combination, since this is the first time the universe has been

electrically neutral. There is no one instant in time when all atoms everywhere become neutral, but if we do want to tie one point in time to recombination, it makes sense to use the point at which ~50% of the atoms were neutral. That occurs when the temperature was ~3700-3800 K, or at a z of ~1370. In terms of age, the universe was ~255,000 years old. Over the course of about 100,000 years, the universe goes from being ~90% ionized to being ~10% ionized. Note that in terms of particle numbers (rather than mass), 90+% of particles are hydrogen. It's also easier to ionize hydrogen than helium (less positive charge to attract electrons to the nucleus) so it suffices to calculate the percentage of hydrogen ionization when accounting for free elections and not worry about the fact that the helium became neutral a bit earlier, on average.

As the hydrogen ionization fraction falls below ~10% the average time between photon scatterings rises to the point where it becomes longer than the age of the universe. At that point we can say that the photons decouple from matter because they stop interacting, much, with matter. (We used the same terminology, above, to describe the point at which the neutrinos stopped interacting, much, with matter.) As with recombination, there's no one exact point at which the universe became transparent, although we can say that the process of decoupling happens more rapidly than the process of becoming neutral. The universe at this point is about 380,000 years old, or at a z of ~1100. Transparent means that the photons are now free to travel across the universe until they hit our radio telescopes and are recorded as the cosmic microwave background radiation.

Prior to the universe becoming neutral, the baryonic matter and the photons were interacting all the time, regularly exchanging energy, and were in thermal equilibrium. The universe was a black body. Recall that there were anisotropies because some regions were more dense than others. Following the decoupling, there are a few processes that will have modified the energy spectrum of the photons slightly, but not by much, so it makes sense that the spectrum we observe today for the CMB is almost exactly a blackbody spectrum. It has been redshifted by ~1100 from the spectrum we would have observed at the time of the decoupling. The spectrum is remarkably isotropic, even to the extent that the anisotropies, in size and amplitude, are themselves isotropically distributed around the sky.

Structure formation. The dominant reasons for the observed variation in the anisotropies in the CMB, gravitational redshift and regions of different temperature, are both related to the fact that the densities of the various regions from which the light was being emitted at the time of decoupling are not all exactly the same. The temperature variations in the CMB are on the order of a few 10^{-5} K and to first order the temperature varies linearly with the density, meaning that the early density variations were also on the order of a few 10^{-5}. In other words, the average value of $[(\rho_{region} - \rho_{average}) / \rho_{average}]$ was $\sim 10^{-5}$, where $\rho_{average}$ represents the overall mean matter (dark + baryonic) density of the universe. That's not a very large difference and modeling how we get from these early density fluctuations to the observed large-scale structure of the universe is an active topic of research.

The imbalance of pressure and gravity is critical to the formation of any structure that forms by collapse, whether it's a star or a globular cluster or a supercluster of galaxies. The basic question is as follows: gravity will encourage an overdense region to collapse but the collapsing region will heat up and experience an outward pressure that will oppose further collapse; under what conditions will collapse proceed and structure form? When gravity and pressure are in balance, as for instance in the Sun today, the condition is called hydrostatic equilibrium. Collapse requires disequilibrium.

First, consider how long it would take for an overdense sphere to collapse under the influence of gravity and in the absence of outward pressure. It takes a bit of calculus to get there, but the result is that this dynamical timescale is

$$t_{\text{dynamical}} \sim \left(\frac{c^2}{4\pi G \bar{\varepsilon}} \right)^{1/2},$$

where the overbar on the energy density signifies the average value. We expect once the collapse starts, though, that a pressure gradient would build up (it's a gradient because the pressure would be greater in the center of the sphere) in response. The critical question for collapse is how long it would take for the pressure to change significantly. The information that collapse is happening propagates at the local speed of sound. Mathematically, we could say that

$$t_{pressure} \sim R / c_s,$$

where c_s represents the sound speed. The sound speed depends on the relationship between the pressure and the energy density; if we have an equation of state in the form $P = w\varepsilon$, with $w > 0$, then the sound speed $c_s = (w)^{1/2} \cdot c$. If we only had radiation, $w = 1/3$; if we have non-relativistic matter that doesn't interact much, w will approach 0.

If hydrostatic equilibrium is going to be achieved, the pressure timescale must be less than the collapse timescale; if collapse is going to continue, the collapse timescale must be less. For the pressure in a region to build up fast enough for the region to be stable against gravitational collapse, the region must be *smaller* than the Jeans length, named for British astrophysicist Sir James Jeans (1877 – 1946):

$$\lambda_{\text{Jeans}} = c_s \cdot \left(\frac{\pi c^2}{G\bar{\varepsilon}} \right)^{1/2}.$$

An overdense region smaller than the Jeans length will oscillate, i.e., there will be acoustic waves, at least for a while, even if it doesn't collapse. (The factor of π does belong in the numerator; we had made some simplifications in the expressions for the timescales, above.) If the universe is flat, the dynamical collapse time is very close to the characteristic expansion time, given by the inverse of the Hubble coefficient:

$$\frac{1}{H} = \left(\frac{3c^2}{8\pi G\bar{\varepsilon}} \right)^{1/2}.$$

In other words, we could express the Jeans length as

$$\lambda_{\text{Jeans}} = 2\pi\sqrt{2/3} \cdot c_s / H \cdot$$

We noted above that the sound speed is related to the pressure. If, for instance, we only had radiation in our early universe, $P = 1/3\ \varepsilon$ and $c_s = (1/3)^{1/2} c$. That's fast. And it gives a Jeans length $\sim 3c/H$, meaning that nothing smaller than about 3 times the Hubble distance could collapse. The actual early universe was not just radiation, but this calculation does show that there had to have been some non-relativistic component for structure to have formed and that it's unlikely for structure to start developing while the universe was still dominated by radiation. That may seem obvious, given that galaxies clearly contain non-relativistic matter, but getting to galaxies isn't totally straightforward.

The CMB tells us that prior to ~380,000 years the baryonic matter was still coupled to, and still being pushed around by, the photons. In fact, at the point of decoupling, the energy density in the baryons was ~70% the energy density in the photons. Baryons could not realistically start collapsing until after they decoupled from photons because no clump could possibly have had enough mass (calculated from $\rho \cdot (4\pi/3) \cdot (\lambda_J)^3$). After the decoupling, the sound speed for the baryons is

$$c_{s,\,baryons} = \left(\frac{kT}{mc^2} \right)^{1/2} \cdot c\,,$$

which, using $T \sim 3000$ K and an average particle mass, is $\sim 1.5 \cdot 10^{-5} \cdot c$, considerably less than the ~0.6 c we calculated for a photon fluid. The mass of a clump that might collapse drops in proportion to the drop in sound speed, down to about 10^5 solar masses, or about the mass of a globular cluster.

We have another wrinkle, though: remember that our model of density perturbations is that they will grow in amplitude as long as they are not supported by pressure. We were not considering the fact that the universe is expanding. Structure clearly does form and galaxies happen. It isn't enough, though, to have a gravitational collapse timescale that is faster than the timescale for pressure to build up; it also must be fast enough to keep the expansion of the universe from decreasing the density of a region to so low a level that collapse is no longer possible. When we add this piece to the puzzle what we find is that the baryons are not dense enough by themselves to produce the first galaxies that we observe by the time the universe is only a few hundred million years old. We need dark matter, and, in addition, we need most of that dark matter to be non-relativistic, or "cold", so that it will clump. The energy

density in the dark matter exceeds the baryonic matter, which helps structure form. The dark matter also doesn't interact with the photons, meaning that cold dark matter could have started to collapse about the time that the energy density in the radiation fell below the energy density in matter. That occurred at $z \sim 3600$, at an age of ~ 45,000 years, significantly before the baryons could have started collapsing.

At this point it might appear that we've hit a dead end. It would be reasonable to ask how, in the absence of more information about what dark matter is, we could continue to develop a meaningful model for the formation of structure. Think back, though, way back, back to inflation, to why we think there were any density fluctuations in the first place. Quantum mechanics tells us that the early universe cannot have been totally smooth and we might have some expectations about the nature of those density enhancements regardless of what they are made of. Were there more big ones or more little ones? Were they all overdense by the same percentage or did the amplitude of the density enhancement vary with size? Were some fluctuations so extreme that they created primordial black holes? And when we get to the formation of structure, which scales will collapse first, objects the size of stars or the size of globular clusters or the size of galaxy superclusters? In other words, can we describe the density fluctuations statistically?

Math alert: the expectation is that the components of the Fourier transform of the density field will be Gaussian (i.e., the phases of the components will be uncorrelated) and described by a power spectrum of the form $P(k) \propto k^n$, where k is the wavenumber and the index n is expected to be very close to 1 (called a Harrison – Zel'dovich spectrum). This has more power on smaller length scales that we would expect if particles were randomly distributed around the universe (for a Poisson distribution, $n = 0$). If you want to pursue this in more depth, consult a more advanced cosmology textbook; e.g., Barbara Ryden's *Introduction to Cosmology*, chapter 12.

The result from the math is that we expect the smaller, globular cluster-sized, density fluctuations to collapse first, forming galaxies, which then later form clusters of galaxies and, still later, superclusters. That's helpful, since it agrees with observations that show that the first galaxies are quite old and that superclusters, by contrast, are still forming today.

Reionization. The time period between the decoupling at ~380,000 years and the first early galaxies, in existence by ~380 million years, is called the "dark ages" because it was, quite literally, dark. The background radiation had gotten redshifted into the infrared and without any stars or quasars having formed yet there would have been almost no visible or ultraviolet light. Once we have the earliest stars and galaxies, soon to be active galaxies, that all changes. The gas, hydrogen and helium, had been neutral since the decoupling because it takes ultraviolet energies (13.6 eV for H and 24.6 and 54.4 eV for the 1st and 2nd He electrons) to ionize them. As soon as there is significant production of UV photons, the gas, especially the hydrogen, around those sources of UV will get ionized. Remember the H II regions or Strömgren spheres around hot blue stars? Those emission nebulae are approximately spherical bubbles of hydrogen around a source of UV photons; the UV means that the hydrogen atoms spend a significant proportion of their time being ions rather than neutral atoms. We expect to see the same thing in the early universe, just on a larger scale because our clouds of gas are surrounding more than just one or two stars; there's more UV available, and thus a larger amount of gas can get ionized. This process is called the reionization of the universe.

We can probe the history of the reionization because the expansion of the universe means that the source of the UV radiation and the clouds of gas along our line of sight to that source are not necessarily at the same redshift. The result is that a cloud of gas sees the UV redshifted compared to what the source originally emitted. For example, imagine that we see a background source of UV, perhaps a quasar, with a $z = 7$. Atoms of neutral hydrogen right next to that $z = 7$ source, i.e., not redshifted with respect to the source, will absorb Ly-α at the expected 121.57 nm wavelength. Imagine that there is also a cloud of gas along the line of sight with a $z = 6.5$. That means there is space between this cloud and the source and that space has expanded during the time that the light has been traveling from the source to the cloud. Atoms in this cloud of gas see the UV from the source slightly redshifted. Atoms in the $z = 6.5$ cloud would see the light the source emitted at 121.57 nm to have been redshifted and they won't absorb it. But

the source is normally going to be emitting UV at a whole range of wavelengths, not just at Ly-α. The atoms in the z = 6.5 cloud *will* absorb light that the source emitted at a *shorter* wavelength that's gotten redshifted just enough to be seen by these atoms as now being at the correct wavelength of 121.57 nm. If we add a third cloud, say at z = 6.0, that cloud will absorb what it sees as 121.57 nm from light emitted by the source at an even shorter wavelength. In other words, unlike the Strömgren spheres where the UV gets used up by the immediately surrounding hydrogen, clouds at different z will absorb light the source emitted at different wavelengths. This means that when we observe the spectrum of that background source we will see multiple absorption lines. Let's do some arithmetic on our hypothetical clouds; we will need $z + 1 = \lambda_{obs} / \lambda_{rest}$:

z of the gas cloud	z +1	Ly-α wavelength as seen by the object:	Ly-α wavelength as observed by us:
7	8	121.57 nm	972.56 nm
6.5	7.5	121.57 nm	911.78 nm
6	7	121.57 nm	850.99 nm

For this quasar we would see three absorption lines due to the hydrogen in the three separate clouds.

In reality, there are lots of clouds and lots of lines. Up to a point; we started this example by talking about the history of the reionization. There will be some lowest z for the absorption lines and then at lower z there will be no more lines. That doesn't say that there's no hydrogen any closer to us but rather that there are no more clouds of substantially *neutral* hydrogen. Hydrogen nearer to us is mostly ionized and thus not capable of producing Ly-α absorption lines. (Why didn't the universe become opaque again after reionization? It's too large and low density for today's free electrons to scatter very many photons.)

A series of absorption lines like this is called the Lyman alpha "forest". Right next to the source wavelength for Ly-α the absorption can be so total that there is no space between the absorption lines and we see in the spectrum what's called the Gunn-Peterson trough, first predicted in the 1960s but not observed until 35 years later. Away from the source, closer to us, we see many distinct lines. Here's a sketch of the geometry and the resulting spectrum:

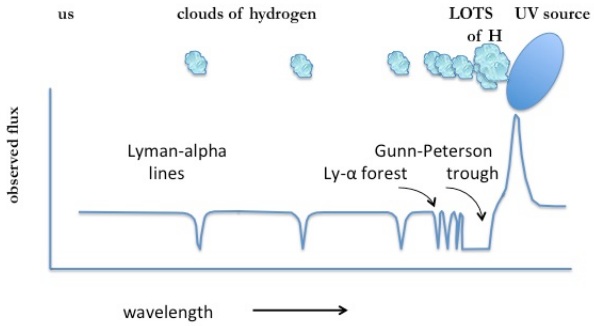


Figure 19.17:
Geometry for Lyman-α forest

The next figure shows spectra of two quasars, one at a z = 0.158 and the other at 3.62. There is additional structure in the spectrum longward of the quasar's Ly-α emission, but for our purposes here, look at the amazing number of absorption lines on the blue side of the Ly-α line in the spectrum of the more distant quasar. Both spectra have been shifted back in wavelength so that the Ly-α emission features line up. In the right-hand panel is a plot, that hasn't be un-redshifted, showing unambiguous evidence for the existence of the Gunn-Peterson trough in the bottom spectrum, a quasar at z = 6.28. Along the line of sight toward this quasar the hydrogen near the quasar was still totally neutral. Here we are seeing reionization just starting. The other quasar spectra in this plot, with z values just under 6, are already showing some structure on the blue side of the quasar's Ly-α emission, even if they don't look quite like the z = 3.62 quasar from the left panel. Roughly speaking, though, by a z ~6, the universe is showing substantial reionization; if it weren't, these other spectra would be totally flat on the blue side of the quasar's Ly-α

emission. But it's also clear that reionization does not happen instantaneously or we wouldn't see the absorption lines of the Lyman alpha forest in the $z = 3.62$ spectrum. In terms of age, by the time the universe is about 1.1 billion years old, reionization is complete.

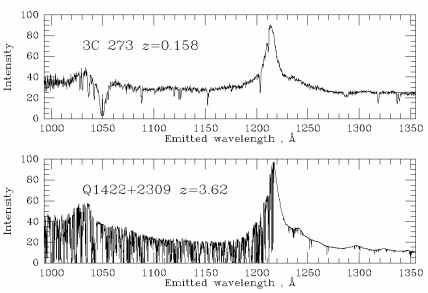


Figure 19.18: quasar spectra from Bill Keel's set of AGN slides.
http://www.astr.ua.edu/keel/agn/forest.html

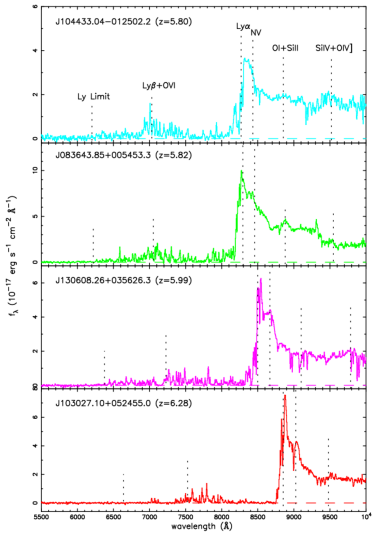


Figure 19.19: Gunn-Peterson trough;
R.H. Becker et al. 2001, *A.J.* **122**, pp. 2850-2857.

We use ionization and spectral lines of excited hydrogen to probe the structure of our galaxy and others, so it shouldn't be surprising that we can also use these indicators to probe the structure of the distant universe. We also use the 21-cm emission from neutral hydrogen in studies of galaxies and it makes sense to ask whether it could be helpful, as well, in understanding the universe around the time when reionization was beginning. One potential obstacle would be the very long wavelengths. To probe the conditions at $z \sim 6 - 12$, i.e., during the era when the hydrogen goes from being ~90% neutral to being ~90% ionized, means observing emission that's been redshifted to 150 cm and greater and, from such large distances, is very faint. It would also be interesting to ask about using 21-cm emission to investigate the clumps of still-neutral hydrogen during the period of early galaxy formation, after the majority of the reionization has taken place. Several large radio array telescopes are now or soon to be making observations of 21-cm emission at large z possible. The Square Kilometer Array, by way of example, is an international collaboration aiming to construct the world's largest radio telescope, with a collecting area of literally one square kilometer. The SKA's main locations are in regions in South Africa and Australia, chosen because of their relative lack of interference from other radio sources. (The project is described in detail at https://www.skatelescope.org .)

Here is an artist's illustration of the history of the universe, emphasizing the period of reionization and early galaxy formation.

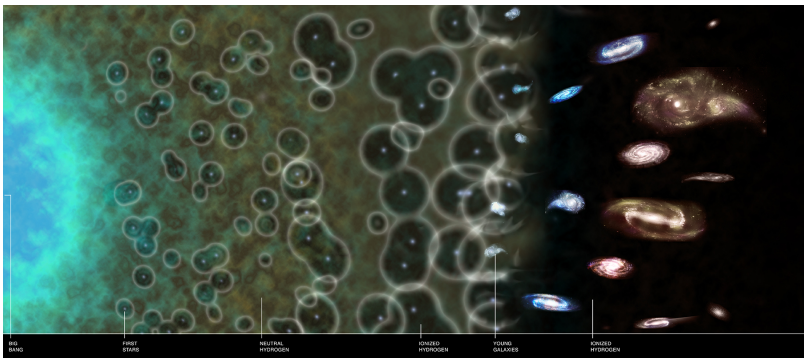


Figure 19.20: Reionization history; credit: NASA/CXC/ M.Weiss;
http://chandra.harvard.edu/photo/2011/cdfs/cdfs_ill_ionlabel.jpg

In the above illustration the early proto-galaxies are small and blue-ish, presumably containing massive Population III stars, those first generation, ~zero metals, stars. We don't yet have the observational capability to study very distant galaxies in detail. We do have, though, at least one relatively nearby galaxy that seems to have gotten a late start on most of its star formation. The following image is of the very low metallicity galaxy IZw18. Recent observations have shown the presence of ionized helium gas, which, recall, takes more energy to ionize than does hydrogen. The presence of ionized helium could suggest the presence of very massive, very hot stars – possibly Population III stars – as the source of the amount of ultraviolet light needed. At only 18 Mpc, this galaxy is clearly worthy of more study.

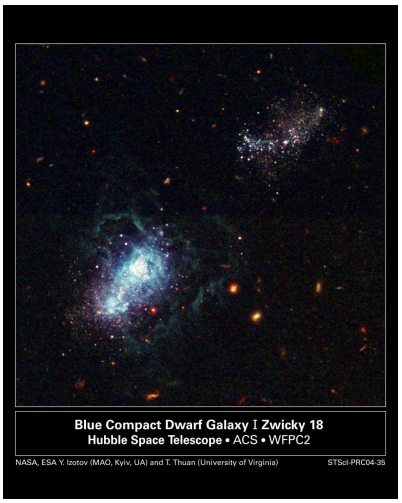


Figure 19.21: low-metallicity galaxy;
http://hubblesite.org/image/1621/news_release/2004-35

More structure, more hot gas. Our story of the history of the universe is approaching the current era, and $z \sim 0$. Once galaxies form, they collect in clusters. And clusters of galaxies will have influences on our observations of earlier, more distant events. Let's consider two of these, slightly esoteric, effects.

In the chapter on galaxies we saw that a significant fraction of the baryons in a cluster are in the intracluster / intergalactic medium, rather than confined to the galaxies themselves. This gas is hot, often over 10^7 K. In part it was heated as the material that made the cluster fell together; recall that the virial theorem tells us that half the released gravitational potential of a collapsing system will go into the kinetic energy of its components. In part the gas is heated by interactions with the galaxies. Left to itself, the gas would cool on timescales that are shorter than the lifetimes of galaxy clusters, so there must be a mechanism for maintaining its energy. Remember that even around our relatively inactive Milky Way we observe hot bubbles of plasma; more active galaxies, with supermassive black holes and active jets, can inject a substantial amount of energy into the surrounding gas because of the turbulence that they create. This may be enough to keep the intergalactic medium heated.

We know that the gas is hot because it emits in the x-ray. This particular type of emission is called free-free or thermal *bremsstrahlung*, German for "braking radiation". The intergalactic medium is low density and hot. In these conditions a free electron can often interact with a positive ion without being captured. The electron is much less massive, so for all practical purposes we can consider that it is only the electron's path that is bent and the ion's position is stationary. The electron is decelerated a bit by the interaction and that means it will radiate. The energy of the photon emitted depends on the deceleration and that depends on the kinetic energy of the electron. Since we are really looking at a large ensemble of electrons, the average kinetic energy of the electrons will be given by the temperature of the gas and will determine the overall spectrum emitted. That's why this type of radiation is called "thermal". For the IGM, at temperatures of several 10^7 K, the photons will be in the x-ray.

The following image, for example, is a Chandra x-ray image of the center of the Perseus cluster (= Abell 426), one of the brightest x-ray objects in the sky. NGC 1275 (= Perseus A), the massive galaxy near the core of the cluster, has a $z = 0.0176$, meaning the cluster is at a distance of about 70 Mpc.

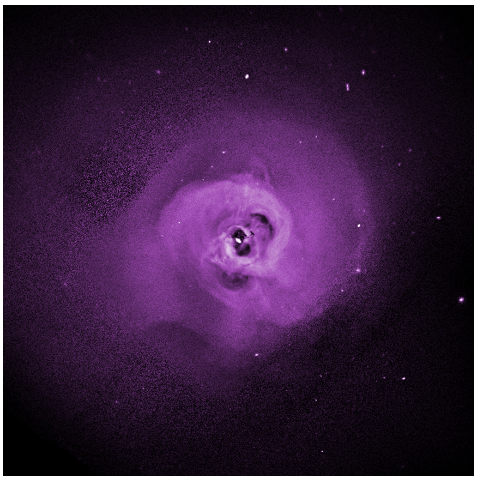


Figure 19.22: Perseus A

This image is 20 arcmin across, or ~ 1.5 million light years. It is a composite of multiple observations, made with Chandra's Advanced CCD Imaging Spectrometer (ACIS). Credit: NASA / CXC / Stanford / I. Zhuravleva et al.
http://chandra.si.edu/photo/2014/perseusvirgo/perseus.jpg

Example: what is the energy corresponding to $10^7 – 10^8$ K?

$E = kT$; in this case let's use Boltzmann's constant k in units of eV / K, giving

$E = (8.61 \cdot 10^{-5}\ \text{eV/K}) \cdot (10^{7-8}\ \text{K}) \cdot (1\ \text{keV} / 1000\ \text{eV}) = 0.9 – 9\ \text{keV}$.

In other words, a plasma of $10^7 – 10^8$ K should emit in the ~1 – 10 keV range.

Will this hot plasma have an effect on the light coming through it from farther away? Yes. The hot electrons can interact with photons of the CMB, boosting some of them to higher energies. This type of interaction is called the inverse Compton effect, where the Compton effect generally refers to interactions between photons and charged particles. In the specific context of the CMB interacting with the hot intergalactic gas, this is called the *Sunyaev-Zel'dovich Effect*, after the two researchers who pointed out in 1972 that this effect should be observable. It can create a noticeable distortion, less than ~1 mK, in the otherwise smooth blackbody spectrum of the cosmic background radiation. Early observers attempting to confirm the S-Z effect looked for decreases in the CMB temperature at radio wavelengths, i.e., wavelengths from which photons should have been removed as they were boosted to higher energies, along the line of sight toward known clusters of galaxies. By the mid-1980s there was convincing evidence that this effect is measureable. Twenty years later, astronomers had turned that project on its head and were actively searching for new galaxy clusters that could be discovered because of the S-Z effect, using, for example, the 8-dish Sunyaev-Zel'dovich Array at the Owens Valley Radio Observatory in California. On one hand, the S-Z effect gives us another tool for probing the development of structure in the universe; on the other hand, it also gives us a distortion to the CMB, something for which you must account if what you really want is to study the background radiation.

A second effect on the CMB due to galaxy clusters arises from the fact that mass is clumped along the path that the CMB photons have taken. We met the Sachs-Wolfe effect, above, in the context of the large-scale anisotropies in the CMB at the time of decoupling. Photons being emitted from regions of slightly higher density will experience a slightly larger gravitational redshift than photons from regions of lower density. As those photons are traveling toward us they will fall into and climb out of additional regions of greater than average density, namely clusters, and superclusters, of galaxies. The light of the CMB should be slightly blueshifted as it falls in and slightly redshifted as it climbs out of the gravitational potential well that is a galaxy cluster. If galaxy clusters weren't very large or if the universe weren't expanding very fast we we could argue that these effects, the blueshift and the redshift, should cancel. But the universe is likely to expand noticeably over the time it takes a CMB photon to traverse a galaxy cluster, decreasing the matter density and thus decreasing the gravitational tug that the photon feels

as it exits the cluster. It could also be the case that the galaxy cluster was still forming, increasing the matter density along the photon's path. Those two scenarios vary with time. Galaxy clusters might have formed not too long after the first galaxies whereas the expansion starts to speed up when the dark energy begins to dominate in the more recent few billion years of the history of the universe. In other words, a CMB photon is likely to have had a somewhat complicated individual history! It does seem to be the case that in the era since the dark energy took over, a path through a supercluster will result in slightly warmer photons and a path through a huge void will result in slightly cooler photons.

The future of the universe; string theory; a few unanswered questions.

At the moment the evidence suggests that we live in an expanding, flat universe now dominated by dark energy. That suggests that the future will be cold and dark. The Milky Way will merge with Andromeda and the Sun will burn out, but those events happen in the near future, say, 5-ish billion years from now. It's the very much more distant future that should be cold and dark. The brief picture is the following: Accelerating expansion carries apart objects that aren't gravitationally bound together. Galaxies run out of star-forming material and eventually evaporate approximately half of their stars / stellar remnants, with the other half losing enough energy to fall into the growing black hole at the galaxy's center. The same could be said for clusters or superclusters; evaporate a few galaxies and the rest collect in a super-supermassive black hole. Whether those evaporated stellar remnants decay or just get cold depends on whether protons decay, and that's still an open question. The black holes should evaporate and the photons that are produced in that process will themselves eventually get redshifted into oblivion. By the time the universe is 10^{100}-ish billion years old, about the time it should take a galaxy-sized black hole to evaporate, it's cold and dark. For fans of large numbers, 10^{100} is called a googol, and is larger than the estimated number of all the atoms in the observable universe, which, if you do the arithmetic, is ~10^{80}.

Example: How many baryons are there in the observable universe?

The radius of the observable universe is ~14,100 Mpc, so the volume ~$3.45 \cdot 10^{80}$ m^3. The energy density of baryons is ~210 MeV/m^3; converting that to kg and assuming that most baryons are protons, gives ~$8 \cdot 10^{79}$ baryons in the observable universe:

$(4/3)\pi \cdot (14{,}100 \text{ Mpc})^3 \cdot (3.086 \cdot 10^{22} \text{ m/Mpc})^3 = 3.45 \cdot 10^{80} \text{ m}^3$.

$210 \text{ MeV/m}^3 \cdot (1.602 \cdot 10^{-13} \text{ J/MeV}) / (3 \cdot 10^8 \text{ m/s})^2 = 3.74 \cdot 10^{-28} \text{ kg/m}^3$

$(3.45 \cdot 10^{80} \text{ m}^3) \cdot (3.74 \cdot 10^{-28} \text{ kg/m}^3) / (1.67 \cdot 10^{-27} \text{ kg/proton}) = 7.7 \cdot 10^{79} \text{ protons}$

Let's look at a few of the assumptions that have gone into the current standard model of the universe.

One of the most recent additions to our model is dark energy. Understanding dark energy better is going to require better calibration of distances to our "standard candles", including a better understanding of Type Ia supernovae. We do not yet have a satisfactory answer to the question of what percentage of Type Ia supernovae are doubly degenerate, i.e., are due to the collision of two degenerate stellar remnants, as opposed to the traditional model of a single white dwarf of ~1.4 solar masses exploding. It may also be possible to calibrate distances to one or more additional classes of very bright objects, such as Gamma Ray Bursts (GRBs). GRBs are thought to be associated with neutron star collisions (short, < 2 seconds, GRBs) or the collapse of the most massive stars at the ends of their lives (longer GRBs). Trying to get a handle on how to determine the absolute magnitude of any given GRB is an important area of current research because these explosions are even more energetic than Type Ia SNe and thus visible to even larger distances. The distances to which we can see Type Ia SNe can be improved: in 2017 researchers reported observations of a gravitationally lensed Type Ia supernova (iPTF16geu). The explosion, 4.3 billion years ago, appeared 50 times brighter than it might have been because of the gravitational lensing effects of a foreground galaxy. Concerted efforts to find more such lensed supernovae could significantly increase the range of distances over which Type Ia SNe provide useful measures of the expansion history of the universe.

Another observational enigma still awaiting an adequate explanation are the events dubbed Fast Radio Bursts (FRBs). FRBs are intense bursts lasting milliseconds. The first was discovered in 2007 and enough have been detected (over 60) to be fairly sure that they are isotropically distributed and come from extragalactic sources rather than anthropogenic sources of radio-frequency interference. Like pulsar signals, FRBs exhibit frequency dispersion (due to the frequency dependence of the index of refraction of plasma in the interstellar or intergalactic medium); the amount of the dispersion implies extragalactic distances. FRBs also have polarization signatures that suggest that they originate in regions of strong magnetic fields. One (FRB 121102) is known to repeat and seems to orginate from a dwarf galaxy with a redshift $z \approx 0.19$. Another nine repeating FRBs were reported in 2019 by researchers with CHIME (the Canadian Hydrogen Intensity Mapping Experiment), a radio telescope in British Columbia; ongoing observations of these objects in coming years should help immensely in the quest to understand their nature. If the distances implied by the frequency dispersion are correctly understood, the energy output of an FRB is comparable to a supernova, although several orders of magnitude less energetic than a Gamma Ray Burst. Still, once we understand them well enough, FRBs may serve as another form of standard candle. At the other end of the electromagnetic spectrum are Soft Gamma Repeaters (SGR), very energetic outbursts, lasting roughly 0.1 seconds, whose energy semms to arise in a neutron star magnetosphere. Neutron stars with an extremely strong magnetic fields, ranging from 10^8 to 10^{11} tesla, have been dubbed magnetars. Known magnetars seem to rotate even more rapidly than pulsars; it may be that being a magnetar is a normal phase in the lives of many neutron stars or that these objects result from the merger of binary neutron stars (whose total mass isn't enough to make the resulting object collapse to a black hole). There are also some hints that there's a connection between magnetars and Fast Radio Bursts, meaning that it's possible that these several phenomena are related.

In other words, we could use better standard candles, or standard rulers, or standard sirens (as gravitational wave events have been dubbed). A better characterization of the expansion history of the universe depends on having better distance determinations to distant bright objects for which we can also obtain redshifts.

Speaking of gravitational waves: It's expected that there should be a low-level / low-frequency gravitational wave background. Quite a few types of events, including many that would pre-date the CMB, are expected to produce low-frequency gravitational waves. The phase transition associated with inflation might be one of those. So might the formation or mergers of primordial black holes. On a larger / later scale, we might also expect to detect two super-massive black holes beginning to merge in the center of a galaxy collision. Low frequencies, on the order of 10^{-9} to 10^{-6} sec^{-1}, corresponds to wavelengths that are too long to be detected by Earth-based systems such as LIGO. Pulsar Timing Arrays are attempting to get around this problem by looking for evidence that gravitational waves are altering the pathlength traveled by radio signals from pulsars. Remember that the blips from pulsars arrive with a frequency corresponding to the rotation frequency of the neutron star. Millisecond pulsars are incredibly stable clocks (older pulsars may experience glitches or star quakes that could alter their rotation periods). If the radio waves from a pulsar happen to cross a gravitational wave, i.e., cross a region where spacetime is shrinking and expanding, then the timing of the arrival of the pulsar signal will change, slightly, but detectably. In 2020 researchers reported the first preliminary detection of gravitational waves using Pulsar Timing Arrays, demonstrating that this method works. Additional reports in 2023, tracking doezens for pulsars over more than two decades, solidified these results. The supermassive black holes in merging galaxies orbiting each other are the likely, but not definitive, culprit. At this point, there are too many waves coming from too many directions to pinpoint individual sources. One day, though, we may have enough data to clearly be able to add low-frequency gravitational waves to our box of tools for understanding the history of the expansion of the universe.

We are not, today, able to characterize that expansion history well enough to say with confidence that the dark energy density is actually constant with time and place. If dark energy is not constant, there are several possible alternative models for the timeline of the universe. First, it's possible that the dark energy density increases with time. In this case the rate of the acceleration of the expansion increases, ultimately exceeding the binding energy of any composite object. It would overcome the gravitational attraction that holds together everything from superclusters down to planets, it would overcome the electromagnetic attraction that holds together atoms and molecules, and ultimately even the strong force holding nucleons together. Everything would get pulled apart, leaving nothing but elementary particles and radiation. This infinite rate of expansion is called the "Big Rip". A

second possibility would be that the dark energy density decreases with time or is otherwise rendered meaningless. A prominent variant of this type of model is called the cyclic model, described in a bit more detail below.

A set of related models considers the possibility that the dark energy, or, more generally, accelerated expansion, is not constant as a function of location. Accelerated expansion describes both dark energy, now, and the early period of inflation, although the magnitude of the acceleration was very much greater during inflation. The dominant models for inflation say that the state of the whole universe changed from inflation to the state of much more sedate expansion everywhere and all at once. But suppose that it didn't. Suppose that there were regions where inflation stopped later, or hasn't stopped yet. Our "universe" could be a bubble of what we think of as normal spacetime within a larger "Universe" containing regions with very different physics. If those other regions are inherently unobservable, then effectively they are other universes and we occupy one universe within a larger "multiverse".

Another option related to these descriptions of the universe changing phase is the possibility that we might change phase again. Inflation models might be correct in stating that the energy in the vacuum dropped from the very high level during inflation to the low level we see today and yet wrong in asserting that today's energy level is actually the lowest possible energy level. It's possible that we are still in a "false vacuum" state and that at some unknown point in the future the universe could suddenly decay into an even lower energy state, presumably with devastating effects on the contents of the current universe. There is a theoretical link between the masses of the Higgs boson and the top quark and the possibility that the universe is in such a "metastable" (i.e., stable but not forever) state. At the moment the measured masses of the Higgs and the top quark are in agreement with the assertion that we are in a metastable universe, one that will decay to a lower energy level someday. But 1) there are still large uncertainties in the measurements and 2) the probability that the decay would occur any time soon (like, within a trillion years) is very low.

Better standard candles can help us better understand dark energy and the current period of accelerating expansion. Different techniques are needed if we want to see whether inflation might have left evidence we could observe today. Inflation does provide an answer to the questions of why the universe we observe is isotropic, flat, and lacking in magnetic monopoles. But inflation is not the only model to try to explain these observations. Is there some positive signature of inflation that would provide support for the assertion that such an event happened? One tactic that some researchers are currently pursuing is to examine the polarization of the CMB looking for the signature of quantum fluctuations in the gravitational field that might date from the period of inflation. Quantum fluctuations occur at the time of inflation and promptly get stretched to macroscopic scales. They are stretched to the point of being larger than the visible horizon, meaning that they get "frozen" until universe expands enough that the fluctuations re-enter the horizon, at which point they can develop further. We clearly observe the evidence of *density* fluctuations in the anisotropies in the CMB. Those density perturbations induce one sort of polarization, called E-mode, that was first reported by the DASI telescope team (the Degree Angular Scale Interferometer, at the South Pole) in 2002. Density enhancements in the early universe don't just sit still, they make sound waves, mentioned above in connection with the formation of structure. A density wave means that moving free electrons, of which we have still have plenty just prior to decoupling, see slightly hotter and slightly cooler radiation coming from different directions and, as the radiation is scattered by those electrons, it will become polarized. There won't be much polarization because this effect will only get imprinted on the radiation right at the time of decoupling and because the waves are moving in lots of different directions. Detecting it does tell us something about the wavelengths and velocities of those density waves right at the time of decoupling. There's another kind of polarization, called B-mode, that has a twist to it that isn't present in the E-mode polarization (physics note: the B-mode has a non-zero curl). *Gravitational* perturbations, making gravitational waves (as opposed to the *density* fluctuations making sound waves), are expected to produce this twisted polarization pattern. There's a wrinkle here in that gravitational perturbations actually produce both kinds of polarization patterns and there are some other mechanisms for producing B-mode polarization, like gravitational lensing. Still, researchers are actively working to detect the B-mode polarization in the CMB in the hopes of detecting the signature of primordial gravitational waves dating from the period of inflation. A successful detection would not only provide support for inflation, it would also shed light on the way quantum mechanics (the fluctuations) interacts with general relativity (the gravitational waves).

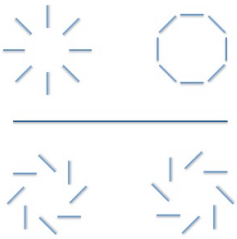


Figure 19.23: E-mode (top two) and B-mode (bottom two) polarization patterns.

This sort of observation is tough. In the spring of 2014 the BICEP2 team reported a detection of the B-mode polarization, but so far this has failed to stand up to scrutiny by other researchers who have suggested that foreground dust could create some of the signal detected by the BICEP2 instruments. The following graphic illustrates these two types of waves imprinting their polarization signals on the CMB:

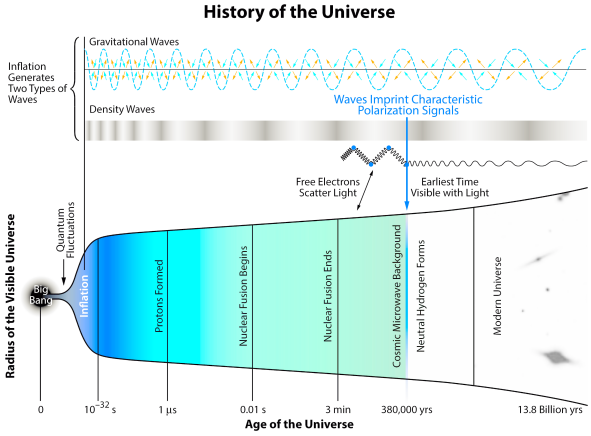


Figure 19.24; BICEP2 collaboration
http://bicepkeck.org/visuals.html

On the scale of the large and massive the universe is well described by general relativity. Likewise, on the scale of the small and energetic the universe is well described by quantum mechanics. We have problems both with elementary particles and singularities and with gravity. Why do particles have the masses they do? Are they really point particles? Why is gravity so weak? String theory is one attempt to deal more effectively with particles and gravity. There are several approaches to string theory but all of them suggest that elementary particles are not points but very tiny vibrating strings; the particle properties we observe would be due to different amounts of vibration. The most success has been with a theory that suggests that the universe has not three but nine spatial dimensions. The extra dimensions would be curled up – "compactified" – such that we don't observe them. As an analogy, imagine that you see a stretched out garden hose from a large distance. From your vantage point, it looks like a line. An ant crawling on the hose, though, can move in two dimensions, both along it lengthwise and around it. At the ant's scale, the hose has an extra curled up dimension that is not apparent to you.

String theory might be able to provide a quantum theory that successfully includes gravity and avoids nasty singularities. When distances in a theory go to zero, other quantities tend to become infinite, suggesting that there's a problem with the theory; if, on the smallest scales, the universe is just tiny curled up strings, we avoid singularities. In addition to one-dimensional strings, the theory permits higher-dimensional branes (short for membrane), which are relevant for some of the cyclic models, to which we now turn.

Inflation has its detractors. Why, for instance, should it have begun when it did and ended when it did? In other words, why should the universe have expanded by just the right amount at just the right time to drive the universe toward flat and isotropic, with a scale-invariant spectrum of density fluctuations, and any topological defects conveniently carried far away and out of view? Why is the energy difference between that early period of exponential expansion and the current period of dark energy-dominated exponential expansion so great? Inflation does not arise naturally from the known conditions of the universe and its properties require quite a bit of fine tuning. Cyclic models suppose that space and time did not originate in the big bang but that big bangs recur; a long period of contraction before the next bang sets the stage for the large scale structure of the universe that follows, doing away with the need for inflation.

The prolonged period of contraction in such a cyclic model would reheat the universe, leading to the term *ekpyrotic* (out of the fire) to describe this phase of such a model. In a cyclic model proposed by Paul Steinhardt and Neil Turok the universe as we know it could be considered to reside on a three-dimensional (+ time) brane, separated by some small distance in a fourth spatial dimension from another brane. Within our universe, now, expansion is accelerating, galaxies are being carried apart, and energy is being diluted. In the future, perhaps 10^{15} years from now, the universe will be cold, empty, and parallel to the other brane. The two branes are attracted to each other; as they get very close, quantum fluctuations create small wrinkles, amplified by the presence of the other brane. The branes collide and bounce apart, ever so slightly nonuniformly because of the wrinkles, filling the universe with hot plasma. This looks very like the traditional big bang model, only without any possibility of infinite initial temperatures and densities and without inflation. The branes separate for a time and then start slowly to come back together; within our brane, the universe is expanding, dominated first by radiation, then matter, and finally by dark energy. After another 10^{15} years, repeat. This sort of model makes different predictions for primordial gravitational radiation than inflationary models, suggesting that in the future, if we can make progress in detecting B-mode polarization in the CMB or gravitational waves themselves, we should be able to distinguish cyclic from inflationary universes.

Another possibility is that the universe we observe is a hologram. Entropy is commonly described as a measure of the disorder in some system; more technically, it is the log of the number of distinct microstates available to the particles in the system while still occupying a given macrostate. We introduced this concept, above, in considering why black holes should have a temperature and should radiate. The number of possible microstates for a system is related to the question of how many bits of information it would take to specify one of those state. In other words, entropy is intimately related to information. The second law of thermodynamics states that the entropy of a system can't decrease with time. A related idea from quantum mechanics is that information isn't destroyed. So when an object, with its associated entropy or information, falls into a black hole, where does the entropy or information go? Theoretical considerations, related to the determination that black holes should evaporate, have demonstrated that the entropy of a black hole is related to the area (*not* the volume) of the event horizon of a black hole. If the information that falls into a black hole is in some sense located on the event horizon, then we've got something very like a hologram. A hologram records information about a three-dimensional scene on a two-dimensional surface. Shine light (often a laser, related to the way the hologram was recorded) through the hologram and the three-dimensional scene is recreated. The holograms we are familiar with are two-dimensional images of three-dimensional scenes; is it possible that our four-dimensional (space + time) universe is a holographic projection of information stored on a three-dimensional "boundary" (whatever that might mean)? Maybe.

If you look at a projected hologram it appears a bit fuzzy; how fuzzy depends on the pixel size of the recorded image on the two-dimensional surface, i.e., how much information is contained in the image. If the universe is a hologram, there should be an inherent fuzziness or jitter, a pixelation, at the smallest possible length scales, the Planck scale, $\sim 10^{-35}$ m, where gravity and quantum mechanics must meet. A team at Fermilab is using a very sensitive interferometer (which they are calling a "holometer") in an attempt to measure that tiny amount of jitter in the fabric of spacetime and determine whether it matches the expected pixelation that the universe would have if it were a hologram.

Whether the universe really is a hologram or not, the holographic principle that says there is this relationship between space in n dimensions and surfaces in $n - 1$ dimensions can provide a useful tool for tackling

tough physics problems, in the sense that calculations that are exceedingly complex in our four-dimensional world might be more tractable, and provide useful results, in a different number of dimensions.

Is our universe the only one or is it part of a larger, possibly higher-dimensional, multiverse? We have seen a couple of possible versions of this question already; e.g., are there bubbles in which inflation occurs at a different time than it did in our universe, or the model in which our universe exists on a brane within a higher-dimensional space. The holographic universe could be one of a number of universes.

The many worlds interpretation of quantum mechanics, one we haven't considered before, asserts that every time a quantum mechanical observation is made – e.g., did this radioactive nucleus decay or not? – the universe splits, with each universe following a possible outcome of the observation.

One philosophical reason for asking the question is rooted in an extension of the Cosmological Principle: perhaps it is not simply that there is nothing special about our location within the universe, but, further, that there is nothing special about our universe. If that's the case, then why shouldn't we simply be one of possibly an infinite number of universes with varying properties? On the other hand, science is about making models, predictions, and observations. If our universe equals all that exists, and if other universes are not normally expected to produce observable effects, then they are not necessarily a proper subject of scientific inquiry.

If the Universe is infinite, then all possible permutations occur an infinite number of times and every possible event, as well. That would mean an infinite number of iterations of you reading these words and contemplating whether there are infinitely many universes.

Sample questions

1. Relative to the current length scale, how large was the universe when light from a galaxy with $z = 7$ left on its way toward us?

2. Explain why the energy density in matter falls as $1/a^3$ while the energy density in radiation falls as $1/a^4$.

3. Explain why we can't see to a time before the universe was ~ 380,000 years old.

4. Explain why primordial nucleosynthesis doesn't produce heavy elements.

5. On what scale can we say that the universe is isotropic and homogeneous?

6. If the sum of the angles in a triangle is *less* than 180° what can you say about the curvature of space where that triangle is located?

7. Reading carefully? Briefly explain / define:
   a) equivalence principle
   b) cosmological principle
   c) gravitational lensing
   d) Hawking radiation
   e) Casimir effect
   f) critical density
   g) Ω_Λ
   h) Planck time

i) the horizon problem
j) the flatness problem
k) Higgs field
l) *u* and *d* quarks
m) baryonic matter
n) lepton
o) decoupling
p) cold dark matter
q) reionization
r) Lyman-α forest
s) Gunn-Peterson trough
t) Sachs-Wolfe effect
u) Sunyaev-Zel'dovich effect
v) matter - anti-matter asymmetry
w) proton decay
x) ekpyrotic universe
y) dark energy
z) and *z*