1

Setting the Stage

Star Formation and Hydrogen Burning in Single Stars

1. Introduction

We look up on a dark night and wonder at the stars in their brilliant isolation. The stars are not, however, truly isolated. They are one remarkable phase in a web of interconnections that unite them with the Universe and with us as human beings. These connections range from physics on the tiniest microscopic scale to the grandest reaches in the Universe. Stars can live for times that span the age of the Universe, but they can also undergo dramatic changes on human time scales. They are born from great clouds of gas and return matter to those clouds, seeding new stars. They produce the heavy elements necessary to make not only planets but also life as we know it. The elements forged in stars compose humans who wonder at the nature of it all. Our origin and fate are bound to that of the stars. To study and understand the stars in all their manifestations from our life-giving Sun to black holes is to deepen our understanding of the role of humans in the unfolding drama of nature.

This book will focus on the exotica of stars, their catastrophic deaths, and their transfigurations into bizarre objects like white dwarfs, neutron stars, and black holes. This will lead us from the stellar mundane to the frontiers of physics. We will see how stars work, how astronomers have come to understand them, how new knowledge of them is sought, how they are used to explore the Universe, and how they lead us to contemplate some of the grandest questions ever posed.

We will begin by laying out some of the fundamental principles by which stars and, indeed, the Universe function.

2. Background

2.1. The Basic Forces of Nature

The nature of stars is governed by the push and pull of various forces. The traditional list of the basic forces of nature is as follows:

· *Electromagnetic Force* – long-range force that affects particles of positive (+) and negative (–) electrical charge, as shown in Figure 1.1 (top). *Protons* (p) are examples of positive charges, and *electrons* (e–), negative charges.

· *Strong or Nuclear Force* – short-range force that affects heavy (high-mass) particles such as protons (p) and *neutrons* (n). The strong force binds protons and neutrons together in the atomic nucleus, as shown in Figure 1.1 (middle). The strong force turns repulsive at very small distances between the particles.

· *Weak Force* – short-range force that affects interactions between light (low-mass) particles such as electrons (e–) and *neutrinos* (&ngr;). The weak force converts one light particle into another and one heavy particle into another; for instance,

as shown in Figure 1.1 (bottom).

· *Gravity* – long-range force that affects all matter and is only attractive.

The particle known as the neutrino is a special one with no electrical charge. It interacts only by means of the weak force (and gravity), that is to say scarcely at all. Its properties and its role in nature will be explained in more detail below and in later chapters.

The results of theoretical work in the 1960s by Steven Weinberg, Abdus Salaam, and Sheldon Glashow followed by experimental verification in the 1970s and 1980s by a large team led by Carlo Rubbia and Simon Van Der Meer showed that the electromagnetic and weak forces are actually manifestations of the same basic force, which has come to be called the *electroweak force*. This unification is analogous to the recognition based on the work of Thompson and Maxwell in the nineteenth century that electrical effects and magnetic effects are actually intimately interwoven in what we now call the electromagnetic force. Nobel Prizes are only the celebrated tip of the ferment that leads to scientific progress; however, their winners deserve their credit, and the prizes are signposts of major progress. Weinberg, Salaam, and Glashow won the 1979 Nobel Prize in Physics for their work; Rubbia and Van Der Meer, for theirs in 1984.

Current research is aimed at the goal of showing that the strong force is also related to the electroweak force, and that both are manifestations of some yet more fundamental force. Definite progress has already been made toward this goal of constructing a *grand unified theory*. Another dream is to show how gravity may also be understood as intrinsically related to the other forces. The story of gravity is a complex one at the heart of modern physics, and even its role in the pantheon of forces requires some interpretation. Newton interpreted gravity as a force, but, as will be elaborated in Chapter 9, Einstein's theory leads to the interpretation that gravity is a property of curved space and time, that there is no "force of gravity" in the sense that Newton conceived it. Recent dramatic progress has been made toward a unified picture of gravity and the other forces by envisaging particles as one-dimensional strings, rather than as points, as we will see in Chapter 12. In this evolving theory, gravity is again interpreted as a force, but one Newton would scarcely recognize. In practice, we will often refer to these forces in their four traditional categories, as given earlier, with emphasis where appropriate on the interpretation of gravity as a property of curved space.

2.2. Conservation Laws

To a physicist, conservation does not mean careful use to ensure future supplies, but that some quantity is constant and does not change during an interaction. Physicists have learned to make powerful use of principles of conservation, which are stated in roughly the following manner: "I don't care what goes on in detail; when all is said and done, quantity X is going to be the same." Conservation laws do not help to untangle the details of a given physical process; rather, they help to avoid complex details. Conservation laws are of great help exactly when the details are complicated because one can proceed with confidence that certain basic quantities are known and unchanging, despite the details. How this works will be more clear when we see how these conservation laws are used in various ways. They are employed to help understand why stars get hotter when energy is radiated away, the nature of nuclear reactions that power the stars, why stars become red giants and white dwarfs, the very existence and role of the elusive neutrino, how stars circle one another in binary orbits, why disks of matter form around black holes, and why some supernovae shine by radioactive decay. For now we will describe some of the conservation laws most frequently used in the astrophysics of stars.

One of the most fundamental conservation laws is the *conservation of energy*. Energy can be converted from one form to another so understanding energy conservation can sometimes be tricky, but, for all physical interactions, energy is conserved. The energy can be converted from energy of directed motion to random thermal energy and from, or to, gravitational energy. Even mass can be converted to energy and energy to mass according to Einstein's most famous formula, $E = mc^2$. Despite all these potential conversions in form, the energy of a physical system is conserved. When you drop a piece of chalk, it shatters with a small crash, as illustrated in Figure 1.2 (top). The potential gravitational energy goes first into the kinetic energy of falling, then into the energy of breaking electrical bonds among the particles of chalk, and even into the energy of the sound waves of the noise that is made. Despite the complicated details, the total energy of everything is conserved.

Momentum is a measure of the tendency of an object to move in a straight line. The measure of the momentum is not which team scored the last touchdown or goal, a common usage of the phrase in a sports context, but the product of the mass of an object with its velocity. The *mass* is a measure of the total amount of stuff in an object. The *velocity* is the speed in a given direction. Momentum characterized as mass times velocity is also conserved. A mass moving with a certain speed in a certain direction will continue to do so unless acted upon by a force. A given mass may be sped up or slowed down by the action of a force, but the agent supplying the force must suffer an equal and opposite reaction so as to conserve the momentum as a whole. Try jumping suddenly out of a boat (Figure 1.2; middle) and ask your companions if they appreciate the overwhelming verity of the principle of conservation of momentum. If you leap out one side, the boat must react by moving in the opposite direction with the same momentum as your leap. The boat will inevitably tip and leave everyone in the drink.

Angular momentum is a property related to ordinary momentum, but it measures the tendency of an object of a given mass to continue to spin at a certain rate. The measure of the angular momentum is the mass times the velocity of spin times the size of the object. A popular demonstration of conservation of angular momentum is an ice skater. When a spinning skater draws his arms in closer to his body, his “size” gets smaller. Because his mass does not change, his rate of spin must increase to ensure that his total angular momentum will be constant. In detail, this is a complex process involving the contraction and torsion of muscles and ligaments. You do not have to understand the details of how muscles and ligaments work, however, to see that the skater must end up in a dizzying spin when he pulls his arms in, and that he will slow again by simply extending his arms (Figure 1.2, bottom).

Other conservation laws are important to physics but are not reflected so easily in everyday life. An especially powerful example is that of *conservation of charge*. Electrical charge, the total number of positively and negatively charged particles, is conserved. Physical processes can cancel charges, a positive charge against a negative one, but the net positive or negative charge cannot change in a physical process. Neither positive nor negative charges can simply appear or disappear. In a reaction involving a bunch of particles, the total charge at the end of the reaction must be the same as at the beginning of the reaction. Here is an example:

Elementary particles have other properties, akin to electrical charge, that are conserved. The heavy particles like protons and neutrons that constitute atomic nuclei are called *baryons* (from the Greek "bary" meaning heavy). In a nuclear reaction, the number of baryons is conserved. The baryons may be changed from one kind to another, protons to neutrons for instance, but the number of baryons does not change. If there were four baryons at the start, there will be four at the end. The same example applies to baryons:

There are other elementary particles that do not belong to the baryon family. The ones in which we will be especially interested are the low-mass particles known as *leptons*. Electrons and neutrinos are members of this class. As for baryons, nuclear reactions conserve the total number of leptons, even though individual particles may be created or destroyed. Common reactions will involve both baryons and leptons, and both classes of particles are separately conserved. That is true in our sample reaction:

These last two conservation laws, of baryon number and lepton number, are highly accurate. These laws were once thought inviolate. Recent theoretical developments have suggested that this is not strictly true. One of the suggestions arising from the work of constructing a grand unified theory of the strong and electroweak forces is that baryons may not be completely conserved. The big bang itself may depend on the breakdown of these conservation laws. On time scales vastly longer than the age of the Universe, baryons, including all the protons and neutrons that make up the normal matter of stars, may decay into photons and light particles. For all "normal" physics, and hence for all practical purposes, baryons and leptons are conserved, and we will use these conservation laws to understand some of the reactions that are crucial to understand the nature of stars.

An important offshoot of the ideas of conservation of energy, charge, baryon number, and lepton number is the existence of matter and antimatter. For all ordinary particles – electrons, neutrinos, protons, and neutrons – there are antiparticles – antielectrons, antineutrinos, antiprotons, and antineutrons. These are not fantasy propositions; they are made routinely in what are loosely called "atom smashers," and more formally, particle accelerators, and they rain down continually on the Earth in the form of cosmic rays. The connection to the conservation of charge is that antiparticles always have the opposite charge of the "normal" particle. The antielectron, also called a *positron*, has a positive electrical charge. An antiproton has a negative charge. Because neutrinos and neutrons have no electrical charge, neither do their antiparticles, but they have other complementary properties. For instance, to make sense of the way physics works, it is necessary to consider an antielectron to count as a "negative" lepton and an antiproton to count as a "negative" baryon. In that sense, assigning the property of "leptonness" or "baryonness" to a particle is like assigning an electrical charge; it can be positive or negative and is opposite for particles and their antiparticles.

A remarkable property of particles and antiparticles is that they can be produced from pure energy and can annihilate to produce pure energy. Carl David Anderson won the Nobel Prize in Physics in 1934 for the discovery of positrons. Positrons were first created in a laboratory by applying a very strong electric field, the energy source, to an empty chamber, a vacuum. When the electric field reached a critical value, out popped electrons and positrons. You can see the connection with conservation of energy, charge, and leptons here. The energy of the electric field must be strong enough to provide the energy equivalent of the mass of an electron and a positron, twice the mass of a single electron. Because the original vacuum, even with the imposed electrical field, had no net electrical charge, the final product, the electrons and positrons, also must have no net electrical charge. For every negatively charged electron that is created in this way, there must be a particle with the opposite electrical charge, an antielectron, a positron. Likewise, the original apparatus had no "leptons," just the electrical field and vacuum. When an electron and positron appear, the electron must count as plus one lepton, and the positron as minus one lepton, so that the net number of leptons is still zero, in analogy with the way one keeps track of electrical charge. Here is a schematic reaction:

This experiment can also be run backward. If an electron and positron collide, they annihilate to produce pure energy – photons of electromagnetic energy – with no net electrical charge and no net number of leptons. The same is true of any particle and antiparticle. When they collide, they annihilate and produce pure energy; all the mass disappears. This is a very dramatic example of conservation of energy and of Einstein's formula, $E = mc^2$; pure energy can be converted into matter, and matter can be converted into pure energy. In the process, the total number of electrical charges, the total number of leptons, and the total number of baryons does not change. The total of each is always zero.

You might wonder, if antiprotons annihilate protons on contact and hence are antimatter, do they antigravitate? If I make an antiproton in a particle accelerator, will it tend to float upward? The answer is no. Energy is directly related to mass by the formula $E = mc^2$. One implication of this relation is that because mass falls in a gravitational field, energy also falls in a gravitational field. Because particles and antiparticles annihilate to form a finite, positive amount of energy that will fall in a gravitational field, so the individual particles and antiparticles must fall. An antigravitating particle might annihilate with a gravitating particle to produce no energy, but we do not know of any such particles. Current physics does give some hints of the existence of antigravity which we will discuss in Chapter 12.

2.3. The Energy of Stellar Contraction

We can now apply these various conservation laws to stars. We will start with the principle of conservation of energy. The result is a little surprising at first glance, but crucial to understanding the way in which stars evolve.

Let us first consider the nature of a star. A star is a hot ball of gas in *dynamic equilibrium*. This means that a pressure of some kind pushes outward and balances the gravity that pulls inward. The Sun does not have the same size day after day because there are no forces on it that might alter its size; rather there are great forces both inward and outward at every point in the Sun. The structure of the Sun has adapted so that the forces just balance. The equilibrium is such that the pressure force keeps gravity from collapsing the star, and gravity keeps the pressure from exploding the star. We will see in Chapter 6 that this condition of delicate balance can be interrupted and either collapse or explosion can result, depending on the circumstances. The mass and size of a star determine the gravity and hence the pressure and heat needed to arrange the balance of forces.

The Sun and most stars we see scattered in the night sky are supported by the pressure of a hot gas. The pressure, in turn, is directly related to the thermal energy in the star. At the same time, the star is held together by gravity. As the star radiates energy into space, it loses a net amount of energy. What happens to the temperature in the star? The answer is dictated by the principle of conservation of energy.

If the star were like a brick, the answer would be simple. As energy is radiated away, a brick just cools off. Gravity plays a crucial role in the makeup of a star, however. If the star were to cool, the pressure would tend to drop, and then gravity would squeeze the star, compressing and heating it. A star responds to a loss of radiant energy in just this paradoxical way. As the star loses energy, it contracts under the compression of gravity and actually heats up! This process, illustrated in Figure 1.3, is completely in accord with the conservation of energy. One must remember only that the squeezing by gravity is an important energy source that cannot be ignored when counting up all the energy, just as the energy of falling breaks the chalk in Figure 1.2.

If nuclear reactions happen by accident to momentarily put more energy into the star than it radiates, the star gains energy. What happens to its temperature in this case? If you were bitten in the first case, you should be wise by now. As shown in Figure 1.3, if the temperature were to go up, the pressure would rise and push outward against gravity. The expansion would cause the star to cool. That is just what a star does; if you put in an excess of energy, it expands and gets cooler.

This apparently contradictory behavior of a star to heat up when it loses energy and cool off when it gains energy is a direct application of the law of conservation of energy. This behavior is crucial for the evolution of stars as various nuclear fuels flare up and burn out.

2.4. Quantum Theory

Things work differently in the microscopic world of atoms and elementary particles than would seem to be "normal" from our everyday experience. On the scale of very small things, behavior is described by *quantum theory*. On this scale, changes do not occur smoothly, but in jumps. The behavior of matter on the quantum level does, however, have important implications for big things like stars.

In our ordinary macroscopic world, the old argument about the impenetrability of matter is approximately true; you cannot put your fist through a concrete wall. Your fist and the wall cannot occupy the same volume. The notion of impenetrability is very different in the microscopic world of the quantum theory. According to the quantum theory, elementary particles are not hard little balls, but also have wavelike qualities to them. Particles can, in principle, occupy exactly the same position in the same way that two ripples on a pond can occupy the same position momentarily as they pass through one another. Another aspect of the wavelike nature of particles is that their position cannot be specified. Think of the task of specifying where an ocean wave is: is it where the surface starts to curl upward, where the froth breaks on the crest, or in the wake? According to the *uncertainty principle* of the quantum theory, the positions of particles cannot be specified exactly. More precisely, there is complementary uncertainty between the position and the momentum of a particle. If the location of a particle is limited in some way, for instance by being confined in an atom, the momentum and the energy become very uncertain. If the momentum is made more certain, you do not know where the particle is. According to the quantum theory, the position of a particle is the place where it might be and the volume it occupies is a measure of the uncertainty of its position. Rather than hard spheres, particles are more like little fuzzy balls or collections of waves. This property of uncertain position, momentum, and energy allows more than one of them to occupy the same volume in the right circumstances.

There are particles in the quantum world, however, that in special situations possess a property of absolute impenetrability. Among the particles that possess this property are familiar ones – electrons, protons, neutrons, and neutrinos. Particles of this class cannot occupy the same little smeared-out uncertain region of space if they have the same momentum, or, rather loosely, the same energy. This property is known formally in the quantum theory as the *exclusion principle*. Curiously, these particles can occupy the exact same volume as long as they have different momentum or energy. Two electrons, for instance, cannot occupy the same place if they have the same momentum, but they can if they have different momentum, as shown in Figure 1.4 (top). A common particle that does not obey the exclusion principle is the photon; two photons of the same energy can occupy the same volume at the same instant.

The uncertainty and exclusion principles determine the structure of atoms. The electrons exist in a smeared volume surrounding the atomic nucleus. The size of this volume is in accord with the uncertainty principle and the fact that electrons are wavelike and their positions cannot be specified precisely. The electrons are confined into a restricted volume by the positive attraction of the protons in the nucleus. The electrons can all occupy nearly the same volume because some have higher energy, thus satisfying the constraint of the exclusion principle.

These quantum properties of particles come into play in a very important way as stars evolve. Normally the particles in a star are spread out in space and in energy, as shown in Figure 1.4 (bottom left). In this situation, the gas exerts a *thermal pressure* as the particles randomly collide and bounce off one another and generally tend to move apart. This thermal pressure associated with a hot gas supports the Sun and stars like it.

As the stars burn out their nuclear fuels, they contract and become very dense. The electrons in the stars are squeezed tightly together. The electrons get compacted into a state where the volume of quantum uncertainty occupied by each electron is bumping up against that of its neighbor. Electrons of the same energy would then absolutely resist any more compaction. That state of the star would be the maximum compression allowed according to the exclusion principle if no two electrons could occupy the same volume. Many electrons can, however, occupy the same volume if some of the electrons have extra energy. Extra energy does arise in this circumstance as a result of the compaction by gravity and the action of the uncertainty principle. As the space that the electrons occupy becomes more confined, their positions becomes more “certain.” To satisfy the uncertainty principle, the energy (strictly speaking the momentum) must become more uncertain. As the uncertainty in the electron energy becomes higher, the effective average energy of the electron increases. Thus the compaction squeezes the electrons together, the exclusion principle prevents two electrons with the same energy from occupying the same volume, and the restricted volume gives the electrons more energy in accord with the uncertainty principle. With more energy, some electrons can now occupy the same volume as illustrated in Figure 1.4 (bottom right). The fact that electrons can gain energy and hence overlap in the same volume allows greater compaction of the star.

The net effect is that the squeezing of the electrons gives them an energy that derives purely from quantum effects. The “quantum energy” that results from stellar compaction depends only on the density and is completely independent of the temperature. This quantum energy can exceed the normal thermal energy due to random motion of gas particles by great amounts. The electrons that acquire this quantum energy can also exert a *quantum pressure*. This quantum pressure can provide the pressure to hold the star up even when the thermal pressure is insufficient.

The fact that the quantum pressure is independent of the temperature has major implications for the thermal behavior of compact stars for which this pressure dominates. When a star is supported by the quantum pressure, it does not contract upon losing energy by radiating into space. The reason is that as the temperature drops, the quantum pressure is unaffected and remains constant. A star supported by quantum pressure behaves like "normal" matter; when it radiates away energy, it cools off. In this sense, such a star is more like a brick that just cools off when it radiates its heat, as illustrated in Figure 1.3.

Stars supported by the quantum pressure of electrons are known as *white dwarfs*. They will be discussed in more detail in Chapters 5. Only so much mass can be supported by the quantum pressure of electrons. This limiting mass is called the *Chandrasekhar mass* after the Indian physicist, Subramanyan Chandrasekhar, who first worked out the concept, shortly after the birth of the quantum theory. Chandrasekhar did this work as a very young man and was finally awarded the Nobel Prize for it in 1983. Chandrasekhar and his work have been honored once again by naming a major NASA orbiting observatory, the *Advanced X-ray Astronomy Facility*, the *Chandra Observatory*. The maximum mass a white dwarf can have for an ordinary composition is 1.4 solar masses, not much more massive than the Sun. If mass were to be piled onto a white dwarf so that its mass exceeded that limit, the white dwarf would collapse, or perhaps explode if it were composed of the right stuff. That notion will be explored in Chapter 6.

3. Evolution

The mass of a star sets its fate. The structure and evolution of a star of typical composition follow from the mass with which it is born. The mass determines the pressure required to hold the star up. The condition that the pressure balances gravity determines the temperature and the temperature determines the rate of nuclear burning and hence the lifetime of the star. For much of a star's life, the pressure to support it comes from the thermal pressure of a hot gas. This means that when a star loses a net amount of energy it heats up and when it gains a net amount of energy it cools off, as described in Section 2.3. This fundamental property controls the development of the star.

3.1. Birth

Stars first come into existence as protostars. Protostars are thought to form by some sort of intrinsic instability in the cold molecular gas that pervades the interstellar medium. Sufficiently massive clumps of this matter have an inward gravity that exceeds the pressure they can exert, so they contract and become ever more dense and hot until nuclear reactions start and the clump becomes a star. Alternatively, there are processes involving energetic shock waves that may cause the matter floating through space to clump together. The shocks may come from the passage of the interstellar gas through the spiral arm of a galaxy, from the explosion of a supernova, or from the flaring birth of another nearby star.

When a protostar forms, it is not yet hot enough to burn nuclear fuel. To burn nuclear fuel, the protostar must get hotter. The wonderful property of stars, even as protostars, is that if they must become hotter to yield nuclear input, they will automatically do so. That is the nature of the star machine, a machine controlled by conservation of energy under the influence of gravity. For the protostar, this works because the protostar is warmer than the cold space around it. Under this circumstance, the protostar will radiate energy into space. Because a protostar has no energy input from nuclear burning, it loses a net amount of energy into space. This is exactly the circumstance in which a star will heat up! As shown in Figure 1.5, the protostar will continue to lose energy and heat until it becomes hot enough to ignite its nuclear fuel. At this point, the protostar becomes a real star, shining with its own nuclear fire.

3.2. The Main Sequence

If you point at a person in a crowded shopping mall, the probability is that the person is middle aged, neither an infant nor very aged. The stars about us in space have a similar property. If you pick a star in the night sky at random and ask what it is doing, the probability is that it will be in the phase where stars spend most of their active lives. When stars were first categorized, most were empirically found to fall in one category in terms of the basic observable criteria of temperature and luminosity. This category is called the *main sequence*. We now understand the physical meaning of the main sequence. Stars are composed mostly of hydrogen, and the main sequence represents the phase of the thermonuclear burning of that hydrogen. Hydrogen burns for a long time compared to other elements. For this reason, stars spend most of their active lifetimes not as protostars or highly evolved stars but as hydrogen-burning stars, just as humans spend most of their lifetime as adults not as infants or octogenarians.

The Sun is in the main sequence hydrogen-burning phase. It is about halfway through its allotted span of 10 billion years. Stars more massive than the Sun burn hydrogen for a shorter time. This may seem strange because massive stars contain more hydrogen fuel to burn. The reason is that massive stars require a greater pressure to support them and hence have a higher temperature. This causes them to burn their extra fuel at a far more prodigious rate than the Sun and so spend their extra fuel in a very short time. Likewise, stars with less mass than the Sun have lower pressure and temperature. They burn their smaller ration of fuel very slowly and live on it far longer than even the Sun will. Stars with less than about 80 percent of the mass of the Sun that were born when the Galaxy first formed have scarcely begun to evolve; the Universe is not old enough yet.

A given star burns its hydrogen at a very steady rate. This is because the star acts to regulate its burning to a very precise level, using the same principle of energy conservation that ignites the fuel in the first place. If the nuclear furnace belches slightly and puts forth a little more heat than can be carried off by the radiation from the star, the excess heat increases the pressure and causes the star to expand. The excess energy is spent in making the star expand. More energy goes into the expansion than was produced in the nuclear belch, and the star actually ends up slightly cooler as explained in Section 2.3 (Figure 1.3). The nuclear reaction rates are sensitive to the temperature, and so the nuclear burning slows as the expansion occurs and the temperature drops. The net effect is a highly efficient process of negative feedback. If the star temporarily produces an iota too much heat, the nuclear fires are automatically damped a bit by the expansion to restore equilibrium. The opposite is also true. If the nuclear burning should fail to keep up with the energy radiated away for an instant, the heat would be insufficient, the pressure would drop, the star would contract, and the temperature would rise. The result is that the nuclear burning would be increased to the equilibrium value. A star burning hydrogen on the main sequence thus works in a manner similar to the thermostat and furnace in a house. If the temperature drops, the furnace kicks in and restores the lost energy. If the house gets too hot, the furnace turns off temporarily until the desired temperature is restored.

The process of hydrogen burning on the main sequence is one of thermonuclear fusion. Nuclei of hydrogen atoms, protons, are fused together to make the nucleus of the heavier element helium, which consists of two protons and two neutrons. Burning hydrogen to helium depends primarily on the nuclear force. The role of the nuclear force is to bind the four particles in the helium nucleus. The energy left over from combining the particles is available as heat. This process is not different in principle from ordinary burning where chemical forces bind the combined products together and liberate the energy of combining the molecules as heat. Chemical forces are based on the electrical force. The reason that nuclear burning is so much more powerful than chemical burning is because the nuclear force is so much stronger than the electrical force. The energy released in the fusion of hydrogen into helium is an appreciable fraction, about 1 percent, of the maximum amount of energy that could be released if all the mass of hydrogen were turned into pure energy in accordance with $E = mc^2$. That very high efficiency of energy release is why thermonuclear bombs are such a fearful weapon and why the promise of controlled thermonuclear fusion is so enticing as an ultimate energy source.

Look more closely at the process of turning hydrogen into helium. There are many ways in which this can be done in practice, but they all have a common link. The process of thermonuclear fusion consists of combining four protons to make helium. Of necessity, some step in this process requires that two of the protons be converted into two neutrons. Protons are converted into neutrons (and vice versa) by the influence of the weak force. To understand how this process works, and to reveal an important practical consequence, we must also invoke the laws of conservation of charge and of baryons and leptons as introduced in Section 2.2.

The conversion of two protons into two neutrons during hydrogen fusion conserves the number of heavy, baryon, particles; there are two to start and two in the end. That process cannot occur alone, however, because charge is not conserved; the charge on the protons cannot just disappear. One way to get around this is to produce two positively charged particles to balance the charge on the protons and to give no net change in the electrical charge. These positive particles cannot be baryons of any kind because the number of baryons in the reaction is already balanced. Nature solves this problem by providing leptons in the form of positrons. If two protons are converted into two neutrons and two positrons by the weak force, we have no net charge. Now, however, we are making two new leptons, and to conserve the lepton number, the reaction must spit out two other leptons along with the two neutrons. Recall from Section 2.2 that positrons have the opposite charge and the opposite leptonness from electrons. Algebraically, they each count as “minus one” lepton in the exit channel. The other leptons coming out of the reaction must carry no charge, because the charge is already properly balanced, but must count as “plus one” in terms of leptons in order to offset the positrons. To balance charge, baryons, and leptons all at once in this reaction, nature provides the neutrino!

The fact that the neutrino was needed to conserve all the relevant quantities in certain nuclear reactions was first realized by the Italian physicist, Enrico Fermi. It was Fermi who gave the particle its name, meaning little neutral one. Fermi was awarded the Nobel Prize for this and related work in 1938 as he prepared the world's first nuclear reactor and took seminal steps that would lead to the Manhattan Project in World War II. The neutrino was not directly detected until after the war in the 1950s when Fred Reines and colleagues registered neutrinos coming from a nuclear reactor. Reines was given the Nobel Prize for this discovery in 1995.

Figure 1.6 summarizes the essential processes that occur when hydrogen undergoes thermonuclear fusion to make helium. In that conversion, a neutrino must be made for every neutron that is produced in order to conserve baryons, leptons, and electrical charge simultaneously. For every atom of helium produced, two neutrinos must be generated. That fact represents both an opportunity and a challenge to astronomers and physicists.

3.3. The Solar Neutrino Problem

Hydrogen burns and neutrinos are produced in the centers of stars because that is where the temperature is the highest. Because neutrinos interact only by the weak force, normal stellar matter is virtually transparent to them. The neutrinos that are produced in the central hydrogen-burning reactions immediately flow out of the star at nearly the speed of light, as shown in Figure 1.7. They carry off a small amount of energy that would otherwise be available to heat the star, but this energy is not of great import. The importance of the neutrinos to astronomers is that they come directly from the center of the star, carrying information about conditions in the stellar core. Otherwise, astronomers are limited to studying photons of light that come only from the outer surface of the stars. Study of these photons is a powerful tool to deduce the nature of the inner portions of a star, but that is no substitute for being able to directly “see” inside. Neutrinos from the Sun provide that opportunity.

The problem with observing the heart of the Sun by means of neutrinos is that the neutrinos will stream through any detector unimpeded, for the same reason that they stream freely out of the star. Detection of the neutrinos depends on amassing a huge detector and then waiting for that rare time when the weak force causes a reaction within the detector. This process is totally impractical for any star but the Sun, because the great distance dilutes the neutrino “brightness” from a distant star as rapidly as it does visible photons.

The first successful effort to detect neutrinos from the Sun was the result of a multidecade effort by Ray Davis and his collaborators (see Figure 1.7). This work has not yet won a Nobel Prize, but it should. The detector consists of a hundred thousand gallons of chlorine-rich cleaning fluid. The chlorine undergoes an interaction with a neutrino by means of the weak force. This interaction turns a neutron within a chlorine nucleus into a proton, just the opposite of the reaction that produced the neutrino in the Sun. Changing a neutron in chlorine into a proton converts an atom of chlorine into an atom of radioactive argon. The argon can be collected efficiently because it is a noble gas and does not combine chemically. The tank containing the cleaning fluid is at the bottom of the Homestake gold mine in Lead, South Dakota. The underground operation is necessary to screen out cosmic ray particles that could induce spurious transitions of the chlorine to argon. The mine was vacant until the price of gold soared to astronomical highs several years ago. The Homestake company reactivated it, and for a while the scientists had to work to the sound of dynamite explosions as new veins were developed.

At first, the solar neutrino experiment gave no signal at all above the background “noise” of extraneous reactions. This caused a great deal of anguish in the astronomical community because the first opportunity to peer directly inside the Sun gave a result inconsistent with apparently straightforward theoretical predictions. With patience, a positive signal was detected. A few hundred atoms of argon are collected each month from the hundred thousand gallons of fluid! Detection of some neutrinos is more reassuring than detection of none at all, but a new serious problem still arose. The most careful analysis of a standard computer model of the Sun predicts several times more neutrinos than are observed.

The discrepancy could lie in several areas. The nuclear reactions could proceed in a different manner than we envisage. The structure of the Sun could be somehow different. Perhaps the composition, particularly the heavy elements, is not spread uniformly through the volume, as assumed. Perhaps the fundamental properties of the neutrinos themselves are different. The gold mine experiment is looking for the particular type of neutrino produced when protons change to neutrons. There are (at least) two other kinds of neutrinos. If the neutrinos have undergone a Jekyll and Hyde transformation in flight and are one of the other types when they arrive at Earth, they would not induce the desired transformation of chlorine to argon and would go undetected.

Recent developments may have given the key to this mystery. One reassuring result came from an underground neutrino detector constructed in Kamioka, Japan, called Kamiokande (Figure 1.7). This detector is a massive vat of water. Unlike the chlorine experiment, it can see neutrinos in real time and can tell the direction in which the neutrinos are moving and hence the direction from which they came. The neutrinos can trigger the conversion of a neutron to a proton in the oxygen in the water or collide with one of the electrons in the water. In either case, the particle that is hit is given substantial energy and flies rapidly through the water in the direction that the neutrino was traveling. The recoil particles give a flash of blue light known as Cerenkov radiation in the direction in which they are moving. From this flare of light in the detector, the direction of the neutrinos can be tracked. The Kamiokande experiment saw the same kind of neutrinos as the chlorine experiment and at the same low rate, but, to everyone's great relief, the neutrinos were definitely coming from the direction of the Sun! Without that confirmation, there was a small probability that the Homestake detection was some local contamination and not solar neutrinos at all. That would have made the problem even worse.

The second development may have given the real answer. The Homestake and Kamiokande experiments detect only the stream of the few high-energy, relatively easy to detect neutrinos that come from a rare version of the hydrogen-burning process. That rare process might be affected by subtle changes in the interior of the Sun that would not affect the overall power output. The chlorine and water experiments cannot detect the far more numerous neutrinos that must be produced in the basic reaction by which a proton is turned into a neutron at a rate that is directly proportional to the power that flows in radiation from the surface of the Sun. Another experiment, carefully planned for a decade in collaboration between Ray Davis and Russian physicists, uses the element gallium as a detector. This substance is sensitive to the basic flood of low-energy neutrinos that must be there because the Sun, after all, is shining. The gallium experiment also failed to see the predicted rate of neutrinos! The only remaining conclusion is that something is omitted from our simplest physical picture of the neutrinos.

As mentioned earlier, there are three different types of neutrinos, each with their antineutrinos. That there are three types of neutrinos is related to the fact that there are three types of quarks that make up other particles like protons and neutrons. When neutrinos were first discovered, it was suspected that they had no mass. If that were the case, each type of neutrino would always be the same. The fledgling grand unified theory combining the strong and electroweak forces suggests that neutrinos must have a small mass. In that case, the theory predicts, there are circumstances in which one type of neutrino can be converted to another type. If this happens round and round and back and forth among the three types of neutrinos, then by the time the neutrinos arrive at the Earth there might be roughly equal amounts of all three. In this case, only one-third of the type originally produced in the Sun that the experiments were specifically designed to register would reach the detectors. The fact that about one-third of the expected rate is observed is consistent with this notion.

This interpretation of the solar neutrino experiments strongly suggests that we not only have at last the solution to the solar neutrino problem but also have strong evidence for the grand unified theory of elementary particles. This is probably the answer, but it also raises the challenge of building more experiments to test the hypothesis.

A major step in this saga was announced in the summer of 1998 by the teams of scientists working on the new, larger underground experiment in Japan known as Super Kamiokande. This experiment found evidence that neutrinos do shift from one type to another as they interact with the Earth's atmosphere, and hence that they must have a mass, as expected from theory. The mass is not measured directly, only the difference in the masses, but this is a major breakthrough. On the other hand, to account for all the data from all the experiments, there is some discussion of the need to introduce yet another type of neutrino called a “sterile” neutrino that interacts only with neutrinos and with no other particles at all. This seems a step backward. Study of solar neutrinos still has much to teach us. We will return to neutrinos in another context in Chapters 6 and 7.