

Relativity, Spacetime and Gravity

John R. Boccio
Professor of Physics
Swarthmore College

February 14, 2011

Contents

1 Special Relativity	1
1.1 Definitions	1
1.2 Galilean Relativity	11
1.3 Special Relativity	13
1.3.1 Review of Wave Properties	15
1.3.2 Interference between Waves	16
1.4 Spacetime Diagrams	18
1.5 Radar Method	19
1.6 Special Relativity	20
1.6.1 Features of the Theory	24
1.6.2 Minkowski Spacetime Diagrams	27
1.6.3 General Spacetime Diagram Construction Procedure . . .	32
1.7 The Strange World of Special Relativity	33
1.7.1 Relationships between Events	33
1.7.2 light Cones	37
1.8 Measurements in Special Relativity	40
1.8.1 Using Lorentz Transformations	43
1.9 The Doppler Effect	45
1.9.1 Sound and the Acoustic Doppler Effect	45
1.9.2 Light and the Relativistic Doppler Effect	46
1.10 How Do We Talk to Each Other in this New Relativistic World?	49
1.11 The Famous Paradoxes	49
1.11.1 The Twin Paradox	49
1.11.2 The Pole in the Barn Paradox	51
1.11.3 Signals faster than Light Paradox	53
1.12 Basic Ideas of Classical Kinematics and Dynamics (A Quick Tour)	55
1.12.1 Kinematics (or the study of motion in time)	55
1.12.2 Newtons Laws (the crowning achievement of classical physics) .	57
1.12.3 Energy	58
1.13 Some First Thoughts about General Relativity	65
1.13.1 Black Holes	68
1.14 Digression to 4-Vectors	69
1.14.1 The Standard Language of Vectors	70

2 Notes on Mermin It's About Time	83
2.1 The Principle of Relativity	83
2.2 Combining(Small) Velocities	88
2.3 The Speed of Light	89
2.4 Combining (Any) Velocities	91
2.5 Simultaneous Events; Synchronized Clocks	96
2.6 Moving Clocks Run Slowly; Moving Sticks Shrink	101
2.7 Looking at a Moving Clock	107
2.8 The Interval between Events	110
2.9 Trains of Rockets	114
2.10 Space-Time Geometry	117
2.11 $E = Mc^2$	128
2.12 A Bit About General Relativity	128
3 Notes on Geroch General Relativity from A to B	131
3.1 Events in Space-Time: Basic Building Blocks	131
3.1.1 Events	131
3.1.2 1st Try - Everyday Experience = Aristotelian View	131
3.2 The Aristotelian View: A <i>Personalized</i> Framework	134
3.2.1 Geometrical Objects in Space-time	135
3.2.2 What about Light?	140
3.2.3 Discussion	141
3.2.4 Final Thoughts	143
3.3 The Galilean View: A <i>Democratic</i> Framework	144
3.4 Difficulties with the Galilean View	151
3.5 The Interval: The Fundamental Geometrical Object	154
3.6 The Physics and Geometry of the Interval	171
3.7 Einstein's Equation: The Final Theory	183
3.8 An Example: Black Holes	191
4 Notes on Weinberg The First Three Minutes	209
4.1 Introduction: The Giant and the Cow	209
4.2 The Expansion of the Universe	211
4.3 The Cosmic Microwave Background	227
4.4 Recipe for a Hot Universe	238
4.5 The First Three Minutes	243
4.6 A Historical Diversion	248
4.7 The First One-Hundredth Second	248
4.8 What Lies Ahead	248

5 Notes on Inflationary Universe	249
5.1 The Standard Model of Particle Physics: 1970's	249
5.2 Grand Unified Theories (GUTs)	253
5.3 The Magnetic Monopole Problem	260
5.4 The Inflationary Universe	265
5.5 Implications and Remaining Problems of the Inflationary Theory	274
5.6 A New Inflationary Theory	278
6 Latest Developments	281
6.1 Cosmic Background Radiation	281
6.2 Dark Matter	289
7 WormHoles and Time Machines	307

Chapter 1

Special Relativity

1.1 Definitions

An event intuitively means something happening in a fairly limited region of space and for a short duration in time. Mathematically, we idealize this concept to become a point in space and an instant in time. In the universe, as we understand it at this time, it requires 4 numbers to specify an event, namely, three numbers to describe spatial position and one number to describe time. This is called *4-dimensional spacetime*. We will modify this later on.

Everything that happens in the universe is either an event or a collection of events. Events are independent of observers. The four numbers describing an event are not independent of observers, as we will see.

Spacetime is the collection of *all possible events*.

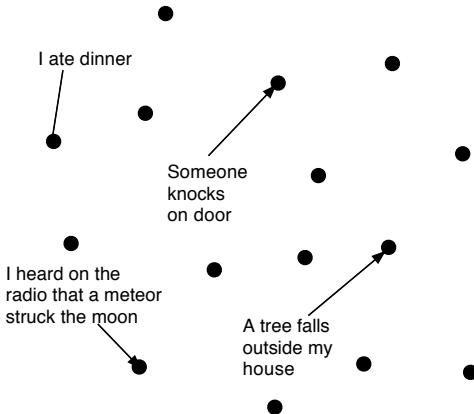


Figure 1.1: Events in Spacetime

How do we measure the *coordinates or where/when* of an event?

One method is the so-called *many-observer model*. It works as follows:

- (1) synchronize clocks ahead of experiment
- (2) measure and label grid locations ahead of experiment
- (3) observers move clocks to grid locations (assume this process has no effect on synchronization)
- (4) throw eraser into the air
- (5) if eraser passes an observers location then observer records local time
- (6) the collection of such *where and when* information gives the *set of events* representing the motion being observed

This *operational definition* of each event is simply one possible prescription for assigning numbers to the associated *where and when* in a precise and reproducible way.

For our purposes, we will assume a two-dimensional spacetime consisting of one spatial dimension and one time dimension. All the physics that we derive in this restricted universe is easily extended to the real 4-dimensional universe.

A particular set of coordinate axes and associated scales are chosen inside spacetime at our convenience and only so that we can relate the events to measured quantities in experiments, i.e., so that the theorists can talk to the experimentalists.

We represent events using a spacetime diagram (as shown below in the 2-dimensional case).

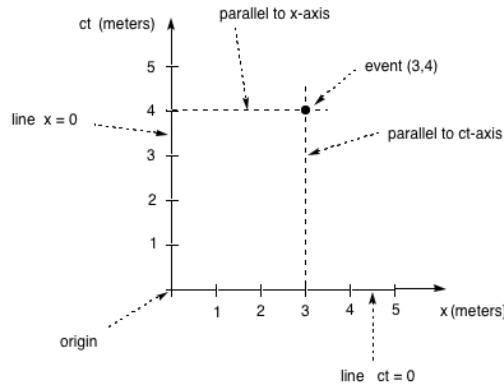


Figure 1.2: Spacetime Diagram

Note that we use ct rather than t for the vertical axis, where c is the speed of light ($c = 3.0 \times 10^8 \text{ m/sec}$). This is just a change in scale for the vertical axis and the reason for this will become clear later. Note also this is a parallel line definition of coordinate values rather than a perpendicular definition (they are different as we shall see). The definition of the coordinate axes is given by

$$ct = 1 \text{ meter} \rightarrow t = \frac{1 \text{ meter}}{3.0 \times 10^8 \text{ m/sec}} = \frac{1}{3} \times 10^{-8} \text{ sec} = 3.33 \text{ nanosecond} = 3.33 \text{ ns}$$

Many texts use a different scheme for labeling the axes as shown below:

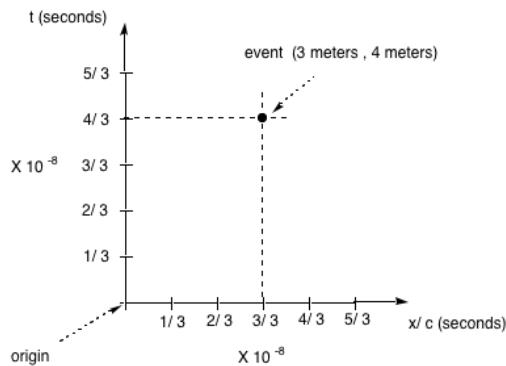


Figure 1.3: Alternate Axes

Here we have:

both t and x are measured in seconds instead of meters

$$ct = 1 \text{ meter} \rightarrow t = \frac{1}{3} \times 10^{-8} \text{ sec} \text{ and } x = 1 \text{ meter} \rightarrow \frac{x}{c} = \frac{1}{3} \times 10^{-8} \text{ sec}$$

I think it is better to use x and ct axes as we shall see. Since both of these schemes are extensively used in physics it is best that you understand both.

A *collection* of related events is called a *worldline*. Some examples of worldlines and other things are shown below:

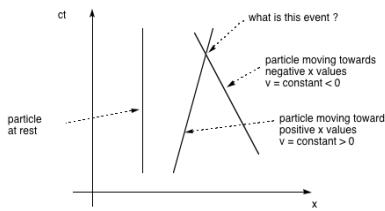


Figure 1.4: Examples of Worldlines

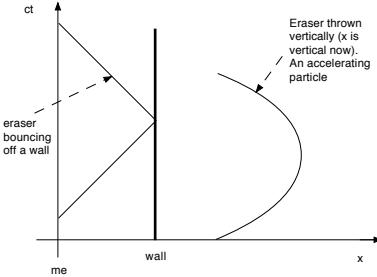


Figure 1.5: More Examples of Worldlines

Two very important lines on a spacetime diagram are shown below:

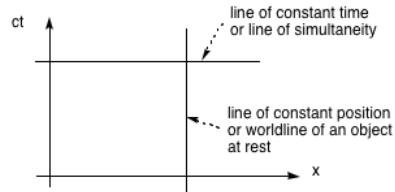


Figure 1.6: Special Lines in Spacetime

Let us now discuss these concepts in more detail. We start with things from everyday experience, which is something we hope that we know something about! Consider the diagram below

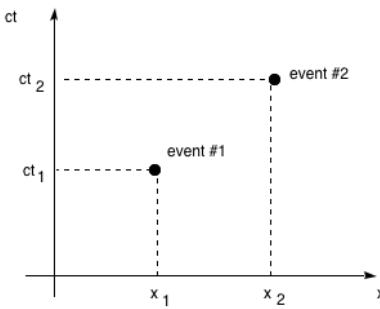


Figure 1.7: A Pair of Events

The quantity $\Delta t = t_2 - t_1$ is called the *time-separation or coordinate time* between the events and the quantity $\Delta x = x_2 - x_1$ is called the *spatial-separation* between the events.

Can we also say at this point that $\Delta t =$ time-interval between events and $\Delta x =$ distance between events?

The answer is NO!

We must be very careful not to make any such assumptions when we cannot prove the statements; a good rule is – **if we do know something is true, then we should not assume it!!**

It is clear, however, that two events on the same vertical line take place at the *same position* and two events on the same horizontal line take place at the *same time* (they are *simultaneous*) as in Figure 1.6.

In the diagram below we have several objects all moving with different speeds.

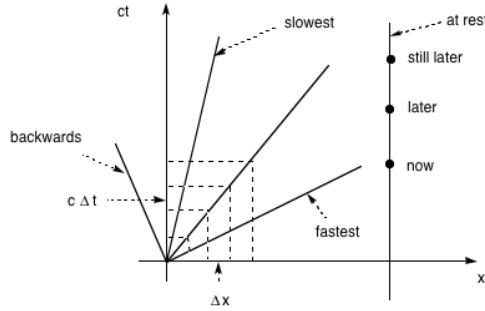


Figure 1.8: Different Speeds

In all cases, during any interval the speed is $v = \Delta x / \Delta t = 1/\text{slope}$.

Now let us imagine a car on a track and create a diagram to represent the motion of the car. The diagram below shows the worldlines corresponding to a car at rest and a car moving with constant speed in the positive x -direction.

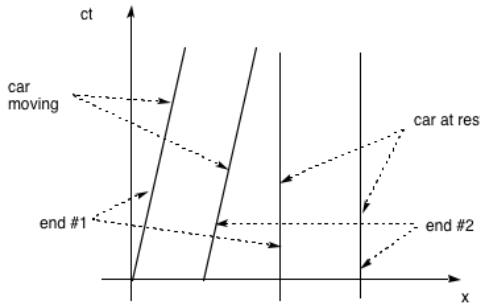


Figure 1.9: Cars at Rest and Moving

Now let us attempt to measure the length of the car. First we consider the car at rest. The diagram below represents me walking (in your frame of reference) first to one end of the car and recording its position (x_1) and then walking to

the other end and recording the position x_2).

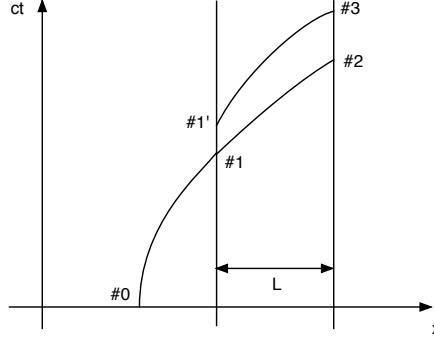


Figure 1.10: Measuring Length of a Car at Rest

At event #0, I am standing at rest and talking. I then walk over to one end of the car (event $\#1 = (x_1, ct_1)$). I then walk over to the other end of the car (event $\#2 = (x_2, ct_2)$).

Where are you on this diagram? The answer: You are the ct -axis in your own frame of reference!!

Alternatively, I could have delayed walking over to the other end of the car (event $\#1' = (x_{1'}, ct_{1'})$) and then gone over to the other end (event $\#3 = (x_3, ct_3)$).

The length of the car is then $L = x_2 - x_1$ or $L = x_3 - x_1 = x_2 - x_1$.

For a car at rest, the length measurement is the same no matter how long I delay (or whether I use event #2 or #3).

Notice how the car is just in spacetime. We do not have to be there!! What is actually in spacetime for the car? Look carefully.... all of its past, all of its future — everything about the car is in spacetime!!!!

What is the difference if we use our many-observer model? Two observers are located at the ends of the car. They record the locations and we then calculate the length. There are no gains with this approach for a car at rest!!

At this point we do not seem to have any problems with a length measurement.

Now consider a moving car as shown below:

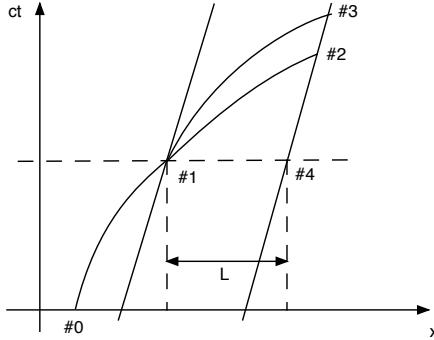


Figure 1.11: Measuring Length of a Moving Car

At event #0, I am standing at rest and talking. I then walk over to one end of the car (event #1 = (x_1, ct_1)). I then walk over to the other end of the car (event #2 = (x_2, ct_2)). Alternatively, I could have walked over more slowly to the other end of the car (event #3 = (x_2, ct_2)).

In this case, $x_3 \neq x_2$. Is the length of the car $L_{12} = x_2 - x_1$ or $L_{13} = x_3 - x_1 > L_{12}$?

As can be seen from the diagram, neither is the correct result L .

Is there an operational procedure that we can use to guarantee that we will always measure the correct length (*defined to be the length measured at rest*)?

The diagram indicates the answer. If we measure the location of the ends of the car at the same time (*simultaneously*), namely, events #1 and #4, then we get the correct length L (or along any other line of simultaneity).

Of course, this is an impossible measurement for a single observer, but not for the many-observer model. We just have all observers close their eyes and when their clock alarms go off (all set to go off simultaneously), then two of the observers will be located at the ends of the car (even if it is moving) and the length is the spatial separation of their grid locations.

So we *define* a length measurement as

**the spatial separation between the endpoints
of the object measured simultaneously**

Philosophers would not let me use the word *define* for this *operational procedure*, but they are not here now to challenge me!

A question: Have we just exchanged the unknown meaning of *length* for a new unknown, namely, *simultaneity*?

The answer is YES!

However, that is what operational procedures are all about and that is why they differ from *definitions*.

We think, at this point, that we will be able to define simultaneity unambiguously and that is better than not knowing how to *measure* length.

What about time intervals? Consider the two events shown below:

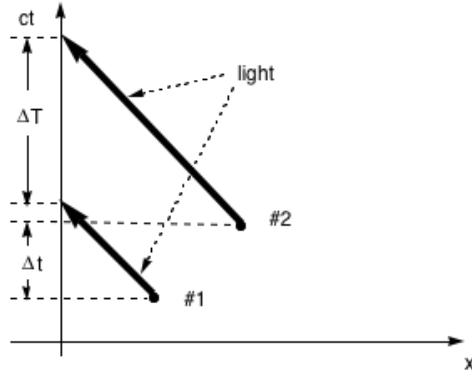


Figure 1.12: What is a Time Interval?

The time-separation between events #1 and #2 is $\Delta t = t_2 - t_1$.

Now my worldline is the ct -axis. Can I measure this quantity? We must also impose the restriction that I can only have confidence in measuring instruments that are always on my worldline(i.e., always with me). Then, the answer is NO!.

I can, however, measure the quantity ΔT , which is the time indicated on the diagram, since that measurement can be made with a clock I am carrying with me on my worldline in the following manner. Event $\#1 = (x_1, ct_1)$ takes place and sends out a light beam towards me. The light beam arrives at me (our worldlines intersect) at time $t_1 + x_1/c$, i.e., the actual time of the event $t_1 +$ the time it takes the light beam to reach me from a distance x_1 . Similarly for event $\#2$. Thus

$$\Delta T = \left(t_2 + \frac{x_2}{c} \right) - \left(t_1 + \frac{x_1}{c} \right) = \Delta t + \frac{\Delta x}{c} \quad (1.1)$$

where $\Delta x = x_2 - x_1$.

It is clear that $\Delta T > \Delta t$. If I independently know the spatial separation Δx

between the two events, then I could infer(calculate) the time-separation Δt , but this is *not* a measurement!

ΔT is the time I *see* between the two events.

Is either of these the time interval? We just do not know!

We must create an operational definition for the time interval. This is done as follows. Looking at the figure below

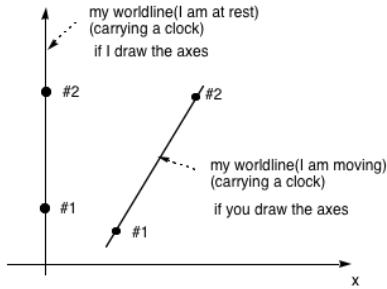


Figure 1.13: Define a Time Interval?

we see that in both cases, the *time interval* between two events is operationally defined as the difference in my clock readings, i.e., the clock and thus me must have a worldline that *passes through both events* in order to define the time-interval between the events in an unambiguous manner.

This prescription assumes that nothing happens to a clock when it moves that changes this result. We do not know that this is true.

Thus, to be safe, we *define*

**time interval between two events = time-separation
when clock is at rest and thus, the two events take
place at the same position according to the observer
carrying the clock**

as shown below:

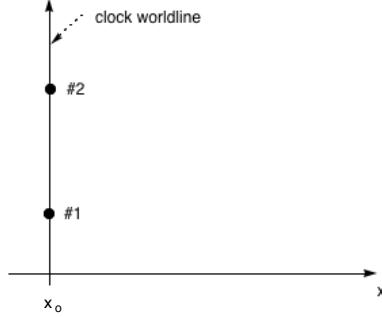


Figure 1.14: The Defined Time Interval?

The two events in this case are

$$\begin{aligned} \text{event } \#1 &= (x_0, ct_1) \\ \text{event } \#2 &= (x_0, ct_2) \end{aligned}$$

and the time interval or elapsed time between events is given by $t_2 - t_1$.

Is this what you actually do? **NO.**

You move between events usually (changing your speed in the process) and assume that this has no effect on the clock or you stay still and infer the time interval by measuring the time that you *see* between the two events. As we shall see from the theory we are developing, this is OK for the everyday world we live in but not in a world where objects move with large speeds.

We are assuming that all of these measurement procedures are objective. Suppose there is a rotten core in the apple of scientific objectivity. Physics as we shall present it works it makes correct predictions. Does it matter if we are really being subjective, i.e., that our entire view of spacetime might be dependent on human observation or that all measurements are *relative*. Philosophers spend a lot of time discussing this sort of stuff!

So we now have operational definitions that allows a *single observer* looking at the universe to describe events, measure distances and time intervals between events, and so on and report on what happened in some experiment.

Our problem arises, however, when a second observer, who is moving relative to the first observer, appears and also tries to describe the experiment using the same procedures.

1.2 Galilean Relativity

Central to any discussion of the relativity that prevailed alongside Newtonian (pre-Einstein) physics is the concept of *absolute time*.

Newton and Galileo assumed that the passage of time was the same for all observers no matter what they were doing. Thus if two observers separately measured the time interval between two events, then it was assumed that $\Delta t = t_2 - t_1 = t'_2 - t'_1 = \Delta t'$.

Suppose that two observers are moving with respect to each other (along a common x -direction) with relative speed u such that their respective origins coincide at $t = t' = 0$. Then, at some time t later, we might have the situation shown below.

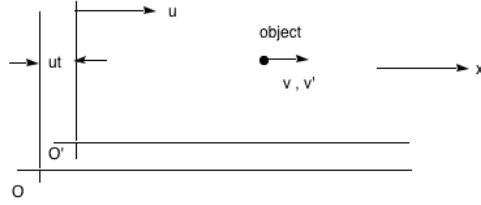


Figure 1.15: Frames in Relative Motion

We know from *everyday experience* that if observer O' measures a velocity v' and observer O measures a velocity v for some moving object, the relationship between these two measured velocities is given by $v' = v - u$.

Now to measure the velocity (assumed to be constant) of an object, each observer must observe two events in its motion. Suppose that has occurred and we have the measured results for the two events:

$$\begin{aligned} \#1 &\rightarrow (x_1, ct_1) \text{ and } (x'_1, ct'_1) \\ \#2 &\rightarrow (x_2, ct_2) \text{ and } (x'_2, ct'_2) \end{aligned}$$

We then have

$$\begin{aligned} v &= \text{velocity measured by } O(\text{frame S}) = \frac{x_2 - x_1}{t_2 - t_1} = \frac{\Delta x}{\Delta t} \\ v' &= \text{velocity measured by } O'(\text{frame S}') = \frac{x'_2 - x'_1}{t'_2 - t'_1} = \frac{\Delta x'}{\Delta t'} \end{aligned}$$

Now the absolute time concept says that $\Delta t = \Delta t'$ and this then implies that

$$\begin{aligned} v' &= v - u \\ \frac{\Delta x'}{\Delta t'} &= \frac{\Delta x}{\Delta t} - u \\ \text{or} \\ \Delta x' &= \Delta x - u\Delta t \end{aligned}$$

Now if we choose the events representing the measurement of the particle velocity to be

$$\begin{aligned} \text{event } \#1 &\rightarrow (x = 0, ct = 0) \text{ and } (x' = 0, ct' = 0) \\ \text{event } \#2 &\rightarrow (x, ct) \text{ and } (x', ct') \end{aligned}$$

which is just a choice of the *origin* values for space and time measurements (always allowed because physical phenomena are not dependent on choice of origin - confirmed by many experiments), we then obtain the equations

$$ct' = ct \quad , \quad x' = x - ut = x - \frac{u}{c}ct$$

or

$$c\Delta t' = c\Delta t \quad , \quad \Delta x' = \Delta x - u\Delta t = \Delta x - \frac{u}{c}c\Delta t$$

as the equations *relating the two sets of observations*. This relationship is shown below:

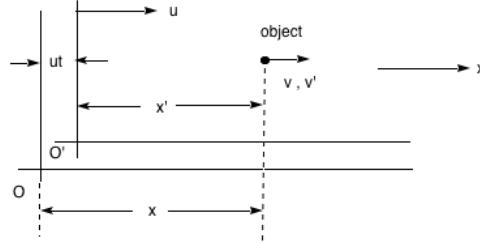


Figure 1.16: Galilean Relationships

These are the equations of *Galilean Relativity* and are called the *Galilean transformation* equations or relations.

They allow two observers in frames of reference moving with constant speed relative to each other to *compare their respective observations* under the assumption that Newtonian/Galilean physics is valid.

Galilean relativity was the basis of Newtonian physics until 1900. You have a deep understanding of Galilean relativity ingrained within your brain. If you

did not, then you would not have survived to be taking this class at Swarthmore College. Galilean relativity accurately describes the everyday world we live in.

The relative velocity formula is one of the signatures of the *old classical physics* and the everyday world.

This is the *Old Original World View* of 19th century physics circa 1900 a product of the finest minds was developed over several centuries. Everyone was comfortable with the theory. It was internally consistent. It worked amazingly well(agreed with all experiments).

And then there was light..... and

1.3 Special Relativity

First, what went wrong?

When measured by an observer at rest relative to the experiment setup, the speed of light is $c = 3.0 \times 10^8 \text{ m/sec} = 186,000 \text{ mi/sec}$, which is very large compared to everyday speeds. What is the fastest we have launched any object?

Now if you do an experiment and measure that I can throw an object with a speed of 20 m/s when I am at rest relative to you, then what speed will you observe me throwing it if I am running at a speed of 10 m/s relative to you?

The Galilean velocity relationship or velocity addition formula tells us the answer is $20 + 10 = 30 \text{ m/s}$.

Suppose instead that I am at rest and throw the object at 20 m/s and you are running in the direction opposite to that of the moving object(towards me) at 10 m/s . What speed will you measure?

Again, the Galilean velocity addition formula tells us the answer is $20 + 10 = 30 \text{ m/s}$.

Finally, suppose instead that I am at rest and throw the object at 20 m/s and you are running in the same direction as the moving object(away from me) at 10 m/s . What speed will you measure?

In this case, the Galilean velocity addition formula tells us the answer $20 - 10 = 10 \text{ m/s}$.

So it is clear that in our everyday experience with objects moving at everyday speeds, that Galilean relativity works(or that classical theory is valid). *The observed speed of objects depends on the motion of the source and observer of*

the object.

Michelson and Morley, two American physicists, did an experiment of this sort with light. They found that the speed of light was always measured to be $c = 3.0 \times 10^8 \text{ m/sec}$ no matter what the source or observer of the light was doing! Their experiments gave the astonishing result that:

$$\begin{aligned} \text{the speed of light - constant} &= c = 3.0 \times 10^8 \text{ m/sec} \\ \textbf{independent of the motion of the source or the observer} \end{aligned}$$

This leads to a direct breakdown of Galilean relativity since Galilean relativity says that for two observers in relative motion both looking at light (the moving object in this case) we must have $c' = c - u \neq c$.

Clearly, a *new theory* was needed. A very careful experiment was forcing us to make a paradigm shift in our theoretical understanding of the world. That is the way physics works! We will derive this new theory assuming one *general principle* and the *results of two experiments*:

(1) **The Principle of Relativity**

**the laws of physics are identical for all
observers in uniform relative motion**

- (2) Experiment #1: The speed of light is a universal constant c independent of the motion of the source or the observer
- (3) Experiment #2: It has been experimentally observed that when a *source* of light and a *detector of light* are moving relative to each other with a speed v the wavelength of the observed light changes with the relative speed. The experimental result is given by the formula

$$\lambda = k(v)\lambda_0$$

where λ = wavelength observed by an observer moving with speed v with respect to the light source, λ_0 = wavelength observed by an observer at rest with respect to the light source ($v = 0$) and

$$k(v) = \sqrt{\frac{c+v}{c-v}}$$

where c = speed of light and $v > 0 \rightarrow$ the source and the observer are moving away from each other ($v < 0 \rightarrow$ the source and the observer are moving towards each other).

This is the famous *galactic red shift* observed by astronomers for light

received on the earth from distant galaxies moving away from the earth.

The wavelength of a light wave is related to frequency and the period of a light wave by the formula

$$\lambda f = c = \frac{\lambda}{T}$$

where f = frequency ($\text{hertz} = \text{Hz} = \text{oscillation/sec}$), T = period (sec) and $f = 1/T$.

Other physicists might derive these results with a smaller number of assumptions. For clarity, however, at the level we are working, the derivations will be clearer if we use an extra experimental result. With a lot more work we could do the same derivation leaving out (3).

Before proceeding let us review some properties of waves.

1.3.1 Review of Wave Properties

Waves are periodic phenomena in space and time. A sinusoidal wave illustrates a typical wave ... but we really *only need periodicity*. referring to the figure below

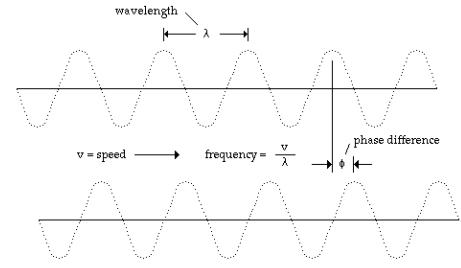


Figure 1.17: Wave Relationships

Wavelength = distance between like points

Frequency = $1/(\text{time for a point to repeat}) = 1/\text{period}$

Amplitude = maximum displacement.

Wave energy (intensity) is related to square of amplitude. The energy content of a classical wave is proportional to ($A = \text{Amplitude}$)² and is independent of frequency.

Example:

$$\begin{aligned}
 y &= A \cos(kx - \omega t) \\
 k &= \frac{2\pi}{\lambda} , \quad \omega = 2\pi f = \frac{2\pi}{T} \\
 y &= A \cos\left(\frac{2\pi}{\lambda}x - \frac{2\pi}{T}t\right) = A \cos 2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right) \\
 v &= \text{wave speed} = \lambda f
 \end{aligned}$$

Fix $t \rightarrow$ photograph of waveform in space \rightarrow wavelength.

Fix $x \rightarrow$ oscillation in time \rightarrow frequency or period.

1.3.2 Interference between Waves

Consider two waves(same amplitude, same frequency, same wavelength) which start out at the same time and propagate in this room. We assume that they travel over different paths and eventually arrive at the same point say on a screen (we assume that the time of arrival is $t = 0$ for simplicity). We then have two different (because they have traveled different distances) waves arriving at the same point with different waves given by

$$y_1 = A \cos kx_1 , \quad y_2 = A \cos kx_2$$

where x_1 and x_2 represent the *total* distances traveled. The effect of the waves at this point is given by the sum of the wave (that is the way nature works). Thus

$$y = y_1 + y_2 = A \cos kx_1 + A \cos kx_2$$

The quantity kx for each wave is called the *phase* of the wave.

We can visualize what happens at the point on the screen by looking at a different experiment. Suppose that we have two waves both traveling along the same line with different starting points. Assume that they travel the distances

$$x_1 = x \text{ and } x_2 = x + \delta$$

to get to the common point. The two traveling waves and their sum look like

$$\begin{aligned}
 y_1 &= A \cos kx , \quad y_2 = A \cos k(x + \delta) \\
 y &= y_1 + y_2 = A \cos kx + A \cos k(x + \delta)
 \end{aligned}$$

So if the waves are in phase (max to max and min to min), which means that they have traveled the same distance ($\delta = 0$) or

$$\begin{aligned}
 x_1 &= x_2 = x \\
 y_1 &= A \cos kx , \quad y_2 = A \cos kx \\
 y &= y_1 + y_2 = 2A \cos kx
 \end{aligned}$$

which corresponds to a bright spot (maximum intensity) on the screen. In the same way, if the distances differ by an integral number of wavelengths ($\delta = n\lambda$) we have

$$\begin{aligned}x_1 &= x \quad , \quad x_2 = x + n\lambda \\y_1 &= A \cos kx \quad , \quad y_2 = A \cos kx + nk\lambda = A \cos kx + 2n\pi = A \cos kx \\y &= y_1 + y_2 = 2A \cos kx\end{aligned}$$

which also corresponds to maximum intensity.

But if the waves get out of phase (if the path lengths do not differ by an integral number of wavelengths or zero) then we get smaller total amplitudes and less bright spots. In particular, if the path difference is exactly $1/2$ wavelength, then the waves cancel, that is, we have

$$\begin{aligned}x_1 &= x \quad , \quad x_2 = x + \lambda/2 \\y_1 &= A \cos kx \quad , \quad y_2 = A \cos kx + k\lambda/2 = A \cos kx + \pi = -A \cos kx \\y &= y_1 + y_2 = 0\end{aligned}$$

so that we have zero intensity or a dark spot.

When we add waves, it turns out to be just a simple algebraic sum of their amplitudes at each space-time point. This is called the *principle of superposition*.

Interference Types

constructive → phase difference = 0 or peaks line up with peaks

destructive → phase difference = $1/2$ wavelength or peaks line up with valleys

Mathematically this looks like

$$\begin{aligned}y &= A \cos kx + A \cos k(x + d) \\&= 2A \cos k \left(x + \frac{d}{2} \right) \cos \frac{kd}{2}\end{aligned}$$

Now

$$\begin{aligned}\frac{kd}{2} &= \pi \frac{d}{\lambda} \\d &= \lambda \rightarrow \frac{kd}{2} = \pi \rightarrow \cos \frac{kd}{2} = -1 \rightarrow \text{maximum} \\d &= \frac{\lambda}{2} \rightarrow \frac{kd}{2} = \frac{\pi}{2} \rightarrow \cos \frac{kd}{2} = 0 \rightarrow \text{minimum}\end{aligned}$$

Returning to our discussion of the assumption (3), we can write

$$c = \frac{\lambda}{T} \text{ (for an observer at rest wrt source)}$$

$$c' = c = \frac{\lambda'}{T'} \text{ (for an observer moving wrt source)}$$

$$T' = T_{\text{observer moving wrt source}} = k(v)T = k(v)T_0 = T_{\text{observer at rest wrt source}}$$

Here we have explicitly assumed the result of experiment (2) and experiment (3) in writing this formula, namely, that c = speed of light = constant for all observers and the red-shift relation between time intervals. The last equation follows as below.

$$c' = \frac{\lambda'}{T'} = c = \frac{\lambda}{T} \rightarrow \frac{T'}{T} = \frac{\lambda'}{\lambda} = \frac{k(v)\lambda}{\lambda} = k(v)$$

Using these results as our theoretical assumptions, we can now derive special relativity.

1.4 Spacetime Diagrams

We consider two observers A and B. Observer B is moving away from observer A with constant speed v (we consider only 1-dimensional motion for simplicity). This is represented by the spacetime diagram shown below:

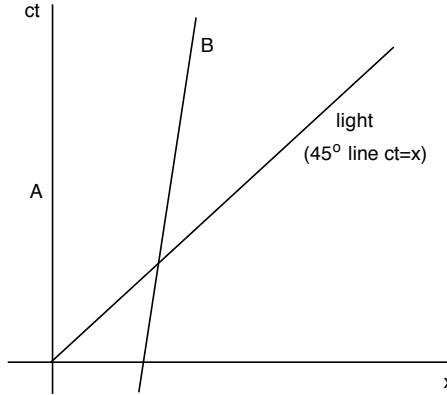


Figure 1.18: Two Observers in Spacetime

We have also included the worldline of a light beam that started at $(0, 0)$.

It is now clear why we choose the vertical axis to be ct rather than just t - the world line of light is then always a 45° line! We assume that each observer carries their own clock.

1.5 Radar Method

As we shall see from the results of our derivation, each observer will need to determine the events on the worldline of the other observer using only measurements available on their own worldline. We will not be able to trust any information that is not recorded by instruments moving with us (on the same worldline). This means that we must figure out how A (or B) can measure the (x, ct) values for an event not on their own worldline.

Remember, on my own worldline, it is true that the time interval between events that I experience (my worldline passes through them) is directly measured by the clock that I carry and my position is constant (usually assumed to be zero).

The method we now develop is called the *radar* method. Now consider the diagram below:

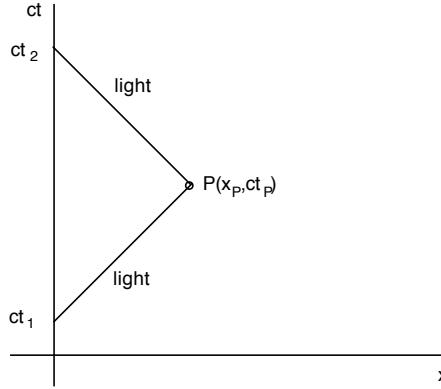


Figure 1.19: Radar Method

Observer A assigns coordinates to the event P by bouncing a light signal off of whatever is occurring at P. The light signal is sent out at the event $(0, ct_1)$ and received back at the event $(0, ct_2)$. Note the important fact that both of these events are on As worldline. We then have (using $\Delta x = c\Delta t$ for light)

$$(x_P - x_{me}) = (x_P - 0) = x_P = (ct_P - ct_1) \quad , \quad (x_P - x_{me}) = (x_P - 0) = x_P = (ct_2 - ct_P)$$

or

$$ct_P - ct_1 = ct_2 - ct_P \rightarrow ct_P = \frac{c(t_2 + t_1)}{2}$$

which is the *average* of sending and receiving times (makes sense). Then, substituting, we obtain

$$x_P = (ct_2 - ct_P) = \frac{c(t_2 - t_1)}{2}$$

Thus, any observer (a particular worldline) can determine the coordinates of an event off that worldline by only using light, which has constant speed, by assumption, for all observers and only measuring time values on their own clock (clock is on same worldline). I emphasize again that this is crucial..... we must only use information about events we actually experience (that are on our worldline), otherwise we cannot be certain of their validity.

1.6 Special Relativity

We now use this procedure and our assumptions to derive a new theory called Special Relativity (this was done by Einstein in 1905).

Consider the experiments represented by the worldlines in the spacetime diagrams below. In each case, observers A and B are assumed to be moving away from each other with speed v .

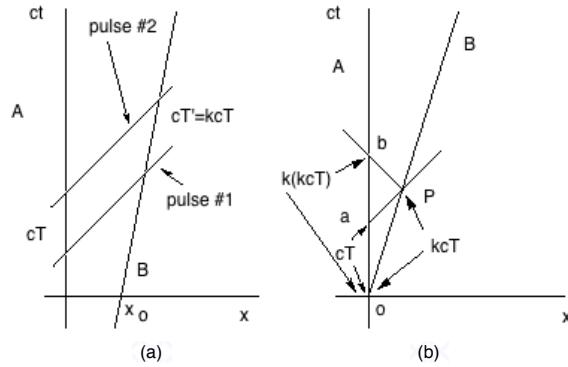


Figure 1.20: Two Experiments

In part(a), two pulses are sent from A to B. In part(b), two pulses (one is sent at $t = 0$) are sent from A to B and then B sends each of them back to A.

In part(a), B's worldline is given by the equation

$$x = x_0 + vt = x_0 + \frac{v}{c}ct \quad (\text{B is at } x_0 \text{ at } t = 0)$$

and in part(b) B's worldline is given by the equation

$$x = \frac{v}{c}ct \quad (\text{B is at } x = 0 \text{ at } t = 0)$$

In both of these cases, we are assuming the light being sent out consists of a series of pulses separated by a time T in the frame of the source (A) as shown below.

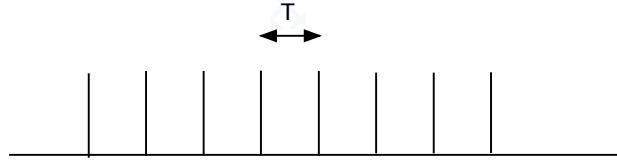


Figure 1.21: Light as Pulses

For the first experiment, our assumptions say that the interval between reception of the two signals by B (according to a clock that is traveling with B), is cT' and that this interval is proportional to cT (see diagram) with the proportionality factor $k(v)$ that depends only on the relative velocity between A and B, that is,

$$cT' = k(v)cT \quad , \quad v = \text{velocity of B wrt A}$$

In the second experiment, we see two pulses separated by T sent out by A (the first when they are at same spacetime point) and received by B separated by kT and then sent back to A and received separated by $k(kT)$.

We have used the fact that the physical laws are independent of the relative motion (assumption (1)), which requires that the relationship between A and B be reciprocal, so that, if B emits two signals separated by an interval cT (according to B's clock), then A must receive them with an interval kcT (according to A's clock). Therefore the intervals go like

$$cT \rightarrow kcT \rightarrow k(kcT)$$

as shown in the diagram.

Now Consider the experiment shown below:

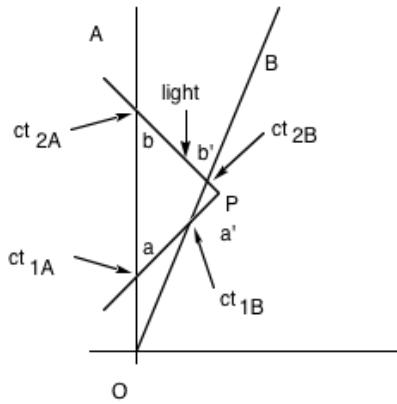


Figure 1.22: Two Observers Measuring Same Event

Here is what is happening in this diagram.

- (1) A and B synchronize their clocks to zero when their worldlines cross at event O.
- (2) After a time T (according to A) A sends a light signal to P - this is event a (a is on A's worldline).
- (3) B receives the light signal at event a' (a' is on B's worldline)
- (4) The signal is reflected back to A from event P.
- (5) B receives the reflected signal at event b' (b' is on B's worldline).
- (6) A receives the reflected signal at event b (b is on A's worldline).

For event P observer A says(using the radar method) that

$$x_P = \frac{c(t_{2A} - t_{1A})}{2}, \quad ct_P = \frac{c(t_{2A} + t_{1A})}{2}$$

and observer B says(using the radar method) that (same experiment and same equations for both A and B)

$$x'_P = \frac{c(t_{2B} - t_{1B})}{2}, \quad ct'_P = \frac{c(t_{2B} + t_{1B})}{2}$$

It is clear, using $\Delta x = c\Delta t$ and $\Delta x' = c\Delta t'$ that

$$c(t_{2A} - t_P) = x_P = c(t_P - t_{1A}) \text{ and } c(t_{2B} - t'_P) = x'_P = c(t'_P - t_{1B})$$

or

$$\begin{aligned} ct_{2A} &= ct_P + x_P \text{ and } ct_{2B} = ct'_P + x'_P \\ ct_{1A} &= ct_P - x_P \text{ and } ct_{1B} = ct'_P - x'_P \end{aligned}$$

Our earlier experimental results(Figure 1.20 (b)) now imply that

$$ct_{1B} = kct_{1A} \text{ and } ct_{2A} = kct_{2B}$$

as shown in the diagram below

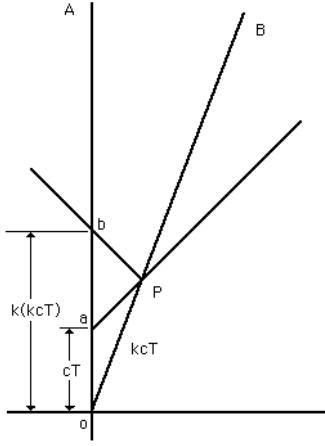


Figure 1.23: Using Assumption (3)

i.e., for observer B, the interval $OP = kcT$ (according to B's clock) and for observer A, the interval $Ob = k(kcT)$ (according to A's clock). Therefore, A has sent out a signal to event P at $ct_{1A} = cT$ and received it back at $ct_{2A} = k^2cT$.

Putting everything together and doing some algebra we get

$$\begin{aligned} ct'_P + x'_P &= \frac{ct_P + x_P}{k} \\ ct'_P - x'_P &= k(ct_P - x_P) \end{aligned}$$

Further algebra then gives(dropping the subscript P since there is nothing special about that particular spacetime point) using the value of k from assumption (3) we get

$$ct' = \gamma(ct - \beta x) \text{ and } x' = \gamma(x - \beta ct)$$

where

$$\beta = \frac{v}{c}, \quad \gamma = \sqrt{\frac{1}{1 - \beta^2}}$$

These are the so-called **Lorentz transformations**. They allow the two observers to relate their experimental results. They are *translators* between experiments done in different frame moving relative to each other with constant velocity in the common $x(x')$ direction.

We note that for relative motion in the x --direction (as above) the y and z coordinates are unchanged,i.e.,

$$y' = y, \quad z' = z$$

so that we have the relations

$$\begin{aligned}ct' &= \gamma(ct - \beta x) \\x' &= \gamma(x - \beta ct) \\y' &= y \\z' &= z\end{aligned}$$

We first note that as $v \rightarrow 0$, $k \rightarrow 1$ which implies that there is no difference between A and B (which is correct because they will then be at rest relative to each other).

Note the *mixing* of space and time so that neither is any longer independent of the other. *A very dramatic occurrence.*

So the principle of relativity together with two experimental results allows us to derive these new relations which constitute basic equations of the theory of special relativity.

This is the way theoretical physics works.

We take a mixture of general principles (things that no one can argue with) and experimental results and create a set of assumptions about the way the world works. We then derive the consequences of these assumptions, in this case, the Lorentz Transformations.

We then have a theory that agrees with our assumptions (we will show that shortly). If the theory represents a new paradigm in physics then we should be able to make new predictions not related to our assumptions that agree with all future experiments.

We can make the immediate prediction that nothing can travel faster than light. Look at the form of the γ -factor. If it were possible for $v > c$, then one observer could measure two events separated by *real* time and space intervals while a second observer would have to measure *imaginary* intervals. Since this has never been observed to happen, we can confidently predict that all objects must have $v < c$ so that γ is always real. This is corroborated by all known experiments.

This encourages the theorist to proceed further and see what other interesting features are lurking about.

1.6.1 Features of the Theory

Now that we are confident about our theory, let us work out some other features it predicts.

Suppose that we now have three observers A, B and C, such that velocity of B relative to A is $v_{BA} > 0$ and velocity of C relative to A is $v_{CA} > 0$ and velocity of C relative to B is $v_{CB} > 0$.

A then sends out two light signals, separated by interval cT (according to A) that are received by both B and C (as shown in the diagram below).

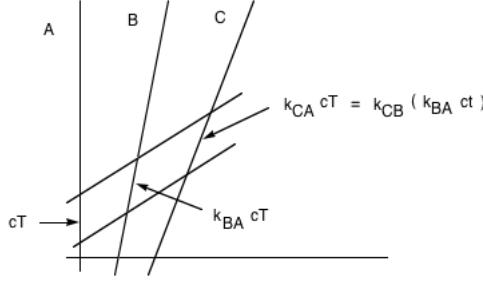


Figure 1.24: Three Observers in Relative Motion

We know from previous discussions that B thinks the interval between signals is $k_{BA}cT$ and C thinks it is $k_{CA}cT$, where

$$k_{BA} = \sqrt{\frac{c + v_{BA}}{c - v_{BA}}} , \quad k_{CA} = \sqrt{\frac{c + v_{CA}}{c - v_{CA}}}$$

In a similar manner, C could assume that the signals came from B and not A and therefore would think the interval is $k_{CB}(k_{BA}cT)$, where

$$k_{CB} = \sqrt{\frac{c + v_{CB}}{c - v_{CB}}}$$

But these two results must be identical(according to C) which means we must have

$$k_{CA}cT = k_{CB}(k_{BA}cT) \rightarrow k_{CA} = k_{CB}k_{BA}$$

This is the *relativistic velocity addition formula*. Converting to velocities we have

$$v_{CA} = \frac{v_{CB} + v_{BA}}{1 + \frac{v_{CB}v_{BA}}{c^2}}$$

This reduces back to the Newton-Galileo result for $c \ll c$, as it must, i.e.,

$$\begin{aligned} v_{CA} &= v , \quad v_{CB} = v' , \quad v_{BA} = u \\ v_{CA} &= \frac{v_{CB} + v_{BA}}{1 + \frac{v_{CB}v_{BA}}{c^2}} \rightarrow v_{CB} + v_{BA} \\ v &= v' + u \rightarrow v' = v - u \end{aligned}$$

Finally, if $v_{CB} = c$ (B is looking at a light signal) and $v_{BA} = u$ (B is moving relative to A), then we find that

$$v_{CA} = \frac{v_{CB} + v_{BA}}{1 + \frac{v_{CB}v_{BA}}{c^2}} = \frac{u + c}{1 + \frac{uc}{c^2}} = c$$

and we have the *prediction* (or *verification of our assumption*) that if *one observer* measures something moving with the speed of light c , then *all observers* will also measure its speed to be c .

In this new picture, space and time merge into a new *4-dimensional continuum*.

The most important variables in any theory are those that are unchanged for different observers. Such objects are called *invariants*.

The speed of light is such an invariant.

Another invariant is the so-called *spacetime interval*, which is constructed as follows.

Observers A and B can independently measure the spacetime coordinates for two events

$$\begin{aligned} \text{Observer A: } & (ct_{A1}, x_{A1}, y_{A1}, z_{A1}) \text{ and } (ct_{A2}, x_{A2}, y_{A2}, z_{A2}) \\ \text{Observer B: } & (ct_{B1}, x_{B1}, y_{B1}, z_{B1}) \text{ and } (ct_{B2}, x_{B2}, y_{B2}, z_{B2}) \end{aligned}$$

The Lorentz transformations relate these coordinates by

$$\begin{aligned} ct_{B1} &= \gamma(ct_{A1} - \beta x_{A1}), \quad x_{B1} = \gamma(x_{A1} - \beta ct_{A1}), \quad y_{B1} = y_{A1}, \quad z_{B1} = z_{A1} \\ ct_{B2} &= \gamma(ct_{A2} - \beta x_{A2}), \quad x_{B2} = \gamma(x_{A2} - \beta ct_{A2}), \quad y_{B2} = y_{A2}, \quad z_{B2} = z_{A2} \end{aligned}$$

Now the spacetime interval for an observer, is *defined* in general for any two events by

$$(\Delta s)^2 = c^2(\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2$$

It is then easy to show using the Lorentz transformations that the corresponding spacetime intervals for any two observer for the two events above

$$\begin{aligned} (\Delta s_A)^2 &= c^2(t_{A2} - t_{A1})^2 - (x_{A2} - x_{A1})^2 - (y_{A2} - y_{A1})^2 - (z_{A2} - z_{A1})^2 \\ (\Delta s_B)^2 &= c^2(t_{B2} - t_{B1})^2 - (x_{B2} - x_{B1})^2 - (y_{B2} - y_{B1})^2 - (z_{B2} - z_{B1})^2 \end{aligned}$$

are invariant, i.e.,

$$(\Delta s_A)^2 = (\Delta s_B)^2$$

We will investigate the powerful consequences of this result shortly.

1.6.2 Minkowski Spacetime Diagrams

We can visualize the Lorentz transformation by superposing the (x, ct) and (x', ct') planes into a *common* diagram called a *Minkowski or spacetime diagram* by following these steps:

- (1) Choose the (x, ct) axes to be perpendicular (we are always free to do this for one set of axes).
- (2) Calibrate these axes (arbitrary choice).
- (3) Locate the x' and ct' axes within the framework of the (x, ct) axes.

The x' axis is the line $ct' = 0$ and the ct' axis is line $x' = 0$.

From the Lorentz transformations these lines (axes) correspond to the equations:

$$x' = \gamma(x - \beta ct) = 0 \rightarrow ct = \frac{1}{\beta}x \rightarrow ct' - axis$$

$$ct' = \gamma(ct - \beta x) = 0 \rightarrow ct = \beta x \rightarrow x - axis$$

Thus, the x' -axis is a straight line with slope $1/\beta$ in the (x, ct) plane and the ct' -axis is a straight line with slope β in the (x, ct) plane as shown in the diagram below for the case $\beta = 3/4$:

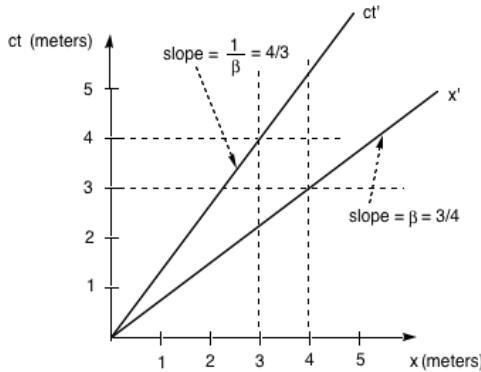


Figure 1.25: Primed Axes in Spacetime

Thus, we can only choose one set of axes as perpendicular!! We have no choice for the second set of axes if we want them to coexist on the same diagram!!

Now we see why we had to make the correct choice about parallel versus perpendicular for determining the coordinates of an event. They give very different results for non-perpendicular axes.

- (4) Calibrate the primed axes using the invariance of the interval as follows:

Consider two events, namely $(0, 0)$ and (x, ct) such that

$$(\Delta S)^2 = c^2(\Delta t)^2 - (\Delta x)^2 = c^2t^2 - x^2 = -1$$

For the second observer, these events are $(0, 0)$ and (x', ct') such that

$$(\Delta S)^2 = c^2(\Delta t')^2 - (\Delta x')^2 = c^2t'^2 - x'^2 = -1$$

where

$$\begin{aligned} ct' &= \gamma(ct - \beta x) \\ x' &= \gamma(x - \beta ct) \end{aligned}$$

and we have used the invariance of the spacetime interval. The set of all events that satisfy these equations is a curve on the spacetime diagram

This curve is a hyperbola (see the diagram below).

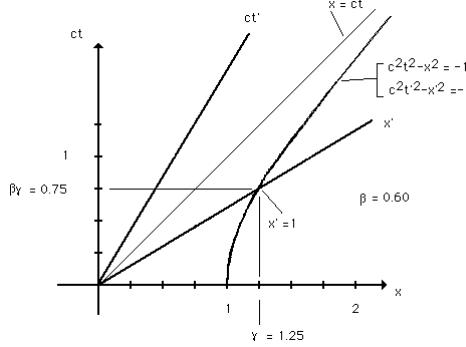


Figure 1.26: Calibration Curve in Spacetime

It intersects the x -axis at $x = 1$ (when $ct = 0$) and the x' -axis at $x' = 1$ (when $ct' = 0$) and allows us to calibrate the x' -axis once we have calibrated the x -axis (or vice versa). For diagram construction convenience we note that the point $(ct = \beta\gamma, x = \gamma)$ corresponds to the intersection determining the point $x' = 1$ as shown on the diagram.

In a similar manner, the ct' -axis is calibrated in terms of the ct -axis using the curves

$$\begin{aligned} (\Delta S)^2 &= c^2(\Delta t)^2 - (\Delta x)^2 = c^2t^2 - x^2 = +1 \\ (\Delta S)^2 &= c^2(\Delta t')^2 - (\Delta x')^2 = c^2t'^2 - x'^2 = +1 \end{aligned}$$

It intersects the ct -axis at $ct = 1$ (when $x = 0$) and the ct' -axis at $ct' = 1$ (when $x' = 0$) and allows us to calibrate the ct' -axis once we

have calibrated the ct -axis (or vice versa). For diagram construction convenience we note that the point $(ct = \gamma, x = \beta\gamma)$ corresponds to the intersection determining the point $ct' = 1$.

Note that light rays are 45° lines on the Minkowski diagram (because of our scale choice).

Using Experiment to Calibrate the Axes

Alternatively, we can use an experimental result to calibrate the time axis and then assume by symmetry that the space axis calibrates in the same manner. This experiment involves the decay of an elementary particle called a mu-meson.

A mu-meson is a short-lived elementary particle that is produced in large numbers at the top of the atmosphere when the atmosphere is struck by a high-energy cosmic ray particle. Mu-mesons can also be produced in large numbers at any accelerator laboratory.

Experimentally, if the mu-mesons are produced in the laboratory at rest ($v = 0$), then they only live for a very short time of about $\tau_0 = 2 \times 10^{-6} \text{ sec} = 2 \text{ microseconds} = 2 \text{ ms}$ = their lifetime at rest. Since we have already decided that no object can have a speed greater than $c = 3 \times 10^8 \text{ m/sec}$, the maximum distance the mu-mesons can travel during their lifetime before they decay into an electron and a neutrino is about $c\tau_0 = 600 \text{ m}$. In this calculation, we have explicitly assumed absolute time, which says that the lifetime of a moving mu-meson is the same as that of a mu-meson at rest (we now know this is not true).

The first experimental indication that absolute time was a false concept came from these mu-mesons produced by cosmic rays. Since they are produced at the top of the atmosphere and can only live to travel a maximum of 600 m and since the atmosphere is about 10000 m thick, no mu-mesons should be observed on the ground (certainly only a small number compared to the number at the top of the atmosphere).

Experimentally, however, the number at the top is the *same* as the number at the bottom. So something is extending the lifetime of the mu-mesons.

In the laboratory we can do this experiment with precision. The setup is as shown below:

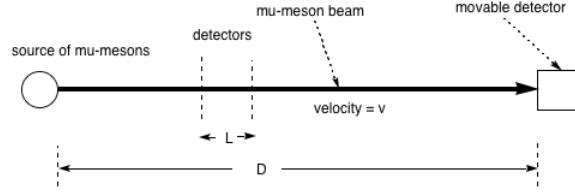


Figure 1.27: Mu-Meson Experiment

A beam of mu-mesons is sent from the source to a movable detector a distance D away. Along the way two detectors a distance L apart measure the time Δt it takes the mu-mesons to travel the distance L . This determines their velocity

$$v = \frac{L}{\Delta t}$$

If absolute time were correct, then after a distance $d = v\tau_0$ all the mu-mesons should decay and none should be seen in the movable detector if $D > d$. The experimental result is that the mu-mesons travel a maximum distance $= v\tau$ where τ is the lifetime of the moving mu-meson. The experiments found that

$$\tau = \gamma\tau_0 = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}\tau_0$$

A plot of this result looks like:

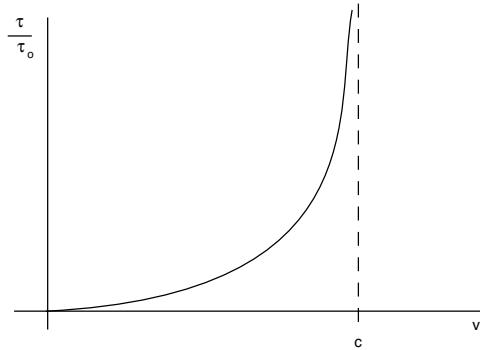


Figure 1.28: Experimental Result

The lifetime gets larger and larger the closer the velocity approaches the speed of light!!!!

If we let $\tau_0 = 1$ tick of a clock (the clock vanishes after a single tick!) and let the mu-meson travel with the primed observer, then these experimental results are represented by the diagram below:

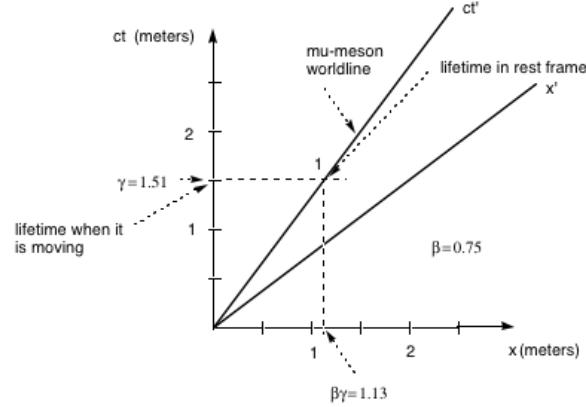


Figure 1.29: Experimental Results in Spacetime

So we see that the calibration procedure using the invariance of the spacetime interval agrees with this experimental result as shown below.

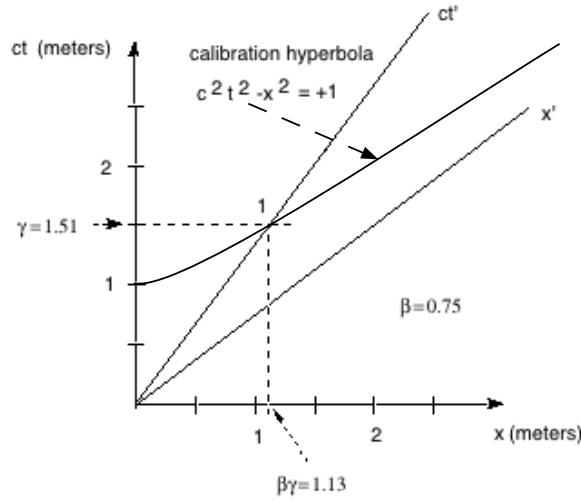


Figure 1.30: Spacetime Interval Calibration

1.6.3 General Spacetime Diagram Construction Procedure

As shown in the diagram below we carry out these steps:

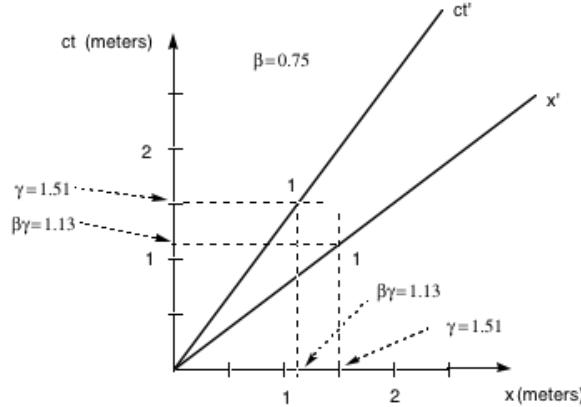


Figure 1.31: Spacetime Diagram Construction Procedure

- (1) Set up orthogonal(perpendicular) x - and ct -axes.
- (2) Choose identical scales for these axes (units(meters, light-seconds, light-years, etc) are chosen appropriate to the problem at hand)
- (3) Locate the point $(ct = \beta\gamma, x = \gamma)$ on these axes.
- (4) Draw a line from the origin $(0,0)$ through this point. That is the x' -axis.
The point $(ct = \beta\gamma, x = \gamma)$ is the point $(x' = 1, ct' = 0)$ so this calibrates the x' -axis.
- (5) Locate the point $(ct = \gamma, x = \beta\gamma)$ on these axes.
- (6) Draw a line from the origin $(0,0)$ through this point. That is the ct' -axis.
The point $(ct = \gamma, x = \beta\gamma)$ is the point $(x' = 0, ct' = 1)$ so this calibrates the ct' -axis.

The diagram is just a visual representation of the Lorentz transformation equations. It is a view of all spacetime (past, present and future).

To see that it agrees with the Lorentz transformations let us do an example.

Suppose that $\beta = 0.75$. Then we have $\gamma = 1.51$ and $\beta\gamma = 1.13$. The spacetime diagram looks like Figure 1.31 above. Now consider an event $(x = 2.0, ct = 1.75)$. The Lorentz transformations say that the other observer sees the event

$$x' = \gamma(x - \beta ct) = 1.51(2.0 - 0.75(1.75)) = 1.04$$

$$ct' = \gamma(ct - \beta x) = 1.51(1.75 - 0.75(2.0)) = 0.38$$

This result is confirmed by the diagram below:

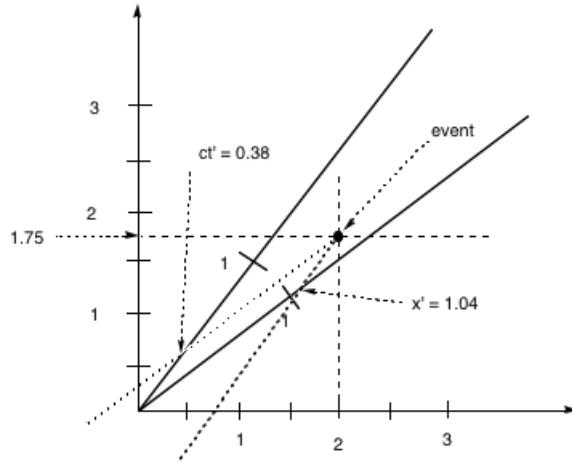


Figure 1.32: Events Coordinates in Spacetime

Note that in order to find the primed coordinate values we must draw lines *parallel* to the primed-axes.

We now have a theory called Special Relativity. We can represent it either by the Lorentz transformations, the invariance of the interval or the Minkowski spacetime diagram. All representations of the theory are equivalent. It was discovered by Albert Einstein in 1905.

What is a theory? A theory is a set of assumptions that

- (1) agree with a set of known experiments (three in our case)
- (2) lead to predictions (correct) for all *new* experiments

Newton-Galileo Relativity lasted over 250 years before any experiment was sophisticated enough to show that it was invalid. Special Relativity has now lasted 100 years.

It has been subjected to significantly more experiments than was case for the Newton-Galileo theory. These experiments are significantly more sophisticated and more precise also.

What are the new predictions of the theory?

1.7 The Strange World of Special Relativity

1.7.1 Relationships between Events

From the diagram below it is clear that:

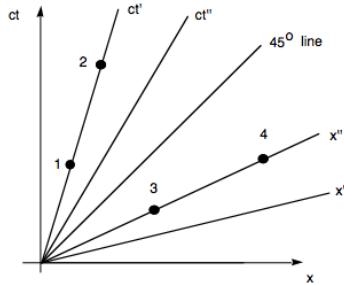


Figure 1.33: Related Events in Spacetime

For any timelike pair of events (1 and 2) it is always possible to find some observer (corresponding to a new ct -axis) such that the two events takes place at the same location and hence represent a pure time interval. Hence the name *timelike*.

For any spacelike pair of events (3 and 4) it is always possible to find some observer (a new x -axis) such that the two events takes place simultaneously and hence represent a pure space interval. Hence the name *spacelike*.

Timelike and spacelike events are radically different. As the diagram below clearly shows:

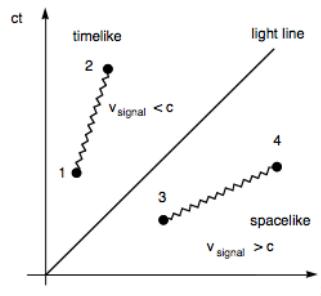


Figure 1.34: Timelike and Spacelike Related Events

Event #2, which is timelike relative to event #1, is in the future(later time) of event #1.

Event #4, which is spacelike relative to event #3.

Events 1 and 2 can be connected with a signal traveling with a speed less than that of light.

Events 3 and 4 require a signal speed greater than that of light.

Now consider the events labeled O, A, B, C, D, E, F, and G on the spacetime diagram below:

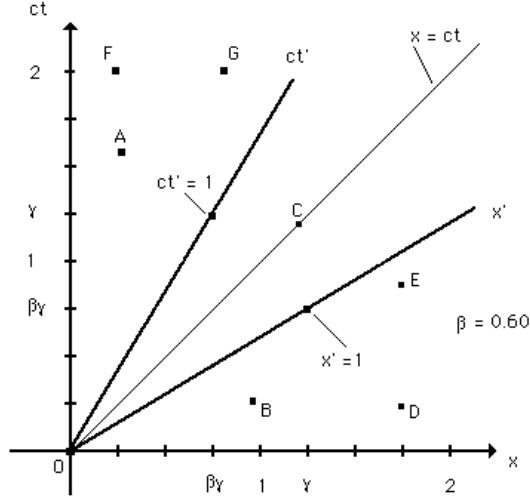


Figure 1.35: Events in Spacetime

The corresponding intervals have the following properties:

$$(\Delta S)_{AO}^2 = c^2(t_A - t_O)^2 - (x_A - x_O)^2 > 0 \rightarrow \text{a timelike interval}$$

$$(\Delta S)_{DO}^2 = c^2(t_D - t_O)^2 - (x_D - x_O)^2 < 0 \rightarrow \text{a spacelike interval}$$

$$(\Delta S)_{CO}^2 = c^2(t_C - t_O)^2 - (x_C - x_O)^2 = 0 \rightarrow \text{a lightlike or null interval}$$

$$(\Delta S)_{FG}^2 = c^2(t_F - t_G)^2 - (x_F - x_G)^2 < 0 \\ \rightarrow F \text{ and } G \text{ are simultaneous in the } (x, ct) \text{ frame}$$

$$(\Delta S)_{ED}^2 = c^2(t_E - t_D)^2 - (x_E - x_D)^2 > 0 \\ \rightarrow E \text{ and } D \text{ are at the same place in the } (x, ct) \text{ frame}$$

Now we consider these same events from the viewpoint of the (x', ct') frame. Look at the diagram on the next page where we have marked off all of the coordinate values.

Looking carefully at this diagram we can draw the following conclusions:

- (1) events that are simultaneous in one frame are not simultaneous in other frames (see events F and G) - **simultaneity is a relative concept!**
- (2) events occurring at the same place in one frame do not occur at the same place in other frames (see events E and D)
- (3) the time order of timelike events (events with a timelike interval) does not change between frames (see events O and A)

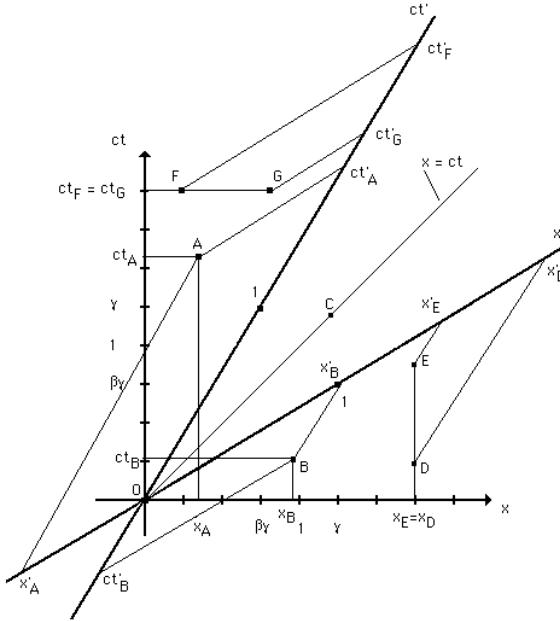


Figure 1.36: Coordinates in Different Frames

- (4) the time order of spacelike events (events with a spacelike interval) *can have their time order reversed* (see events O and B); in the (x, ct) frame B occurs after O, but in the (x', ct') frame O occurs after B.
- (5) numerical values of spatial separations and time separations are different in different frames
- (6) note that the line $x = ct$, which represents a light ray starting at the origin in the unprimed frame is also the line $x' = ct'$, which represents a light ray starting at the origin in the primed frame - **Light is the only physical object that both observers see in identical fashion.**

Let us consider in more detail the *reversal in time order* of two events.

This seems to be a very serious problems since it could possibly lead to a violation of the idea of *causality*. The concept of *causality* is connected with the idea of *cause and effect*, i.e., that an event should not occur *before* its own cause, for example, a firecracker should not explode *before* we light the fuse!

Suppose that we have two events in the (x, ct) frame with coordinates (x_1, ct_1) and (x_2, ct_2) and suppose, in addition, that

$$\Delta x = (x_2 - x_1) > 0 \quad , \quad \Delta t = (t_2 - t_1) > 0$$

so that event 2 comes *after* event 1 in the unprimed frame. Then the Lorentz transformations give the result (in the (x', ct') frame) that

$$\begin{aligned}\Delta t' &= \frac{1}{c}(ct'_2 - ct'_1) = \frac{1}{c}(\gamma(ct_2 - \beta x_2) - \gamma(ct_1 - \beta x_1)) \\ &= \gamma\left((t_2 - t_1) - \frac{\beta}{c}(x_2 - x_1)\right) = \gamma\left(\Delta t - \frac{\beta}{c}\Delta x\right)\end{aligned}$$

It is easy to see that $\Delta t'$ can be negative, which means that the time order of the two events is reversed, if the two events are related such that we have

$$\Delta t - \frac{\beta}{c}\Delta x < 0 \text{ or } \frac{\Delta x}{\Delta t} > \frac{c}{\beta} > c$$

or the events must be connected by a signal with $v > c$, which means that they are spacelike separated!

Now, for all timelike related pairs of events we have

$$\frac{\Delta x}{\Delta t} < c$$

and thus we *cannot reverse their time order*.

It is important to note that it is only for timelike related events that can event #1 cause event #2 (since all signals must have $v < c$). Thus, all cause/effect related events cannot have their time order reversed, preserving the idea of causality.

Special relativity is consistent with causality without us having to impose the consistency!

On the other hand, all spacelike related pairs of events have

$$\frac{\Delta x}{\Delta t} > c$$

and thus, their time order might be reversed in different frames.

Since they cannot be cause/effect related, this does not affect the idea of causality. It does, however, lead to a number of strange *paradoxes*, as we shall discuss later.

1.7.2 light Cones

Another way to look at these ideas is via the concept of the *light cone*.

Since the maximum allowed speed for any physical object is the speed of light c , we can use the world lines of light emanating from an event to delineate distinct

regions of spacetime for any object having that event on its worldline.

Consider the diagram below:

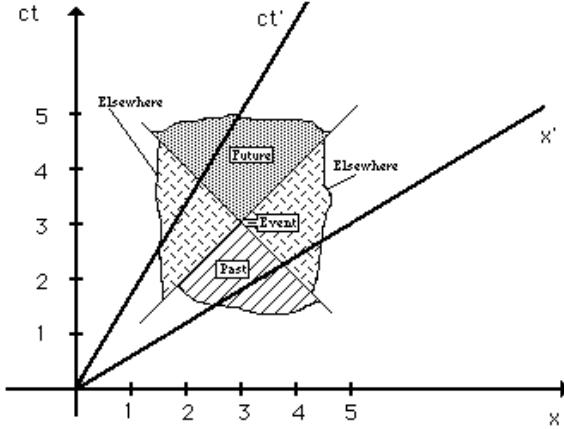


Figure 1.37: Structure of Spacetime - Light Cone

Suppose that you experience the event (that is, it is on your worldline) as indicated on the diagram above. Since neither you nor any signals you send or receive can travel faster than the speed of light and since the light ray worldlines containing this event are 45° lines as shown, the region labeled *future* represents all the events that you can either experience or influence with a signal at a later time(all events in this region are timelike separated from the event you experienced), the region labeled *past* represents all the events you could have experienced or that could have influenced you(all events in this region are timelike separated from the event you experienced). The regions labeled *elsewhere* are such that you can neither experience them nor influence them with any signal (all events in these regions are spacelike separated from the event you experienced).

If we draw this picture in a 3-dimensional world (x, y, ct) then the corresponding regions would look like:

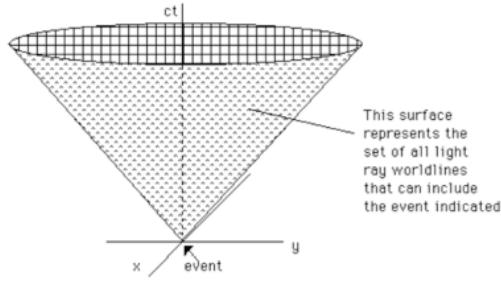


Figure 1.38: Light Cone in 3 Dimensions

and hence the name **light cone**.

What has happened to your possible future while we have been discussing these light cones?

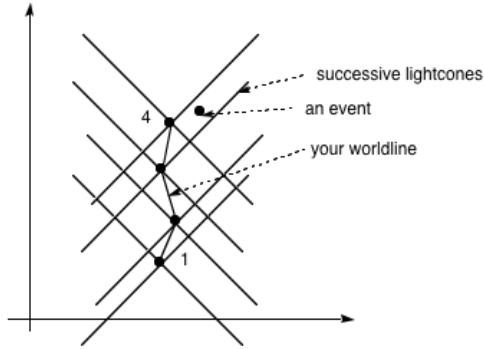


Figure 1.39: Disappearing Future

The event labelled above was in your possible future when you were experiencing event #1, but is no longer in your possible future when you are experiencing event #4. So be careful about wasting time doing nothing!!

Let us explicitly show the invariance of the spacetime interval. Suppose that we have two events with unprimed coordinates

$$x_1 = 2.0, \ ct_1 = 1.0 \quad ; \quad x_2 = 4.0, \ ct_2 = 2.0 \\ \Delta x = x_2 - x_1 = 2.0 \quad ; \quad c\Delta t = c(t_2 - t_1) = 1.0$$

and we assume that

$$\beta = 0.8 \rightarrow \gamma = \frac{1}{\sqrt{a - \beta^2}} = 1.67$$

Using the Lorentz transformations we have

$$\begin{aligned}
 x'_1 &= \gamma(x_1 - \beta ct_1) = 1.67(2.0 - 0.8(1.0)) = 2.00 \\
 ct'_1 &= \gamma(ct_1 - \beta x_1) = 1.67(1.0 - 0.8(2.0)) = -1.00 \\
 x'_2 &= \gamma(x_2 - \beta ct_2) = 1.67(4.0 - 0.8(2.0)) = 4.00 \\
 ct'_2 &= \gamma(ct_2 - \beta x_2) = 1.67(2.0 - 0.8(4.0)) = -2.00 \\
 \Delta x' &= \gamma(\Delta x - \beta c\Delta t) = 1.67(2.0 - 0.8(1.0)) = 2.00 \\
 c\Delta t' &= \gamma(c\Delta t - \beta \Delta x) = 1.67(1.0 - 0.8(2.0)) = -1.00
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 (\Delta S)^2 &= (c\Delta t)^2 - (\Delta x)^2 = 1.00 - 4.00 = -3.00 \\
 (\Delta S)^2 &= (c\Delta t')^2 - (\Delta x')^2 = 1.00 - 4.00 = -3.00
 \end{aligned}$$

The interval has the same numerical value, even though the time order between the two events is reversed!!!!

1.8 Measurements in Special Relativity

Now let us turn to the measurement properties of spacetime, in particular, the measurement of length and time. First, we need to restate the definitions we decided on earlier:

Length of an object = spatial separation of the two events representing the endpoints of an object measured *simultaneously* (the two events are on a line of simultaneity in a given frame).

Time interval between two events = time separation of the two events measured by a clock at *rest* with respect to the two events (the two events are on the worldline of the clock).

With these definitions we can represent these measurements as follows. Suppose we have two events (ct_1, x_1) and (ct_2, x_2) that correspond to the events on the worldlines of the endpoints of the object being measured, crossing a line of simultaneity (see diagram below).

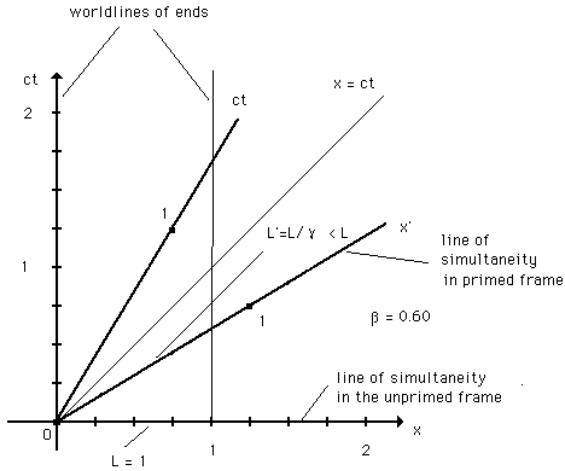


Figure 1.40: Measuring Length

Then the length of the object is given by $L = x_2 - x_1$. We note as shown in the diagram above that this is not the length as measured in the other reference frame. In fact, $L' = x'_2 - x'_1 = L/\gamma < L$, which is the famous *length contraction* (we will discuss this in more detail shortly). Do not be deceived by it looking longer, remember the scales on the different axes are not the same.

The **proper length** of an object is the length measured in the objects rest frame (the unprimed frame in this case, because that is where the endpoint worldlines are parallel to the time axis, which is the definition of being at rest). The *proper length* is the *maximum measured length*.

We note that we have not said that any object has physically *contracted*, but instead, we have said its measured length is less!

The measured length is less because the two observers *do not agree about simultaneity*, i.e., they have different lines of simultaneity. So even though we use the word *contraction*, we must understand that the effect is due to a disagreement about simultaneity and *no physical contraction has actually occurred*.

If the object is at rest in the primed frame, then we get an identical result just exchanging the roles of the two frames. As can be seen from the diagram below, in this case, $L = x_2 - x_1 = L'/\gamma < L'$.

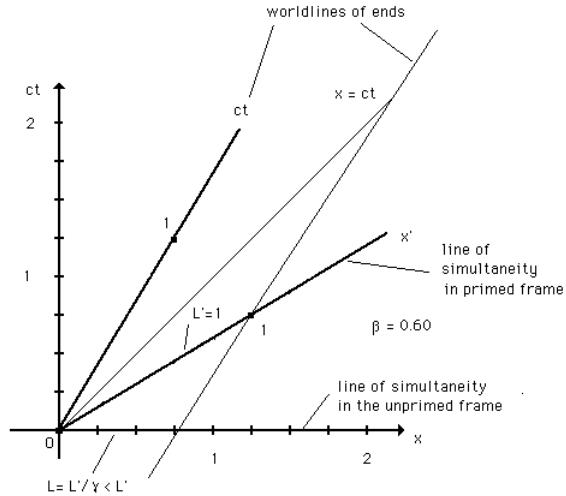


Figure 1.41: Measuring Length

Time measurements and *time dilation* is handled in the same way. Consider the diagram below representing a system that is at rest in the unprimed frame and only lives for a finite amount of time (like mu-mesons).

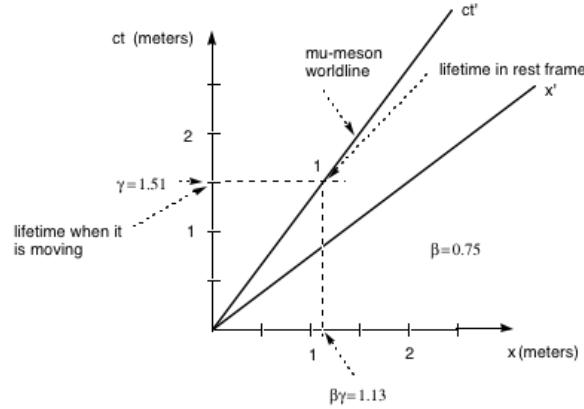


Figure 1.42: Measuring Time

The **proper time** interval for this system is the time separation T between the events (its birth(event #1) and its death(event #2)) as measured by a clock at rest with respect to the system or, in this case, at rest in the unprimed frame.

As can be seen from the diagram the time separation for an observer in the primed frame is $T' = \gamma T > T$.

The **proper time** is the *shortest time interval*.

This result is identical to the mu-meson experiment we discussed earlier, which was just an example of time-dilation as can be seen from the diagram above, where in this case $T' = 1$ and $T = \gamma T' = \gamma$. We can also see both of these results directly using the Lorentz transformations or the invariance of the interval.

1.8.1 Using Lorentz Transformations

Length Contraction

The relevant events representing on the worldlines of the ends of an object are

$(x_1, ct_1) = (0.0, 0.0)$ and $(x_2, ct_2) = (1.0, 0.0)$ for the unprimed observer

$(x_1, ct_1) = (0.0, 0.0)$ and $(x_3, ct_3) = (1.0, \beta) = (1.0, 0.6)$ for the primed observer

These events on the diagram below.

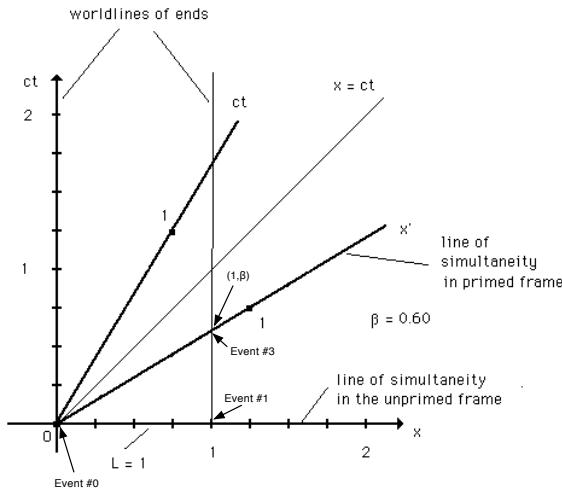


Figure 1.43: Length Contraction

Then the length this object, by definition, is the spatial separation along a line of simultaneity for the unprimed observer $L = x_2 - x_1 = \Delta x = 1$.

For the other observer, the length is the spatial separation is along a line of simultaneity for the primed observer

$$L' = x'_3 - x'_1 = \gamma((x_3 - x_1) - \beta c(t_3 - t_1)) = 1.25(1.0 - 0.6(0.6)) = 0.8 = \frac{L}{\gamma} = \frac{1}{\gamma}$$

What is the spatial separation between events 1 and 2 for the primed observer? Does it have any physical meaning?

Time Dilation

Suppose the clock is at rest in the primed frame. Then the relevant events representing on the worldlines of the ends of an object are

$$(x'_1, c'_1) = (0.0, 0.0) \quad , \quad (x'_2, c'_2) = (0.0, 1.0)$$

as shown in the diagram below.

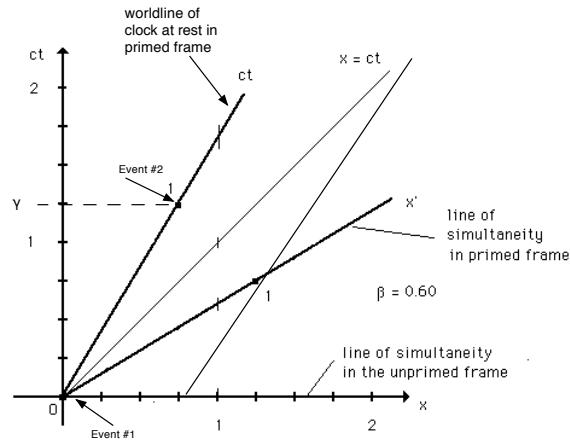


Figure 1.44: Time Dilation

Then for unprimed observer we have

$$\Delta x = \gamma(\Delta x' + \beta c \Delta t') = \gamma\beta \quad , \quad c \Delta t = \gamma(c \Delta t' + \beta \Delta x') = \gamma$$

Note the change in signs in the Lorentz transformations when we go from the primed to the unprimed coordinates. Why?

Let us now return to the k -factor. Our original k -factor assumption says that, if the unprimed observer is sending out signals every T seconds and the primed observer is receiving them every T' seconds where $T' = kT$, then we have the relationship

$$f' = \frac{1}{T'} = \frac{1}{kT} = \frac{f}{k} = \sqrt{\frac{c-v}{c+v}} f$$

between the frequency f as measured in the unprimed frame and the frequency f' as measured in the primed frame. The case above corresponds to the two observers moving away from each other. In this case $f' < f$ and hence the primed observer sees the wavelength increase (wavelength = c/f), which is the famous *red shift*. If they move towards each other, then $v \rightarrow -v$ or $k \rightarrow 1/k$ and the frequency increases (wavelength decreases) and we get a *blue shift*. This is called the relativistic *Doppler effect* for light. Let us look at the important Doppler effect in more detail.

1.9 The Doppler Effect

1.9.1 Sound and the Acoustic Doppler Effect

Sound travels through a medium such as air with a speed v . This speed is determined by the properties of the medium and is independent of the motion of the source. We consider a source of sound that is moving with velocity w through the medium towards an observer at rest. We assume for simplicity that the observer (detector) lies along the line of motion of the source. See diagram below.

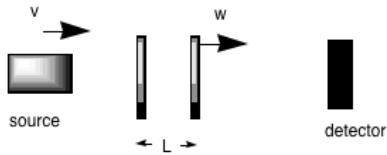


Figure 1.45: Acoustic Doppler Effect Experiment

As shown in the diagram, we represent the sound wave as a regular series of pulses. These pulses are separated in space by a distance L and in time by an amount $\tau_0 = 1/f_0$ where f_0 is the frequency of the sound from the source. In a time T the sound travels a distance wT , and if the pulses are separated by a distance L , the number reaching the detector is wT/L . The rate at which pulses arrive is

$$\frac{w}{L} = \text{frequency } \left(\frac{\text{number}}{T} \right) \text{ of sound at the detector} = f_0$$

To determine L , we consider a pulse emitted at $t = 0$ and a second pulse emitted at $t = \tau_0$. During the interval τ_0 the first pulse travels a distance $w\tau_0$ in the medium and the source travels a distance $v\tau_0$. As shown in the figure below

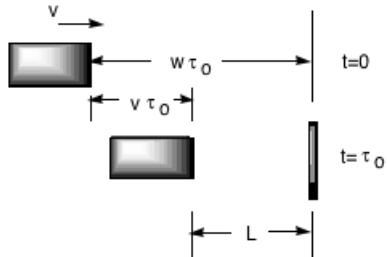


Figure 1.46: Associated Separations

the distance between the pulses is given by

$$L = w\tau_0 - v\tau_0 = (w - v)\tau_0 = \frac{(w - v)}{f_0}$$

and

$$f_D = \text{frequency at detector} = f_0 \frac{1}{1 - \frac{v}{w}} \text{ for a moving source}$$

For an approaching source, $v > 0$ and thus $f_D > f_0$. For a receding source, $v < 0$ and thus $f_D < f_0$.

If the source is at rest and the detector is moving (as shown below) the situation is different.

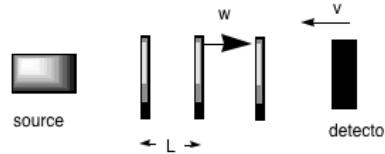


Figure 1.47: Detector Moving

The speed of the pulses relative to the detector is $w + v$. The rate at which the pulses arrive is

$$f_D = \frac{w + v}{L}$$

Since the source is at rest, $L = w\tau_0 = w/f_0$ and thus

$$f_D = f_0 \left(1 + \frac{v}{w}\right) \text{ for a moving detector}$$

The two results are not symmetric. They are approximately the same for small v/w . If we know f_D , then we can tell whether it is the source or the detector that is moving!!

This is so because the speed of sound is not a universal constant, but only has a definite value relative to the medium where it is propagating.

1.9.2 Light and the Relativistic Doppler Effect

Suppose a light source flashes with period $\tau_0 = 1/f_0$ in its rest frame and that the source is moving towards the observer(detector) with velocity v as shown below.



Figure 1.48: Doppler Effect with Light

Due to time dilation, the period in the detector rest frame is $\tau = \gamma\tau_0$. Since the speed of light is a universal constant, the pulses arrive at the detector with speed c . As shown in the diagram below

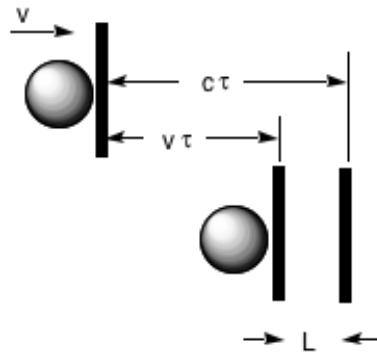


Figure 1.49: Associated Separations

the frequency of the pulses is $f_D = c/L$, where L is the pulse separation in the detector frame. Since the source is moving towards the detector we have (as shown in the diagram above)

$$L = c\tau - v\tau = (c - v)\tau = (c - v)\gamma\tau_0 = \gamma \frac{(c - v)}{f_0}$$

and

$$f_D = f_0 \sqrt{\frac{1 - \frac{v^2}{c^2}}{1 - \frac{v}{c}}} = f_0 \sqrt{\frac{c + v}{c - v}}$$

Here f_D is the frequency in the detector frame and v is the relative velocity of the source and the detector. **It does not matter which one is actually moving!!**

This result is just the red shift formula we started with earlier, as expected.

Now consider the spacetime diagram below. We have two observers in relative motion and the unprimed observer is sending signals to the primed observer at regular interval (separated by a time T).

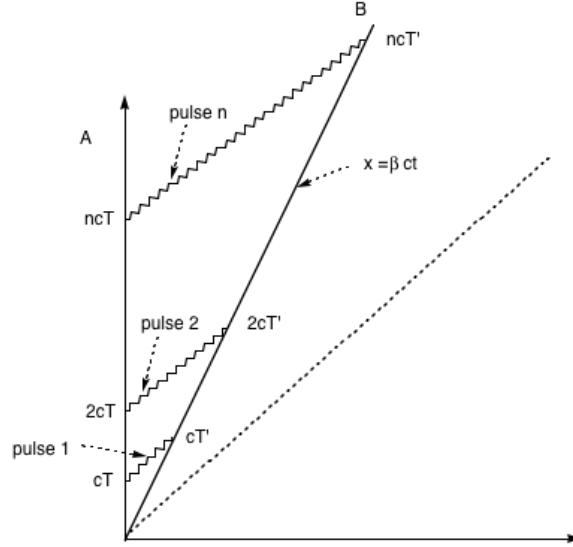


Figure 1.50: Doppler effect in Spacetime

The reception of the last pulse occurs at the point of intersection of the lines

$$x = c(t - nT) \text{ and } x = \beta ct$$

(as shown in diagram) or at the event

$$ct = \frac{cnT}{1 - \beta} , \quad x = \frac{\beta cnT}{1 - \beta}$$

n pulses are sent out by the unprimed observer in nT seconds and thus the period is T seconds and the frequency is $1/T$.

n pulses(same number) are received by the primed observer in nT' seconds and thus the period is T' seconds and the frequency is $1/T'$.

Now, the reception point also corresponds to

$$\begin{aligned} ct' &= \gamma(ct - \beta x) = \gamma \left(\frac{cnT}{1 - \beta} - \frac{\beta^2 cnT}{1 - \beta} \right) \\ &= \gamma cnT \frac{1 - \beta^2}{1 - \beta} = ncT' = \gamma cnT(1 + \beta) \end{aligned}$$

Using

$$\gamma = \frac{1}{\sqrt{1 - \beta^2}}$$

we get

$$T' = \sqrt{\frac{1+\beta}{1-\beta}}$$

which is the standard Doppler effect formula for light.

1.10 How Do We Talk to Each Other in this New Relativistic World?

In this new world what happens if we try to tell a story?

In particular, these are some of the words are no longer usable?

where, when, speed, distance, time interval,
simultaneous, same place, length, etc ...

If we want to use such words, then each reader (other observers) must first use the Lorentz transformations to translate the story before trying to read it!

The only words (concepts) that we are allowed to use if we do not want to do any translations are

interval, c, number of events

Not having grown up in this new world, we would find it very difficult to tell such a story.

1.11 The Famous Paradoxes

1.11.1 The Twin Paradox

Let me state this problem in a *bad* way, i.e., a way that leads to the so-called paradox. Then we will state it correctly and the paradox will disappear and we will be able to draw the correct conclusions. This might be a lesson for life also!

Statement #1 - Two twins are traveling relative to each other with speed [image.pict]. Time dilation says that the clock of the *moving* twin should tick slower (the time between ticks is larger). Since each twin considers herself to be at rest, the other twin should have a clock that runs slower and hence the other twin should be younger. Which twin is younger? There is no definite answer to the question as posed since we do not know which twin is moving (changed reference frames - has accelerated) and hence we have a supposed paradox.

Statement #2 - Two twins have been together since birth (they have been on the same worldline - in the same frame of reference). At one point in time, one of the twins, gets into a rocket ship and changes her frame of reference - changes her velocity - experiences a period of acceleration). The twin in the rocket ship travels to a distant star and then changes her frame of reference again (reverses her velocity - accelerates for a period of time). The twin in the rocket ship travels back to the earth and then changes her frame of reference again (come to rest on the earth - accelerates for a period of time). Finally, the two twins remain together again (in the same frame of reference - on the same worldline).

Which twin, if any, is younger? This description is represented by the spacetime diagram below:

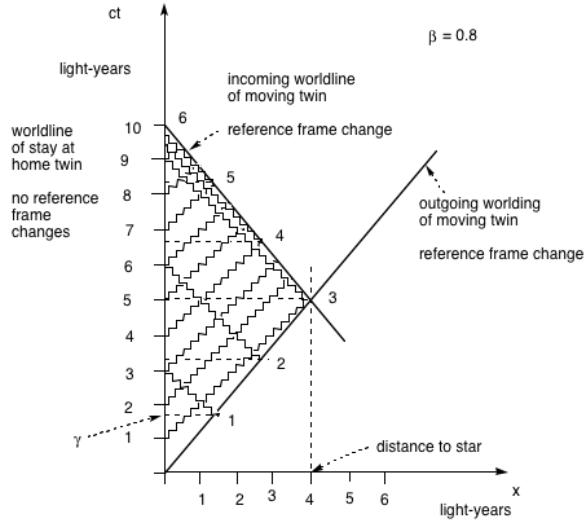


Figure 1.51: Twin Paradox in Spacetime

On this diagram $\beta = 0.8$ and $\gamma = 1.67$. The respective time axes have been calibrated. Each twin sends one signal per year (by their own clock) to the other twin.

While they are separating, the k -factor says that

$$f_{observed} = f_{reduced} = \sqrt{\frac{1-\beta}{1+\beta}} 1(\text{per year}) = \frac{1}{3} \text{ per year}$$

While they are coming back together, the k -factor says that

$$f_{observed} = f_{increased} = \sqrt{\frac{1+\beta}{1-\beta}} 1(\text{per year}) = 3 \text{ per year}$$

It is clear from the diagram that both twins see these different rates during the designated periods.

For both twins the reduced rate starts immediately.

However, the switch over to the increased rate takes place at different times according to each observer. They are not identical observers and thus we should not expect identical results from their measurements.

For the moving twin, the switchover takes place exactly at the midpoint of the trip or at year 3 (as can be seen in the diagram). For the stay-at-home twin, however, the switchover take place at year 9.

Thus the moving twin sees $3 \times 1/3 + 3 \times 3 = 10$ signals from the stay-at-home twin and thus knows that the stay at home twin is 10 years older and she is only 6 years older

The stay-at-home twin sees $9 \times 1/3 + 1 \times 3 = 6$ signals from the moving twin and thus knows that the moving twin is 6 years older and that she is 10 years older

Both agree and this there is no paradox. The traveler ages less because moving clocks(clocks that have changed frames of reference) run slower.

1.11.2 The Pole in the Barn Paradox

In this case we have the following situation - two farmers have a barn which is 10 meters long in their rest frame (unprimed). The farmers are standing at the left and right doors of the barn (the doors are open).

A pole carrier has a pole of length 12 meters in her rest frame and is carrying it horizontally while she runs towards the barn with a speed given by $\beta = 0.8$ and $\gamma = 1.67$.

If we believe all this relativity and length contraction stuff, then the farmers think the pole is

$$L'_{pole} = \frac{L_{pole}}{\gamma} = 9.8 \text{ meters}$$

However, the pole carrier thinks the barn is only

$$L'_{barn} = \frac{L_{barn}}{\gamma} = 8.0 \text{ meters}$$

This means that, according to the farmers, the pole should be able to fit into the barn. The pole carrier, however, say no way, the barn is much too small.

A possible spacetime diagram for this experiment is shown below.

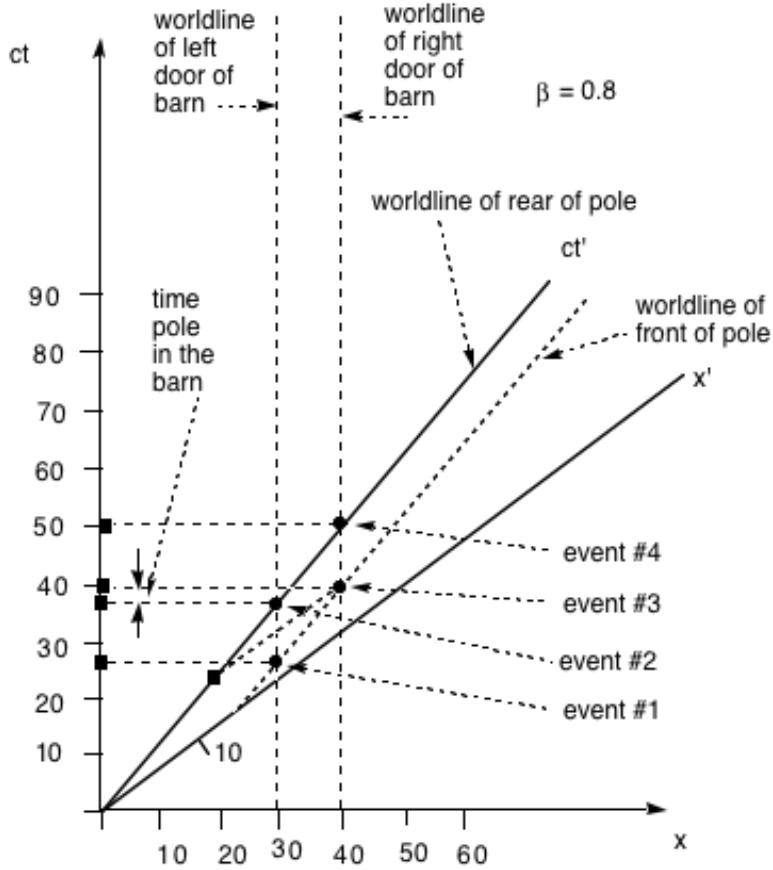


Figure 1.52: Pole in Barn Paradox in Spacetime

Is there any correct answer to this dilemma? To answer the question, we label 4 crucial events:

- Event #1: front of the pole enters the barn
- Event #2: ear of the pole enters the barn
- Event #3: front of the pole leaves the barn
- Event #4: rear of the pole leaves the barn

These events are clearly shown on the diagram.

Now if $t_3 > t_2$, then the pole is completely within the barn for the period of time $t_3 - t_2$.

It is clear from the diagram, that according to the farmers the pole is within

the barn for a short period of time!

The pole carrier disagrees, however. For the pole carrier, $t'_2 > t'_3$ and therefore the pole is never completely within the barn.

There is a disagreement between the two sets of observers because the *time order of the two crucial events (namely 2 and 3) has reversed*.

Thus, both are correct.

The pole is within the barn and not within the barn depending on your frame of reference. Relativity is subjective, that is, dependent on the observer information in certain cases. Relativity allows different observers to tell differing stories like this when time order reverses. The time order reversal is OK in this case because events 2 and 3 are spacelike separated and thus reversing their time order cannot upset causality. There is no paradox!

1.11.3 Signals faster than Light Paradox

What happens if we allow some signal to go faster than the speed of light?

Consider the following story.

Sam is walking down the path towards Sharples. As he passes near Clothier tower a stone block falls off the tower and lands on his head, killing him. So Sam is now lying in heap at the base of Clothier tower. Soon after that incident, Sally comes along. Sam is Sally's good friend and she is distraught when she sees Sam lying in a heap. Sally is walking past Sam with some speed u (she is in a different frame of reference). Now, Sally understands Special Relativity. Sally has in her possession a special device that can send a signal to someone on the other side of the universe at a speed $> c$ if they are in the same frame of reference. So Sally sends out a signal indicating what happened to Sam. The signal is received on the other side of the universe by George(in the same frame of reference as Sally). He is now desperate to tell Sam so he can avoid the stone block, but Sam is in a different frame and cannot receive his signal. So he tells the story to someone in Sam's frame, namely, Samantha. Samantha also happens to have one of those devices that sends the speedy signal and she sends a signal to Sam.

The entire sequence of worldlines with the associated events is shown in the diagram below:

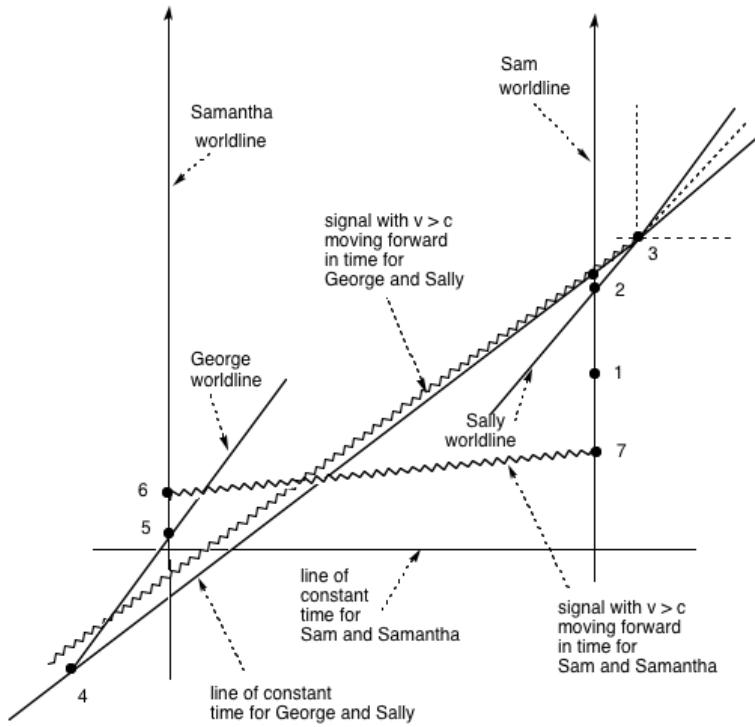


Figure 1.53: Faster than Light Paradox in Spacetime

The important events are:

- Event #1: Sam gets killed
- Event #2: Sally sees Sam
- Event #3: After patiently waiting Sally sends a $v > c$ signal to George
- Event #4: George receives the signal
- Event #5: George tells Samantha
- Event #6: Samantha patiently waits and then sends a $v > c$ signal to Sam
- Event #7: Sam receives the signal from Samantha, realizes he is about to die and stops walking, thus avoiding the block and subsequent death

Questions

If Sam is not dead, why would Sally send any signal?

If Sally does not send a signal making all the other stuff happen, then why would Sam stop?

If Sam has no reason to stop, he then gets killed and Sally has a reason to send the signal.

Which is it?

We have what is called a closed causal loop here. There is no logical way out of this loop.

Does that mean it cannot occur, i.e., that no signal can travel faster than light?

or

Is there some other explanation? What about *free will*?

1.12 Basic Ideas of Classical Kinematics and Dynamics (A Quick Tour)

1.12.1 Kinematics (or the study of motion in time)

Position ($\vec{r}(t)$) is defined as a vector from the coordinate origin to the 3-dimensional point where the object is located. In 1-dimension we have $x(t)$.

The goal of all classical physics is to determine the *position of an object as function of time*.

Position answers the **Where** question for the events we have been discussing.

Velocity ($\vec{v}(t)$) is defined as a vector in the direction of the change of the position vector and having a magnitude given by

$$\vec{v}(t) = \lim_{\Delta t \rightarrow 0} \frac{\Delta \vec{r}}{\Delta t} \text{ or } v(t) = \lim_{\Delta t \rightarrow 0} \frac{\Delta x}{\Delta t} \text{ in 1-dimension}$$

The direction of the velocity is always *tangent* to the path of motion.

Velocity tells us how fast the object is moving and in what direction.

If the velocity is constant this means both its magnitude (speed) and direction are constant because it is vector.

Now for a simple example. Suppose a particle has $v = +10 \text{ m/s}$ (+ means towards $+\infty$) and $x = 2 \text{ m}$. Where will the particle be 1 sec later?

Clearly the answer is $x = 12 \text{ m}$ since $12 = 2 + 10(1) = x(t=0) + v\Delta t$.

If the velocity is not constant the situation is more complicated. If, however, I can tell you that the *average velocity* over the next second = 8 m/s, then the rule $12 = 2 + 10(1) = x(t = 0) + v\Delta t$ still works.

So, in general, we have

$$x(t + \Delta t) = x(t) + v(t)\Delta t \text{ in 1-dimension}$$

where $v(t)$ = the average velocity during the interval Δt . In 1-dimension direction is indicated by \pm signs.

Acceleration is defined as a vector in the direction of the change of the velocity vector and having a magnitude given by

$$\vec{a}(t) = \lim_{\Delta t \rightarrow 0} \frac{\Delta \vec{v}}{\Delta t} \text{ or } a(t) = \lim_{\Delta t \rightarrow 0} \frac{\Delta v}{\Delta t} \text{ in 1-dimension}$$

It is a generalization of previous equation for velocity. Physicists (along with everyone else) like to generalize ideas as long as they can get away with it ... it is easy...and you do not have to think up anything new.

We then have (as with velocity)

$$v(t + \Delta t) = v(t) + a(t)\Delta t \text{ in 1-dimension}$$

Now, suppose I interact with 2 different bodies in *same* manner, i.e., hang the same object over a pulley and attach it to the 2 bodies with a string.

We define the *stuff* in each body by seesaw balancing such that the amount of stuff in 2 bodies is identical if the seesaw balances and ratio of the stuff in two bodies is given by the inverse ratio of the distance from the pivot when the seesaw balances. **Stuff = mass = m!**

We note that(experiment) says that when I interact with a body (mass) or *exert a force on it* it will accelerate (change its velocity and hence change its position) and we find that when I interact with 2 different bodies in *same* manner (exert the same force)

$$\frac{a_1}{a_2} = \text{constant} = \frac{\text{stuff in 2}}{\text{stuff in 1}} = \frac{m_2}{m_1}$$

Whenever simple results like that come out of an experiment, physicists(Newton and Galileo in this case) say that something profound must be going on here.....

In this case, they turned the equations around and said

$$m_1 a_1 = m_2 a_2 \rightarrow \text{something to do with my } \textit{same} \text{ interaction!!}$$

So given the acceleration, we can calculate the velocity and then calculate the position and get the answer we are looking for the process uses calculus

that is why Newton invented it.

But how do we find the acceleration from first principles remember that is what theorists do ! This leads us into a discussion of **Dynamics**

1.12.2 Newtons Laws (the crowning achievement of classical physics)

A body at *rest* is not moving! There is no difference between a body at rest and a body moving with constant velocity since we can always change our frame of reference and then the body with constant velocity looks like it is at rest (and the body that was at rest now looks like it has a constant velocity).

A body is *interacting* with its surroundings in some manner when we see a *changing* velocity or that it has an acceleration.

Now push(or pull) on object and watch it accelerate. It is clear that I can make the body have a smaller or larger acceleration depending on the strength of my interaction with it. It is also clear that my interaction is directional ... it produces a directional or vector quantity ... the acceleration.

This leads to the concept of a *force*. Force is a vector quantity that somehow represents and quantifies my interaction with the body. Since in the earlier experiment, my interaction in the two cases was the *same*, I must have been exerting the same force.

This led Newton to postulate the relationship

$$\vec{F} = m\vec{a}$$

so that in the earlier experiment I was exerting the same force!

Be careful here! Is there any new physical content to the introduction of the concept of force or is all the physics contained in the acceleration? I can measure the acceleration! Can I measure the force or do I just infer it from a measured acceleration?

Newton's laws are:

- (1) an isolated body has no acceleration (Any real content?)
- (1') a body at rest or moving with constant velocity remains at rest or moving with constant velocity unless it interacts with something (Any real content?)
- (2) $\vec{F} = m\vec{a}$ (Is this simply a definition of the force?)
- (3) If body A exerts a force on body B , then body B exerts an equal and opposite force on body A (here is real content!)

1.12.3 Energy

Most dynamics problems of the everyday world can be solved using Newtons laws. But they are not suitable for generalization beyond the realm of everyday experience.

In order to find the rules and laws that are appropriate in other regimes of interest like very high speeds (Special Relativity(SR)) we must find a different way of thinking about the universe. This new way is based on Newtons laws so there is no new physical content, but it will be possible to extend to meaning of the new laws so that new physical content and thus new physical theories can be formulated.

Energy is one of these new concepts that allows generalization.

We first define *kinetic energy or energy due to motion* as

$$K = \frac{1}{2}mv^2$$

Now we do an experiment. We take any object raise it up to some height h above the ground and then release it. We find the following relationships:

$$\begin{aligned} v_{\text{ground}}^2 &= 2gh \quad g = 9.8 \text{ m/s}^2 \\ v(t) &= gt \\ y(t)h - \frac{1}{2}gt^2 & \\ v^2(t) + 2gy(t) &= \text{constant} \end{aligned}$$

This last result is the key. As we said earlier, much of theoretical physics is a search for invariants. We saw a couple in SR. When studying dynamics, invariants are quantities that are constant in time. The last experimental relation can be written

$$\begin{aligned} \frac{1}{2}mv^2(t) + mgy(t) &= \text{constant} \\ K + V &= E \end{aligned}$$

where we have defined two new energies

$$\begin{aligned} V &= mgy(t) = \text{potential energy} \\ E &= K + V = \text{total energy} \end{aligned}$$

The last experimental result then allows us to postulate

The total energy is a constant of the motion

The kinetic energy K and the potential energy V are not constant during the

motion. In fact, they are constantly changing into one another, i.e., there is an exchange between K and V during the motion.

The law that energy is a constant is an example of a conservation law for some invariant quantity. We derive conservation laws from simple experiments and then generalize their validity to a much wider range of phenomena. **Momentum** is another one of the new concepts. It is a vector quantity.

It turns out that velocity is not the important dynamic variable. How can we see this?

Suppose we have a hill with two dump trucks at the top. One of the dump trucks is filled with sand and the other is empty.

We know from experiment that the velocity is the same for different trucks when they reach the bottom of the hill or they have the same acceleration ... the acceleration seems to be independent of the amount of stuff in the trucks (a property of the gravitational interaction).

Now we ask this question?

Which of these two trucks would you want to attempt to stop at the bottom of the hill?

Clearly the answer is the truck with the least stuff or the smaller mass.

So we define a new dynamical quantity

$$\vec{p} = m\vec{v} = \text{linear momentum}$$

Newton's second law then becomes

$$\vec{F} = \lim_{\Delta t \rightarrow 0} \frac{\Delta \vec{p}}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{\Delta(m\vec{v})}{\Delta t}$$

which for a constant mass system becomes

$$\vec{F} = \lim_{\Delta t \rightarrow 0} \frac{\Delta \vec{p}}{\Delta t} = \lim_{\Delta t \rightarrow 0} m \frac{\Delta \vec{v}}{\Delta t} = m\vec{a}$$

as before. Therefore we now restate Newton's laws as:

- (1) The linear momentum is conserved for an isolated body; $\Delta \vec{p} = 0$.
- (2) $\vec{F} = \lim_{\Delta t \rightarrow 0} \frac{\Delta \vec{p}}{\Delta t}$

- (3) The total momentum of an isolated system is a constant

$$\begin{aligned}\vec{p}_1 + \vec{p}_2 &= \text{constant} \\ \Delta\vec{p}_1 + \Delta\vec{p}_2 &= 0 \\ \frac{\Delta\vec{p}_1}{\Delta t} &= -\frac{\Delta\vec{p}_2}{\Delta t} \\ \vec{F}_{12} &= -\vec{F}_{21}\end{aligned}$$

So for an isolated system we have a *conservation law* for linear momentum, which we can generalize to a much wider range of phenomena.

So, here is the way scientists of that day thought....

The classical universe followed well-defined laws. Everything was, and is, predictable. If we only find the force, know the masses, positions, and velocities of all the objects under consideration at one single time, then all is predictable from then on!!

The universe is a gigantic Newtonian clockwork. Cause and effect rule. Nothing is by chance. Everything is ultimately accountable.

Perfect determinism. The laws of physics are to be obeyed, because it is impossible to disobey them. There is no room for free will, salvation and damnation, or love and hate. Even the most trifling thought has been determined long ago. You might have imagined that you are a free-thinking person, but even that imagination is nothing but the universal clockwork turning in some yet-to-be-discovered way. So now you are probably thinking...glad they found out those ideas were wrong and got rid of them! Just remember it is always dangerous to make quick judgements like that, especially when you are not sure what will come along to replace it.

And then there was light..... and Special Relativity.

Our derivation of SR has shown that:

- (1) We lose position and time as separate quantities, which is an indication that everyday experience may not carry over into these new realms.

Why didn't physicists notice before? It generally is simply a matter of the accuracy and precision available to experimentalists, i.e., prior to this century, experimental measurements of the speed of light could not say that it was not infinite. If it were infinite, then SR would reduce to GR and Newtonian physics would still be valid.

- (2) We must choose our observables with some care.....

- (3) We must use conservation laws to give us the physical quantities that represent really what we can know about systems.
- (4) We can fully extend classical physics validity to all speeds.
- (5) We must rethink our world view (happens all the time in physics)

Everyday experience cannot be our guide

Everyday experience is fine for world of everyday objects

We must be prepared to give up preconceived ideas because they are based on our experiences

We must trust measurements to tell us what is going on but we must define them carefully

But classical physics still hangs on, albeit modified.....

Everything works so well

What does that statement really mean to a physicist? We must adopt an *only know what we measure* philosophy.

In this world motion is a continuous blend of changing positions. The object moves in a flow from one point to another.

Science is a reasonable, orderly process of observing nature and describing the observed *objectively*.

There is a conviction that whatever one observed as being out there was really out there. The idea of objectivity being absent from science is abhorrent to any rational physicist.

One firmly believes in the passive(non-disruptive) observer. Humans are creatures of the eye. They believe what they see.

So summarizing, classically

- (1) Things moved in a continuous manner.
- (2) Things move for reasons. The reasons are earlier causes and all motion was determined and predictable.
- (3) All motion could be analyzed or broken down into its component parts. Each part played a role in the giant machine called the universe. The complexity of this machine could be understood in terms of the simple movement of its various component parts.

- (4) The observer observed and never disturbed. All experimental errors could be analyzed and understood.

All of these ideas turn out to be false in modern theoretical physics!!!!

Now back to SR.

What happens to momentum and energy when we enter the realm of SR?

At this level we must rely on experiments to point the proper way to proceed. In a more mathematical physics class one can derive these results from first principles using the interval and linear algebra.

The following result has been confirmed by experiment.

The force felt by a charged particle in electric and magnetic fields is given mathematically by the Lorentz force law

$$\vec{F} = q \left(\vec{E} + \frac{\vec{v} \times \vec{B}}{c} \right)$$

where \vec{v} is the particle velocity and \vec{E} and \vec{B} are the electric and magnetic fields, respectively.

Consider the experimental setup below.

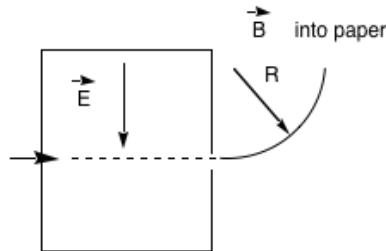


Figure 1.54: Measuring Momentum and Velocity

In the box region the electric and magnetic fields are adjusted so that $\vec{F} = 0$ for a particle moving along the dotted line with a definite velocity. The electric force always points downward and the magnetic force is always perpendicular to the velocity direction (upward in the box for a particle moving along the dotted line). This means that particles with a particular velocity, namely,

$$q \left(-E + \frac{v}{c} B \right) = 0 \rightarrow \frac{v}{c} = \frac{E}{B}$$

pass undeflected through the box. The box is called a velocity selector. Outside the box there is no electric field, so the particle moves on a circular path (force

always perpendicular to the velocity) with a radius of

$$R = \frac{pc}{qB}$$

So that measuring the radius corresponds to measuring the relativistic momentum. Thus, in the same experiment we can measure *both* the velocity and momentum *independently* and thus determine the relationship between them.

A plot of the experimental results looks like

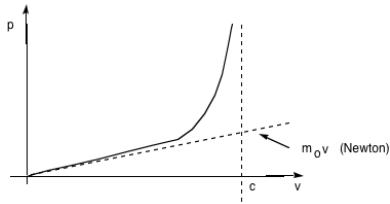


Figure 1.55: Experiment Result: p versus v

This corresponds to the result

$$p = \gamma m_0 v \text{ where } \gamma = \frac{1}{\sqrt{1 - \beta^2}}$$

instead of the Newtonian assumption that $p = m_0 v$, where m_0 = the so-called rest mass. It is the only valid mass for a particle since we measure mass when a body is at rest. Any measurement of mass when a particle is moving is really a measurement of its momentum and thus it would be incorrect for us to assume that any different mass value can be used for a moving object. Thus, there is no such thing as the *relativistic mass*.

Now what about relativistic energy? What is the relativistically correct form of the energy of a particle?

One way to generalize the concept of energy is to use the Newtonian definition of kinetic energy in conjunction with the relativistically correct definition of momentum. The derivation that follows uses calculus. Do not worry if you cannot follow all of the mathematical steps. For this course only the results are important. We proceed as follows.

The formal definition of kinetic energy is given as

$$\Delta K = K - K_0 = \text{work done by force} = \int_{\vec{r}_0}^{\vec{r}} \vec{F} \cdot d\vec{r} = \int_{\vec{r}_0}^{\vec{r}} \frac{d\vec{p}}{dt} \cdot d\vec{r}$$

We found that $\vec{p} = m_0\gamma(v)\vec{v}$ where $\gamma(v) = (1 - \beta^2)^{-1/2}$ and $\beta = v/c$. Therefore we have

$$K - K_0 = \int_{\vec{r}_0}^{\vec{r}} \frac{d(m_0\gamma(v)\vec{v})}{dt} \cdot \vec{v} dt = m_0 \int_0^v \vec{v} \cdot d(\gamma(v)\vec{v})$$

Since the kinetic energy is zero when the velocity is zero we finally have

$$K = m_0 \int_0^v \vec{v} \cdot d(\gamma(v)\vec{v})$$

Now since

$$d(\gamma v^2) = d(\gamma \vec{v} \cdot \vec{v}) = \vec{v} \cdot d(\gamma \vec{v}) + \gamma \vec{v} \cdot d\vec{v}$$

we can write

$$\begin{aligned} K &= m_0 \int_0^v (d(\gamma v^2) - \gamma \vec{v} \cdot d\vec{v}) = m_0 \int_0^v d(\gamma v^2) - m_0 \int_0^v \gamma \vec{v} \cdot d\vec{v} \\ &= m_0 \int_0^v d(\gamma v^2) - \frac{1}{2} m_0 \int_0^v \gamma d(v^2) = m_0 \gamma v^2 - \frac{1}{2} m_0 c^2 \int_0^{v^2/c^2} \frac{du}{\sqrt{1-u}} \\ &= m_0 \gamma v^2 + m_0 c^2 \left(\frac{1}{\gamma} - 1 \right) = m_0 c^2 \left(\frac{1}{\gamma} + \gamma \beta^2 \right) \\ &= m_0 c^2 (\gamma - 1) \end{aligned}$$

The first thing we should do is check that this makes sense. What is the low velocity limit of this expression?

Using

$$\gamma(v) = (1 - \beta^2)^{-1/2} \approx 1 + \frac{1}{2}\beta^2 = 1 + \frac{1}{2}\frac{v^2}{c^2}$$

we have

$$K = m_0 c^2 (\gamma - 1) \approx m_0 c^2 \frac{1}{2} \frac{v^2}{c^2} = \frac{1}{2} m_0 v^2$$

as expected.

If we rearrange this result we have

$$\gamma m_0 c^2 = K + m_0 c^2 = \text{Energy(motion)} + \text{Energy(rest)} = \text{Total Energy} = E$$

It is only the total energy that is conserved!

We thus obtain Einsteins famous relation $E_{rest} = m_0 c^2$.

What is the connection to momentum? Some algebra gives the following results for relativistic objects:

$$\frac{pc}{E} = \frac{m_0 \gamma v c}{\gamma m_0 c^2} = \frac{v}{c} = \beta \text{ and } \left(\frac{E}{c} \right)^2 - p^2 = (m_0 c)^2$$

Some questions arise

How come we do not notice the rest energy in everyday experience?

Some numbers:

$$\begin{aligned}\text{typical kinetic energy} &= 0.5(1)(1)^2 \approx 1 \text{ Joule} \\ \text{typical rest energy} &= (1)(3 \times 10^8)^2 \approx 10^{17} \text{ Joule}\end{aligned}$$

We typically ignore the significantly larger quantity!! The reason for this is that in everyday situations the rest energy does not change; all the same mass is remains in the system at all times. Thus, the rest energy is not a source of possible energy to do other things.

However, in microscopic systems like atoms and nuclei, etc, the rest mass changes in many interactions and thus this energy becomes available for other purposes. Two examples are nuclear fission and fusion.

Are there any new predictions we can make from these results?

The two relations above make the following interesting prediction:

$$\begin{aligned}v = c \rightarrow \beta = 1 \rightarrow E = pc \\ \left(\frac{E}{c}\right)^2 - p^2 = 0 = (m_0c)^2\end{aligned}$$

or the only objects that can travel at the speed of light must have a rest mass equal to zero! However, even though they have a zero rest mass, they still possess energy and momentum defying the classical equations!

Such a particle has been observed it is the **photon** or the particle of light.

1.13 Some First Thoughts about General Relativity

In special relativity we found that the spacetime interval or just *interval* between two events

#1: (x, y, z, ct)

#2: $(x + \Delta x, y + \Delta y, z + \Delta z, c(t + \Delta t))$

is given by

$$\Delta s^2 = c^2 \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$$

We can generalize this result to

$$\Delta s^2 = \sum_{i=0}^3 \sum_{j=0}^3 g_{ij} \Delta x_i \Delta x_j$$

where

$$\begin{aligned} x_0 &= t, x_1 = x, x_2 = y, x_3 = z \\ g_{00} &= 1 = -g_{11} = -g_{22} = -g_{33} \\ g_{ij} &= 0 \quad \text{if } i \neq j \end{aligned}$$

In the case of special relativity, this is just a change in notation and all the g_{ij} (called the *metric components*) are constants.

The equation for the light cone for the event (x, y, z, ct) can be expressed as

$$\Delta s^2 = c^2 \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2 = 0$$

This says that spacetime itself is *flat*. This means that the shortest distance between two points in space is a straight line.

Now what happens if the metric components g_{ij} are not constants but are functions of space and time? First, the Lorentz transformations are no longer valid. Second, the *light cone* will look different at different points in spacetime. Third, spacetime itself is *curved*. This means that the shortest distance between two points in space is not a straight line.

Einstein in his theory of gravitation (called General Relativity) proposed that

$$F(g_{ij}) = G(\text{Energy, Momentum})$$

i.e., that the metric components g_{ij} , which determine the shape of the light cones in spacetime, are determined by the distribution of energy and momentum in spacetime or the very structure of spacetime is determined by the energy density in spacetime.

A result of the theory is that light would be affected by gravitational fields. The following predictions were made and have been confirmed experimentally :

- (1) If we place a light source at the top of a tower and shine the light downwards, then the change in the strength of the gravitational field as we go from the top to the bottom of the tower causes a gravitational redshift such that

$$\frac{f_{h+\Delta h} - f_h}{f_h} \approx \frac{GM}{R_{\text{earth}}^2} \Delta h \approx 10^{-14} \text{ for a 10 meter tower}$$

- (2) If we place a clock at the top of a tower and a clock at the bottom, then because of the difference in the strength of the gravitational field between the top and the bottom of the tower the clocks run at different rates - called the gravitational time dilation

$$\frac{\tau_{h+\Delta h} - \tau_h}{\tau_h} \approx \frac{GM}{R_{earth}^2} \Delta h \approx 10^{-14} \text{ for a 10 meter tower}$$

- (3) If light passes by a large mass (like a star) it does not travel in a straight line but is bent. The amount of bending has two observable consequences
- a) if a star is observed when the sun is not in the way and then when the light would just pass by the sun, the observed difference in direction to the star is about 1.75 seconds of arc
 - b) if a signal is sent from Earth to Venus with the sun in between, there is a time delay due to longer(bent) paths of motion of about $1.1 \times 10^{-4} \text{ sec}$
- (4) Galaxies can cause gravitational lensing which results in double images for distant stars
- (5) The long axis of the planetary orbit ellipse in the solar system precesses - for mercury this is about 43 seconds of arc per century

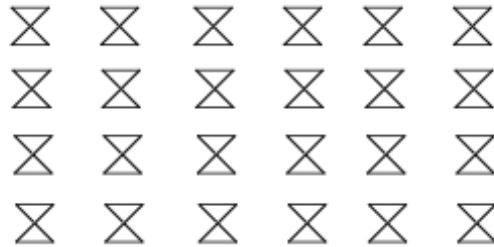
All particles in this theory are free particles, i.e., there are no forces. All particles move along geodesics, which are the path of shortest interval in spacetime. The geodesics for a given spacetime are determined by the metric components. So the distribution of energy determines the metric components which in turn determines the geodesics and particles move on geodesics. In flat spacetime (think of a plane in space) the geodesic is a *straight* line. In fact, the geodesic is always the *straightest* line in a given spacetime. In curved spacetime (think of the surface of a sphere) the geodesic is not a straight line(great circle on the sphere). If we move a vector *parallel* to itself over a closed curve in flat spacetime it does not change its direction. If we move a vector *parallel* to itself over a closed curve in curved spacetime it does change its direction.

If I turn off gravity and throw an eraser, then it follows the geodesic in this *flat* spacetime which is a straight line. If gravity is present, then it follows the geodesic in this *curved* spacetime which is a parabola. The planetary elliptical orbits in space are the geodesics for the 4-dimensional spacetime near the sun.

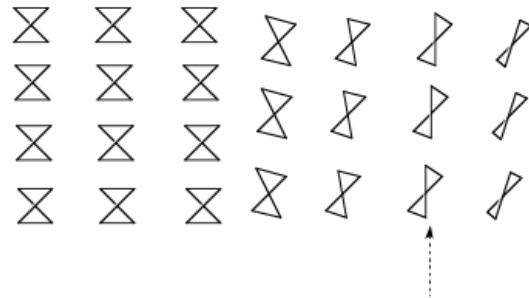
All of these result have been confirmed experimentally.

1.13.1 Black Holes

A few thoughts at this point about Black Holes. We will elaborate greatly when we discuss Black Holes in Geroch. In special relativity the light cone structure of spacetime looks like



This what we mean by *flat space*the light cones are the same everywhere. What happens, however, if we observe a light cone structure of spacetime that looks like



Then, to the left we have flat spacetime, but to the right something strange is happening. The left side of the light cone is rotating clockwise. This means that access to regions to the left is being restricted (takes longer to get there).

As we go further to the right, we reach a point (arrow) where the left side of the cone is vertical and all of space to the left is no longer accessible. This point is on a surface called an *event horizon*. Once some observer crosses this surface we can no longer see them (there is an infinite redshift) and they (and light) can no longer get to the back across the surface (hence the name *black hole*). The observer can only proceed (remember must stay inside forward cone) to the right where the light cones tilts even further. The end result is the light cone being a single line and the observer having no choice about future motion. This point is called a *singularity*. Long before reaching the singularity, the tidal forces become so large that any object is torn apart.

These radical solutions to Einstein's equation have now been confirmed experimentally (via the radiation coming from matter falling into the black hole) and

are thought to exist at the center of all galaxies.

In a *static* black hole as just described, the event horizon and the infinite red shift surface are the same surface and energy can only pass through in one direction. If the black hole is rotating, however, the event horizon and the infinite red shift surface are not necessarily the same surface. The regions between an infinite red shift surface and a event horizon is called the *ergosphere*. It is possible to extract energy from the ergosphere as follows:

1. a spaceship falls from infinity into ergosphere along an orbit with positive energy
2. once there, using a spring-loaded device we eject a brick into an orbit with negative energy
3. the spaceship recoils into a new orbit with larger positive energy
4. energy is constant(conserved) so the spaceship emerges with more energy than it went in, but the black hole + brick have lost energy

A very tricky and dangerous maneuver.

Now it is possible to follow a worldline that is everywhere timelike(allowed) such that one passes through the ergosphere and the particle emerges before it entered ($\Delta t < 0$). The time change can be made arbitrarily large by completing orbits inside the ergosphere this is a model of a time machine for travel to the past!! This violates causality, however, and results in a logical contradiction.

Consider the particle to be a signal (a signal rocket) that is emitted at $t = 0$ by an apparatus located far from the rotating black hole (where spacetime is flat), but is received by this same apparatus at an earlier time, say $t = -2$. Suppose the apparatus is programmed with the following instructions:

1. emit a signal if the signal is not received before $t = 0$
2. do not emit a signal if the signal is received before $t = 0$

This implies a logical contradiction with emission at $t = 0$ and reception at $t = -2$!!

So something will have to give!!!! Very fundamental and exciting stuff..... More later.

1.14 Digression to 4-Vectors

Let us now digress to study a little bit of mathematical physics so that some of the ideas we have been talking about can be linked to each other in ways not previously imagined.

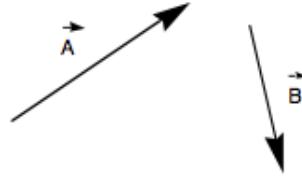
What is an ordinary vector in 3-dimensional space?

A vector has many levels of complexity and is a very abstract mathematical object. A vector is a mathematical(geometrical) object that is representable by two numbers in two dimensions, three numbers in three dimensions, and so on. One characterization is to specify its magnitude or length and orientation or direction - imagine that it is a directed line segment. As we shall see, quantum mechanics will be formulated in terms of vectors, but they will not be directed line segments.

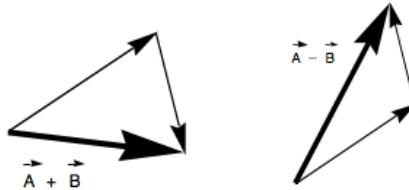
1.14.1 The Standard Language of Vectors

As we said, in ordinary space, we can represent a vector by a directed line segment(an arrow). A straightforward property of a vector is multiplication of the vector by a scalar (a real number) α . In this case the magnitude of the vector changes and the direction stays the same (it might reverse if $\alpha < 0$).

Now given two vectors as shown below



we define the sum and difference of the two vectors or the general property *vector addition* by the diagrams shown below:

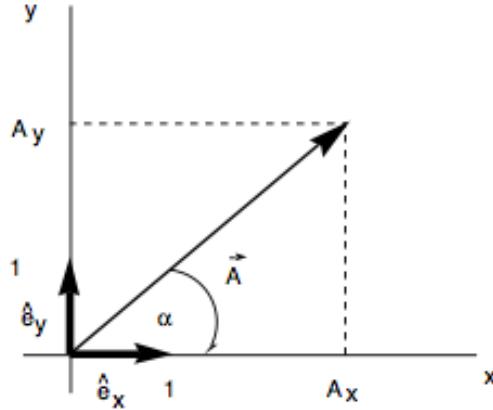


Clearly vector addition as defined above, i.e.,

$$\begin{aligned}\vec{C} &= \vec{A} + \vec{B} \\ \vec{D} &= \vec{A} - \vec{B} = \vec{A} + (-\vec{B})\end{aligned}$$

yields a new vector in each case. This new vector can have both a different direction and a different magnitude than either of the two vectors that are used to create it. These two properties allow us to define a linear combination of vectors $\vec{C} = \alpha\vec{A} + \beta\vec{B}$, which is also a well-defined vector. Although this is

a perfectly good way to proceed, it will not allow us to generalize the notion of a vector beyond ordinary space, which is an arena that will turn out to be much too confining in our effort to understand theoretical physics. We need to formulate these same concepts in another way. Consider the vector shown below:



In this figure, we have also defined two special vectors, namely,

$$\begin{aligned}\hat{e}_x &= \text{unit (length = 1) vector in x-direction} \\ \hat{e}_y &= \text{unit (length = 1) vector in y-direction}\end{aligned}$$

In terms of these unit vectors we can write

$$\vec{A} = A_x \hat{e}_x + A_y \hat{e}_y$$

where

$$\begin{aligned}A_x \hat{e}_x &= \text{vector of length } A_x \text{ in x-direction} \\ A_y \hat{e}_y &= \text{vector of length } A_y \text{ in y-direction}\end{aligned}$$

and the sum of these two vectors equals \vec{A} because of the rule for adding vectors that we stated earlier.

We now define

$$\begin{aligned}A_x &= \text{component of } \vec{A} \text{ in x-direction} \\ A_y &= \text{component of } \vec{A} \text{ in y-direction}\end{aligned}$$

From the diagram it is also clear that

$$A_x = A \cos \alpha \quad \text{and} \quad A_y = A \sin \alpha$$

where

$$A = \text{length of vector } \vec{A} = \sqrt{A_x^2 + A_y^2} \text{ by the Pythagorean theorem}$$

We can then redefine vector addition in terms of components and unit vectors as follows:

$$\begin{aligned}\vec{A} &= A_x \hat{e}_x + A_y \hat{e}_y \\ \vec{B} &= B_x \hat{e}_x + B_y \hat{e}_y \\ \vec{A} + \vec{B} &= (A_x + B_x) \hat{e}_x + (A_y + B_y) \hat{e}_y \\ \vec{A} - \vec{B} &= (A_x - B_x) \hat{e}_x + (A_y - B_y) \hat{e}_y\end{aligned}$$

i.e., we can just add and subtract components.

We now define an important new mathematical object using unit vectors. It is the *scalar or inner product* and its symbol is a . (dot). We *define* this operation with a set of rules involving the unit vectors:

$$\begin{aligned}\hat{e}_x \cdot \hat{e}_x &= 1 = \hat{e}_y \cdot \hat{e}_y \\ \hat{e}_x \cdot \hat{e}_y &= 0 = \hat{e}_y \cdot \hat{e}_x\end{aligned}$$

or

$$\hat{e}_i \cdot \hat{e}_j = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

The inner product satisfies the following relations:

$$\begin{aligned}(\alpha \hat{e}_i) \cdot (\beta \hat{e}_j) &= \alpha \beta \hat{e}_i \cdot \hat{e}_j \\ (\alpha \hat{e}_i + \gamma \hat{e}_k) \cdot (\beta \hat{e}_j + \eta \hat{e}_m) &= \alpha \beta \hat{e}_i \cdot \hat{e}_j + \alpha \eta \hat{e}_i \cdot \hat{e}_m + \gamma \beta \hat{e}_k \cdot \hat{e}_j + \gamma \eta \hat{e}_k \cdot \hat{e}_m\end{aligned}$$

Using these defining relations we can determine the scalar product of any two vectors as follows

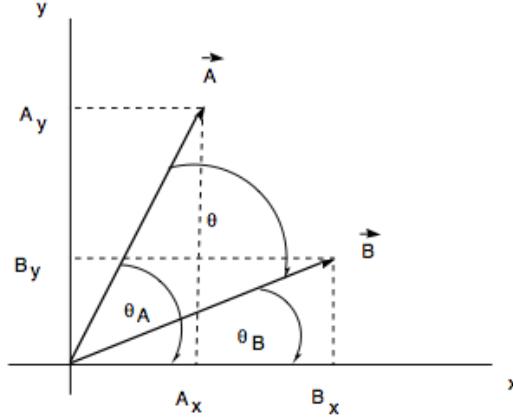
$$\vec{A} = A_x \hat{e}_x + A_y \hat{e}_y , \quad \vec{B} = B_x \hat{e}_x + B_y \hat{e}_y$$

$$\begin{aligned}\vec{A} \cdot \vec{B} &= (A_x \hat{e}_x + A_y \hat{e}_y) \cdot (B_x \hat{e}_x + B_y \hat{e}_y) \\ &= A_x B_x \hat{e}_x \cdot \hat{e}_x + A_x B_y \hat{e}_x \cdot \hat{e}_y + A_y B_x \hat{e}_y \cdot \hat{e}_x + A_y B_y \hat{e}_y \cdot \hat{e}_y \\ &= A_x B_x (1) + A_x B_y (0) + A_y B_x (0) + A_y B_y (1) \\ &= A_x B_x + A_y B_y\end{aligned}$$

We note that

$$\begin{aligned}\vec{A} \cdot \vec{A} &= A_x B_x + A_y B_y = A_x^2 + A_y^2 = A^2 \\ A &= \sqrt{\vec{A} \cdot \vec{A}} = \text{length of the vector } \vec{A}\end{aligned}$$

Now looking at the diagram below we can derive another important result.



We have

$$\begin{aligned}\vec{A} \cdot \vec{B} &= A_x B_x + A_y B_y = AB(\cos \theta_A \cos \theta_B + \sin \theta_A \sin \theta_B) \\ &= AB \cos(\theta_A - \theta_B) = AB \cos \theta\end{aligned}$$

so that

$$\begin{aligned}\vec{A} \cdot \vec{B} &= AB \cos \theta \\ &= (\text{length of } \vec{A})(\text{length of } \vec{B}) \cos(\text{angle between } \vec{A} \text{ and } \vec{B}) \\ &= (\text{length of } \vec{A})(\text{length of } \vec{B} \text{ in the direction of } \vec{A}) \\ &= (\text{length of } \vec{A})(\text{projection of } \vec{B} \text{ onto the direction of } \vec{A})\end{aligned}$$

Therefore, we have

$$\vec{B} = \vec{A} \rightarrow \theta = 0 \rightarrow \vec{A} \cdot \vec{A} = A^2 \text{ as before}$$

$$\vec{B} \text{ perpendicular(orthogonal) to } \vec{A} \rightarrow \theta = \pi/2 = 90^\circ \rightarrow \vec{A} \cdot \vec{B} = 0$$

or vice versa

$$\text{If } \vec{A} \cdot \vec{B} = 0, \text{ then } \vec{A} \text{ is orthogonal to } \vec{B}$$

If two vectors satisfy $\vec{A} \cdot \vec{B} = 0$, then they are said to be orthogonal. If the vector satisfies $\vec{A} \cdot \vec{A} = 1$, then it is said to be normalized to one.

We also have for any vector

$$\begin{aligned}\vec{A} &= A_x \hat{e}_x + A_y \hat{e}_y \\ \vec{A} \cdot \hat{e}_x &= (A_x \hat{e}_x + A_y \hat{e}_y) \cdot \hat{e}_x = A_x = \text{x-component} \\ \vec{A} \cdot \hat{e}_y &= (A_x \hat{e}_x + A_y \hat{e}_y) \cdot \hat{e}_y = A_y = \text{y-component} \\ \vec{A} &= (\vec{A} \cdot \hat{e}_x) \hat{e}_x + (\vec{A} \cdot \hat{e}_y) \hat{e}_y\end{aligned}$$

Generalizing to 3 dimensions we have

$$\vec{A} = A_x \hat{e}_x + A_y \hat{e}_y + A_z \hat{e}_z$$

where the set of three orthonormal vectors $\{\hat{e}_x, \hat{e}_y, \hat{e}_z\}$ are called a *basis* for the vector space (any vector can be written as a linear combination of the basis vectors) and we have

$$\hat{e}_i \cdot \hat{e}_j = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

for $i, j = x, y, z$. The number of required basis vectors is the number of numbers needed to characterize a general vector = the *dimension* of the space. The entire collection of vectors we can generate from a basis set is called a *vector space*. So in this room, I would need 3 numbers to characterize each vector. This room is a small part of a 3-dimensional vector space, which is called the *universe* at an instant of time.

Completely removing (x, y, z) from our notation (because it limits us to a maximum of 3 dimensions) we have

$$\begin{aligned} \vec{A} &= \sum_{j=1}^3 A_j \hat{e}_j \\ \hat{e}_k \cdot \vec{A} &= \hat{e}_k \cdot \sum_{j=1}^3 A_j \hat{e}_j = \sum_{j=1}^3 A_j \hat{e}_k \cdot \hat{e}_j = \sum_{j=1}^3 A_j \delta_{ij} = A_k = k^{th} - \text{component} \end{aligned}$$

so that

$$\vec{A} = \sum_{j=1}^3 A_j \hat{e}_j = \sum_{j=1}^3 A_j \hat{e}_j (\hat{e}_j \cdot \vec{A})$$

Returning to the Discussion of 4-Vectors

Consider a vector \vec{A} representing some physical variable. Using cartesian unit vectors we can write

$$\vec{A} = \sum_{i=1}^3 A_i \hat{e}_i$$

The components of the vector A_i $i = 1, 2, 3$ are its representation in a given coordinate system. We must choose a coordinate system in order to define the unit vectors. The coordinate system is not an essential part of the physics however. We can just as well use any other coordinate system to define unit vector and the vector \vec{A} .

In particular, we consider another coordinate system with the same origin, but rotated from the first system. In another coordinate system we would write

$$\vec{A} = \sum_{i=1}^3 A'_i \hat{e}'_i$$

Note that the vector \vec{A} has not changed; only its representation (components) in the new system (new basis) has changed. We relate the two representations (components) as follows:

$$\begin{aligned}\sum_{i=1}^3 A_i \hat{e}_i &= \sum_{i=1}^3 A'_i \hat{e}'_i \\ \hat{e}'_j \cdot \sum_{i=1}^3 A_i \hat{e}_i &= \hat{e}'_j \cdot \sum_{i=1}^3 A'_i \hat{e}'_i = \sum_{i=1}^3 A'_i \hat{e}'_j \cdot \hat{e}'_i = \sum_{i=1}^3 A'_i \delta_{ij} = A'_j \\ A'_j &= \sum_{i=1}^3 A_i (\hat{e}'_j \cdot \hat{e}_i)\end{aligned}$$

The coefficients $(\hat{e}'_j \cdot \hat{e}_i)$ are numbers that are determined by the specific rotation. They are independent of the vector \vec{A} . We now redefine a vector as follows:

A vector in 3 dimensions is a set of 3 numbers $\{A_i\}$ (components) which transform under a rotation of the coordinate system according to

$$A'_j = \sum_{i=1}^3 A_i (\hat{e}'_j \cdot \hat{e}_i)$$

Any quantity which is unchanged by a coordinate transformation is called an **invariant** of the transformation. Since the principle of relativity requires that the results of physical theories (physical laws) be independent of the choice of coordinate system (must be inertial however), all physical laws must involve *only* invariants.

The dot product of two vectors is a scalar. Scalars are numbers that are independent of our choice of coordinate system. This gives us a method for *constructing* invariants. We can show that the dot product produces an invariant as follows:

$$\begin{aligned}\vec{A}' \cdot \vec{B}' &= \left(\sum_{i=1}^3 A_i (\hat{e}_i) \right) \cdot \left(\sum_{j=1}^3 B_j (\hat{e}_j) \right) = \sum_{i,j} A_i B_j \hat{e}_i \cdot \hat{e}_j \\ \sum_{i,j} A_i B_j \delta_{ij} &= \sum_{i,j} A_i B_i = \vec{A} \cdot \vec{B}\end{aligned}$$

In particular, the norm or length-squared of a vector, $A^2 = \vec{A} \cdot \vec{A}$ is a scalar invariant. We now define a rotation.

A rotation is any transformation which leaves $r^2 = \vec{r} \cdot \vec{r} = x^2 + y^2 + z^2$ invariant

In Minkowski 4-dimensional spacetime we define vectors in a different manner. Both the ordinary space 3-dimensional and the Minkowski 4-dimensional

vector definitions are special cases of a more general definition. The ordinary 3-dimensional definition corresponds to Euclidean geometry.

In Minkowski 4-dimensional spacetime we write the spacetime 4-vector(using BOLDFACE) in this way

$$\mathbf{s} = (ct, x, y, z)$$

and the scalar product of the vector with itself (its norm) as

$$\mathbf{s} \cdot \mathbf{s} = c^2 t^2 - x^2 - y^2 - z^2$$

where we note the appearance of minus signs. This is a scalar invariant under Lorentz transformations(it is the spacetime interval). In fact, any set of 4 numbers $\mathbf{A} = (A_0, A_1, A_2, A_3)$ represents a Minkowski 4-vector if its norm defined by

$$\mathbf{A} \cdot \mathbf{A} = c^2 A_0^2 - A_1^2 - A_2^2 - A_3^2$$

is a scalar invariant. In addition, if a set of 4 numbers is a 4-vector then the components transform between frames via the Lorentz transformations as

$$\begin{aligned} A'_0 &= \gamma(A_0 - \beta A_1) \\ A'_1 &= \gamma(A_1 - \beta A_0) \\ A'_2 &= A_2 \\ A'_3 &= A_3 \end{aligned}$$

for relative motion along the 1-axis.

It is in this sense that spatial and time variables are *not distinct entities* but are simply *different components* of the same vector and transform into each other under Lorentz transformations. This corresponds to a non-Euclidean geometry.

Another 4-vector is $d\mathbf{s} = (cdt, dx, dy, dz)$ since it is the difference of two 4-vectors. Hence, its norm

$$ds^2 = c^2 dt^2 - dx^2 - dy^2 - dz^2$$

is a Lorentz invariant. A related quantity of great importance is $d\tau^2 = ds^2/c^2$ (dividing an invariant by an invariant means that we still have an invariant). In particular,

$$d\tau^2 = dt^2 - \frac{1}{c^2}(dx^2 + dy^2 + dz^2)$$

Now consider a displacement $d\mathbf{s}$ between two events on the worldline of a moving particle. In the rest frame of the particle, $dx = dy = dz = 0$ and hence $d\tau = dt$ in the particle rest frame (the events are separated only by time). $d\tau$ is the time interval between the two events measured in the rest frame and is thus the *proper time*. It is a *Lorentz invariant*.

Time Dilation (the easy way)

Consider an observer at rest in the (x', y', z', ct') system. In this system the proper time between two events is $d\tau = dt'$. In the (x, y, z, ct) system moving with velocity v relative to the first frame, the time interval between the same two events is given by

$$dt^2 - \frac{1}{c^2}(dx^2 + dy^2 + dz^2)$$

But $d\tau$ is an invariant or its value is the same in all frames. We therefore have

$$dt'^2 = dt^2 - \frac{1}{c^2}(dx^2 + dy^2 + dz^2)$$

or

$$\begin{aligned} \left(\frac{dt'}{dt}\right)^2 &= 1 - \frac{1}{c^2} \left(\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 \right) \\ &= 1 - \frac{v^2}{c^2} = \frac{1}{\gamma^2} \end{aligned}$$

Therefore, $dt = \gamma dt'$, which is the time dilation formula. We did not need to introduce hypothetical experiments or discussions of simultaneity to obtain this result. That is an example of the power of using 4-vectors.

Other 4-Vectors

Using $d\mathbf{s} = (cdt, dx, dy, dz)$ and dividing by the Lorentz invariant $d\tau$ yields another 4-vector

$$\frac{d\mathbf{s}}{d\tau} = \left(c\frac{dt}{d\tau}, \frac{dx}{d\tau}, \frac{dy}{d\tau}, \frac{dz}{d\tau} \right) = \mathbf{u} = \text{4-vector velocity}$$

Its norm is an invariant so it can be calculated by in any frame. We pick the rest frame where

$$\mathbf{u} = (c, 0, 0, 0) \rightarrow u^2 = c^2 = \text{invariant}$$

For a moving particle where the (x, y, z, ct) system moves with velocity $-v$ relative to the rest frame of the particle we have $dt = \gamma d\tau$ and thus

$$\mathbf{u} = \left(c\frac{dt}{d\tau}, \frac{dx}{d\tau}, \frac{dy}{d\tau}, \frac{dz}{d\tau} \right) = \gamma(c, \vec{v}) \quad \text{where} \quad \gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Since the rest mass m_0 of any particle is a Lorentz invariant, is a 4-vector with dimensions of momentum. We define the 4-momentum as

$$\boldsymbol{\rho} = m_0 \mathbf{u} = m_0 \gamma(c, \vec{v}) = \left(\frac{E}{c}, \vec{p} \right)$$

We already saw that

$$\rho^2 = \left(\frac{E}{c}\right)^2 - \vec{p}^2 = m_0^2 c^2 = \text{invariant}$$

Since the variables E and \vec{p} are components of a 4-vector the must obey the Lorentz transformations

$$\begin{aligned}\frac{E'}{c} &= \gamma \left(\frac{E}{c} - \beta p_x \right) \\ p'_x &= \gamma \left(p_x - \beta \frac{E}{c} \right) \\ p'_y &= p_y \\ p'_z &= p_z\end{aligned}$$

These relations are used to prove that a magnetic field is observed in frames moving relative to fixed charged particles whereas only electric fields are observed in the rest frame of the charged particles. Magnetic fields are a consequence of special relativity!!

Finally we confirm our identification of the energy. Using $E = \gamma m_0 c^2$, we define the 4-vector Minkowski force as

$$\phi = \frac{d\rho}{d\tau} = \left(\frac{d(\gamma m_0 c)}{d\tau}, \frac{d\vec{p}}{d\tau} \right)$$

If dt is the time interval in the observers frame corresponding to the interval of proper time $d\tau$, then $dt = \gamma d\tau$ and we get

$$\phi = \gamma \left(\frac{d(\gamma m_0 c)}{d\tau}, \vec{F} \right) , \quad \vec{F} = \frac{d\vec{p}}{dt}$$

With this construction, the 4-momentum is conserved(constant) when the 4-force is zero. This corresponds to energy and momentum conservation. If the 4-force is zero in one frame then it is zero in all frames and hence if energy and momentum are conserved in one frame they are conserved in all frames. In Newtonian physics

$$\vec{F} \cdot \vec{v} = \frac{dE}{dt}$$

where E = total energy. Let us look at the corresponding quantity in 4-dimensions

$$\phi \cdot \mathbf{u} = \left(\frac{d(\gamma m_0 c)}{d\tau}, \vec{F} \right) \cdot \gamma(c, \vec{v}) = \gamma^2 \left[\frac{d(\gamma m_0 c^2)}{d\tau} - \vec{F} \cdot \vec{v} \right]$$

Now the scalar product is an invariant and thus we can evaluate it in the rest frame of the particle. In this frame $\vec{F} \cdot \vec{v} = 0$ since $\vec{v} = 0$. We also have

$$\frac{d(\gamma m_0 c^2)}{d\tau} = \gamma m_0 v \frac{dv}{dt} = 0$$

since $v = 0$. Therefore

$$\phi \cdot \mathbf{u} = 0 = \gamma^2 \left[\frac{d(\gamma m_0 c^2)}{d\tau} - \vec{F} \cdot \vec{v} \right]$$

or

$$\frac{d(\gamma m_0 c^2)}{d\tau} = \vec{F} \cdot \vec{v} \rightarrow E = \gamma m_0 c^2$$

as we indicated earlier. In this sense the momentum and energy variables are *not distinct entities* but are simply *different components* of the same vector and transform into each other under Lorentz transformations.

A Further Generalization

We can generalize the scalar product to any number of dimensions and any type of geometry. We have

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^n \sum_{j=1}^n g_{ij} A_i B_j$$

where g_{ij} is the so-called metric object. We can represent it by a matrix. In ordinary 3-dimensional space we have

$$[g] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \rightarrow g_{ij} = \delta_{ij}$$

and hence in three dimensions we have

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^n A_i B_i = \vec{A} \cdot \vec{B}$$

In Minkowski 4-space we have

$$[g] = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

In the theory of gravitation (general relativity) we have

$$[g] = \begin{pmatrix} g_{00}(x, y, z, t) & g_{01}(x, y, z, t) & g_{02}(x, y, z, t) & g_{03}(x, y, z, t) \\ g_{10}(x, y, z, t) & g_{11}(x, y, z, t) & g_{12}(x, y, z, t) & g_{13}(x, y, z, t) \\ g_{20}(x, y, z, t) & g_{21}(x, y, z, t) & g_{22}(x, y, z, t) & g_{23}(x, y, z, t) \\ g_{30}(x, y, z, t) & g_{31}(x, y, z, t) & g_{32}(x, y, z, t) & g_{33}(x, y, z, t) \end{pmatrix}$$

The metric not constant but is dependent on where you are and what time it is! Clearly the world is considerably more complicated as we will see.

Now back to special relativity

Let us return to the relations

$$\left(\frac{E}{c}\right)^2 - \vec{p}^2 = m_0^2 c^2 \quad , \quad \frac{pc}{E} = \frac{v}{c} = \beta$$

Notice that if $v = c$, then $E = pc$ and $m_0 = 0$. Therefore, particles with zero rest mass exist. They always move with the speed of light. Even though they have no mass they do have energy and momentum! An example of such a particle is the photon, the particle of light.

Radiation Pressure

Here is an interesting application. When light(photon) which carries momentum and energy reflects off of a surface it transfers momentum and energy to the surface. Since a change in momentum corresponds to a force and a force on a surface area corresponds to pressure, light exerts radiation pressure on any reflecting surface. If we have normal incidence on the surface, then the total change in the photon momentum is

$$\Delta p = 2p = 2\frac{E}{c}$$

If there are n photons per unit area per second, then the total momentum change per second per unit area is

$$\text{pressure} = 2n\frac{E}{c} = 2\frac{I}{c}$$

where $I = nE$ is the intensity of the light (the power per unit area). The average intensity of sunlight falling on the earth surface is $\approx 1000 \text{ W/m}^2 = 1000 \text{ J/m}^2 \cdot \text{sec}$. The radiation pressure on a mirror is then

$$\text{pressure} = 2\frac{I}{c} = 7 \times 10^{-4} \text{ N/m}^2$$

This is very small(atmospheric pressure is 10^6 N/m^2). On a cosmic scale, however, this radiation pressure is large, that is, it is able to help keep stars from collapsing under their own gravitational forces.

We now ask this question. How big must the sail of a light-sail starship be to work effectively? Suppose that the sail material has the property *mass per m²* = $\rho \text{ kg}$ and that the ship has a mass of $M \text{ kg}$. A crude calculation goes like this

$$\text{pressure at distance } r \text{ from the sun} = 7 \times 10^{-4} \left(\frac{r_{\text{earth}}}{r}\right)^2 \text{ N/m}^2$$

$$\text{force on sail} = \text{pressure} \times \text{area} = 7 \times 10^{-4} \left(\frac{r_{\text{earth}}}{r}\right)^2 A$$

$$\text{acceleration } = a = \frac{\text{force}}{\text{total mass}} = \frac{7 \times 10^{-4} \left(\frac{r_{\text{earth}}}{r}\right)^2 A}{M + \rho A}$$

Suppose we have a sail with an maximum area = USA = $10^{13} m^2$. For $r = nr_{\text{earth}}$, $\rho = 10^{-8}$, $M = 10^5$, α = fraction of area used, we get

$$a = \frac{7 \times 10^4 \left(\frac{1}{n}\right)^2 10^{13} \alpha}{10^5 + 10^5 \alpha} = 7 \times 10^{-4} \frac{\alpha}{1 + \alpha} \left(\frac{1}{n}\right)^2 \text{m/sec}^2$$

where $0 \leq \alpha \leq 1$ and $n \geq 1$. The acceleration will drop below $0.0001 g$ for $\alpha = 1$ when $n = 20000$ or we are at a distance of 20000 earth radii or about $2 \times 10^{12} \text{ miles}$ from the sun. This is about 0.03 light-year . Depending on what we did earlier we could have a sizable speed by this point!

Chapter 2

Notes on Mermin

It's About Time

2.1 The Principle of Relativity

The *principle of relativity* is the first postulate due to Einstein for his theory of (special) relativity. It is an example of an *invariance principle*. All invariance principles begin with the phrase *All other things being the same* and then go on to say:

1. it does not matter where you are. (*Principle of translational invariance in space*)
2. it does not matter when you are. (*Principle of translational invariance in time*)
3. it does not matter how you are oriented. (*Principle of rotational invariance*)

The principle of relativity follows the same pattern.

All other things being the same it does not matter how fast you are going if you are moving with fixed(constant)speed along a straight line(constant or uniform velocity).

Invariance principles are useful because they allow us to extend our knowledge to new situations. In particular, the principle of relativity tells us that no experiments that we do can enable us to distinguish between our being in a state of rest or a state of uniform motion. Thus, we can infer what happens in all uniformly(fixed speed and direction) moving frames(systems in which we have chosen to describe things) if we know what is happening in one of them. Uniformly moving frames are equivalent to non-accelerating frames. How do we know if there is acceleration? Simple example: hang any object on a string

from the ceiling. If it hangs vertically downwards, then there is no acceleration!

The principle of relativity does not say that physical properties cannot depend on velocity. It only requires that if an object has certain properties when it is in a frame where it is at rest, then if the same object moves uniformly, it will have the same properties *in a frame that moves uniformly with it*. A non-trivial example, which we discuss in detail later, is the *Doppler effect*.

If yellow light moves away from you at a very large speed, the color you see changes from yellow to red; if it moves towards you the color changes from yellow to blue. Thus, the color of an object in a fixed frame can depend on whether it is moving or at rest and in what direction it is moving. The principle of relativity says that if the light is seen as yellow when it is stationary, then when it moves with uniform velocity it will still be seen as yellow *by an observer who moves with that same velocity*.

The way we will use the principle of relativity is the following: *Take a situation which you do not fully understand. Find a new frame of reference in which you do understand it. Examine it in that new frame of reference. Then translate your understanding in the new frame back into the language of the old frame.*

Examples: Newton's first law of motions states that in the absence of an external force a uniformly moving body continues moving uniformly. It turns out that this law follows from the principle of relativity and a much simpler law. The simpler law states that in the absence of an external force a stationary body remain stationary. Let us see how this works.

Suppose that we only know the simpler law. The principle of relativity tells it must be true for all uniformly moving frames of reference. If we want to learn about the subsequent behavior of a ball initially moving at 50 ft/sec in the absence of any external force, all we have to do is find a uniformly moving frame in which we can apply the simpler law. Clearly, the frame we need is the one that moves uniformly at 50 ft/sec in the same direction as the ball. In that frame the ball is stationary. Thus, by the simple law it remains at rest in this frame and thus in our frame we think it continues moving uniformly at 50 ft/sec . Thus, the principle of relativity works. This, however, is not an impressive example, since you could have easily guessed this result. Let consider an example, where the result is not obvious.

Suppose we have two identical perfectly elastic balls. Identical elastic balls have the property that if you shoot them directly at each other with the same speed, then after they collide each bounces back in the direction from which it came with the same speed that it had before the collision. Question: What happens if one of the balls is at rest and you shoot the other one directly at it?

The upper part of figure below illustrates the rule we know: when the balls move

at each other with equal speeds, they simply rebound with the same speeds.

	Before	After
Known		
Unknown		?

Figure 1.1

The lower part of the figure illustrates the new situation (what we are trying to figure out). We will assume that the moving ball has speed 10 ft/sec . What should we draw in the box with the question mark in it?

Assume now that we are in a train moving at 5 ft/sec in the same direction as the moving ball. This is illustrated in the upper part of the figure below.

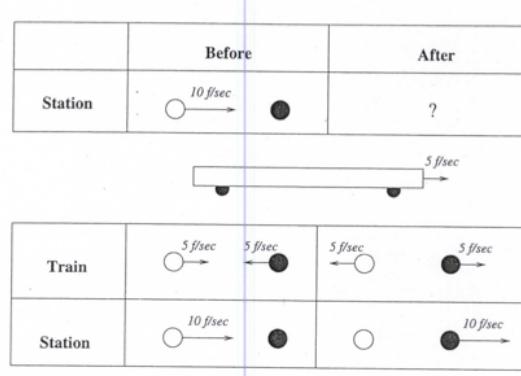


Figure 1.2

What would this look like from the frame of the train? This is illustrated in the middle part of the figure (labelled TRAIN). This corresponds to the statement of the law; thus, in this frame the two balls rebound with the same speeds as shown. Now we have to translate back to the original frame (STATION). This is illustrated in the lower part of the figure.

So we have used the principle of relativity to learn something new about elastic balls - if one is at rest and is hit by a moving ball, then the original moving ball ends up at rest and the ball that was originally at rest is now moving with the same speed as the originally moving ball.

This is a dramatic illustration of the power of the principle of relativity.

Another example is shown in the figure below.

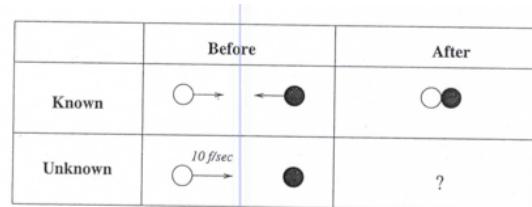


Figure 1.3

Here two sticky balls have the property that if they are fired directly at each other with equal speeds, then they stick together upon collision and the resulting compound ball is at rest. Question: What happens if one sticky ball is sent at 10 ft/sec directly at another sticky ball that is at rest as shown above - with what speed and in what direction will the compound ball move after the collision?

The principle of relativity easily deals with this case also. We view the initially moving ball and the initially stationary ball from a train from which each is moving with a speed of 5 ft/sec but in opposite directions. The train must be moving at 5 ft/sec in the same direction as the initially moving ball as shown below.

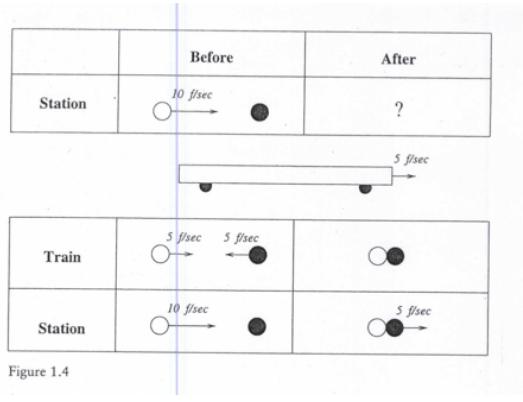


Figure 1.4

We already know what happens in this frame - the two balls stick together and are at rest. Since the train is moving at 5 ft/sec in the direction shown, in the station frame the compound ball will be moving at 5 ft/sec in the direction shown. That solves the problem.

Another example is as follows: Suppose we have two elastic balls, but one of them is very big and the other is very small. If the big ball is stationary and the small ball is fired directly at it, the small ball simply bounces back in the direction it came from with the same speed and the big ball stays at rest. Question: With what speed will each ball move after the collision if the small ball is stationary and the big ball is fired directly at it with a speed of 10 ft/sec ? This is shown in the figure below.

We want to examine the initial situation in a frame where the big ball is at rest (that is where we know what will happen). This means we must ride on a train moving with 10 ft/sec to the left as shown in the figure below.?

In that frame the small ball will move at 10 ft/sec to the right (as shown) and the situation before the collision is the one we understand. So in the train frame the small ball will move at 10 ft/sec to the left and the big ball will remain stationary. Returning to the station frame, the big ball moves with the train at 10 ft/sec to the left. The little ball, however, moves at 20 ft/sec to the

	Before	After
Known		
Unknown		?

Figure 1.5

	Before	After
Station		?
Train		
Station		

Figure 1.6

left; it now moves at twice the speed of the big ball.

Clearly, the principle of relativity is very powerful.

2.2 Combining(Small) Velocities

During our discussions of the principle of relativity we made use of a rule for finding relative velocities that no one challenged because it seems to make sense and using it we have survived all the years of our lives,i.e., it seems to work in the world of everyday experience. Let us elaborate somewhat on this rule since, as we will find out, it is not correct in general and understanding it better will help us to eventually see the problems.

Let us call this rule the *nonrelativistic(velocities very small compared to light) velocity addition law*. We will only consider one-dimensional motion for simplicity. In this case we can deal with velocity directions very easily. There are only two possible direction. Assuming that the direction of motion is east-west, we will(arbitrarily) assign a positive velocity for motion towards the east and a negative velocity for motion towards the west. Remember also that a velocity must always be defined with respect to some frame. The rest frame of an object is often called the *proper frame*. If we have objects X and Y, then we label the

velocity of Y with respect to X (with respect to the proper frame of X) as v_{YX} .

The nonrelativistic velocity addition law is then

$$v_{XZ} = v_{XY} + v_{YZ} \quad (2.1)$$

Now we then have $0 = v_{XX} = v_{XY} + v_{YX}$, which implies that

$$v_{XY} = -v_{YX} \quad \text{equal in magnitude and opposite in direction} \quad (2.2)$$

In using this rule, we have to calculate velocities in different frames moving relative to each other. This will involve measuring spatial and time separations between events in different frames. It turns out the that result (2.1) will result only if the time separations in all frames are identical - that clocks run at the same rate in different frames. Of course this is true, isn't it? Einstein did not think so as we will see. It assumed that this assumption of *absolute time* was false and this leads to many other *obvious* assumptions to fail as we will see.

2.3 The Speed of Light

It is straightforward to measure the speed of light. Send a light beam somewhere, have it reflect back to you and measure the distance(D) to the other place and the time(T) for the entire trip and we have

$$c = \frac{2D}{T} \quad (2.3)$$

The speed of light is $3 \times 10^8 \text{ m/sec}$ in a vacuum (empty space). The first question that arises is - *with respect to what?*

First Obvious Answer: This is the speed with respect to the source of the light(just like a bullet from a gun). There is no theoretical or experimental evidence for this assumption so we must not accept it.

Second Obvious Answer: This is the speed of light with respect to the *medium* the light is propagating in (like waves in water or sound in air). The medium was called *the ether*. It was what was left in a vacuum after you removed everything it is possible to remove. Sounds silly and it was - it was out of desperation - no one knew what to do! No experimental evidence for existence of the ether. So this assumption cannot be accepted.

It turns out that Maxwell's equations of electromagnetism(1854) predicted the existence of light and predicted that the speed of light would not depend on anything - that it would be a universal constant. Einstein latched on to this fact and proposed the principle of relativity applied to electromagnetism also and this then says there can be no such thing as the ether! This then says that the speed of light, which follows from Maxwell's equations, must be the same in

all frames moving uniformly with respect to each other. If you send out a light beam when at rest relative to some frame, then its speed will be measured as c in that frame. If you send out a light beam when you are moving, say a speed $c/2$ with respect to that same frame, then its speed will still be measured as c in that frame! This result certainly violates the *common sense* nonrelativistic velocity addition law (2.1). The constancy of the speed of light is *counterintuitive* or some would even say *impossible*. However, it is experimentally verified and therefore we must change our theory so that it agrees with these experiments.

Something is fundamentally wrong with the way we are thinking about *having a speed with respect to a particular frame*. Maybe measured distances and/or measured time intervals are different in different frames. We must look carefully at how one does these measurements to see where we messed up.

Let P be a valid procedure for time and distance measurements so that we can determine the speed of an object in a given frame. Let Bob carry out this procedure in a space station frame and measure the speed of light as it zooms off into space. He measures $299,792 \text{ km/sec}$. Alice is flying swiftly after the light at a speed that Bob determines is 792 km/sec . Bob then would say that each second the light gets $299,792 \text{ km}$ away from him and that Alice gets an additional 792 km away, so that Bob says the distance between Alice and the light is growing at only $299,000 \text{ km/sec}$. However, if Alice carries out the same procedure P in her rocket frame she finds that the speed of light is $299,792 \text{ km/sec}$ so that in her own frame, the distance between her and the light is still growing at the full $299,792 \text{ km/sec}$.

Why is there a discrepancy? They are using exactly the same procedures. Or are they? What do we mean by the word *exactly*? If Bob, for example, uses clocks that are stationary in his frame to measure times, then, if Alice must use clocks that are stationary in her frame. That is what *exactly* means. In Bob's frame Alice's clocks are moving and in Alice's frame Bob's clocks are moving. We can say similar thing about any meter sticks they might use to measure distances. Thus, Alice's procedure as described in Bob's frame is not exactly the same as Bob's procedures as described in Bob's frame.

This difference allows either Bob or Alice to account for any discrepancy in an entirely rational way.

Before 1905 all physicists made these assumptions about how things work:

1. The procedure that Alice uses to synchronize all the clocks in her frame gives a set of clocks that Bob agrees are synchronized when he tests them against a set of clocks that he has synchronized using the same procedure in his frame.
2. The rate of a clock, as determined in Bob's frame is independent of how fast that clock moves with respect to Bob.

- The length of a meter stick, as determined in Bob's frame, is independent of how fast the meter stick moves with respect to Bob.

If any of these assumptions is false, then the nonrelativistic velocity addition law (2.1) will need to be reexamined. It is now known that all three are false. The theory of relativity that we will develop will tell us how they fail and how, once corrected, we find a simple and coherent picture of space and time measurements that agrees with the experimental fact that the speed of light is a universal constant.

The way we will proceed is to accept the idea that the speed of light is a universal constant and also insisting that the principle of relativity remains valid. This will allow us to figure out how to modify the three assumption above. Once we understand the new assumptions, the universality will not seem strange anymore. This universality goes by the name - *the principle of the constancy of light*. The theory of relativity rests on these two principles or postulates.

2.4 Combining (Any) Velocities

Earlier we said that if Alice, a passenger on a train moving at speed v , can throw a ball with speed u , then if she throws the ball toward the front of the moving train, its speed w with respect to the tracks will be

$$w = u + v \quad (2.4)$$

in the same direction as the train. This was called the *nonrelativistic velocity addition law*. Evidently, it fails to work when $u = c$ (if Alice instead sends out a light beam) for experiment says that $w = c$ and not that $w = c + v$ as would be the case if (4.1) were valid.

So the velocity addition formula fails for light; it turns out it also fails for all other objects moving at any speed. the correct formula turns out to be

$$w = \frac{u + v}{1 + \frac{uv}{c^2}} \quad (2.5)$$

If u and v are small compared to c , then (4.2) reduces to (4.1), which is why observers never noticed that (4.1) was wrong.

The correct result is a direct consequence of the principle of the constancy of the speed of light and the principle of relativity as we now show. The direct way to get (4.2) is to use the fact that we know the speed of light. We need to use light to measure the speed of other things in a way that makes no use of either clocks or meter sticks which we suspect behave differently in frames that are moving relative to each other. We let the moving object (a ball) run a race with a pulse of light (a photon). By comparing respective distances traveled we can determine the speed of the ball, i.e., if the photon moving at speed c covers

twice the distance covered by the ball, then the speed of the ball is $c/2$.

We immediately run into difficulty, however. Although the photon and the ball start the experiment in the same place, they are in different places when the experiment ends. In order to compare the distances traveled we must be able to determine the distance the ball traveled at the exact time the photon distance is measured. This requires two synchronized clocks - one where the photon is measured and one with the ball. Thus, we still have the same problem, i.e., we need to know whether two clocks located at different places are synchronized.

We can deal with this problem by letting the photon reflect off a mirror and go back towards the moving ball and then end the experiment when the photon and the ball are at the same place again. In this case, we can determine the distances traveled at the same time without using any clocks.

Suppose all of this is done on a train. First describe the experiment in the train frame. The experiment starts at the rear of the train. When the photon reaches the front of the train it reflects back toward the rear. Suppose the photon meets the ball a fraction f of the way from the front of the train back to the rear.

Therefore, during the experiment the photon has traveled a distance $L(1 + f)$, where L = length of train and the ball has traveled a distance $L(1 - f)$. The ratio of the distances traveled is

$$R = \frac{1 - f}{1 + f} \quad (2.6)$$

as shown below.

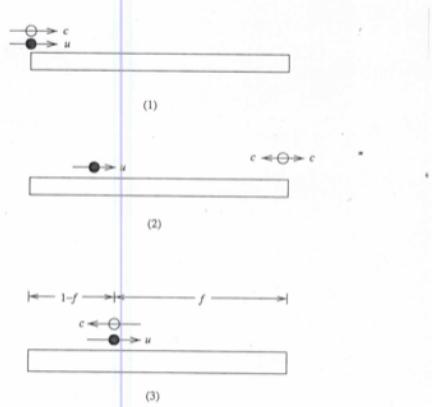


Figure 4.1. A photon (white circle, speed c) runs a race with a ball (black circle, speed u) in a stationary train (long rectangle). The race is pictured at three different moments. (1) At the start of the race the photon and the ball are together at the rear of the train, moving with speeds c and u . (2) The photon reaches the front of the train and bounces back toward the rear (whence the two-headed arrow). (3) At the conclusion of the race the photon reencounters the ball a fraction f of the way back from the front of the train.

Since the photon and the ball were moving for the same time (no clocks needed), this ratio corresponds to the ratio of their speeds. Thus

$$\frac{u}{c} = \frac{1-f}{1+f} \quad (2.7)$$

where u = velocity of the ball. Therefore, the observers on the train can measure the speed of the ball without using clocks and without needing to know the length of the train. We can rearrange this equation to determine the fraction f

$$f = \frac{c-u}{c+u} \quad (2.8)$$

Now let us redo the calculations from the standpoint of the track frame where the train has a velocity v and the ball has a velocity w . We assume that all the velocities are positive, i.e., the ball moves to the right in the train frame and the train and the ball move to the right in the track frame (means velocities and speeds are the same). Carry out same experiment. The analysis now goes as follows.

The photon moves with speed c in both directions in the track frame. Consider the figure below.

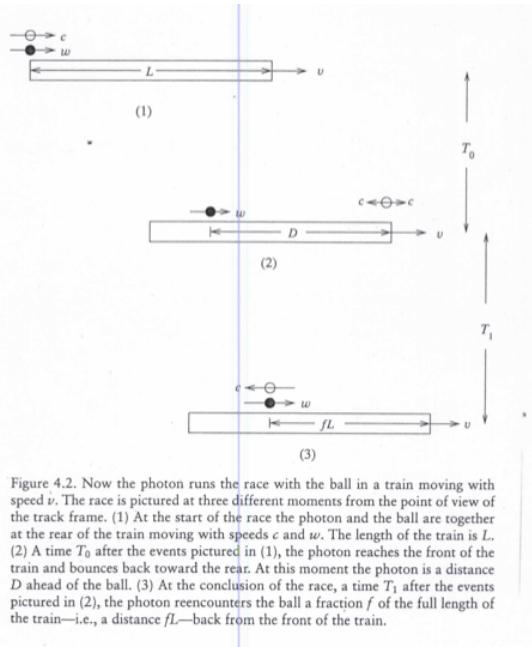


Figure 4.2. Now the photon runs the race with the ball in a train moving with speed v . The race is pictured at three different moments from the point of view of the track frame. (1) At the start of the race the photon and the ball are together at the rear of the train moving with speeds c and w . The length of the train is L . (2) A time T_0 after the events pictured in (1), the photon reaches the front of the train and bounces back toward the rear. At this moment the photon is a distance D ahead of the ball. (3) At the conclusion of the race, a time T_1 after the events pictured in (2), the photon reencounters the ball a fraction f of the full length of the train—i.e., a distance fL —back from the front of the train.

Suppose it takes a time T_0 for the photon to get from the rear to the front of the train and a time T_1 for the reflected photon to get from the front to the point a fraction f of the way back. Let L be the length of the train and D be the

distance between the front of the train and the ball at the moment the photon reaches the front of the train. As we will see, all unknown quantities like D , L , T_0 and T_1 will disappear in the final result.

We now have D is the difference between photon and ball distances in time T_0 , i.e.,

$$D = cT_0 - wT_0 \quad (2.9)$$

and D is also the sum of the distance traveled by the ball and the photon in time T_1 , i.e.,

$$D = cT_1 + wT_1 \quad (2.10)$$

We eliminate D to get $cT_0 - wT_0 = cT_1 + wT_1$ or

$$\frac{T_1}{T_0} = \frac{c-w}{c+w} \quad (2.11)$$

Now let us compare the motion of the photon with that of the train instead of the ball. Since T_0 is the time it takes the photon to get a distance L ahead of the rear of the train(speed v), we have

$$L = cT_0 - vT_0 \quad (2.12)$$

Also T_1 is the time it takes the photon, now moving towards the rear to meet a point on the train(moving with v) originally a distance fL away. Therefore,

$$fL = cT_1 + vT_1 \quad (2.13)$$

Eliminating L we have $cT_1 + vT_1 = f(cT_0 - vT_0)$ so that

$$\frac{T_1}{T_0} = f \left(\frac{c-v}{c+v} \right) \quad (2.14)$$

We then get

$$f = \left(\frac{c+v}{c-v} \right) \left(\frac{c-w}{c+w} \right) \quad (2.15)$$

We have now calculated the fraction f value in two frames and since there is no disagreement between observers in different frame about the point where the ball meets the photon, the fraction value should be the same in both frames (even if length are not the same). Thus we have

$$\left(\frac{c+v}{c-v} \right) \left(\frac{c-w}{c+w} \right) = \frac{c-u}{c+u} \quad (2.16)$$

We can now do some algebra to put this result in a more useful form.

$$\begin{aligned} \left(\frac{c-w}{c+w} \right) &= \left(\frac{c-u}{c+u} \right) \left(\frac{c-v}{c+v} \right) \\ c-w &= (c+w) \left(\frac{c-u}{c+u} \right) \left(\frac{c-v}{c+v} \right) \\ w \left(1 + \left(\frac{c-u}{c+u} \right) \left(\frac{c-v}{c+v} \right) \right) &= c \left(1 - \left(\frac{c-u}{c+u} \right) \left(\frac{c-v}{c+v} \right) \right) \end{aligned}$$

$$w = c \frac{\left(1 - \left(\frac{c-u}{c+u}\right) \left(\frac{c-v}{c+v}\right)\right)}{\left(1 + \left(\frac{c-u}{c+u}\right) \left(\frac{c-v}{c+v}\right)\right)}$$

$$w = c \frac{(c+u)(c+v) - (c-u)(c-v)}{(c+u)(c+v) + (c-u)(c-v)} = c \frac{2c(u+v)}{2(c^2 + uv)}$$

or

$$w = \frac{u+v}{1 + \frac{uv}{c^2}} \quad (2.17)$$

This result is truly amazing. It says that nothing can travel faster than light, i.e., if the train moves at $v = 0.9c$ and the ball moves at $u = 0.9c$, then the speed of the ball in the track frame w , which we would have thought is $u+v = 1.80c$ before, is now

$$w = \frac{u+v}{1 + \frac{uv}{c^2}} = c \frac{0.9+0.9}{1+(0.9)^2} = \frac{1.80}{1.81}c < c$$

No material object can travel faster than c !

For convenience later on let us remove references to trains, balls and tracks as follows. Let A = track, B = train and C = ball. Then, $u = v_{CB}$, $v = v_{BA}$ and $w = v_{CA}$ so that we have, in general,

$$\left(\frac{c - v_{CA}}{c + v_{CA}}\right) = \left(\frac{c - v_{CB}}{c + v_{CB}}\right) \left(\frac{c - v_{BA}}{c + v_{BA}}\right) \quad (2.18)$$

and

$$v_{CA} = \frac{v_{CB} + v_{BA}}{1 + \frac{v_{CB}v_{BA}}{c^2}} \quad (2.19)$$

What happens now to our earlier predictions about various collisions using the principle of relativity and the old velocity addition formula? Consider the collision of the big and little elastic balls we discussed earlier. In the frame in which the big ball is at rest, the little ball simply bounces back in the direction it came from with the same speed (the big ball remains stationary). Suppose that this holds even if the speed of the little ball is comparable to the speed of light c . We then ask what happens to the little ball if it is originally stationary and the big ball is fired at it with speed u . Earlier, we found that the little ball moved with speed $2u$ after the collision. Clearly, this is a problem if $u > c/2$. What happens when we use the new relativistic velocity addition formula?

Suppose the big ball moves to the right with speed u in the rest frame of the little ball. In the rest frame of the big ball, the little ball moves to the left with speed u , so in that frame the little ball bounces back with speed u . To get back to the original frame, we use the relativistic velocity addition formula where we have u is the after-collision velocity of the little ball in the rest frame of the big ball and v if the velocity of the big ball in the rest frame of the little ball, which

is also u in this case. Therefore, we get for w which is the after-collision velocity of the little ball in the frame in which it was initially at rest, the result

$$w = \frac{u + u}{1 + \frac{u^2}{c^2}} = \frac{2u}{1 + \frac{u^2}{c^2}} \quad (2.20)$$

We see that for $u \ll c$ we get our old answer $w = 2u$. However, if $u = c/2$, we get $w = 4c/5 < c$. If $u = 3c/4$ we get $w = 24c/25$. We have $w < c$ no matter what the speed of the little ball!

2.5 Simultaneous Events; Synchronized Clocks

All of these seemingly strange results we have been discovering only seem strange because we have a deeply ingrained misconception about the fundamental nature of time. We implicitly believe that there is an absolute meaning to the simultaneity of two events that happen in different places independent of the frame of reference in which the events are being described. Our ordinary everyday language will not even allow us to contemplate a different situation!

How can we decide whether two events, happening in different places, that are simultaneous in the train frame are also simultaneous in the track frame? Let us choose a specific example. Suppose the two events consists of making a mark on the tracks as they speed by from the rear of the train and doing the same from the front.

How does Alice (in the train frame) convince herself that the marks were made at the same time? She could provide both ends with synchronized clocks and then each end just makes a mark at an agreed time. The problem is that to confirm that the clocks are synchronized require her to be able to confirm that two events in different places are simultaneous which is what we are trying to prove in the first place!

Another try. Alice brings the two clocks together, confirms they are synchronized when they are in the same place and then carry them to the ends of the train. What if clocks moving at different speeds do not increment time at the same rate? Then to convince herself that they are still synchronized after reaching their destinations we could compare them with synchronized clocks that have been at the ends of the train all along. But how do we guarantee these stationary clocks are synchronized? Same problem!

Suppose now that we have two clocks at the middle of the train, synchronize them, then carry them to the ends of the train in exactly the same way (one to the front and the other to the rear). In this case, no matter what happens to their rates along the way, they should still be synchronized. This method works!

But now we have another problem. All may be OK for Alice, but Bob in the track frame might not agree that it took an identical amount of time for the clocks to reach their respective ends of the train. Let us figure out what Bob would determine.

We use a method in the train frame to check simultaneity of two events in different places without using clocks(as earlier). This method can be analyzed in the track frame also. The method relies on the fact the the speed of light is independent of frame. Basically, we are assuming the constancy of the speed of light and then seeing what it says we must conclude about the simultaneity of events. Now Alice places a lamp at the center of the train. Light from the lamp goes towards both ends of the train at the same speed and since it has to travel the same distance in both directions, it arrives at both ends at the same time. At that time marks are made on the tracks(made at the same time). So Alice has produced a pair of events at different places that are simultaneous. This procedure could be accomplished with any two signals (not light) the calculation would require use of the relativistic velocity addition formula and are complicated (see later in notes).

How does Bob interpret this experiment carried out by Alice? Bob agrees that the lamp is at the center of the train(just count meters sticks laid down on floor). In the track frame, however, when the lamp is turned on and light starts to move towards the ends, the rear of the train moves towards the place where the light originated and the front of the train moves away from it. Clearly then, since the speed of light is the same in both direction in the track frame also, it will take less time to reach the rear of the train than to reach the front of the train. Thus, Bob concludes that the mark made on the track at the rear is made before the mark made at the front. So, using the same experimental evidence, Alice say they are simultaneous and Bob says they are not!

Whether or not two events at different places happen at the same time has absolutely no meaning - it depends on the frame of reference in which the events are described.

Let us find a quantitative measure of this disagreement by analyzing the experiment in detail. Let L be the length of the train and v be the speed of the train *in the track frame*. Consider the figures shown below.

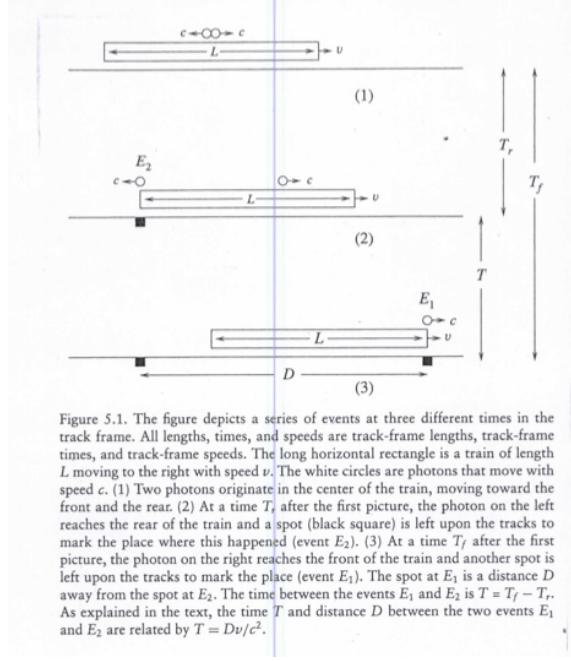


Figure 5.1. The figure depicts a series of events at three different times in the track frame. All lengths, times, and speeds are track-frame lengths, track-frame times, and track-frame speeds. The long horizontal rectangle is a train of length L moving to the right with speed v . The white circles are photons that move with speed c . (1) Two photons originate in the center of the train, moving toward the front and the rear. (2) At a time T_r after the first picture, the photon on the left reaches the rear of the train and a spot (black square) is left upon the tracks to mark the place where this happened (event E_2). (3) At a time T_f after the first picture, the photon on the right reaches the front of the train and another spot is left upon the tracks to mark the place (event E_1). The spot at E_1 is a distance D away from the spot at E_2 . The time between the events E_1 and E_2 is $T = T_f - T_r$. As explained in the text, the time T and distance D between the two events E_1 and E_2 are related by $T = Dv/c^2$.

In part 1 of the the figure, the light is turned on in the middle of the train and two pulses of light (photons) move towards the front and rear from the center. Part 2 shows the same things a time T_r later, just as the photon moving towards the rear meets the rear of the train, which has been moving towards the photon. At that instant, a mark is made on the tracks (at the meeting place). During the time T_r the photon has covered a distance cT_r . That distance equals $1/2$ the length of the train minus the distance the rear of train has moved in the same time, so that

$$cT_r = \frac{1}{2}L - vT_r \quad (2.21)$$

Part 3 of the figure shows things a (longer) time T_f after the light was turned on. At this time the photon moving towards the front of the train reaches the front of the train, which has been moving away from it. In this case the distance covered by the photon cT_f is equal to $1/2$ the length of the train plus the distance the front has moved, so that

$$cT_f = \frac{1}{2}L + vT_f \quad (2.22)$$

We are interested in the time $T = T_f - T_r$ between the making of the two marks. Using (5.1) and (5.2) we have

$$cT = v(T_f + T_r) \quad (2.23)$$

Note that the unknown length L has dropped out of the calculation. Now we

also have the total distance D along the track between the marks given by

$$D = c(T_f + T_r) \quad (2.24)$$

Thus, we finally obtain the relation between the time T between the making of the two marks and the track-frame distance between them as

$$T = \frac{Dv}{c^2} \quad (2.25)$$

We can change this into a general rule by eliminating all talk of Alice, Bob, trains, tracks, and marks: *If events E_1 and E_2 are simultaneous in one frame of reference, then in a second frame of reference that moves with speed v in the direction pointing from E_1 to E_2 , the event E_2 happens at a time Dv/c^2 earlier than the event E_1 , where D is the distance between the events in the second frame.*

This rule for simultaneous events gives rise to a rule for synchronized clocks. Suppose the times of the two markings are recorded in the track frame and attached to the track at the places where the marks were made. How do people in the train frame, who think that the two marks were made simultaneously, account for the fact that the track-frame clocks read times that differ by Dv/c^2 ? They would say that the reason the track-frame clocks differ, i.e., the rear mark was made a time Dv/c^2 before the forward mark, is that the track-frame clock that recorded the time the rear mark was made is actually *behind* the track-frame clock that recorded the time the forward mark was made by exactly that amount: Dv/c^2 . The clocks are not synchronized! We thus have this rule:

If two clocks are synchronized and separated by a distance D in their rest frame, then in a frame in which the clocks move along the line joining them with speed v , the reading of the clock in front is behind the reading of the clock in the rear by Dv/c^2 .

Let us now do in detail the case when the signals being sent to the rear and front are not light. These signals will be assumed to have train-frame speed u and all other aspects of the experiment remain the same. Now suppose that the signal speed u is greater than the speed of the train v in the track frame so that the signal to the rear moves in the direction opposite to the signal to the front, even in the track frame. Let the track frame speeds of Alice's signals to the front and rear be w_f and w_r . The figure below illustrates this experiment.

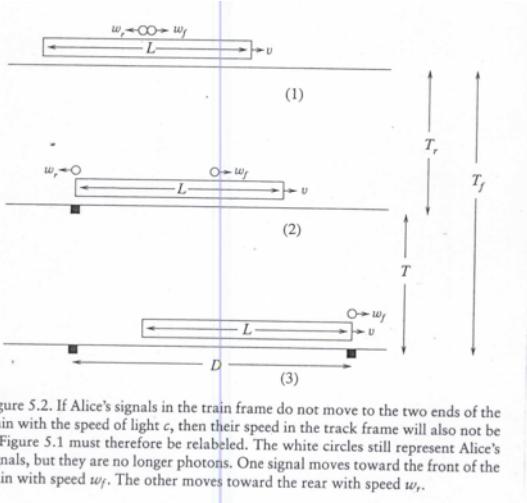


Figure 5.2. If Alice's signals in the train frame do not move to the two ends of the train with the speed of light c , then their speed in the track frame will also not be c . Figure 5.1 must therefore be relabeled. The white circles still represent Alice's signals, but they are no longer photons. One signal moves toward the front of the train with speed w_f . The other moves toward the rear with speed w_r .

Using the same arguments as earlier we now have (with the slower than light signals)

$$w_r T_r = \frac{1}{2}L - vT_r \quad (2.26)$$

and

$$w_f T_f = \frac{1}{2}L + vT_f \quad (2.27)$$

and the total distance D along the tracks between the two marks is now

$$D = w_f T_f + w_r T_r \quad (2.28)$$

We first solve these equations for T_f and T_r

$$T_r = \frac{\frac{L}{2}}{w_r + v} \quad (2.29)$$

$$T_f = \frac{\frac{L}{2}}{w_f - v} \quad (2.30)$$

so that we get

$$T = T_f - T_r = \frac{L}{2} \left(\frac{1}{w_f - v} - \frac{1}{w_r + v} \right) \quad (2.31)$$

We also have

$$D = w_f T_f + w_r T_r = \frac{L}{2} \left(\frac{w_f}{w_f - v} + \frac{w_r}{w_r + v} \right) \quad (2.32)$$

Dividing (5.11) by (5.12) we get

$$\frac{T}{D} = \frac{2v - (w_f - w_r)}{2w_f w_r + v(w_f - w_r)} \quad (2.33)$$

Some comments about this result. If the signals were photons and we set $w_f = w_r = c$ we get the same result as earlier. If the nonrelativistic velocity addition law were valid, i.e., if $w_f = u + v$ and $w_r = u - v$ so that $w_f - w_r = 2v$, then we get $T/D = 0$ so that we do not have any simultaneity or clock synchronization problems. Finally, if we use the correct relativistic velocity addition formula so that

$$w_f = \frac{u + v}{1 + \frac{uv}{c^2}} , \quad w_r = \frac{u - v}{1 - \frac{uv}{c^2}} \quad (2.34)$$

we get

$$\frac{T}{D} = \frac{v}{c^2} \quad (2.35)$$

which is the correct result.

2.6 Moving Clocks Run Slowly; Moving Sticks Shrink

We just concluded that if two clocks are synchronized and separated by a distance D in a frame in which they are both at rest, then in a frame in which they move with speed v along a line joining them, they are not synchronized; the reading of the clock in front lags behind the reading of the clock in the rear by an amount T given by

$$T = \frac{Dv}{c^2} \quad (2.36)$$

How did we get this result? Alice(train-frame) thinks Bob(track-frame) does not know how to synchronize clocks. However, Bob thinks that Alice does not know how to do things simultaneously. The principle of relativity says that (6.1) must be valid in all frames. If Alice thinks Bob's clocks are not synchronized and the result is (6.1), then Bob must think that Alice's clocks are not synchronized and the result must also be (6.1).

Let us now investigate further the consequences of (6.1) concerning rates at which clocks run and lengths in moving frames.

Let Alice's two synchronized clocks be placed at the ends of her train. This is shown in the right half of the figure below. Both clocks read the same time 0 - they are synchronized in the train frame. The length of the train in its rest frame - its *proper length* - is L_A . Also shown are two clocks attached to the track, which move to the left with the track at speed v . The clocks on the track have been synchronized in the track frame. This means that in the train frame the track clock in the front lags behind the track clock in the rear by some amount which we label T_B .

Alice's clocks, although synchronized in the train frame, are not synchronized in the track frame. This is shown in the left half of the figure below.

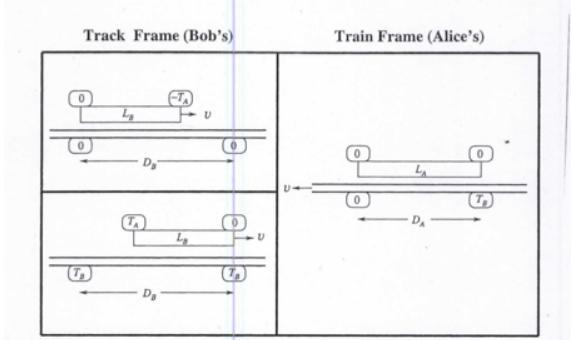


Figure 6.1. Each of the three pictures shows four clocks, a train, and a track. The train is the long rectangle. Two of the clocks are attached to it, one at the front, the other at the rear (the small rounded rectangles just above the front and rear of the train). The tracks are the two long parallel lines below the train. The other two clocks are attached to the tracks (the two small rounded rectangles shown below the tracks). The clocks attached to the train are synchronized in the train frame; those attached to the tracks are synchronized in the track frame. The time shown by a clock is indicated by the symbol inside it. The picture on the right depicts a single moment of time in the train frame. Both train clocks read the same time 0. The track and its attached clocks move to the left with speed v . The track clocks are not synchronized in the train frame: the clock in the front is behind the clock in the rear by a time T_B . The length of the train in the train frame is its proper length, L_A . The two clocks attached to the track are right next to the two clocks attached to the two ends of the train. It is evident from the figure that L_A is the same as D_A , the train-frame length of the segment of (moving) track that stretches between either pair of clocks. The two pictures on the left depict two different moments of time in the track frame. The first picture takes place when both track-frame clocks read 0; the second takes place when both read T_B . The train and its attached clocks move to the right with speed v . Note that the train-frame clocks are not synchronized in the track frame: the clock in front is behind the clock in the rear by a time T_A . The distance between the clock attached to the tracks in the track frame is the proper length D_B of the segment of track between them. The length of the train in the track frame is L_B .

Because the train clocks are *not* synchronized in the track frame, we now must have *two* picture taken at two different track-frame times to depict both of Alice's clocks reading 0. In the upper left picture, the clock at the rear of the train reads 0 and the clock in the front is behind the clock in the rear, reading a negative time that we call $-T_A$. In the lower left picture, the clock at the front has advanced from $-T_A$ to 0, while the clock at the rear has advanced from 0 to T_A . They advanced by the same amount because they are identical clocks moving at the same speed. The track-frame time between the two pictures is the time T_B that the two clocks attached to the track have advanced. Clearly, from the pictures these two clocks are synchronized in the track frame.

Equation (6.1) tells us that T_A (amount two train-frame clocks differ in track frame) is determined by L_A (the train-frame distance between the two clocks) by

$$T_A = \frac{L_A v}{c^2} \quad (2.37)$$

where v is the speed of the train in the track frame. Similarly, T_B (amount two track-frame clocks differ) is determined by D_B (the track-frame distance between

the two clocks) by

$$T_B = \frac{D_B v}{c^2} \quad (2.38)$$

where (the same) v is the speed of the track in the train frame.

Between the two pictures on the left of the figure, both track-frame clocks advance by T_B while both train-frame clocks advance by T_A . Since the track-frame clocks give correct time in the track frame, T_A is the time a train-frame clock advances during a track-frame time T_B . Thus, the slowing-down factor for the two clocks is T_A/T_B .

In a similar manner, the shrinking factor for a moving object is given by the ratio, L_B/L_A , of the track-frame length L_B to its length L_A (proper length) in the train frame (its rest frame). The same shrinking factor is given by the ratio D_A/D_B of the train-frame length D_A of the moving track between the track-frame clocks and the proper length D_B of that same stretch of track. Thus we have

$$\frac{L_B}{L_A} = \frac{D_A}{D_B} \quad (2.39)$$

What is the actual amount of slowing down and shrinking? We need two more facts:

1. the train-frame picture on the right half shows that

$$L_A = D_A \quad (2.40)$$

For this relation to hold, the train-frame synchronized clocks at the two ends of the train must both read the same time implying that the picture represents a single moment of train-frame time.

2. The track-frame pictures on the left imply a more complicated relation between L_B and D_B . According to the pictures, D_B is the track-frame distance between the left end of the train at track-frame time 0 and the right end of the train at track-frame time T_B . This distance is given by the track-frame length L_B if the train plus the distance the train moves between the two pictures (as shown in the figure) and since the train moves with speed v , the additional distance is vT_B so we have

$$D_B = L_B + vT_B \quad (2.41)$$

Now (6.2) and (6.3) tell us that

$$\frac{T_A}{T_B} = \frac{D_A}{D_B} = s \quad (2.42)$$

Then

$$T_A = sT_B \quad , \quad D_A = sD_B \quad (2.43)$$

Then (6.4) says that

$$L_B = sL_A \quad (2.44)$$

Combining (6.6) with (6.3) we have $D_B = L_B + v^2 D_B/c^2$ so that

$$L_B = D_B \left(1 - \frac{v^2}{c^2}\right) \quad (2.45)$$

But (6.9) says $L_B = sL_A$, (6.5) say $L_A = D_A$ and (6.8) says that $D_A = sD_B$. Putting all this together we get $L_B = s^2 D_B$ and thus (6.10) says that

$$s^2 D_B = D_B \left(1 - \frac{v^2}{c^2}\right) \quad (2.46)$$

so that the shrinking factor (or slowing down factor) s is

$$s = \sqrt{1 - \frac{v^2}{c^2}} = \frac{1}{\gamma} \quad (2.47)$$

We now see the first hint that $v < c$ since other wise the factor s would be *imaginary*! The *shrinking* effect is called the *Lorentz contraction*. As we will see later, there is no real physical contraction of anything - it has to do with the definition of length measurement. Similarly, the *slowing down* effect is called *time dilation*. Nothing, as we will see, is happen to time! Again, the effect has to do with the definition of time measurement and the definition of clock synchronization. Since we cannot tell which frame is the *moving* frame, we cannot, in the end, decides which frame lengths *contract* and which clocks *slow down*.

We cannot get rid of this problem because at a fundamental level Alice and Bob disagree on whether two clocks in different places are synchronized or whether two events in different places happen at the same time. Alice and Bob both maintain that the other has determined the rate of moving clocks or the length of moving sticks incorrectly.

To measure the length of a moving stick, one must determine where the two ends are at the same time. If the measurement times are not simultaneous, then the stick movers between measurements and the length will be incorrect. We must therefore be able to judge if spatially separated events are simultaneous. Similarly, to compare how fast a moving clock is running with the rate of a clock at rest, one must be able to compare at least two of the readings of the moving clock with the readings of the stationary clocks that are next to it when it shows those readings. However, a moving clock moves! Therefore we must be able to use two synchronized clocks at rest that are in different places.

Both Alice and Bob use identical procedures in their respective measurement and they disagree about the the results. In the end, as we will see, this is because they disagree about simultaneity. This is not just a matter of convention or

definition, but there are, as we will see, real observable consequences.

Simple examples of such a real manifestations are experiments with unstable elementary particles, in particular, mu-mesons. If these particles are created at rest in the lab, then they typically live for a time τ before they decay into other particles. This represents a *clock with only one tick*. These particles can also be created in the lab moving with any desired speed u . If u is close to the speed of light c , then the particles live for a much longer time than τ , i.e., they are observed to travel a distance much larger than $u\tau$ before decaying. The relativity explanation for this experimental result is that their *internal clocks* that govern when they decay are running much slower in the frame where they are moving with speeds close to c . This is a real effect and allows physicists to build large particle accelerators.

The various explanations for these results involve the use of length contractions and/or time dilations. For example, in the frame that moves with the particles (the frame that is moving in the laboratory), the track is moving with speed u and all distances along the track are reduced by the shrinking factor. Thus, much more of the track can go past the particles in time τ than if the particle was at rest. Both frames agree that the particles are able to cover a greater length of the track $u\tau/s$. In the track frame this is because the particle lives for a time τ/s which is much longer than τ , the time it would survive if at rest. In the rest frame of the particles it is because the length of the track has shrunk by a factor s , so that the length of the moving track that passes the particle in time τ is augmented by the factor $1/s$. Thus, the stories differ, but the results are the same!

We note that when we only describe things that happen both in the same place *and* at the same time - space-time coincidences - both Alice and Bob's picture agree. This shown below where we now emphasize the space-time coincidences. All frames agree in their descriptions of space-time coincidences. Differences of description only arise when we try to describe what is happening everywhere at a given time. This is so because *at a given time* means different things in different frames. We will learn later how to translate the different stories between frames and see that they are really telling the same story.

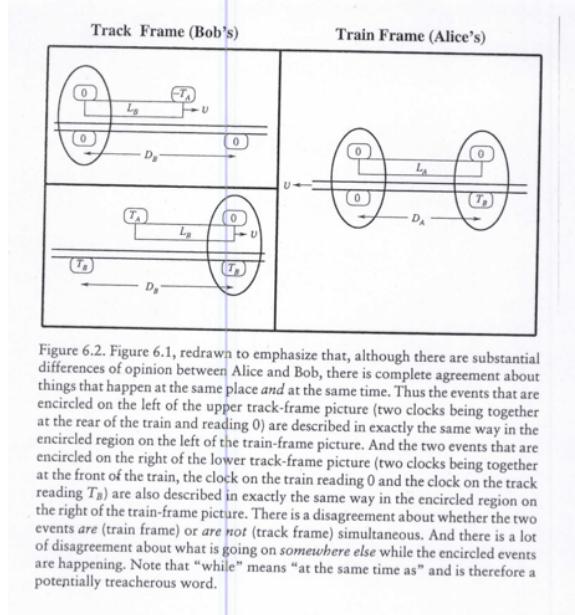


Figure 6.2. Figure 6.1, redrawn to emphasize that, although there are substantial differences of opinion between Alice and Bob, there is complete agreement about things that happen at the same place *and* at the same time. Thus the events that are encircled on the left of the upper track-frame picture (two clocks being together at the rear of the train and reading 0) are described in exactly the same way in the encircled region on the left of the train-frame picture. And the two events that are encircled on the right of the lower track-frame picture (two clocks being together at the front of the train, the clock on the train reading 0 and the clock on the track reading T_B) are also described in exactly the same way in the encircled region on the right of the train-frame picture. There is a disagreement about whether the two events *are* (train frame) or *are not* (track frame) simultaneous. And there is a lot of disagreement about what is going on *somewhere else* while the encircled events are happening. Note that "while" means "at the same time as" and is therefore a potentially treacherous word.

Now the slowing-down factor s for moving clocks must be the same as the shrinking factor s for moving sticks, which makes no use of the Dv/c^2 rule for simultaneous events. If they were not the same, then the behavior of a moving particle would be different, depending on whether you calculated it in the rest frame of the particle or the rest frame of the laboratory. To see this consider the figure below.

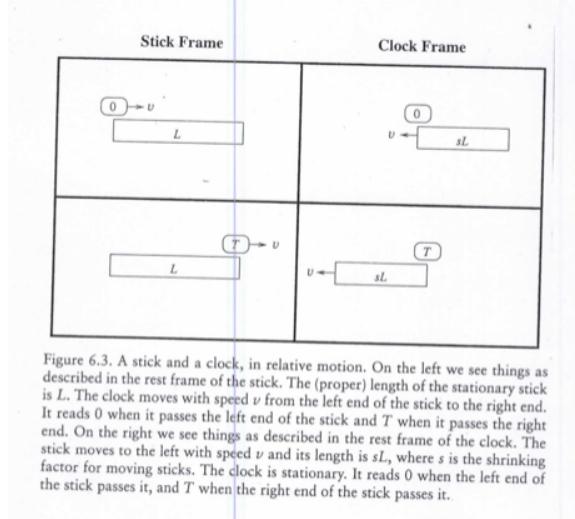


Figure 6.3. A stick and a clock, in relative motion. On the left we see things as described in the rest frame of the stick. The (proper) length of the stationary stick is L . The clock moves with speed v from the left end of the stick to the right end. It reads 0 when it passes the left end of the stick and T when it passes the right end. On the right we see things as described in the rest frame of the clock. The stick moves to the left with speed v and its length is sL , where s is the shrinking factor for moving sticks. The clock is stationary. It reads 0 when the left end of the stick passes it, and T when the right end of the stick passes it.

Here we have a stick of proper length L along which a clock moves to the right with speed v . Let the clock read 0 when it is at the left end of the stick and T as it passes the right end. See left side of above figure. If s is the shrinking factor for moving sticks, then in the clock frame we have a stick of length sL moving with speed v to the left. The time it takes the full length of the stick to get past the clock is just the time it takes the stick, moving at speed v to go its full length sL : $T = sL/v$. Since the clock tells correct time and reads 0 when the left end of the stick passes it, it must read $T = sL/v$ when the right end passes it.

In the stick frame, however, the time it takes the clock to go from the left to the right end of the stick is L/v , the proper length of the stick divided by the speed of the clock. Since the clock runs slowly in the stick frame, during this time the reading of the clock advances only by $T = s'L/v$, where s' is the slowing-down factor of moving clocks and this is what it reads as it passes the right end of the stick.

Both frames must agree on the time T since that is a judgment about a space-time coincidence. Thus, we must have $s = s'$.

Summarizing, we have these basic relativistic facts about clocks and measuring sticks.

RULE FOR SYNCHRONIZED CLOCKS

If two clocks are stationary, synchronized, and separated by a distance D in one frame of reference, then in a second frame, in which they are moving with speed v along a line joining them, the clock in front lags the clock in the rear by

$$T = \frac{Dv}{c^2} \quad (2.48)$$

RULE FOR SHRINKING OF MOVING STICKS OR SLOWING DOWN OF MOVING CLOCKS

The shrinking (or slowing-down) factor s associated with speed v is given by

$$s = \sqrt{1 - \frac{v^2}{c^2}} \quad (2.49)$$

2.7 Looking at a Moving Clock

We have determined the slowing-down factor s for moving clocks compared to clocks at rest. If you actually *look* at a moving clock would you actually *see* it running slowly?

It turns out to depend on whether the clock is moving towards (actually runs

faster) you or away from you (actually runs slower than the slowing-down factor would indicate). These observed disparities result from the fact that you do not see a clock reading a particular number until the light that leaves the clock when it displays that number has actually traveled from the clock to your eyes. If the clock is at rest in your frame, the delay between the clock displaying a number and you seeing the display does not matter because the extra time delay is the same for each number displayed in this case. Even with the delays, you are still receiving the flashes at the same rate that the clock is emitting them and so you think that the clock is running at its actual rate (even though each flash you see is delayed).

If the clock is moving away from you, the light from each successive flash has further to go before you see it, so you see the clock running more slowly than its actual rate in your frame. If the clock is moving towards you, the light from each successive flash has less distance to cover, so you see the clock running faster than its actual rate in your frame. This effect turns out to be larger than the slowing-down effect due to the fact that the clock is moving and hence you *see* it running fast.

Let us derive some quantitative results for this so-called *relativistic Doppler effect*. Along the way we will also get a second derivation of the value of the slowing-down factor $s = \sqrt{1 - \frac{v^2}{c^2}}$ and the relativistic velocity addition law. We also will not need to deal with synchronized clock problems.

Now we assume a clock that flashes a new number every T seconds in its rest frame. Let $f_a T$ and $f_t T$ be the number of seconds in your frame between the flashes that reach you when the clock moves toward (t) or away (a) from you with speed v . Since the moving clock runs slowly it only flashes a new number every T/s seconds. During that time it gets a distance $v(T/s)$ further from (or closer to) you, so that the light from each successive flash takes a time $v(T/s)/c$ more (or less) to get to you. Thus, the time between light from the flashes reaching you (and therefore the time between your *seeing* successive flashes) is

$$f_a T = \frac{T}{s} + \frac{v T}{c s} = \frac{T}{s} \left(1 + \frac{v}{c}\right) \quad (\text{moving away}) \quad (2.50)$$

$$f_t T = \frac{T}{s} - \frac{v T}{c s} = \frac{T}{s} \left(1 - \frac{v}{c}\right) \quad (\text{moving towards}) \quad (2.51)$$

Thus we have

$$f_a = \frac{1}{s} \left(1 + \frac{v}{c}\right) \quad (2.52)$$

$$f_t = \frac{1}{s} \left(1 - \frac{v}{c}\right) \quad (2.53)$$

Now consider the following idea. Suppose that Alice and Bob are stationary in the *same* frame of reference at different places and Bob holds a clock and Alice

watches. Suppose that Bob's clock flashes every t seconds in its proper frame. In this case every flash takes the same time to reach Alice. Since Alice's clock runs at the same rate as Bob's, Alice sees a flash every t seconds according to her own clock. Now suppose that Carol moves from Bob to Alice at speed v . Each time that Carol sees a new number appear on Bob's clock, she reinforces it with a flash of her own. Since Carol is moving away from Bob she sees a flash from Bob's clock every $f_a t$ seconds. She therefore sends out her own flashes every $T = f_a t$ seconds. Since Carol is moving towards Alice at speed v , Alice sees Carol's flashes arriving every $f_t T = f_t f_a t$ seconds. But since Carol's flashes arrive together with Bob's flashes every t seconds, Alice must also see Carol's flashes every t seconds. So the effects of Carol seeing Bob's clock flash slowly and Alice seeing Carol's clock flash fast must precisely cancel or

$$f_t f_a = 1 \quad (2.54)$$

Combining (7.5), (7.4) and (7.3) we have

$$1 = f_t f_a = \frac{1}{s} \left(1 - \frac{v}{c}\right) \frac{1}{s} \left(1 + \frac{v}{c}\right) = \frac{1}{s^2} \left(1 - \frac{v^2}{c^2}\right) \quad (2.55)$$

(7.3) and (7.4) also tell us that

$$\frac{f_t}{f_a} = \frac{1 - \frac{v}{c}}{1 + \frac{v}{c}} \quad (2.56)$$

Finally we get

$$f_t = \sqrt{\frac{1 - \frac{v}{c}}{1 + \frac{v}{c}}} \quad (2.57)$$

$$f_a = \sqrt{\frac{1 + \frac{v}{c}}{1 - \frac{v}{c}}} \quad (2.58)$$

which corresponds to the relativistic Doppler effect. If the speed $v = 3c/5$, we have the slowing-down factor $s = 4/5$ so that a clock moving at 60% of the speed of light takes $5/4 = 1.25$ seconds to flash each second - it runs at $4/5 = 80\%$ of its normal rate. However, we also have

$$f_t = \sqrt{\frac{1 - \frac{v}{c}}{1 + \frac{v}{c}}} = \frac{1}{2} = \frac{1}{f_a}$$

so if a clock is moving toward you with $v = 3c/5$ you see it flash a new second every half second - you see it running at twice its normal rate; if it moves away you see it flash a new second every 2 seconds - you see it running at half its normal rate.

The relativistic Doppler effect will be the starting point for the development of the theory in the Boccio approach.

2.8 The Interval between Events

We have now identified many things that observers in different frames disagree about: the rate of a clock, the length of a stick, simultaneity, clock synchronization, etc. Observers do agree on space-time coincidences (whether two events occur at the same time and the same place), the speed of light, and the number of events that have occurred. A quantity like the speed of light is an example of an *invariant*. - a quantity that all observers agree on. We can get an idea of new invariants that might exist by first looking more closely at the statement of the constancy of the velocity of light and making a more abstract version of the statement.

Consider two events E_1 and E_2 . Different reference frames do not agree on the particular numbers (position, time) associated with these events. Let D and T be the distance and time between the events in one frame. If the two events happen to be events in the history of a single photon with speed c , then $D/T = c$. Since the speed of light is an invariant, we must have $D'/T' = c$ where D' and T' are the distance and time between the events in another frame. D' does not need to be the same as D and T' does not need to be the same as T . We can then give an alternate statement of the constancy of the velocity of light.

If the time T and distance D between two events are related by $D = cT$ in one frame, then they will be related in the same way in any other frame. Remember that both T and D can be negative (just convention). Therefore, no matter whether they are positive or negative numbers, the alternative statement is equivalent to $(cT)^2 = D^2$ in all frames. Equivalently we can say that the time and distance between two events must satisfy

$$c^2T^2 - D^2 = 0 \quad (2.59)$$

in all frames. Two events that are separated by a time and distance satisfying (8.1) are said to be *lightlike separated*. The two events are connected by a single photon. The constancy of the velocity of light is equivalent to the statement *if two events are light-like separated in one frame, they will be light-like separated in all frames*.

This last statement is a special case of a more general rule. We will now show that if T is the time and D is the distance between *any* two events E_1 and E_2 in a particular frame, then even when $c^2T^2 - D^2$ is not zero, its value is still the same in all other frames, although T and D separately vary from one frame to another. This is called the *invariance of the interval*:

For any pair of events a time T and a distance D apart, the value of $c^2T^2 - D^2$ does not depend on the frame of reference in which T and D are specified.

Proof: Assume T and D are positive. We consider two separate cases: $c^2T^2 - D^2 > 0$ and $c^2T^2 - D^2 < 0$, i.e., either $cT > D$ or $cT < D$.

First, assume that $cT > D$. The $D/T < c$. This means that there exists a frame, moving from the earlier event to the later event at a speed

$$v = \frac{D}{T} \quad (2.60)$$

in which both events happen in the same place. Let T_0 be the time between the two events in this special frame. A clock that is present at both events is stationary in the special frame and therefore records a time T_0 between the two events.

This is illustrated in the figure below.

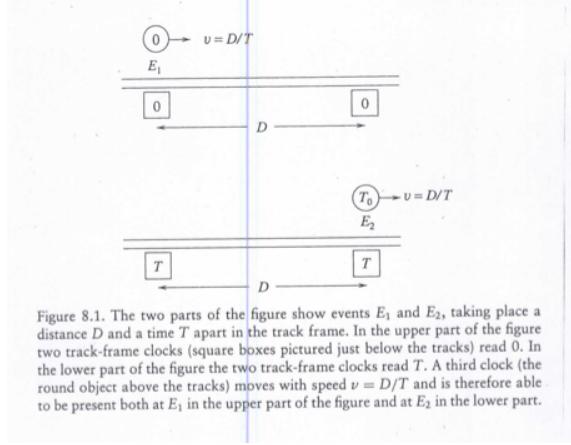


Figure 8.1. The two parts of the figure show events E_1 and E_2 , taking place a distance D and a time T apart in the track frame. In the upper part of the figure two track-frame clocks (square boxes pictured just below the tracks) read 0. In the lower part of the figure the two track-frame clocks read T . A third clock (the round object above the tracks) moves with speed $v = D/T$ and is therefore able to be present both at E_1 in the upper part of the figure and at E_2 in the lower part.

according to the original frame where the events are separated in space and time by D and T . Because the clock moves with speed v in the original frame, the amount T_0 it advances between the events is reduced from time T between the events by

$$T_0 = sT = T\sqrt{1 - \frac{v^2}{c^2}} \quad (2.61)$$

Since $v = D/T$ we get

$$T_0^2 = T^2 - \frac{D^2}{c^2} \quad (2.62)$$

So when the time T and distance D between two events are related by $T > D/c$, then $T^2 - D^2/c^2$ is independent of the frame in which D and T are evaluated. As in (8.4) it is equal to the square of the time T_0 between the two events in the frame in which they happen to be at the same place(at rest).

Now let us consider the other case, $cT < D$. Now $D/T > c$ so no material object ($v < c$) can be present at both events. However, there is now a frame, moving with $v < c$, in which the two events appear to be simultaneous. We can see why as follows. Consider two clocks that are at rest and synchronized in the frame where the two events are separated in space and time by D and T , with one clock present at each event as shown in the figure below.

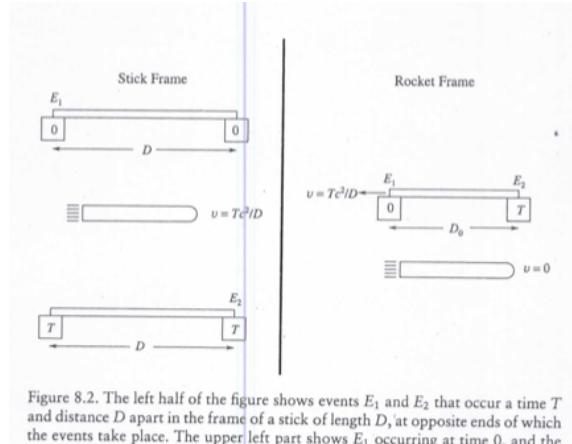


Figure 8.2. The left half of the figure shows events E_1 and E_2 that occur a time T and distance D apart in the frame of a stick of length D , at opposite ends of which the events take place. The upper left part shows E_1 occurring at time 0, and the lower left part shows E_2 occurring at time T , all times being indicated by clocks attached to the two ends of the stick and synchronized in the stick frame (square boxes pictured just below the stick). A rocket (the long object in the middle of the left half of the figure) moves to the right with speed $v = Tc^2/D = c(cT/D)$, which is less than c when $D > cT$. In the rocket frame (right half of the figure) the stick and attached clocks move to the left with speed v . In the rocket frame, when the clock on the left reads 0, the clock on the right reads $vD/c^2 = T$, so the events E_1 and E_2 are simultaneous. Therefore the distance D_0 between the events in the rocket frame is just the shrunken length sD of the moving stick.

If the earlier event clock reads 0, then the later event clock reads T . Since the distance between the clocks is D and they are at rest we can attach them to the ends of a stick of proper(rest) length D also at rest.

In a new frame, moving with speed v along the stick in direction from earlier to later event, the clock at earlier event is behind clock at later event by Dv/c^2 . Thus if we pick v so that $Dv/c^2 = T$, then the two events would be simultaneous in the new frame. We need to have

$$v = \left(\frac{cT}{D} \right) c < c \quad (2.63)$$

Thus, a frame exists in which the two events are simultaneous; the rocket frame in the above figure.

In the rocket frame the events are at opposite end of the stick (proper length D) and speed $v = c^2T/D$. Since events simultaneous in rocket frame, stick does not move during events so that distance D_0 between events is contracted length

$$D_0 = sD = D\sqrt{1 - \frac{v^2}{c^2}} \quad (2.64)$$

Using (8.5) we have

$$D_0^2 = D^2 - c^2T^2 \quad (2.65)$$

Thus, when T and D between two events are related by $D/c > T$, then $D^2 - c^2T^2$ is independent of frame where D and T evaluated (it is equal to square of distance in frame where events are simultaneous).

The two conclusions for $cT > D$ and $cT < D$ are same with space and time interchanged. Thus, we summarize by saying: If D is the distance and T is the time between events, then quantity $D^2 - c^2T^2$ is independent of frame where D and T measured. We distinguish three cases:

1. $D^2 - c^2T^2 > 0$. Events timelike-separated - exists a frame where events happen at same place (separation only in time T_0); $c^2T_0^2 = c^2T^2 - D^2$.
2. $D^2 - c^2T^2 < 0$. Events spacelike-separated - exists a frame where events happen at same time (separation only in space D_0); $c^2T_0^2 = c^2T^2 - D^2$.
3. $D^2 - c^2T^2 = 0$. Events lightlike-separated - events connected by a single photon(present at both events).

The quantity

$$I = \sqrt{|c^2T^2 - D^2|} \quad (2.66)$$

is called the *interval* between two events. If $c^2T^2 - D^2 > 0$, then I/c is time between events in frame where they occur at same place. If $c^2T^2 - D^2 < 0$, then I is distance between events in frame where they occur simultaneously.

2.9 Trains of Rockets

Let us now look at all of this another way to aid in our understanding. We now show how a disagreement about clock synchronization leads to all the relativistic effects we have found so far. We consider two frames from the point of view of a third frame(proper frame of a space station) where the other two are moving in opposite directions with the same speed. The two frames being observed from the third are proper frames for two trains of rockets - gray train moving to left and white train moving to right as shown in figure below.

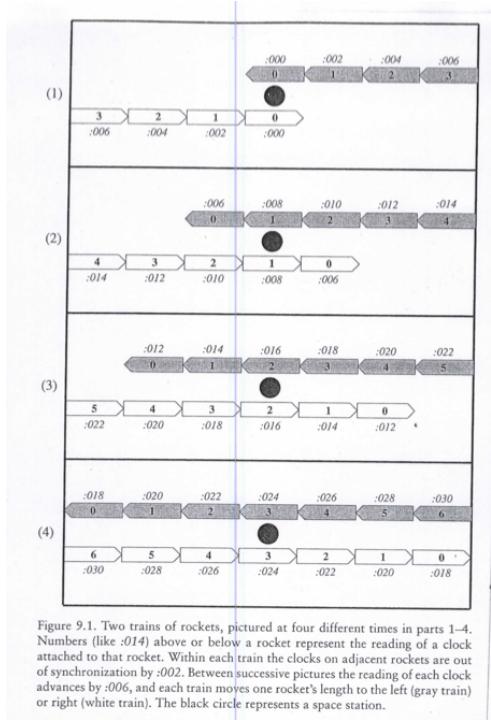


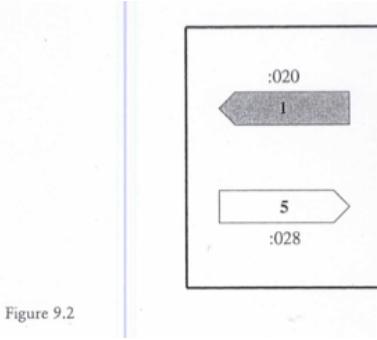
Figure 9.1. Two trains of rockets, pictured at four different times in parts 1–4. Numbers (like $:014$) above or below a rocket represent the reading of a clock attached to that rocket. Within each train the clocks on adjacent rockets are out of synchronization by $:002$. Between successive pictures the reading of each clock advances by $:006$, and each train moves one rocket's length to the left (gray train) or right (white train). The black circle represents a space station.

Space station = black circle. We have two trains of consecutively numbered rockets as observed at four different times. Each train moves one rocket length between observation. The numbers :006 etc near a rocket represent reading of clock carried by that rocket. Each clock is at the center of the rocket. This just reflects the fact that clocks synchronized in a train frame are out of synchronization in the space station frame. We define a time :002 as one tick (successive clocks differ by one tick). Remember a clock in front is behind a clock in the rear by $T = Du/c^2$, where D is the distance between clocks in their rest frame and u is the speed of the train in the space station frame. If we define the unit of length a a rocket length, we then have

$$\frac{u}{c^2} = 2 \text{ ticks per rocket} \quad (2.67)$$

One assumption we might make is to assume all the data presented are genuine relativistic effects due to large relative speeds and all clocks are precise enough to show the effects. Alternatively, we can be conspiracy theorists and say that the speeds involved are not very large, the clocks are not very precise and have been deliberately set out of synchronization by space station observers. The space station observers are interested in what kind of conclusion will be drawn by train observers using unsynchronized clocks if they do not realize they are not synchronized. So we assume that the space station people have given the rocket observer out of synchronization clocks as indicated, but tell the rocket observers they are synchronized. No communication is allowed between different rockets in a train. What happens in this case?

The trains are set in motion. People from either train can only collect information in their immediate vicinity, i.e., in part (1) of the figure white and gray rockets 0 are directly opposite and in part (3) gray rocket 1 is directly opposite white rocket 3 - the occupants of either rocket in both cases can note their own clock reading and the reading of the rocket opposite. As shown below in the segment of part (4), this is the information the two rockets can know.



White train says that at white time 28 ticks gray rocket 1 was opposite white rocket 5 and its clock read 20 ticks; Gray train says that at gray time 20 ticks gray rocket 5 was opposite white rocket 1 and its clock read 28 ticks. Each train

thinks they are telling the correct time and the other is not.

Observers from each train now get together separately and draw conclusions based on the observed data and the assumption that the clocks are synchronized.

Consider the figure below where gray rocket 0 appears in both pictures.

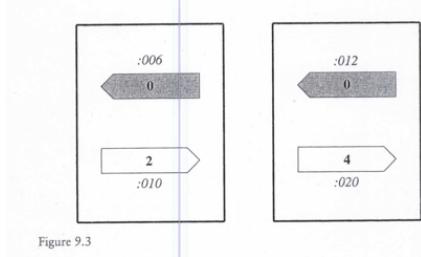


Figure 9.3

On the white train we get this interpretation: The velocity of the gray rocket is opposite to that of the white rocket. The gray rocket 0 took 10 ticks to go 2 rockets or its speed is $1/5$ rockets per tick. The white clock advanced 10 ticks while the gray clock advance 6 ticks. The gray clock is running slowly by a factor of $3/5$. These conclusions require that the white clocks were synchronized (the white observers are using the readings of two different clocks (2 and 4) to draw conclusions about event times.

Since the picture (fig 9.1) is symmetric between gray and white, the gray observers must draw the same conclusions as long as they assume their clocks are synchronized. Thus, each set of observers, thinking their clocks are synchronized believes the other clocks are running slowly!. The space station observers, however, believe that both sets of clocks are running at the same rate and neither set is synchronized.

We now have the speed $v = 1/5$ rocket per tick and slowing down factor $s = 3/5$ either set of observers assign to the other set. However $s = 3/5 = \sqrt{1 - v^2/c^2}$ corresponds to $v/c = 4/5$ or the speed of light is $1/4$ rocket per tick here.

We now look at any pair of pictures that were taken at the same time according to one of the trains. Consider the two pictures taken at gray time of 20 ticks, extracted from parts (3) and (4) of figure 9.1 and shown below.

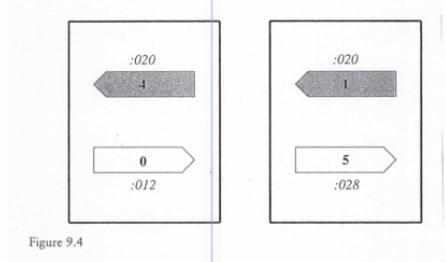


Figure 9.4

Since they are taken at the same (gray) time, the gray observers conclude that the white clocks are not synchronized. They are out of synchronization by $16/5 = 3.2$ ticks per rocket, which is seeming different than the exactly 2 ticks per rocket shown in figure 9.1 - however, remember that picture is as seen in the space station frame where both sets of clocks are out of synchronization and unreliable! The gray observers can also conclude from the last figure that at a single moment of gray time - 20 ticks - 5 white rockets = 3 gray rockets in length. White rockets have shrunk in length by the same factor $3/5$ as the white clocks are running slower! And so on... Mermin goes on with more examples to show that relativistic velocity addition works, if superluminal speed exist, the strange thing happen with time order can causal behavior, the invariance of the interval and clock synchronization actually fails between moving frames.

2.10 Space-Time Geometry

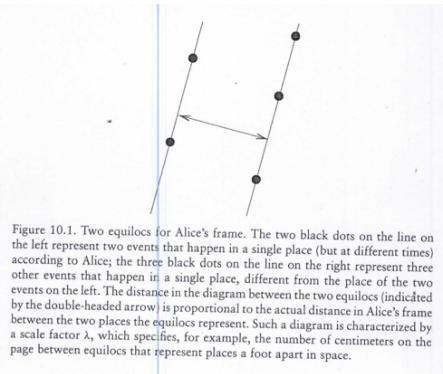
We now develop a more abstract and more powerful generalization of the pictures we have been using to describe frames in relative motion. We will be developing so-called space-time or Minkowski diagrams. They will enable us to see clearly, without most of the complications of the type of diagrams we have bee using up to now, all of the relativistic effects in the new theory. For simplicity we deal with only one spatial dimension - all events are assumed to take place along a single straight track. Not much is gained by the added complications associated with using more than one spatial dimension.

We start with one frame (Alice's) and specify how Alice represents events on the diagram. Everything, at this point, is according to Alice. When the other observer(Bob) in a moving frame appears we will have to more careful about who is saying what.

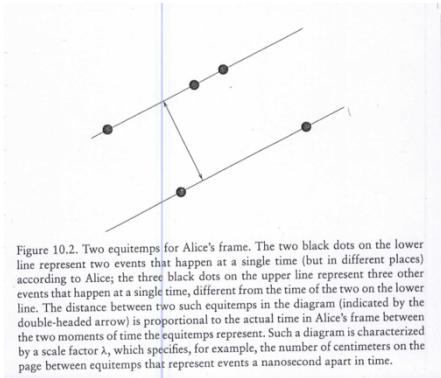
Alice represents an event by a point in her diagram. Events that are coincident (same place at same time) are represented by the same point. Distinct point represent events that have different places, different times or different places and times.

Alice represents several events that happen at the same place by a single straight

line as shown below.



The line is called an *equiloc* or line of constant position. Alice can choose to orient the line in any direction she chooses. Two equilocs representing various events that happen in different places must be parallel; If not, then they intersect somewhere and that point would correspond to a single event that happened in two different places, which makes no sense. We assume the existence of a scale factor λ between actual distances in space and distances on the diagram (specific to observers). In a similar manner Alice defines *equitemps* or lines of constant time or lines of simultaneity corresponding to events all occurring simultaneously as shown below.

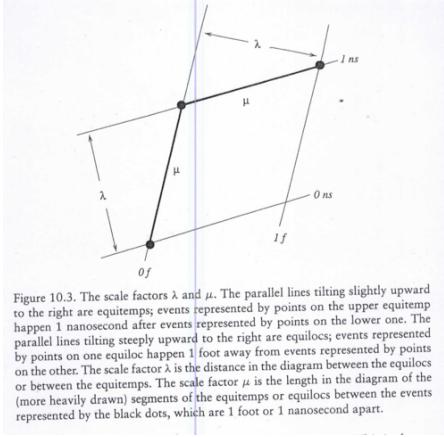


Equitemps make some angle with respect to equilocs and all equitemps are parallel (otherwise would intersect and we would have a single event happening at two different times). Again there could be a proportionality factor between the diagram and real time intervals. Any equiloc intersects any equitimp in only one point (angle between cannot be zero!), which represents those events that happen precisely at that time and in that place.

We now make a convenient choice of orientations and scales as follows. Using the fact that the speed of light is about 1 foot per nanosecond, Alice chooses the separation between two equitemps to be 1 nanosecond. This means that the scale factor λ for equilocs (centimeters of diagram per foot) is numerically the same as the scale factor λ for equitemps (centimeters of diagram per nanosecond). Another convenient scale factor, the distance μ along any equiloc is associated with any two events 1 nanosecond apart. It is exactly the same as the distance along any equitimp associated with two events 1 foot apart in space as shown in the figure below.

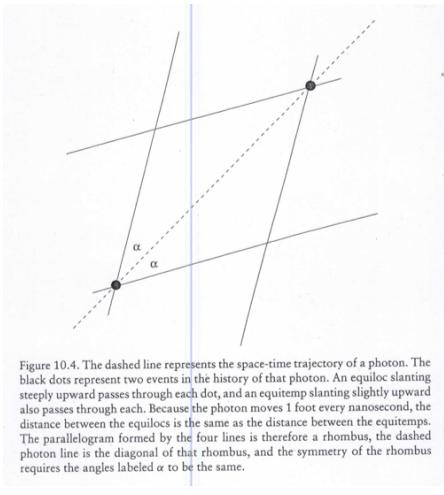
Geometrically, because the separation is the same for equilocs and equitemps, λ as shown, the parallelogram so defined has 4 equal length sides (a rhombus), μ in this case. We note that the diagonals bisect the vertex angles.

Normal $\mu > \lambda$ unless the equitemps are perpendicular to the equilocs, then $\mu = \lambda$ and we have a square.



The set of all events a point object is present at is a continuous line on the diagram - it represents the history of the object (past, present and future) = worldline or space-time trajectory of the object. For example, object at rest in Alice's frame = equiloc for that place, a moving object = straight line not parallel to any equiloc. A wiggly line indicates accelerations.

Different objects moving uniformly with the same velocity are parallel lines. Because of our choice of scales, the worldline of a photon is particularly simple. Any two events on a photon worldline must be as many feet apart in space as they are nanoseconds apart in time = diagonal of rhombus as shown below.



The worldlines of two photons moving in opposite directions are orthogonal on the diagram as shown below.

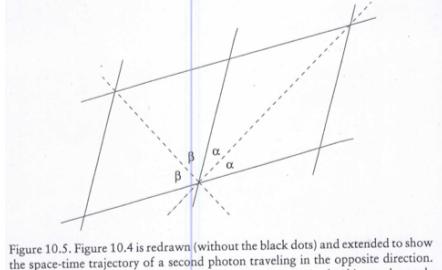


Figure 10.5. Figure 10.4 is redrawn (without the black dots) and extended to show the space-time trajectory of a second photon traveling in the opposite direction. Because the new dashed line is also a photon trajectory it also bisects the angle between the equitemps and equilocs. Since $2\alpha + 2\beta = 180$ degrees, the angle $\alpha + \beta$ between the two photon lines is 90 degrees.

We choose a convention where the photon worldlines make angles of 45° with the vertical as shown above. This means that equilocs are always more vertical than horizontal while equitemps are always more horizontal than vertical. Only a couple of choices remain.

1. Alice is free to choose the numerical values of the scale factor λ .
2. Alice can choose the angle her equilocs make with the photon worldlines.

These two choices fix the other scale factor μ .

All is straightforward and non-controversial as long as we consider only one observer - Alice in the case above. What happens if we add another observer, Bob, to the mix, who is moving uniformly along the track with speed v relative to Alice and Bob wants to describe the same events as Alice, except he wants to use the frame where he is at rest. Bob uses the same set of events and worldlines as Alice since these are intrinsic to space-time and thus independent of observer. He must, however, choose his own set of equilocs and equitemps to describe in own spatiotemporal language what is happening. Remember he will disagree with Alice about general notions like *same place* and *same time*, etc. Let us figure out what Bob says.

If Bob's frame is moving with speed v relative to Alice, then Bob's equilocs must be parallel to a worldline that Alice says is for an object moving with that speed. Thus, Bob's equilocs are parallel straight lines that are not parallel to Alice's equilocs. The faster Bob moves relative to Alice, the more his equilocs tilt away from Alice's. Alice's equilocs and equitemps through any two points on one of Bob's equilocs define a parallelogram, the ratio of whose sides (or the ratio of the distances between the side) is just the velocity v of his frame with respect to hers in feet per nanosecond as shown below.

We now use the principle of relativity or the constancy of the speed of light to determine the orientation of Bob's equitemps. We consider the procedure used earlier to that determines the simultaneity of events taking place at two ends of

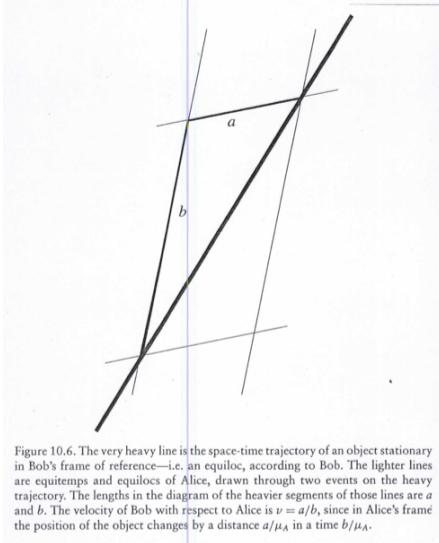


Figure 10.6. The very heavy line is the space-time trajectory of an object stationary in Bob's frame of reference—i.e. an equiloc, according to Bob. The lighter lines are equitemps and equilocs of Alice, drawn through two events on the heavy trajectory. The lengths in the diagram of the heavier segments of those lines are a and b . The velocity of Bob with respect to Alice is $v = a/b$, since in Alice's frame the position of the object changes by a distance a/μ_A in a time b/μ_A .

a train that is at rest in Bob's frame.

Since the train is at rest in Bob's frame, its left end, middle and right end are represented in Alice's diagram by parallel equilocs of Bob. Bob and Alice agree on the point in the train that corresponds to the middle so the three equilocs are equally spaced in Alice's diagram. Now two photons emitted at the middle travel towards the ends. According to Bob, the two photons (having the same speed) arrive at the ends simultaneously. On Alice's diagram, the intersections of the two photon worldlines with Bob's equilocs for the ends are events which are simultaneous according to Bob. This is illustrated below - part (1).

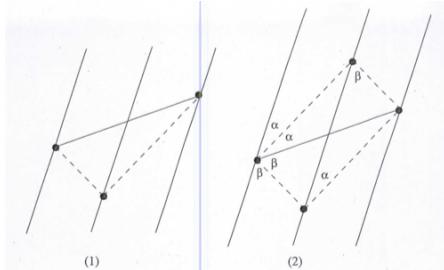


Figure 10.7. The diagram is drawn by Alice. (1) The three equally spaced parallel lines are the two ends and the middle of a train that is stationary in Bob's frame of reference. They establish the direction Bob's equilocs must have in Alice's diagram. The lowest black dot represents the production of two oppositely directed photons at the middle of the train. The dashed lines are the space-time trajectories of the photons. The other two black dots represent the arrival of each photon at an end of the train. Since both photons move at the same speed in Bob's frame of reference and since the train is stationary in Bob's frame, the photons arrive at the ends of the train at the same time in Bob's frame—i.e., the straight line joining the upper two dots is an equitemp for Bob. (2) If the photons are reflected back toward the center of the train when they reach the two ends, they will arrive there at the same time in the event represented by the highest black dot, all four photon lines forming a rectangle. It is evident from the symmetry of the rectangle that the two angles inside the rectangle labeled α are equal, as are the two angles inside the rectangle labeled β . Since the two labeled angles outside the rectangle are just spatial translations of two correspondingly labeled angles within it, it follows that both of the photon trajectories passing through the leftmost black dot bisect the angles between Bob's equitemps and equilocs passing through that dot.

Bob's equitemps and equilocs must make the same angle with the photon worldline (as did Alice's equitemps and equilocs - the principle of relativity). This is most easily seen by letting the photons reflect off the ends and return to the center as shown in part (2) above. Both Alice's and Bob's equitemps and equilocs are symmetrically arranged around photon worldlines as required by the principle of relativity. If we were presented a diagram with both sets of equitemps and equilocs drawn, we could not tell which were drawn first! The fact that both sets are symmetric (45°) has an immediate consequence. As shown below

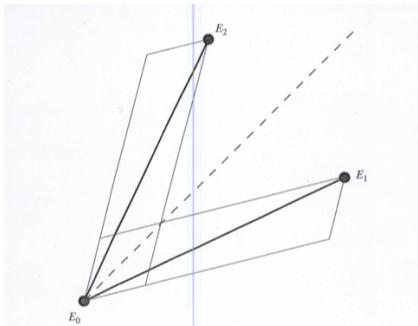


Figure 10.8. The heavy line on which events E_1 and E_2 lie is an equitemp for Bob, and the heavy line on which events E_1 and E_0 lie is an equiloc. The two long, thin parallelograms that overlap in the lower left are made up of segments of equitemps and equilocs for Alice. The entire figure is symmetric when mirrored in the dashed photon line. If Bob moves with speed v in Alice's frame, then since E_2 and E_0 are in the same place according to Bob, according to Alice their separation d in feet is v times their separation t in nanoseconds: $d = vt$. The ratio of d to t is just the ratio of the lengths of the short and long sides of the parallelogram with E_2 and E_0 at opposite vertices. Because of the mirror symmetry in the dashed photon line, v is also the ratio of the short and long sides of the parallelogram with E_1 and E_0 at opposite vertices. But the ratio of the sides of this parallelogram gives the ratio of Alice's time T in nanoseconds separating Bob's simultaneous events E_1 and E_0 to Alice's distance D in feet separating E_1 and E_0 . So $T = \nu D$.

we have the $T = Dv/c^2$ rule for simultaneous events in the form $T = Dv$ that the rule assumes when one measures times in nanoseconds and distances in feet.

So all observers can superimpose their own equitemps and equilocs on the same diagram. Any set is symmetric around the photon worldline. Finally, since when Alice and Bob move away from each other at constant velocity, they must see each other's clock running at the same rate, as measured by their own clock. This will lead, as we will derive later, to a rule relating scale factors (see figure below), namely,

$$\lambda_A \mu_A = \lambda_B \mu_B \quad (2.68)$$

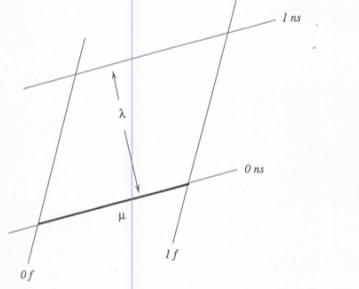


Figure 10.9. The unit rhombus for some frame of reference. The lines labeled 0 ns and 1 ns represent events 1 nanosecond apart and the lines labeled 0 f and 1 f represent events 1 foot apart. Because the distance in the diagram between the two equitemps, regarded as the height of the rhombus, is the scale factor λ , and because the heavier portion of the lower equiloc, regarded as the base of the rhombus, is the scale factor μ , the area of the rhombus—its base times its height—is just the product $\lambda\mu$.

The figure below shows diagrammatically how it is possible for each of two sticks in relative motion to be longer its proper(rest) frame than any other frame.

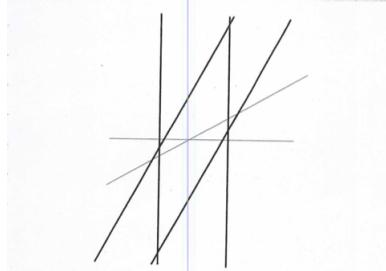


Figure 10.10. The two vertical lines are the left and right ends of a stick. The two parallel lines that tilt upward to the right are the left and right ends of a second stick that moves to the right past the first stick. The horizontal line is an equitemp in the frame in which the first stick is stationary. The line that tilts upward to the right is an equitemp in the frame in which the second stick is stationary. It tilts away from the horizontal by the same amount that the lines representing the ends of the second stick tilt away from the vertical.

The two vertical lines represent the worldlines of the ends of a stick. Since the ends are at rest in this frame these are equilocs. Equitemps in this frame must be horizontal. Any horizontal slice of the figure shows what things are like at that given moment of time in the frame of the stick - what events are simultaneous.

The two parallel lines that slant upward and to the right are the worldlines of the ends of a second stick. They are equilocs in the rest frame of the second stick. Equitemps in that frame make an angle so they are symmetric with respect to photon worldlines (45°). They show events that are simultaneous in the frame of the second stick. The horizontal line is a particular equitemp in the frame of the first stick. Along that line of simultaneity you first encounter the left end of the first stick, then the left end of the second stick, then the right end of the second stick and finally the right end of the first stick. Thus, in the rest frame of the first stick, the two ends of the first stick extend beyond the two ends of the second stick - the second stick is shorter than the first stick. On the other hand, the tilted line is a particular equitemp of the second stick and it is clear that the same conclusions follow from this line - in the rest frame of the second stick, the first stick is shorter - this is just the principle of relativity in action. The reason is that they disagree about simultaneity and length measurement require simultaneity.

Thus, while the worldlines of all parts of the stick (its world-surface) is intrinsic to space-time, the choice of how to slice those worldlines with equitemps is frame-dependent and thus, so is the *stick at a given moment*.

In a similar way, the figure below

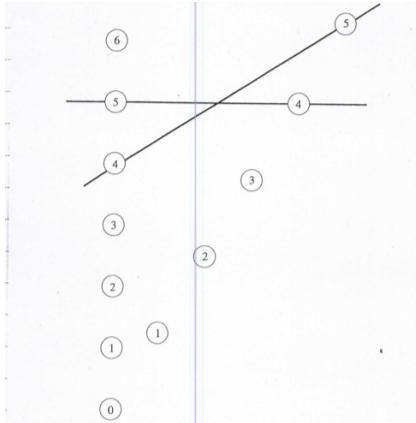


Figure 10.11. Several moments in the histories of two uniformly moving clocks. Each point that represents a clock has been expanded to a circle that displays the reading of that clock. Both clocks read 0 at the same place and time. That event is represented by just a single circle. Subsequent readings of the first clock (1–6) are shown on the set of circles uniformly spaced along a vertical line; subsequent readings of the second clock (1–5) are shown on the set of circles that lie on a line sloping upward to the right. The horizontal line is an equitemp in the frame of the first clock. The slanting line is an equitemp in the frame of the second clock, so it tilts away from the horizontal by the same amount that the line of pictures of the second clock tilts away from the vertical.

shows diagrammatically how it is possible for each of two clocks, in relative motion, to run faster than the other in its proper frame. The vertical row of numbered circles represents seven moments in the history of a clock and the clock readings at those moments. The slanting row represents six moments in the history of a second clock moving to the right relative to the first and its readings at those

moments. Both clocks are coincident when they read 0. Everybody, independent of frame, agrees the clocks read 0 at the same time - because of coincidence. Equitemps for the first clock are horizontal. Since the events for which the first clock reads 5 and the second clock reads 4 are on a line of simultaneity of the first clock, they happen at the same time in that frame - thus, the second clock is running at $4/5$ the rate of the first according to the rest frame of the first clock. However, an equitemp in the rest frame of the second clock is as shown in the figure and we therefore draw the same conclusion in the second clock rest frame - the first clock is running at $4/5$ the rate of the second - again that is the principle of relativity in action. Again the different definitions of what is simultaneous are what is the reason for the differences observed. As long as the two clocks are moving relative to each forever (not clock changes its frame), the which one is *actually* running slower is unanswerable.

Suppose, however, that the second clock suddenly reverse its direction of motion(accelerates) and returns to the first clock. We can then compare the clocks directly when they are once again coincident and see which has advanced by the larger amount. The process of turning around breaks the symmetry between the two clocks. The first clock is at rest in a single frame for its entire history. The second clock changes from one rest frame to another at the moment it turns around. There does not exist any single frame where the second clock remains at rest for its entire history.

As shown in the figure below

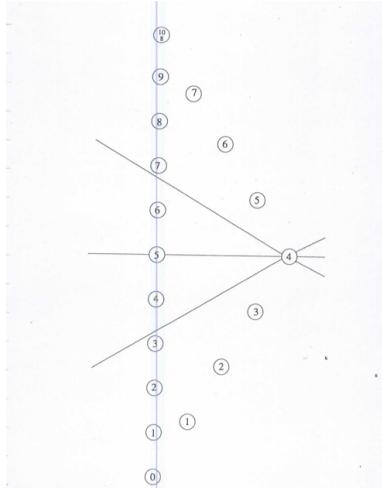


Figure 10.12. Two identical clocks. The first is shown at eleven different moments along a vertical line, as its reading advances from 0 to 10. The second moves uniformly away from the first as its reading advances from 0 to 4; it then moves uniformly back to the first, as its reading advances from 4 to 8. At the bottom and top of the figure, both clocks are at the same place at the same time and are represented by a single circle. The first clock is stationary in a single inertial frame of reference. Since equiloces are vertical in that frame, equitemps are horizontal. The horizontal line is an equitemp in the frame of the first clock. The line slanting upward to the right is an equitemp in the frame of the second clock as it moves away from the first. The line slanting upward to the left is an equitemp in the (different) frame of the second clock as it moves back to the first.

in the frame of the first clock (horizontal equitemps), it is clear that when the trip is over, the second clock, running slower for the entire journey, will have advanced only by 8 (4 out and 4 back) while the first clock has advanced by 10 when they are back together.

Things are trickier, however, from the viewpoint of the second clock. In the frame moving outward with the second clock, the first clock runs slowly (3.2 compared to 4) as can be seen from the lower of the two equitemps in the above figure. Similarly, the first clock runs slowly (3.2 compared to 4) as can be seen from the upper of the two equitemps in the above figure. It still makes sense that when they get back together, the first clock has advanced 10 while the second clock has only advance 8 because the missing time is accounted for by the (instantaneous)shift is the line of simultaneity when the frame shift occurs (see Mermin for calculational details). The abrupt nature of the shift in simultaneity can be made gradual by simply extending the time over which the reversal takes place with the same eventual conclusion!

We can bypass the artificial role played in the discussion by the simultaneity shift by asking a different question. Instead of asking about current clock readings, we ask what do the observers see the clocks doing? Consider the figure below, which is the same as the previous figure without the equitemps but including now photon worldlines emitted when each clock changes its reading.

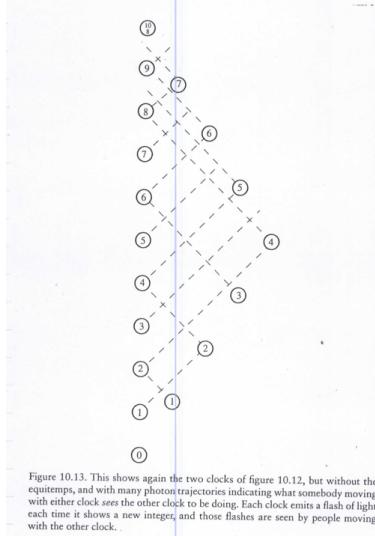


Figure 10.13. This shows again the two clocks of figure 10.12, but without the equitemps, and with many photon trajectories indicating what somebody moving with either clock sees the other clock to be doing. Each clock emits a flash of light each time it shows a new integer, and those flashes are seen by people moving with the other clock.

Since the slowing down factor is $4/5$, the relative velocity of the clocks is $v = 3c/5$ and therefore the Doppler factor is

$$\sqrt{\frac{1 + \frac{v}{c}}{1 - \frac{v}{c}}} = 2$$

People watching a clock moving away from them (or moving away from the clock) at 3/5 the speed of light will *see* it running at half its proper rate. Note how the factors 2 and 1/2 from the Doppler effect emerge naturally from the figure above. The first clock(no frame change) see a shift from slow rate to fast rate only after 8 of the 10 ticks while the second clock(frame change) sees the shift exactly at the half-way point of the trip. Thus, the observers are not identical (no longer symmetric) and the fact that one clock actually runs slower (clock 2) is not surprising anymore.

The fact that two identical clocks, initially in the same place and reading the same, can end up with different readings, if they move apart from each other and then back together = clock paradox or the twin paradox. There really is no paradox in our example since the two clocks moved asymmetrically. The twin paradox version refers to substituting identical twins for the clocks and having them age differently. Mermin now talks about the pole-in-the-barn paradox and faster than light signals, light-cones and invariance of the interval, which we will talk about during a later pass(Boccio notes) through this topic.

2.11 $E = Mc^2$

We will cover this material in the Boccio Relativity Notes.

2.12 A Bit About General Relativity

Let us look only at one aspect of general relativity(GR) that we can connect to our discussions (Geroch) about black holes. In Boccio Notes we saw how GR affects the interval or metric,i.e., in special relativity we have a space-time geometry(flat) given by

$$ds^2 = c^2 dt^2 - dx^2 - dy^2 - dz^2$$

while one solution of the Einstein equations in GR for the case of a point mass is a space-time geometry(curved) given by

$$ds^2 = c^2 \left(1 - \frac{2GM}{rc^2}\right) dt^2 - \frac{1}{\left(1 - \frac{2GM}{rc^2}\right)} dr^2 - r^2(d\theta^2 + \sin^2 \theta d\phi^2)$$

As in our earlier discussions, we can then write

$$d\tau^2 = \left(1 - \frac{2GM}{rc^2}\right) dt^2$$

in the rest frame of the particle ($dr = d\theta = d\phi = 0$). When we did this in special relativity we had the relation $d\tau = dt$ which we called the proper time. $d\tau$ is still the proper time, but now the relationship to dt , the time measure by an outside observer, is more complicated.

The relationship above says that gravity affects the rate at which clocks run. Remember in special relativity that relative motion affected the rate of clocks (the Doppler shift). This *gravitational time dilation* is an additional effect.

This effect leads to the following predictions. Suppose we have two identical clocks separated by a fixed distance D . Let the line directed from one clock to the other define the direction we call *vertical* so we call one clock the upper clock and the other the lower clock. Because of the gravitational time dilation effect the two clocks, one at r and the other at $r + D$ tick at different rates or have different frequencies when they emit signals separated by dt , i.e.,

$$d\tau_{upper}^2 = \left(1 - \frac{2GM}{(r+D)c^2}\right) dt^2$$

$$d\tau_{lower}^2 = \left(1 - \frac{2GM}{rc^2}\right) dt^2$$

The ratio of the frequencies is

$$\frac{f_U}{f_L} = \frac{\left(1 - \frac{2GM}{rc^2}\right)^{1/2}}{\left(1 - \frac{2GM}{(r+D)c^2}\right)^{1/2}}$$

If both clocks are stationary in a given frame and there is no gravity ($M = 0$), then $f_U = f_L$ as we expect from our special relativity discussions (they have no relative motion).

Let us consider the case where $r = R_{earth} \gg D$. We then have

$$f_L = f_L \left(1 + \frac{gD}{c^2}\right)$$

where $g = GM/R_{earth}^2$ = acceleration due to gravity at the earth's surface. This is a very small effect on the earth's surface but as we noted earlier in Boccio Notes, it has been successfully measured and confirmed.

However, note what happens at the horizon of a black hole. There $r = 2GM/c^2$ and thus the factor

$$\left(1 - \frac{2GM}{rc^2}\right) = \infty$$

The rate of signals shifts from the emission rate by an infinite factor! But this is just what we were discussing. Remember that as an observer approached the horizon, another distant observer in the external region, thought that the rate went to zero (divide by infinity). We now have come full circle and can see directly where some of the results we discussed come from mathematically. More than this would require a full blown course in GR with all the mathematics.

One further point we can make is about paths of motion in GR. In GR there

are no forces - there is only curvature due to matter distribution. All particles are free. In our old Newtonian manner of thinking free particles are those experiencing no forces. These particles travel in straight lines with constant speed. In special relativity they also travel on straight worldlines. Both of these cases correspond to the paths of motion being *geodesics* or the *straightest* line in the space or space-time geometry. In GR this still is the case but now the geodesic is not a straight line because of the existence of curvature, i.e., the geodesic on the surface of a 3-dimensional sphere is a great circle and so on.

The orbital motion of a planet around the sun is a geodesic in 4-dimensional space-time (not easy to visualize). The bending of light around the sun is because that path is the geodesic for light in the vicinity of the sun and so on.

Chapter 3

Notes on Geroch

General Relativity from A to B

3.1 Events in Space-Time: Basic Building Blocks

3.1.1 Events

Event = basic building block of theory.

Event = a (where, when) of physical occurrence or potential occurrence ; a point, i.e., hammer head hits nail.

Nail is not event - exists over time period(has extension in time).

Nail also has extension in space.

Two events are the same if coincide - same (where, when).

Are events real? What are events really like? Physics does not ask and certainly cannot answer such questions.

Physics only interested in relationships between events (even if events not understood at a fundamental level). This is what our theories will do!

3.1.2 1st Try - Everyday Experience = Aristotelian View

Event =(position in space, time of occurrence)

Explicitly,

- (1) set up Cartesian coordinate system (x,y,z)
 - (a) define origin and three orthogonal axes (x,y,z)

- (b) each position in space = three real numbers
 - (1) *value of x* = *distance* of event from origin along *x-direction*
 - (2) *value of y* = *distance* of event from origin along *y-direction*
 - (3) *value of z* = *distance* of event from origin along *z-direction*
- (2) At each (x,y,z) point (define a minimum separation between points based on accuracy that we can measure) place an observer (labeled with those (x,y,z) values). Observers remain fixed in space.
- (3) Each observer has clock (all synchronized).

How does this arrangement work? Let event #1 be the explosion of a firecracker. One of the fixed observers will be located in the immediate vicinity of the event. That observer records their (x,y,z) values and a fourth number the time on their clock(when the explosion happened). These four numbers then represent the characterization of this event in the so-called Aristotelian view.

This complicated an explicit procedure is necessary in this view. Does it accurately represent the structure of space and time? We will find that implicit in the procedure are assumptions about the way that space and time operate - these assumption will not turn out to be true! The explicit procedure is required in order for us to figure out what is wrong with the assumptions.

What relationships are implied by this Aristotelian view?

Given two fixed events, we ask whether the follow questions make any sense in the Aristotelian view.

- (1) Do the two events (x_0, y_0, z_0, t_0) and (x'_0, y'_0, z'_0, t'_0) have the same position in space? Make sense with assumed event data. Must have $x_0 = x'_0$, $y_0 = y'_0$, and $z_0 = z'_0$.
- (2) Do the two events occur at the same time? Make sense with assumed event data. Must have $t_0 = t'_0$.

Clearly, two events are the *same* if $x_0 = x'_0$, $y_0 = y'_0$, $z_0 = z'_0$, and $t_0 = t'_0$.

- (3) What is the distance between the two events? Make sense with assumed event data.

Need to calculate $(\text{distance})^2 = (x_0 - x'_0)^2 + (y_0 - y'_0)^2 + (z_0 - z'_0)^2$.

- (4) What is the elapsed time between the two events? Make sense with assumed event data.

Need to calculate $t_0 - t'_0$.

- (5) Is a particle at rest? Make sense with assumed event data. If position is unchanged for different times, then the answer is yes.

- (6) What distance did the particle travel between one time and some later time? Make sense with assumed event data.

Need to calculate $(\text{distance})_1^2 = (x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2$ with $t_0 < t_1$

and $(\text{distance})_2^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2$ with $t_1 < t_2$

and so on and add all the distances for the total distance traveled.

- (7) What is the speed of the particle? Make sense with assumed event data.
Need to calculate total distance traveled divided by total time interval.

All of these ideas make sense in the Aristotelian view and none of them will make sense in relativity theory!

Let M be the set of all possible events(past, present and future); M = space-time. Point in M represents an event. Region of M is a collection of events. Space-time is able to describe more complicated things than just events.

A particle is described by a line(collection of events representing its position for different times) - the line is called a *worldline*. The worldline completely describes everything about the coordinates(space and time) of the particle.

Thus, a particle is not a point from the viewpoint of space-time - it is a line.

Now consider two particles. Each is represented by a worldline. Suppose the two lines intersect at some point p as shown below

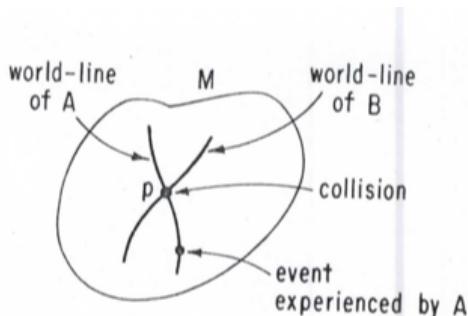


Fig. 2
The representation within space-time of the collision of two particles. The event "collision" is the intersection of their world-lines.

What does this mean physically? Point p is an event which lies on both world-lines - it is experienced by both particle A and particle B - this is a physical collision between the particles.

We want to discover a theory of relationships between events. How is space-time involved? What is its role? Events are points in space-time M . Thus, relationships between events are relationships between points in M . The set of such relationships is an internal structure imposed on M . We need to determine what kind of structures we can impose on space-time.

3.2 The Aristotelian View: A *Personalized* Framework

We have the Aristotelian view and Space-time as just described. The Aristotelian view is a particular way of looking at the relationships between events. Space-time is supposed to be a view-independent collection of events. We now ask the question:

What is the relationship between events according to the Aristotelian view as structure within space-time?

[12pt] If we can figure this out, then we should be able to use this *Aristotelianized space-time* to translate between geometrical constructs in space-time and things happening in the physical world.

The Aristotelian view gives a characterization of each event by four numbers (x,y,z,t). Space-time is the collection of all events.

We incorporate the Aristotelian view into space-time by introducing a coordinate system, which allows us to associate with each point (event in space-time) four real numbers, the values of (x,y,z,t).

What is important about the number of numbers required?

- (1) To locate a point on a line requires 1 number - the line is 1-dimensional.
- (2) To locate a point in a plane requires 2 numbers - the plane is 2-dimensional.
- (3) To locate a point in physical space requires 3 numbers - physical space is 3-dimensional.
- (4) To locate a point in space-time requires 4 numbers - space-time is 4-dimensional.

The figure below shows space-time (rectangular box) with a coordinate system added.

Note there is no z-axis since I cannot draw in 4-dimensions. It will turn out that no physical results are lost by only considering two spatial dimensions plus time. We now want to understand the physical significance of various geometrical objects in space-time.

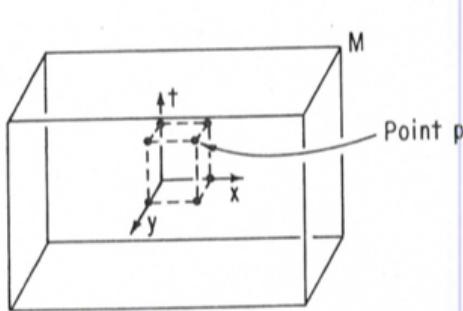


Fig. 3
The location of a point p in space-time by its Aristotelian coordinates.

3.2.1 Geometrical Objects in Space-time

First, consider the t-axis. It is a line in space-time. It represents the collection of events through which the line passes. All points on the t-axis have $(x=0, y=0, z=0)$, i.e., they are the coordinates of a single observer; thus, the t-axis is all events of this observer.

Now consider a different vertical line (parallel to the t-axis). For example, all the events that have $(x=3, y=7, z=-2)$ as shown in the figure below.

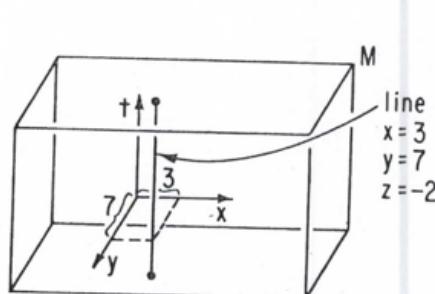


Fig. 4
The world-line of the individual whose badge reads $x = 3$, $y = 7$, $z = -2$ is the vertical straight line with these coordinates.

Again, these events all correspond to a single observer.

Thus, each observer has their own personal vertical worldline. The worldline represents the observer in space-time - it is the observer!. Now consider the plane $t = 0$, i.e., the plane containing all events with t value zero. The values of (x,y,z) on this plane are arbitrary so the plane is 3-dimensional; it is a *3-plane or 3-surface*. (Note that it appears as a 2-dimensional plane in our diagram because we are suppressing the z coordinate). The planes are horizontal planes as in the figure below.

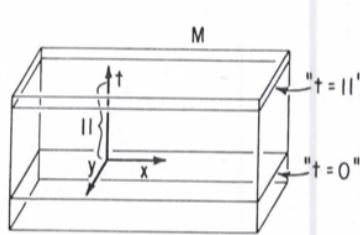


Fig. 5
The horizontal 3-plane given by $t = 11$ in space-time represents "all space at time $t = 11$."

They represent the collection of all events with the same time value. Such a plane corresponds to a *clock reading*. So a vertical line corresponds to a *position in space* and horizontal planes correspond to clock readings or time values. All events on a horizontal plane took place at the *same* time according to all observers - they are *simultaneous* events. We now say - *a point in space-time is characterized by giving the vertical straight line on which it lies together with the horizontal 3-plane on which it lies*. The event is the intersection of the line and the plane.

Let us look back at the relationships between events in the Aristotelian view we discussed earlier and reformulate them geometrically in space-time.

- (1) Do two events have the same position within space-time? This question translates to: Do two points in space-time lie on the same vertical straight line?
- (2) Do two events occur at the same time? This question translates to: Do two points in space-time lie on the same horizontal 3-plane?
- (3) What is the spatial distance between two events? What is the elapsed time? Let p and q be two events as shown in the figure below.
Their positions in space are represented by the vertical straight lines on which they lie.

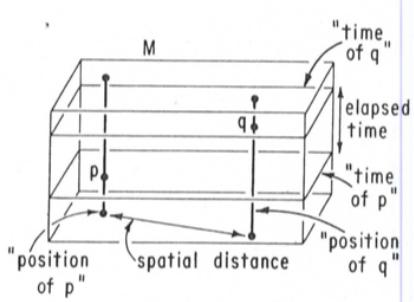


Fig. 6
Points p and q of space-time are described, in the Aristotelian view, by their positions in space (vertical lines) and their times of occurrence (horizontal 3-planes). The spatial distance and elapsed time between the events is then given, respectively, by the distance between the vertical lines and the distance between the horizontal 3-planes.

We now consider only a single particle in the world and assume that we can draw its worldline in space-time as shown below.

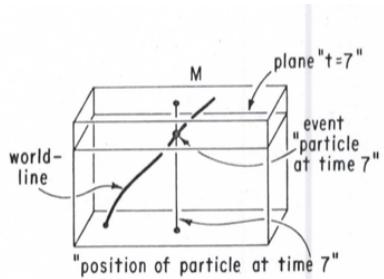


Fig. 7
The dynamics, according to the Aristotelian view, of a particle in space-time.

We ask the following question: How can we find out what the particle is doing physically using only its worldline? Step 1 is to draw one of the horizontal 3-planes, say the one corresponding to $t = 7$ and label the event p where the worldline intersects the plane. This means that p occurred at time $t = 7$ - p represent the event *the particle at time 7*. The *position of the particle at time 7* is represented by the vertical line through p . In a similar manner the spatial position of the particle at other times is given by the intersections of the worldline with other horizontal 3-planes. In this manner, the worldline provides the position of the particle at all times which is everything we need to know.

The figure below shows the worldlines of four different particles. Particle A has a vertical straight worldline which says particle A is at rest. Particle B has a non-vertical straight worldline and thus it is moving (has nonzero velocity) as shown. Particle C has a larger velocity than particle B. Particle D is changing its velocity (accelerating) and thus its world is not a straight line.

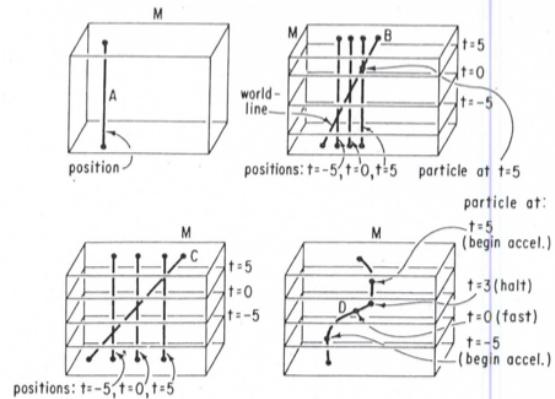


Fig. 8
Four examples of the interpretation of the space-time diagram of a particle. Particle A is at rest. Particle B is moving at a small, constant velocity, and particle C at a larger, constant velocity. Particle D is executing a more complicated motion.

As another example consider the figure below.

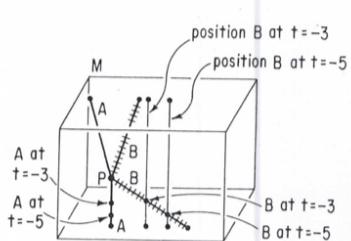


Fig. 9
The space-time diagram of the collision between two particles.

Here we have two particles. Initially A is at rest and B is moving towards A. The intersection p of the two worldlines represents a collision and then the two particles are both moving after the collision.

We note that there is no *dynamics* in space-time - nothing moves in space-time - nothing happens in space-time - nothing changes in space-time. Although the last figure describes a particle collision, this dynamic, ongoing state of affairs is represented , past, present, and future, by a single, unmoving space-time. One does not think of a particle as moving in space-time or as going along their worldline. Particles are just in space-time, once and for all, and the worldline represents, all at once, the complete life history of the particle.

Other objects in space-time are shown in the figures below.

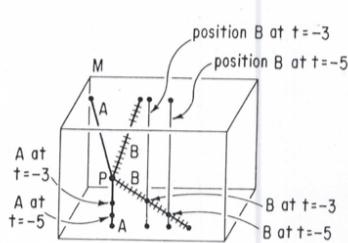


Fig. 9
The space-time diagram of the collision between two particles.

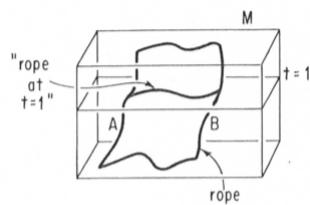


Fig. 10
The space-time diagram of a rope is a two-dimensional surface.

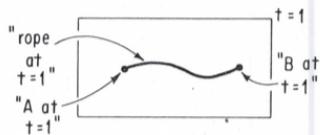


Fig. 11
To obtain the spatial configuration of the rope of figure 10 at a given time, one intersects the world-surface of the rope with the horizontal 3-plane representing that time. Here, then, is that spatial configuration at time $t = 1$.

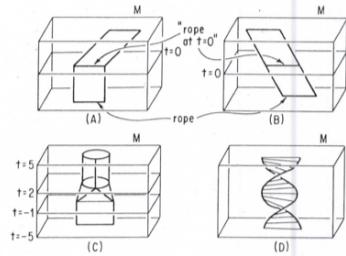


Fig. 12
Four examples of the interpretation, in the Aristotelian view, of the world-surface of a rope. The rope in (A) is initially at rest, but then begins to move in the direction of its extension. The rope in (B) moves at constant velocity orthogonally to the direction of its extension. The rope in (C) forms itself into a circle. Finally, the rope in (D) is rotating about its midpoint.

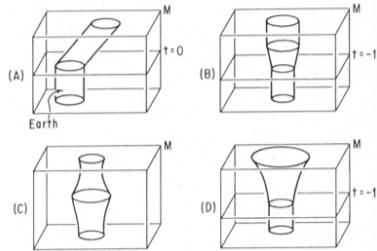


Fig. 13
Four examples of the interpretation, in the Aristotelian view, of the world-region of a planet.

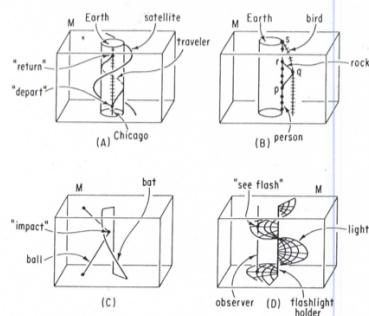


Fig. 14
Four examples of the interpretation, in the Aristotelian view, of more complicated space-time diagrams.

3.2.2 What about Light?

Suppose we send out a light beam (using a laser) as a pulse, i.e., turn it on and quickly turn it off. The result will effectively be a small *particle of light* being emitted from the laser. This is shown in part A of the figure below. Light has the worldline shown which corresponds to all events illuminated by the laser beam. The important event is labeled *laser flashed*. In part B of the figure the

laser is turned on(event p), left on for a while and then turned off(event q). This is represented by a 2-surface which corresponds to all events that are illuminated by the laser beam. Light is described by the collection of all events illuminated.

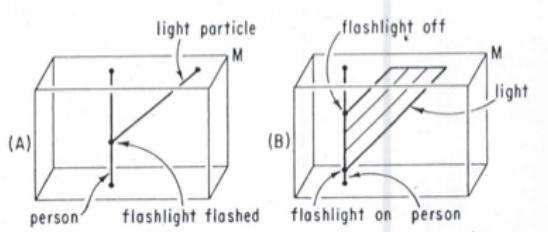


Fig. 15
The space-time diagrams representing (A) the emission of a momentary flash of light in a given direction, and (B) the emission of a beam of light in a given direction.

Now consider a light source which sends light out in all directions (say a light bulb). Keep the source off and then turn it on only for an instant and keep it off thereafter. Again the light emitted will be represented by the collection of all events illuminated. We represent this situation in space-time by a very large number of worldlines (for each light beam going in some direction) all making the same angle (just rotated around the vertical worldline of the person with the light bulb). This look like: events illuminated.

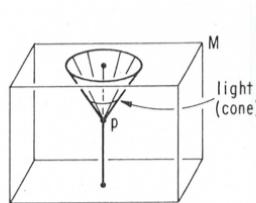


Fig. 16
The light-cone of event p is the locus of all events illuminated by a flash of light emitted in all directions from p .

The surface shown represents all the events that will be illuminated after the light bulb flashes. It is the surface of a cone as shown. This cone represents, physically, a spherical surface of light which is expanding outward with time,i.e., the intersections of the cone surface with horizontal 3-plane times is just a series of circles getting larger as time gets larger. The light bulb flash is the event p

3.2.3 Discussion

- (1) A *view* (of the structure of space-time such as the Aristotelian view) is not a *theory of physics*. A theory tells us the manner in which certain objects operate in space-time while a view generally does not care about *modes of*

operation. The Aristotelian view is something that permits a description of what is happening (has happened, will happen) without placing any restrictions on what happens. A theory takes over at that point and tells us what does happen. The *view* provides a structure on space-time that gives a broad framework within which a theory can be expressed, thought about and tested.

The structure imposed by a view, does, however, have an influence on which theories might be considered (that is why views can be dangerous). A theory must *make sense* within the structure of space-time imposed by a view, i.e., the relationships between events in the theory must be the relationships available within the chosen view. If spatial distance between events makes no sense in a view, then any theory that make reference to spatial distance between events will clearly make no sense in this particular view.

Examples

- (a) Newton's Law of Gravitation. This law makes two reference to space-time, namely, we must be able to specify the position of objects at a given time and we must be able to determine the distance between the objects at that time. Both make sense in the Aristotelian view - so the law of gravitation make sense in this view.
- (b) Law of light. A small pulse of light has a speed of $3 \times 10^{10} \text{ cm/sec}$. The pulse is described by its worldline. Does its *speed* make sense? As we saw earlier, speed makes sense in the Aristotelian view.

None of these results are surprising since the Aristotelian view is that of *everyday experience*.

- (2) What can we say about the relationship between the Aristotelian view and space-time itself. Some statements we will make refer just to the general idea of space-time without reference to any view., a particle is described by its worldline. These qualitative statements are generally about the physical interpretation of geometrical objects in space-time. Other statements, i.e., the particle represented by this worldline is at rest, all require the ability to characterize events by the their location in space together with the time of their occurence and thus need a view.
- (3) Problem with a historical approach. This approach is not historical and therefore it involves a mix of old and new ideas. This can be confusing but will straighten itself out as the old ideas necessarily disappear from the final theory.
- (4) Revolutions or Paradigm Shifts. These two words are most often overused. Major changes in the way we think is a better way to say it. In our case, we have (1) made a decision to introduce events and space-time and describe

the world within that framework; (2) will make a decision that *general relativity* is the appropriate structure to impose on space-time.

3.2.4 Final Thoughts

The Aristotelian view seems simple and comfortable. Is it correct? Consider the following example.

One observer(#1) is far out in space and can maneuver about. A second observer(#2) passes by and explodes two firecrackers at the same place in space (according to #2) separated by 5 seconds in time (according to #2) and then disappears. Can #1 decide whether these two events took place at the same position in space? Of course. #1 just locate the two events relative to his view and then claims they took place at the same position in space if they did so according to the measurements of #1. This is how the Aristotelian view works.

What do your statement depend on?

Suppose at a time earlier than the explosions #1 had accelerated and therefore had a different speed at the time of the explosions than in the first experiment. The statement whether the two events occurred at the same place in space might be quite different now since you will now have moved between the explosions. In addition, another observer moving relative to #1 might have a completely different view of whether the two events occurred at the same place in space.

Our attitude to all these different statements could be well *that is the way it is. Occurred at the sam place* could be regarded as a personal rather than a public issue. Certain relationships between events allowed in the Aristotelian view depend on who is doing the observing and what that observer's past history was, i.e., we just have to live with *many, personalized Aristotelian views*.

Physics tries to make a separation between what observers *see* and an what is, for want of a better word, what is *really* there or what can be attributed to *Nature herself*. Something like *many, personalized Aristotelian views* robs space-time of essentially all *universal structure*. If everything (the whole Aristotelian setup) is to be attributed only to individual observers, then nothing is left which can be regarded as *pure structure on space-time itself* without reference to observers.

We want space-time to retain at least some - preferably, as much as possible, - *universal, observer-independent structure*. Observers will then just be worldlines in space-time (just another geometrical object within the universal, observer-independent structure). The universal structure plus the particular observer's worldline will allow us to recreate that particular observer's experiences

3.3 The Galilean View: A *Democratic* Framework

Let us start on the path towards finding what structure is *intrinsic* to space-time. As a first step, we will develop the Galilean view, discuss what is wrong with it and then be ready to develop special and general relativity.

We start by comparing the characterization of space-time by one observer with that of another observer. What is the basis for this comparison? We must agree on some ideas which will be considered *universal*, common to all observers. These universal ideas will allow us to translate between characterizations from different observers.

The central idea will be the *event*. All observers will agree that there is an event. This implies that the collection of all possible events, namely, space-time is also universal. Observers do not have their own personal set of events or their own personal space-time.

What is the situation from the viewpoint of a single observer (us). Suppose we have constructed the usual Aristotelian setup and can therefore draw our personal picture of space-time (label the events).

How would we now describe another observer(s) setting up their own Aristotelian setup? We assume all are represented by worldlines (parallel) as if they were particles moving with respect to us with some velocity - these world lines will be straight but not vertical as shown in the figure below, which shows three of the other observers. That is our geometrical representation of the other observers.

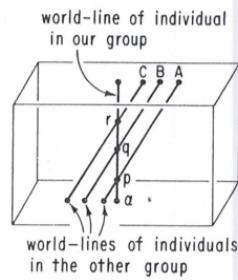


Fig. 17
The description, within our Aristotelian view, of a second Aristotelian setup. The world-lines of the members of the other group are straight and parallel, but are not, in general, vertical.

How do we represent the clocks that each carries? We assume the other observers have been instructed to yell out when their clocks read $t = 0$ (all their clocks are synchronized with each other and also with our clocks). We pass a geometrical object (surface) through the collection of such events - it will be a horizontal 3-plane for synchronized clocks. This could be done for all other times (say separated by 1 second each) and thus we can represent their clocks in space-time.

We have been thinking of the other set of observers as a phenomenon in physical world which can be represented within our space-time diagram. However, the other group could have carried out the same procedures assuming they wanted to figure out their own personal view and we were the phenomenon in physical world being observed. No one group is more important than any other. Now this other group thinks they are all at rest and we are moving. Thus, from their view we have the diagram.

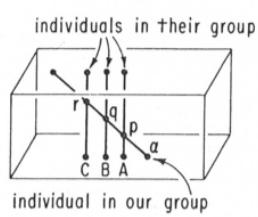


Fig. 18
Figure 17 as drawn by the other Aristotelian group.

where their worldines are now vertical. As can be seen the worldline α representing one of us intersects the worldlines A, B, C of the other observers at three events p , q and r as shown in both figures (representing the two views). Whatever events appear in one view must appear in the other view and the time order of the meetings must be the same, which it is as shown.

There is only one objective world out there with its events. Two sets of Aristotelians look at this world and each describes it in terms of their own space-time diagram or view. Each group can describe, within its own view, any phenomenon taking place in the world. In fact each group can describe itself (vertical straight lines, horizontal 3-planes) and also the other group (parallel non-vertical lines, horizontal 3-planes).

Clearly, the two diagrams are simply related to each other by a geometrical transformation. To obtain one diagram from the other one just *slides* the horizontal 3-planes over each other until the non-vertical lines are vertical and then we get the other diagram! This is shown below.

What about other phenomena? Suppose some event occurs in the world and our two sets of observers represent it within their own space-time diagram. How

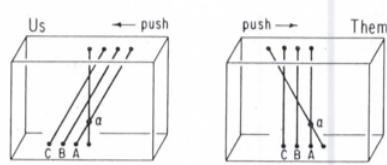


Fig. 19
The geometrical relationship between figures 17 and 18. The two figures are related to each other by sliding horizontal 3-planes over each other.

are these representations related to each other? The geometrical way of finding relationships is to apply the shifting property illustrated in the last figure and below.

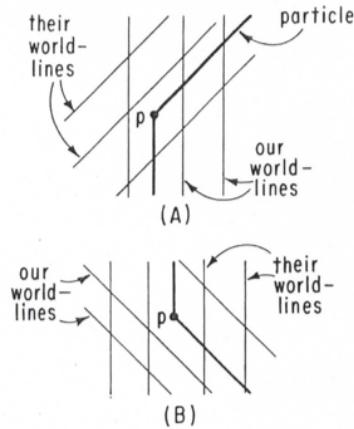


Fig. 20
Space-time diagrams for a particle which initially remains with a member of our group but, at event p , joins their group. Figure (A) is as would be drawn by our Aristotelian setup, (B) as by theirs.

In the figure above the particle is at rest with respect to the first group and then starts moving at event p in such a way that it is at rest with respect to the other group. The first group say the particle was at rest and then started moving at p . The second group say that the particle was moving and then stopped at p .

Two further examples are shown below.

The Aristotelian view, as we have seen, mixes universal structure, intrinsic to space-time itself, with structure particular to individual observers. The translation between space-time diagrams is the first step in separating these two types of structure. This is how we complete the process.

Suppose we consider many Aristotelian setups instead of just two. Now for any given physical phenomenon we draw many different space-time pictures. Any

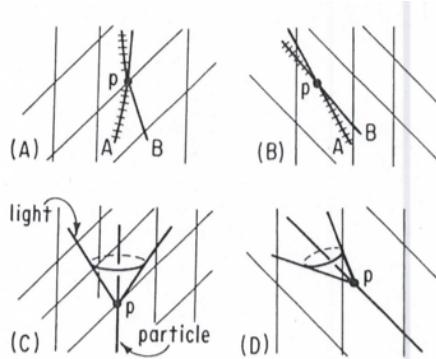


Fig. 21
Two examples of translating space-time diagrams from one Aristotelian setup to another. Figures (A) and (B) show the collision of two particles; (C) and (D) the emission of a flash of light in all directions.

two of these setups would be related as we have described earlier in the notes. There is nothing to distinguish any one of these competing Aristotelian groups from any other. The different pictures are all equally valid. Each group certainly considers their personal assignments of positions and times to be correct and that the other groups are just confused. There is, however, no objective basis for such a statement by any of the groups.

The Galilean view is a way to (or mechanism for) ending this argument between groups (about which is correct). We just *democratize* the situation: all the Aristotelian setups are allowed on equal footing.

What structural features will result? All the groups agree on the family of 3-dimensional surfaces in space-time corresponding to time surfaces. Since there are no disputes these surfaces can appear unchanged in the Galilean view. Each group draws their own position lines (a family of 1-dimensional worldlines in space-time). These families are different for the different groups. In the Galilean view all of these families are admitted with no one family preferred. Thus, we have specified the intrinsic structure of space-time implicit in the Galilean view.

There is, in space-time, a certain family of 3-dimensional surfaces together with an infinite variety of families of worldlines. Space-time consists of an infinite stack of horizontal 3-planes (the time surfaces) pierced by an infinite number of straight lines in every non-horizontal direction through every point of the stack of planes as in the figure below.

An Aristotelian group within space-time is represented by one particular family of mutually parallel worldlines (the position lines of that group). Each group chooses to have (see) its worldlines vertical. Thus, the Galilean view makes no commitment to which family is vertical. The universal, intrinsic structure

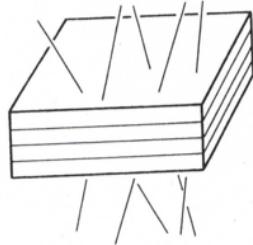


Fig. 22
Space-time, according to the Galilean view. The straight lines represent world-lines of particles, all moving at constant relative velocities. No one family of lines is preferred.

of space-time itself, according to the Galilean view is that which refers only to the 3-dimensional surfaces and to all the lines taken together. The structure particular to one group corresponds to using on particular, preferred family of lines as a reference set.

The Aristotelian view is simple to describe - an event is characterized by position in space plus time of occurrence; the Galilean view, on the other hand is not simple to describe. The only thing we can say is *ignore verticality*.

Remember, however, the thing we must do with a view is determine which relationships between events make sense. We did this earlier for the Aristotelian view and we now do it for the Galilean view.

Since all time surfaces are still valid and no family of vertical lines is singled out, the relationships that will make sense in the Galilean view are those on which all Aristotelian groups will agree. Let us look at some examples.

- (1) These two events occurred at the same time. This is equivalent to statement that two events lie on same horizontal 3-plane. All Aristotelians agree about this statement. This makes sense in Galilean view.
- (2) The elapsed time between these two events is so many seconds. This statement refers to the vertical distance between two horizontal 3-planes. All Aristotelians agree about this statement. This makes sense in Galilean view.
- (3) These two events occurred at the same position in space. Since we do not know which of the families of position lines should be used as the *true position lines*, this will not make sense in the Galilean view. Aristotelians will give different answers.

- (4) The spatial distance between these two events is so many centimeters. The answer to this statement depends on which family of position lines is used. The spatial separation is determined by finding the two position lines that pass through the two events as shown below.

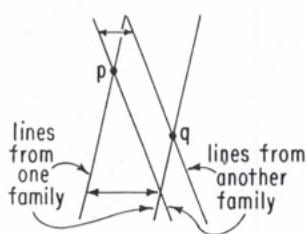


Fig. 23
"Spatial distance between two events" does not make sense in the Galilean view. The value one obtains for the "spatial distance" will in general depend on which Aristotelian group is determining it.

Clearly, the answer depends on which family. Aristotelians will give different answers. This will not make sense in the Galilean view.

There exists one case, however, where all Aristotelians agree. All get the same spatial distance when the two events take place at the same time - on the same horizontal 3-plane as shown below. This case makes sense in the Galilean view.

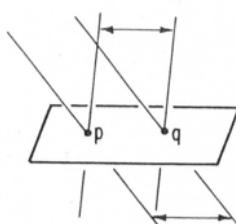


Fig. 24
"Spatial distance" does make sense, even in the Galilean view, for two events which occur at the same time.

Knowing which relationships between events make sense in the Galilean view, we can now decide which other physical notions in space-time make sense.

- (1) This particle is moving a constant velocity. Make sense in Galilean view since geometrically it corresponds to the statement that the worldline is straight on which all Aristotelians agree.
- (2) The particle has a speed of 10 cm/sec . All Aristotelians will agree on the numerical values of the elapsed time interval but will disagree on the numerical value of the spatial distance traveled during the time interval. So does not make sense in Galilean view.
- (3) This particle is at rest. Just a particular value of the speed. This does not make sense in Galilean view.
- (4) This particle collided with that other one. This just means that two worldlines have a common point. This does make sense in Galilean view.
- (5) The rope is straight. This corresponds to the world-surface of the rope intersecting horizontal 3-planes in straight lines. This does make sense in Galilean view.
- (6) The particle traveled 10 centimeters. No agreement on spatial distance. This does not make sense in Galilean view.
- (7) Particle A is moving faster than particle B. The family that corresponds to A's worldline would have A at rest. This does not make sense in Galilean view.
- (8) Particle A is moving at 10 cm/sec relative to particle B. This just means that we choose B's family. This does make sense in Galilean view.

Let us now reconsider the Newtonian law of gravitation and the law of light in the Galilean view. For the Newtonian law of gravitation reference to *one instant of time* is no problem (that is just one of the horizontal 3-planes). The distance between bodies at the same time is also valid in this view (note that general distance between events at arbitrary times does not make sense). So this law is valid. The law of light requires that speed make sense. It does not in the Galilean view. So this law is not valid.

Note that in moving from the Aristotelian view to the Galilean view, fewer things make sense. As we will see, relativity consists essentially in isolating the things in the Galilean view that do not make sense (again fewer things will make sense). Similarly, quantum mechanics simply removes all the things that do not make sense in classical physics and so on.

3.4 Difficulties with the Galilean View

It turns out that the Galilean view also has problems. Various observations cannot be properly accounted for in this view. Before proceeding to develop the relativity view, let us illustrate two of the problems.

Suppose we have two identical guns that can shoot small pellets. Consider the following experiment. Two people, A and B, each moving at uniform speed with a gun, pass each other. At the event of their meeting, event p in the figure below) each person fires their gun in the x -direction as shown.

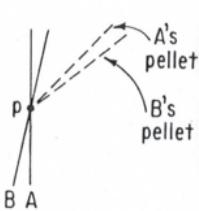


Fig. 25
Space-time diagram for a pellet-shooting experiment. Individuals A and B meet at event p , at which event each shoots a small pellet, using identical spring guns.

The figure shows the resulting space-time diagram. The worldline of A's pellet will be different from the worldline of B's pellet because A and B have different speeds (they are moving relative to each other) and their respective pellets will also have different speeds (even though they leave each gun with the same speed relative to the gun). There is nothing here to offend our common sense and nothing very strange here and experiment confirms the diagram as drawn.

Now, however, we repeat the experiment replacing the guns/pellets with flash-lights(different colors) and pulses of light. We now have a genuine physical question to ask. Will the space-time diagram looks the same or will it look like the figure below?

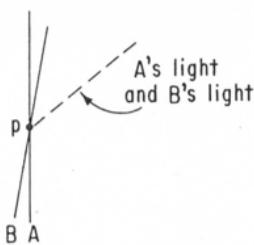


Fig. 26
Figure 25, but now with pulses of light replacing the pellets.

Common sense would force us to say it would be the same! But common sense

is not to be relied on - we can only rely on actual experiments to tell us what happens.

The answer comes from the following experiment. We consider a double-star system (two stars orbiting each other). In this case, where we can see the light emitted by both stars, we have the space-time diagram shown below.

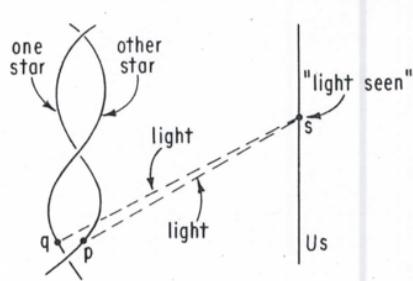


Fig. 27
An experiment to determine whether light behaves as in figure 25 or figure 26. Light from a double-star system is seen by a distant observer (us). This is what the space-time diagram would look like according to the light-propagation of figure 26.

Consider the two beams of emitted light as shown above. The emission of light from one star (event *p*) occurs while the star is moving towards us, while the emission of light from the other star (event *q*) occurs while that star is moving away from us. Note that the two light beams have been drawn as if their speeds relative to us are the same (the new assumption). Thus, in the figure, the two light beams reach us (*are seen*) at event *s* at the *same* time. If we instead use the common sense assumption, then the space-time diagram would look like the figure below

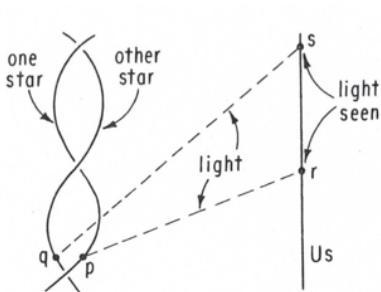


Fig. 28
An experiment to determine whether light behaves as in figure 25 or figure 26. This is what the space-time diagram would look like according to the light-propagation of figure 25.

Under these assumption, what will be the appearance of the double star system - what will we observe? Which assumption is correct?

If the speed of light is same for all sources (the first figure) then we would see two stars orbiting each other. If the speeds were different (the second figure). Although the light beams were emitted simultaneously (events p and q lying on a single horizontal 3-plane) they are received at different times (events r and s). Thus what we see at a single event (time) on our worldline is the light from the two stars emitted at different times. In this case, the stars would be expected to exhibit very complicated motions. The observations confirm that the stars just seems to orbiting each other - no complicated motions! This says that the speeds are the same!

Light behaves completely differently than pellets. Light, once emitted by the flashlight, decides its worldline in space-time will be by reference to space-time itself and that alone - without any reference to what the emitter is doing.

The damaging aspect of all of this to the Galilean view is the idea that light has a speed of $3 \times 10^{10} \text{ cm/sec}$ no matter what - independent of source or detector! Light does not behave as expected in the Galilean view.

A second example involves elementary particles called mu-mesons. In a laboratory, these particle , when produced at rest, decay into other particles in about 10^{-6} sec after they are produced. Mu-mesons are also produced at the top of the atmosphere by collisions with high-energy cosmic rays. This high-energy (fast-moving) mu-mesons shower down on the earth and be detected. However, the mu-mesons would require $10 \times 10^{-6} \text{ sec}$ to reach the detectors. All this is represented by the figure below.

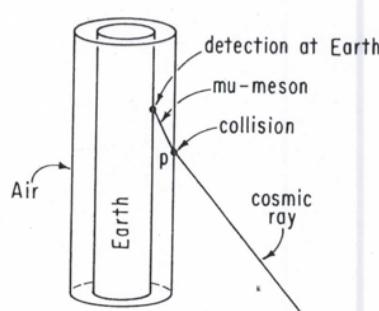


Fig. 29
Space-time diagram for a mu-meson experiment. A cosmic ray collides with the atmosphere at event p , releasing a mu-meson, which is finally detected at the surface of the earth.

We are faced with a question. If these mu-mesons only live for 10^{-6} sec how did they get to the detectors, which requires them to live for 10×10^{-6} sec? Why are they living longer? It seems that mu-mesons live longer if they are moving! This guess can be tested experimentally. The experimental results are shown below.

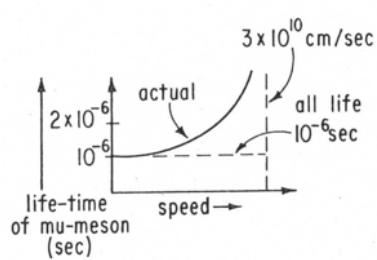


Fig. 30
Graph of the lifetime of a mu-meson as a function of its speed.

Note that no mu-mesons move faster than the speed of light (no matter how much energy they are given or no matter how much force is exerted on them for any period of time). The faster a mu-meson moves, the longer its lifetime, which makes no sense in the Galilean view.

When a theory of physics (the Galilean view in this case) is faced with experimental contradictions like this, then either some modification of the theory can fix things up or maybe the entire theory needs to be discarded in favor of something completely different. The correct theory will be one that does not fail *any* experimental tests. In this case, we will need a completely new view, namely, relativity.

3.5 The Interval: The Fundamental Geometrical Object

We now proceed to find a new view. This will require some bold moves where we jump beyond observations to make new assumptions and it will also require us to carefully examine many of the ideas we already have accepted as being true. The choices we make that lead to the new theory often can not be justified in any way - that is the nature of assumptions or postulates.

We make this choice: we retain the broad, qualitative features of space-time, i.e., we retain the idea of an event and the idea that space-time is the collection of all possible events. Points still represent events, worldlines still represent particles, intersections of worldlines are still collisions of particles and so on. We discard all other things - we do not have

clocks

spatial distance between two events

events occurred at the same time

speed of a particle

Aristotelian setups

Thus, we are discarding all detailed, geometrical, numerical information.

In the transition from the Aristotelian to Galilean views we switched from a rigid picture of space-time to *sliding/shifting* picture. We now go to the extreme in this same direction - imagine space-time drawn on a rubber sheet, which can be stretched, pulled and bent.

We have made an amazing shift here - we introduce events, worldlines, space-time, etc to help us picture the so-called *real geometrical structure* of space and time and to help different observers communicate with each other. Now these ideas are to be treated as *paramount* with the so-called *more natural* ideas such as spatial distances and elapsed times now to be suppressed.

Since the framework we are now considering is immune to conflicts with observations, we cannot call it a view yet. A theory of physics in general, and a world-view in particular, must at least make some commitment about the physical world if it is to be worth anything at all. We must find a way to incorporate hard, numerical, geometrical information about space-time. Since the old spatial distances, elapsed times, etc no longer make any sense, we must find something to replace them.

To proceed, we must re-examine everything. We have been describing space-time using *clocks and meter sticks* without carefully defining how they work. We implicitly assumed we knew how to measure lengths and time intervals which we then thought allowed us to describe space-time around us. Let us now be more careful.

First, we must be very careful about measuring instruments and also try to use as few as possible. Second, all instruments must be representable within our space-time framework. Third, we must explicitly describe and know the properties of our instruments.

We need two types of instruments to obtain geometrical information about space-time. The first instrument must be able to go out into space-time(away from our own worldline), collect information about space-time events and return that information back to our worldline. Otherwise, we could only know about events on our own worldline. The second instrument must be able to record or *make numerical* the information that has been brought back. This allows us

to make definite statements about events off our worldline instead of just being able to sense them. Let us now make some specific choices.

We might use a meter stick. It certainly *sticks out into space-time away from our worldline*. There are problems, however. First, a meter stick is a very complicated object. It corresponds to a 2-surface in space-time and the markings on the meter stick correspond to lines drawn on the 2-surface. Second would we read a marking on the meter stick? The figure below show the 2-surface of the meter stick and the line for the 82 cm mark.

Clearly, the 82 cm mark is out in space-time away from our own worldline. What mechanism will we use to tell us where the 82 cm is? We are in a logical circle of confusion here.

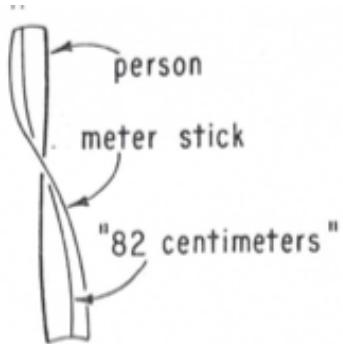


Fig. 31
Space-time diagram of a meter stick.

So meter sticks are poor instruments to use (remember that they were crucial for the Aristotelian setup).

Another possibility is to use particles. We could send particles out into space-time have them *reflect* from an event and return to our worldline as shown below.

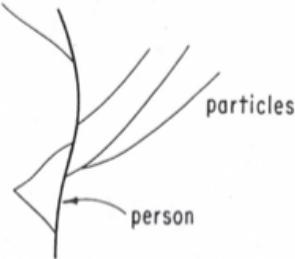


Fig. 32
Space-time diagram of individual using particles to detect the structure of space-time away from his world-line.

The problem for a general particle is that they can have a wide variety of speeds which ends up being a very large complication. It should be obvious now what we should use - light. Light can reflect (use mirrors) off of events in space-time and it has a fixed numerical speed independent of any motions of sources or observers - its speed is intrinsic to space-time. So light will be our instrument for *probing space-time*.

What about the instrument for recording information? In this case, let us just choose the one that works, namely a clock. A clock stays on our worldline and thus can be read locally with no additional constructs. Clocks have no troublesome features - they just generate numbers (the time values) - it a perfect *pure recorder*.

Now let us represent these two instruments in space-time. Light is represented by a worldline. A clock is defined to be a *point* object (can be assigned a worldline), all clocks are identical and time values are monotonically increasing. The time reading of a clock is a function that assigns to each point on the clock's worldline, a number (the time). Each such time value is an event on the clock's worldline.

The number assigned to the point p is referred to as *the time, according to the clock, at event p*. See the figure below. A clock can only assign a time to an events if the clock's worldline passes through the event.

We now need to be explicit about all properties assumed by our two instruments. Light is easy - we assume that two pulses of light, emitted in the same direction from the same event, move together, no matter what the emitters are doing. We

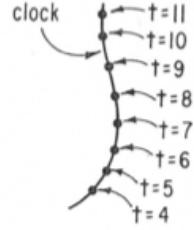


Fig. 33
Representation of a clock in space-time. The clock possesses a world-line, and furthermore assigns a number t , the time-reading of the clock, to each event on that world-line.

consider two clocks with coincident worldlines until some event p after which they are separated (see figure below) and then brought back together at event q and remain together thereafter.

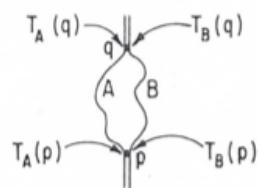


Fig. 34
Space-time diagram of an experiment involving two clocks, A and B . The clocks separate at event p , and return together at event q . Is the elapsed time according to A between p and q the same as that according to B ?

Since p is a point on the worldline of clock A , this clock assigns to p a number, the time according to A of p which we write as $T_A(p)$; since q is a point on the worldline of clock A , this clock assigns to q a number, the time according to A of q which we write as $T_A(q)$. The difference between these times, $T_A(q) - T_A(p)$, represents physically the elapsed time, according to clock A between event p and event q . Similarly, using clock B , we obtain $T_B(q) - T_B(p)$. We cannot, however, claim that $T_A(q) - T_A(p) = T_B(q) - T_B(p)$ since that do not have the same worldlines between p and q . Let us see why.

Let us assume that it is true and see what the consequences would be. Let us fix in space-time some reference event p and arbitrarily assign it the time $t = 0$ as shown below.

Now consider another event q in space-time. We can assign this event q a *time* in this way. Find a clock A , which passes through both events p and q , and thus determine $T_A(p)$ and $T_A(q)$. Using this clock A , we then assign to event q *time* $t = T_A(q) - T_A(p)$. This time assignments is the same no matter what clock we send from p to q (same for clock A and clock B) because we assumed this to be true. Thus, event q is assigned a unique time. Having made this assignment

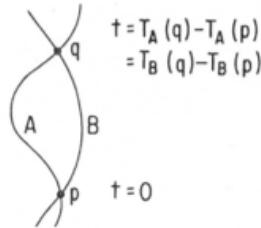


Fig. 35

Under the supposition that the answer to the question of figure 34 is yes, we obtain a universal time in space-time. The measured elapsed time from p to q must be the same no matter what clock measures this time, and so, assigning to p time $t = 0$, a unique time is assigned event q .

for q it is true also for any other event in space-time. Therefore, to each event in space-time we can assign a unique time. Given these unique times we could then draw surfaces corresponding to given times, i.e., the surface corresponding to $t = 3$ which passes through all events that were assigned the time $t = 3$. so we end up with a collection of time-surfaces in space-time.

Thus, the assumptions leads directly to slicing space by time-surfaces, which is the old Galilean view.

The only way to avoid the Galilean view, which is not correct, is to not make the assumption.

Another argument against this assumption comes from observations. Consider the experiment shown below.

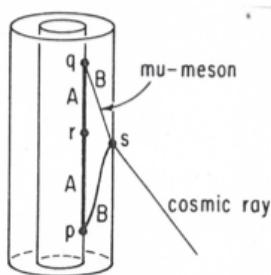


Fig. 36

The mu-meson experiment revisited. Now, however, we imagine introducing two clocks, A and B . The experimental result with mu-mesons suggests that the answer to the question of figure 34 is no.

The mu-meson is a simplified clock; it keeps track of the time since its creation and when that time reaches 10^{-6} sec it decays - it is a clock with one tick! We now insert into the experiment a couple of clocks, one which remains on the surface of the earth (clock A) and another which goes up into the atmosphere experiencing the event *collision of the cosmic ray with air molecule* and then returning to earth with the mu-meson(clock B) as shown in the figure. Now clocks A and B are together at event p and also at event q . One might expect $T_A(r) - T_A(p)$, where r is the event on the earth which Aristotelians would say occurred at the same time as the collision event s , would be the same as the elapsed time $T_B(s) - T_B(p)$. From our earlier discussion of this experiment we know that the actual result is that A's elapsed time is of the order of 10×10^{-6} sec while the elapsed time according to B's clock is of the order of 10^{-6} sec. Clearly, we should not make any assumption of equality of elapsed times for differently moving clocks!

Let us try another idea. Let the two clocks A and B describe the same worldline between events p and q as shown below.

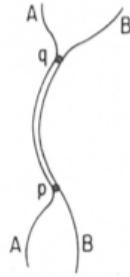


Fig. 37

A second experiment involving two clocks. Now, the clocks are together between events p and q . Will they measure the same elapsed time between event p and event q ?

Again, we can determine the various clock readings and ask the same question about equality of elapsed times. Although, it seems obvious in this case that the elapsed times will be equal, i.e., identical clocks and the same worldline, we must be careful. Looking at the figure, we see that the two clocks had *different histories* prior to event p . Does *past history* of a clock affect its present ticking rate? Since there is no experimental evidence of any effect due to past history, we assume it does not matter. Thus, in this case, the elapsed time will be assumed to be the same - *two clocks having identical worldlines between two fixed events measure the same elapsed times between those events*. This is the way theoretical physics proceed to develop a theory.

Summarizing, this is the *clock property*. Given any worldline of a particle, one acquires an assignment of times to points of that worldline (a clock is carried alongside the particle and clock readings = times). The time assignments are unique as far as elapsed times are concerned - the worldlines thus acquire time functions.

We now have the general framework - events, space-time, worldlines, etc, on the instruments for determining geometrical information - light pulses and clocks along with their descriptions in space-time - light-pulse = worldline and clocks = worldlines with times associated with each point. We also have that the worldline of light is independent of source or observer and two clocks with same worldline record identical elapsed time.

What we are interested in, however, is relationships between events which are equivalent to relationships between physical objects. Let us now determine these relationships using framework, instruments and properties we have just discussed.

We fix two events, p and q in space-time. How do we determine the relationship between them? Since the two events are arbitrary, it will not necessarily be true that any light pulse will pass through both events or that any clock whose worldline meets both events so that we cannot find anything out using only one instrument. Now light pulses convey information and clocks record information. Clearly, we should use both instruments in our procedure. The idea will be to keep the clock near one event, say p , use light-pulses to interact with event q and bring information back to us (traveling with the clock).

Consider then the figure below.

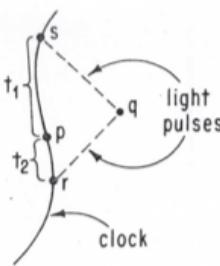


Fig. 38

The construction which expresses numerically the relationship between two events, p and q . A clock is allowed to experience event p , while a light-pulse, leaving the clock at r , reaches q , and another pulse, leaving q , returns to the clock at event s . The clock then measures two elapsed times, t_1 and t_2 .

We send a clock through event p . We send a light pulse from the clock worldline (emitted at event r) to event q . A second light pulse is emitted by q so it reaches the worldline of the clock (arriving at event s). Thus, we (the observers), who travel with the clock and who experience event p cannot experience q , so we let the light pulses experience q and we record the emission (r) and reception (s) of the light. The two events r and s on our worldline represent information of some kind about q that was conveyed by the light.

In this case, the clock we are carrying with us Clocks make information numerical. In this case, the clock we are carrying with us assigns times to the events r , p and s and we use the assigned times to measure two numbers, namely, the elapsed time between r and p is $t_2 = T_A(p) - T_A(r)$ and the elapsed time between p and s is $t_1 = T_A(s) - T_A(p)$. The two events p and q have thus produced two numbers t_1 and t_2 using our two instruments. These tow numbers describe some kind of relations between the events p and q .

Let us now try to understand what the two numbers t_1 and t_2 mean physically and how we will use them. The total time $t_1 + t_2$ is the elapsed time for a light signal to travel from my worldline to event q and back. We might then say that the spatial distance from and event p to event q must be $c(t_1 + t_2)/2$. Similarly, since the light speed going out to q and coming back is always c , event q must have occurred at time $(t_1 + t_2)/2$, i.e., 1/2 the total elapsed time. This time corresponds to event q' , which is halfway between r and s . As we see from the diagram below

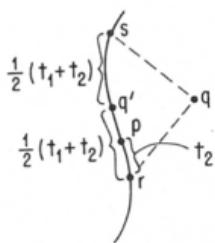


Fig. 39
Aristotelian interpretation of figure 38. An event q' is located halfway (in terms of elapsed time) between events r and s . This q' is interpreted as occurring simultaneously with q . Now, an apparent elapsed time and spatial distance between p and q can be calculated.

the time of event q' (simultaneous with q) is $(t_1 + t_2)/2$ and the time of event p is t_2 . Therefore, the elapsed time between p and q' or q is $(t_1 + t_2)/2 - t_2 = (t_1 - t_2)/2$. Thus, we have found that the spatial distance between event p and event q is $c(t_1 + t_2)/2$ and the elapsed time between event p and event q is $(t_1 - t_2)/2$. These sum and difference combinations are all the information we can extract from the original two numbers t_1 and t_2 and they will represent our physical interpretation of the numbers.

We have managed to re-expressed some ideas which make sense in the Aristotelian and Galilean views in a new and different way - in terms of two numbers t_1 and t_2 with explicit physical significance. The two numbers t_1 and t_2 carry some information which refers to the observer (the choice of the clock worldline) and some which is intrinsic to space-time.

Before looking further at these two numbers, we need to deal with a problem. Both t_1 and t_2 cannot be intrinsic to space-time. Consider this argument. Fix the events p , q , r and s , and consider two clocks A and B each of which experience r , p and s as shown below.

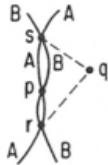


Fig. 40
Two clocks, A and B , both carrying out the measurement of figure 38. In this experiment, the clocks are so adjusted that their events p , q , r , and s are the same.

Let the times be t_1 and t_2 according to A and t'_1 and t'_2 according to B. It is not true in general that $t_1 = t'_1$ and $t_2 = t'_2$ since we already rejected this possibility earlier. So, clock A assigns numbers t_1 and t_2 to events p and q while clock B assigns numbers t'_1 and t'_2 to events p and q and these times can be completely different - seemingly arbitrary. How do we get around this difficulty?

Consider the following situation. Try to determine the spatial distance between two fixed points by walking between the points. Clearly this is an unsatisfactory procedure as shown by the figure below.

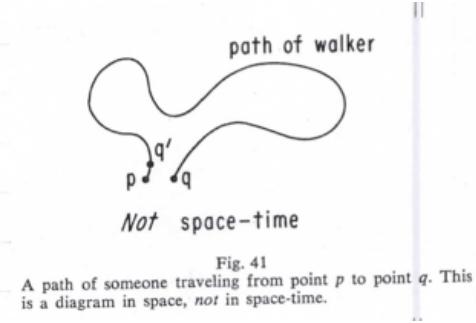


Fig. 41
A path of someone traveling from point p to point q . This is a diagram in space, *not* in space-time.

Clearly, the distance is route-dependent even if the two fixed points are close together. We get around the problem by first requiring that the walker choose a route and then requiring that the second point be *nearby*. We can then avoid the seeming arbitrariness of the times discussed earlier by requiring that we only consider the limit when events p and q are *nearby*. This prevents the clock worldlines from *wiggling* too much between events r and p and between p and s so that we keep the time *distortion* to a minimum.

Continuing our quest to understand the physical meaning of the two measured times t_1 and t_2 we note that there are five separate, distinct cases for the construction depending on the relative locations of p and q , i.e., on the order of r , p and s along the clock worldline. As shown in the figure below we have these cases.

1. Order as shown in (A) below. t_1 = elapsed time between p and s is positive, while t_2 = elapsed time between r and p is negative - p occurs before q
2. Order as shown in (B) below. p and r coincide. Light pulse from clock to q leaves clock at p . t_1 is positive, while t_2 = elapsed time between r and p is zero - since p and r coincide
3. Order as shown in (C) below. Same as earlier case. p occurs between r and s on clock worldline. Both t_1 and t_2 are positive.
4. Order as shown in (D) below. p and s coincide. Light returning from q

meets clock at p . $t_2 = \text{elapsed time between } r \text{ and } p$ is positive, while $t_1 = \text{elapsed time between } p \text{ and } s$ is zero, since p and s coincide

5. Order as shown in (E) below. Both r and s occur before p . $t_2 = \text{elapsed time between } r \text{ and } p$ is positive, while $t_1 = \text{elapsed time between } p \text{ and } s$ is negative.

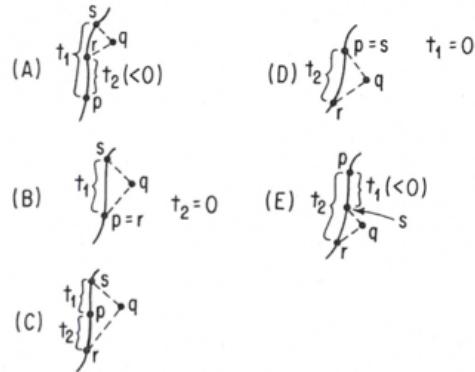


Fig. 42
Five cases of figure 38.

The five cases are dependent on the ordering of r , p and s along the clock worldline. The question of which case one has, however, depends only on events p and q and not on any other details - especially not on the choice of clock worldline.

Consider the figure below.

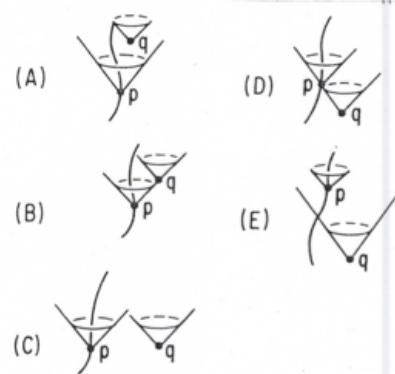


Fig. 43
The five cases of figure 42, now expressed in terms of light-cones.

We see the following.

1. Part (A). r occurs after p on the clock worldline - clock passes from p to r and then light from r to q . All light sent in all directions from p will cover a region of space-time including q - event q will lie inside the forward light-cone of p .
2. Part (B). r and p coincide - light pulse from p reaches q . Event q will lie on the forward light-cone of p .
3. Part (C). Light, to reach q , must be emitted from the clock before p , while light emitted from q reaches the clock after p . p is neither within or on the forward light-cone of q and q is neither within or on the forward light-cone of p
4. Part (D). Light pulse from q reaches p . Event p will lie on the forward light-cone of q .
5. Part (E). Light from q reaches clock before p . p will lie inside the forward light-cone of q .

The important point of this last set of characterizations is that they only involve qualitative features of space-time: light-cones and location of events within, on or outside those light-cones. These features do not refer to any property of the clock worldline. p and q completely determine the particular case. The question of which case we have is *intrinsic to space-time*. The five cases are classified as shown in the table below.

Case	Name	t_1, t_2	Light -Cones
First	Timelike; q future	$t_1 > 0, t_2 < 0$	q in light-cone of p
Second	Lightlike; q future	$t_1 > 0, t_2 = 0$	q on light-cone of p
Third	Spacelike	$t_1 > 0, t_2 > 0$	None
Fourth	Lightlike; q past	$t_1 = 0, t_2 > 0$	p on light-cone of q
Fifth	Timelike; q past	$t_1 < 0, t_2 > 0$	p in light-cone of q

Summarizing, the five cases can be distinguished using the light-cones of p and q or the signs of the numbers t_1 and t_2 . Is there some combination of t_1 and t_2 that reflects this classification and is thus intrinsic to space-time. As is usual in theory, someone makes the right guess (hopefully informed by earlier knowledge gained). The intrinsic quantity is the product $t_1 t_2$. It is called the *interval* between p and q . If it is really intrinsic to space-time, then no matter what clock is chosen, the interval will always be the same number given the events p and q .

We now almost have all the ideas need to formulate general relativity (one minor one still needed). We will proceed to work it all out shortly. Before doing so, let us summarize how we have arrived at this point.

1. Assumed general framework of event, space-time, worldlines, etc.
2. Assume instruments(2) described within space-time framework.
3. Instruments determine relationships between events (two times).
4. Impose geometrical structure on framework - an intrinsic relationship between events - the interval.

The *intrinsic relationship between events* in the relativistic view is the interval associated with any two nearby events - a certain number of seconds squared. It is intrinsic in the sense that it can be determined by any individual observer using their own clock the number obtained will be the same for any other observer using their own clock - that is what *intrinsic* means!

The relativistic view of space-time consists of two catalogs. The first lists all possible events (past, present, future). The second lists all pairs of nearby events, and for each pair, the interval between them. Space-time, in this sense, is a *geometrical entity*. This is enough information to allow us to describe the relationship between non-nearby events - connect with a worldline and relate nearby points on worldline until distant points are related.

What is the physical interpretation of the interval? We can do a simple calculation as follows:

$$\begin{aligned}\text{apparent spatial separation between two events} &= \frac{\Delta x}{c} = \frac{1}{2}(t_1 + t_2) \\ \text{apparent elapsed time between two events} &= \Delta t = \frac{1}{2}(t_1 - t_2)\end{aligned}$$

where *apparent* means *according to the observer with the clock*. We then have

$$\begin{aligned}\frac{\Delta x}{c} + \Delta t &= \frac{1}{2}(t_1 + t_2) + \frac{1}{2}(t_1 - t_2) = t_1 \\ \frac{\Delta x}{c} - \Delta t &= \frac{1}{2}(t_1 + t_2) - \frac{1}{2}(t_1 - t_2) = t_2\end{aligned}$$

Therefore,

$$\begin{aligned}\text{interval} &= t_1 t_2 = \left(\frac{\Delta x}{c} + \Delta t \right) \left(\frac{\Delta x}{c} - \Delta t \right) \\ &= \left[\frac{(\Delta x)^2}{c^2} \right] - (\Delta t)^2\end{aligned}$$

That is the result any observer measuring apparent spatial distance and apparent elapsed time would use to calculate the interval. Different observer will disagree about the values of Δx and Δt but the will agree about the interval value.

Earlier we introduced the distinction between lightlike, timelike and spacelike separated events (or intervals). Using the expression for the interval

$$\text{interval} = \left[\frac{(\Delta x)^2}{c^2} \right] - (\Delta t)^2$$

the observer measuring Δx and Δt says the following.

- Lightlike related events \rightarrow interval $= 0$ or $\Delta x = c\Delta t$. Thus, lightlike related means that light just makes it from one event to the other or one event lies on the light-cone of the other, which, of course, is exactly what we found earlier.
- Timelike related events \rightarrow interval < 0 or $\Delta x < c\Delta t$. Since light can go a distance $c\Delta t$ in time interval Δt we then say that during the elapsed time between the two events, light has enough time to travel further than the spatial distance between the events - the later event is inside the light-cone of the first event as earlier. In the figure below we show the full scenario of this case.

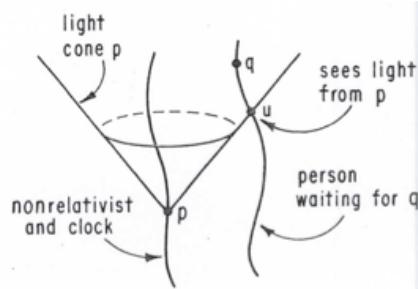


Fig. 47
Timelike-related events p and q , as interpreted by a non-relativist.

Here we have events p and q , the clock worldline passing through p (experiences p) and another observer whose worldline passes through q (experiences q). This second observer experiences the light from p at event u , which is experienced *before* event q . To this observer, the event q occurred too late to be on the light-cone of p - the light had already gone by before q occurred.

- Spacelike related events \rightarrow interval > 0 or $\Delta x > c\Delta t$. Thus, during the elapsed time between the two events, light had enough time to travel a distance less than the spatial distance between the two events. Light emitted from one event, will never make it to the other event, for the other will occur too soon. This is shown in the figure below.

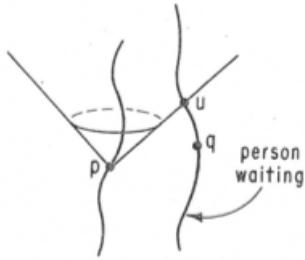


Fig. 48
Spacelike-related events p and q , as interpreted by a non-relativist.

The other observer experiences q and only after that experiences (at u) the light coming from p . The observer would say that q occurred too early - before the light from p was experienced.

Let us look further at the concepts of timelike and spacelike intervals. Consider two nearby events p and q that are timelike related with q within the light-cone of p . We can find a clock in this case that experiences both events as shown in (A) of the figure below.

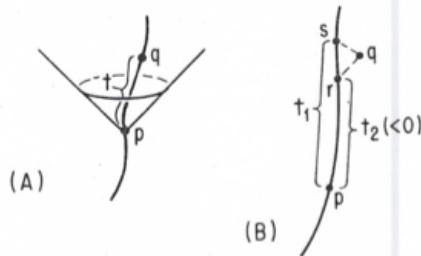


Fig. 49
 A is the limiting case of B as q approaches the world-line of the clock. The interval between p and q in A , therefore, is just $-t^2$.

Suppose such a clock records an elapsed time t between p and q . What is the interval between these two events in terms of t ? Now part (A) of the figure is just the limit of part (B) where q is very close to the worldline of the clock. In this case, r and s are very close together and also close to q . Part (A) just corresponds to q , r and s coinciding. Therefore, $t_1 = t$ (the elapsed time from p to $s = q$) and $t_2 = -t$ (elapsed time from $r = q$ to p). Therefore the interval $= t_1 t_2 = -t^2$. This says that if you experience two nearby events (on your worldline), then the interval between the two events is $-t^2$ where t = elapsed time you measure between the two events.

We can understand the concept of a spacelike interval in a similar way. Now let

p and q be spacelike related. We can send many clocks through p . Some would have $t_1 > t_2$ and vice versa. We choose a clock where $t_1 = t_2$ as shown below.

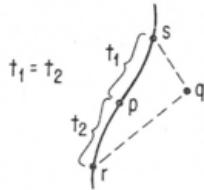


Fig. 50
A clock is so arranged that it experiences event p , and it has $t_1 = t_2$. Then the interval between p and q , according to this clock, is the square of this common time.

In this case, our observer would measure $\Delta x = c(t_1 + t_2)/2$ and $\Delta t = 0$ and the interval is $(\Delta x)^2/c^2 = \text{square of spatial distance between two events divided by } c^2$.

We have now proposed the following way to think about space-time. We think of space-time as consisting of two books, the first listing all possible events and the second listing, for all possible pairs of nearby events, the interval between those events. The two books contain all the spatial and temporal information it is possible to know about the world. The information is used as follows. Physical phenomena are described in terms of collections of events, i.e., worldlines, etc. The intervals in the second book provide information about relationships between events which leads to *spatial and temporal statements* about physical phenomena.

If all of this makes sense then we should be able to describe light-pulses and clocks (our instruments) in terms of intervals. Imagine someone who knows relativity and who has spent their entire life in a closet - no way to observe the outside world. The person has access to the two books we have already described and nothing else. The person in the closet is instructed to figure out, from the books, what observers on the outside already know, because they can shoot light-pulses, find their worldlines, build clocks and associate time-functions with worldlines. Can the person in the closet determine what are the worldlines for light-pulses and what are the worldlines and times for clocks using only the two books.

First, consider light-pulses. A worldline is lightlike if any two nearby points along the worldline are lightlike - this is a valid definition since it only requires the two books, i.e., a worldline is just a collection of events from the first book and whether it is lightlike is determined by whether all nearby points have interval zero, which is information in the second book. Know from knowledge of relativity that given two lightlike events, a single light-pulse can experience both

events. Therefore the actual worldlines of light will be the lightlike worldlines he found from the books. He can correctly predict the worldlines of light knowing only what is in the interval - the behavior of light is carried by the interval.

What about clocks? The person in the closet defines a worldline as timelike if any two nearby points along that worldline are timelike related, which makes sense using the two books. Knowing relativity the person in the closet knows that all clock worldlines are timelike. Therefore, he can predict all possible clock worldlines using only the two books.

What about elapsed time? Can he determine that from the two books? This is how it is done. Take a particular timelike worldline, fix two points p and q on that worldline and predict the elapsed time between the two points using the two books. Choose a sequence of points marked s_1, s_2, \dots, s_n along the worldline from p to q as shown below.

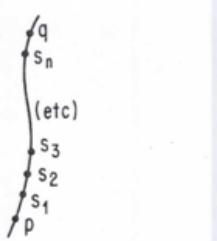


Fig. 51
The computation, using only the interval, of the elapsed time, along a certain world-line, between events p and q . One introduces a succession of closely spaced events along the world-line, computes from the interval the elapsed time between successive events, and adds.

Points chosen so that p and s_1 are nearby, s_1 and s_2 are nearby, and so on until s_n and q are nearby. Look up the intervals for each of the nearby pairs (all will be negative since it is a timelike worldline). For each nearby pair, the square-root of the negative of the interval is the elapsed time between that nearby pair. Add all the elapsed times to find the total elapsed time between p and q . Thus, the elapsed time can also be predicted from the two books. Thus, the information about clocks is already carried by the interval.

Thus, from the interval, we can determine how light goes and how clocks move and tick. What is known about the world is in the two books. This is the *relativity view*.

3.6 The Physics and Geometry of the Interval

We now want to extend our discussion beyond the interval and its meaning to discuss physical phenomena which may be happening in the immediate vicinity

of an observer.

Given an observer in space-time carrying a clock, fix an event p on her worldline. Now ask this question: Can this observer characterize the events near p within an Aristotelian view, i.e., times of occurrence and positions, using our standard instruments? If so, what does this characterization look like?

Let q be some event near p . Our observer can use the standard procedures we have discussed (sending out light-pulses and recording emission and reception times) to determine t_1 and t_2 and thus calculate $\Delta x = c(t_1 + t_2)/2$ and $\Delta t = (t_1 - t_2)/2$ to obtain the apparent spatial distance and the apparent elapsed time between p and q . What if the observer found $\Delta t = 0$, i.e., $t_1 = t_2$? The observer would then say that p and q were *simultaneous* events. Our observer would interpret some events as being simultaneous with p and others not simultaneous with p as shown below.

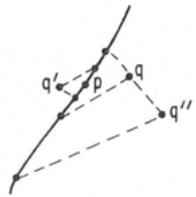


Fig. 52

Given the world-line of an observer, and an event p on that world-line, the observer can decide, for each nearby event, whether or not he regards it as simultaneous with event p .

where the observer might find that q and q' are simultaneous with p and q'' is not.

If the observer considers the collection of all nearby(to p) events (q 's) are regarded as simultaneous with p using the criterion above, one gets a surface in space-time as shown below.

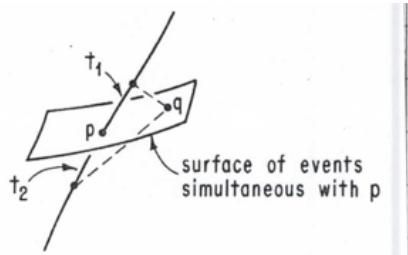


Fig. 53
The locus of events deemed simultaneous with p under construction of figure 52 forms a 3-surface in space-time

All events q on the surface have $t_1 = t_2$ including p ($t_1 = t_2 = 0$). The observer would say that this surface is the locus of all events which are simultaneous with p - it is a *time-surface* - all its events occurred at the same time, namely, the time of event p .

If we now find all events q for which $\Delta t = 1 \text{ sec}$, i.e., $t_1 - t_2 = 2 \text{ sec}$, then the observer would say that q occurred 1 sec after p (the elapsed time between p and q is 1 sec) If we draw a surface through all such events we get the figure below.

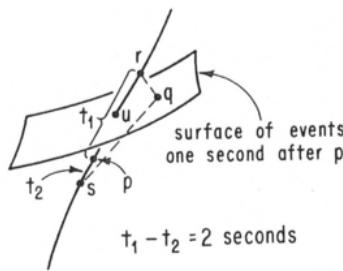


Fig. 54
The locus of events deemed to have occurred one second after event p .

The observer would say this surface represents the locus of all events which are 1 second later than p or the locus of all events which are simultaneous with u which is 1 second later than p on the clock worldline passing through p and

u. This process can be repeated for any other value of elapsed time between events. Each choice produces a corresponding time-surface as shown below.

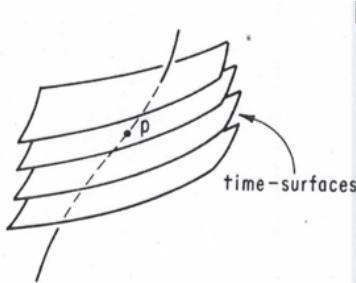


Fig. 55
Local time-surfaces as constructed by observer experiencing event *p*.

As a result we *slice* space-time near *p* into a family of surfaces of simultaneity. We have thus created the old time-surfaces of the Aristotelian view for a particular observer - we have done it, however, with reasonable care, using our instruments.

We can now do the same thing for *position in space*. We consider events *q* for which $\Delta x = 0$, i.e., $t_1 + t_2 = 0$ or $t_1 = -t_2$. These events have zero spatial distance from *p* - they occurred at the same position as *p*. If we draw the locus of all events *q* with $\Delta x = 0$, the result will be some worldline in space-time. The observer say this worldline represents all events at the same spatial position as *p* - this is a *position-line*. What does this line look like in space-time? Now t_1 = elapsed time between events *p* and *s* (fig 38) and t_2 = elapsed time between events *r* and *p*. Thus, $-t_2$ is the elapsed time between *p* and *r*. But $\Delta x = 0$ says that $t_1 = -t_2$ or $\Delta x = 0$ means that elapsed time from *p* to *s* is the same as elapsed time from *p* to *r*. Both *r* and *s* are on the same worldline as *p* and since they have the same elapsed time from *p*, we have *r* = *s*. But *r* = *s* means (fig 49) means event *q* is on the worldline of the clock - the events *q* that have $\Delta x = 0$ (the same spatial position as *p*) are the events *q* which lie on the worldline of the clock.

To find another position-line, we just repeat the above process with Δx having a different value for all events in the collection. We obtain the set of position-lines in this manner. See figure below.

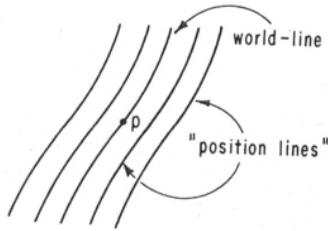


Fig. 56
Local position-lines as constructed by observer experiencing event p .

The result of these two construction procedures is a local(nearby p) Aristotelian setup. The observer then characterizes an event by specifying the time-surface and position line intersection which corresponds to the event. See figure below.

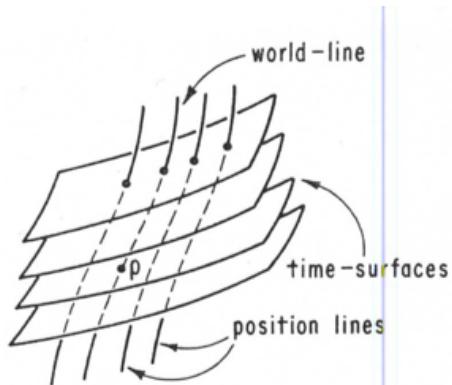


Fig. 57
The local Aristotelian setup.

We note that these constructions were accomplished using only our two instruments. We showed earlier that our two instruments could be completely described in terms of the interval. Thus, knowledge only of the interval is sufficient to construct a local(nearby p) Aristotelian setup.

Each individual Aristotelian observer uses these techniques and intersections of worldlines, etc with time-surfaces to describe what is happening in time. Other observers have their own Aristotelian setups and the own (possibly different) description of events. The link between different observer is the interval since they will all agree on the interval.

Let us now broaden our understanding of this general picture. Fix an event p and let two observers (A and B) experience that event (their worldlines intersect at p) as shown below.

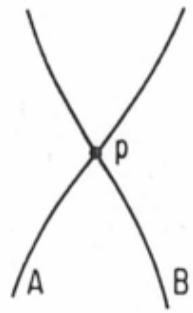


Fig. 58
Two observers, *A* and *B*, meet at event *p*.

Each constructs an Aristotelian setup near *p* such as shown below.

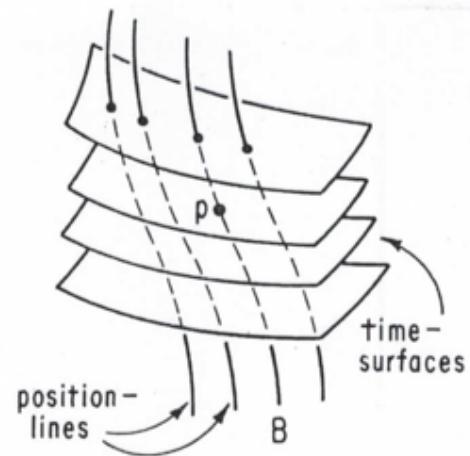


Fig. 59
The local Aristotelian setup of observer *B*.

Thus, we have two setups in the same region of space-time as shown below (last two figures superimposed).

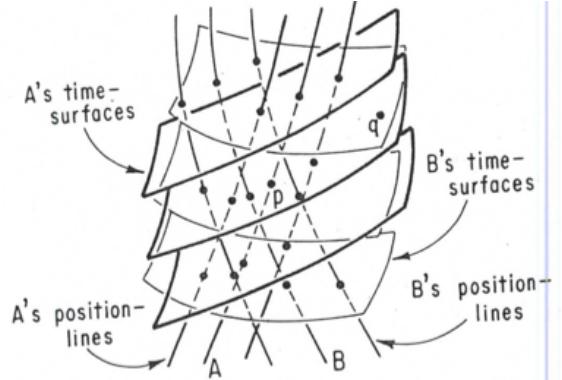


Fig. 60
Figures 57 and 59 superposed.

Here is what happens in the relativity view. The two observers disagree about position-lines and time-surfaces - they will disagree on all spatial and temporal relations for two events like *at the same time; elapsed time; at the same position; spatial distance; etc.* They only agree on the number of events, the speed of light and the interval $[(\Delta x)^2/c^2] - (\Delta t)^2$.

Consider the event q on the diagram. A says p and q on same time-surface - they are simultaneous; B disagrees - in fact B says that q came after p . Another observer might say q came before p . There is no agreement among observers about the time-order of p and q . As we will see later, this is so because p and q are spacelike related. If they were timelike related, then all observers would agree about time-order, although not about the value of the elapsed time. Two points on one of B's position-lines (means B says they are at the same position) clearly are on different position-lines according to A and thus A says they occur at different positions.

The only universal or intrinsic quantity is the interval between events.

Let us now look a some examples to illustrate individual interpretations and their relationships via the interval.

A and B have clocks and while standing around A sees B pass by at event p (of course B would say the same about A). We imagine B's clock reflects light (the image of the clock reading) and A receives that light showing the reading on the clock. We ask the question: What will A observe about the ticking rate of B's clock compared to the clock that A is carrying? This is shown below.

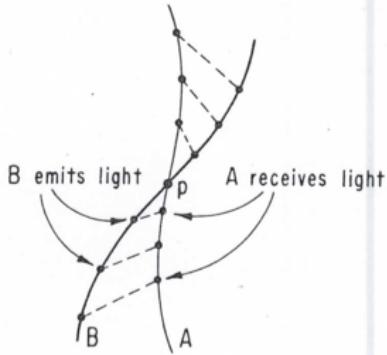


Fig. 61
two observers, A and B , meet at event p . Observer A receives light from B , by means of which B 's clock-readings are transmitted to A .

We now redraw this diagram keeping only the feature we need to do calculations. See figure below.

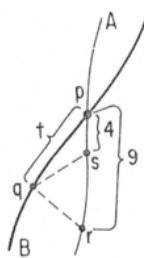


Fig. 62
The space-time diagram for the comparison of B 's clock as seen by A with A 's clock.

Fix an event q on B 's worldline. Event r on A 's worldline where a light pulse was emitted by A that intersected B 's worldline at q . Event s on A 's worldline is the intersection of the return light-pulse from B (from q). We assume as shown some particular elapsed time values, namely, A records 9 seconds between r and p and A records an elapsed time of 4 seconds between s and p .

What does A say? At s A will see whatever B 's clock was reading at q . At p the pass and A can read B 's clock directly. A measures 4 seconds passing between s and p on A 's clock. We denote by t (as shown) the elapsed time on B 's clock. What elapsed time (how many ticks) will A observe occurring on B 's clock between s and p ? Let us calculate the value.

We focus on the interval between p and q . According to A , we have the usual setup. A says that $t_1 = \text{elapsed time between } p \text{ and } s = -4$ and $t_2 = \text{elapsed}$

time between r and $p = 9$. Thus, the interval is $t_1 t_2 = -36 \text{ sec}^2$. For B, both events are on the clock worldline, thus the interval is $-t^2$. Since the interval is universal in the relativity view, we have $t = 6$. Thus, the elapsed time that B measures(directly) between q and p is 6 seconds. However, A says that only 4 seconds have elapsed between s and p . So A sees B's clock as speeded up by 50% (6 ticks in 4 *real* seconds).

What else can we say about this experiment as represented by the above figure? What would B say about the ticking rate of A's clock? We must now let B do the same experiment on A that A just did on B. At q B saw light emitted from A at r . Thus, at q , B sees what A's clock was reading at r . Then at p B can read A's clock directly. B sees A's clock go through 9 ticks while B's clock only ticks 6 times. So B says during 6 seconds, I saw A's clock undergo 9 ticks or I saw A's clock make 9 ticks in 6 *real* seconds - A's clock was speeded up by 50%.

Both observers say the same thing about the other observer! Who is right? Which clock is really speeded up? The answer is that there is not any *really* in relativity or physics in general! Each observer does their own thing (the agreed upon procedures) and they disagree - that is just the way it is in this view.

Using the same figure let us compute the speed of one observer according to the other. A knows that B experienced q and p directly. A then computes the spatial distance and elapsed time between these two events. A knows that $t_1 = -4 \text{ sec}$ and $t_2 = 9 \text{ sec}$ so that $\Delta x = c(t_1 + t_2)/2 = 5c/2$ and $\Delta t = (t_1 - t_2)/2 = -13/2 \text{ sec}$. (it is negative because q occurs before p according to A). Therefore A says B's speed is $\Delta x/\Delta t = 5c/13$.

So we have a scheme and a set of procedures that work. It is all based on the *universality of the interval*. One might argue at this point that we have replaced one baffling experimental observation (mu-mesons) with an equally baffling assumption about then interval. Why don't we *explain* these statements instead of asserting them, i.e., why does physics only explain *how* things work and not *why* things work the way they do? That is how physics works - we never answer the why question about assumptions - we leave that for the philosophers to argue about - usually we no definitive results.

Now for another example.

A emits a light-pulse and wants to measure its speed as the light goes away from her. Here are the details of the experiment she uses. At event p she will emit a light-pulse. A second event q along the worldline of the pulse is chosen. A has an Aristotelian setup nearby p and can determine the elapsed time and spatial distance between p and q . Dividing the spatial distance by the elapsed time then gives the speed of light. Considering the figure below

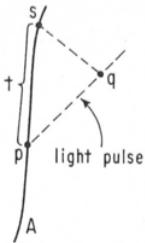


Fig. 64
Space-time diagram of an experiment in which A determines the apparent speed of light.

a retrun light signal from q reaches A 's worldline at s . Let the elapsed time, according to A , between p and s on her worldline be t as shown in the figure. In this case event r is the same as event p and thus $t_1 = t$ and $t_2 = 0$ (because $r = p$). We thus find $\Delta x = ct/2$ and $\Delta t = t/2$ so that the speed of light is $\Delta x/\Delta t = c$; This turns out to be a silly experiment, i.e., ask a silly question and you get a silly answer - A effectively put the speed of light c into the experiment and then pulled it out. Is there a more reasonable experiment?

Since our two measuring instruments incorporate the speed of light c in their operation, we must use another device in order to measure the value of c directly. Consider the following. We build standard meter sticks. Two individuals meet at the center of the meter stick and synchronize their clocks. they then separate, one going to each end of the meter stick. Reaching the ends, one of them emits a light-pulse and records the time of emission while the individual at the other end record the time of reception. They, then return to the center and compare their clocks. They divide 100 cm, the total light travel distance by the elapsed time in seconds and get the numerical value of the speed of light, i.e., $c = 100/(t - t')$ as shown below.

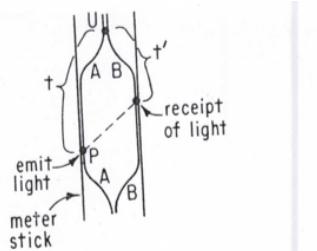


Fig. 65
Space-time diagram for an experiment in which the speed of light is measured directly, using a meter stick and clocks. Two observers, A and B , synchronize their watches, separate to opposite ends of the meter stick, and time the passage of light across the stick.

We now turn to measurements of lengths of objects. Consider the following situation. Observer A is standing around with a clock. A pole approaches her head on. First, the front of the pole passes her and then the rear of the pole passes her as shown below.

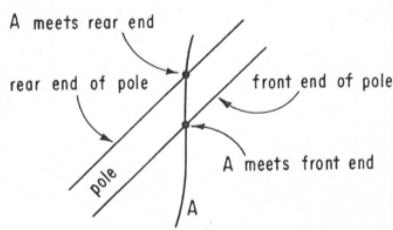


Fig. 66
Space-time diagram of observer *A* being passed by long pole.

The pole in space-time is defined by a world-surface bounded by the worldlines of the front and rear of the pole as shown. What does A measure for the length of the pole? This is the apparent length, according to A. We now detail several experiments that A might use to measure the length of the pole.

Experiment #1: First, A computes the apparent speed of the pole(say of the the rear) using the method discussed earlier. A space-time diagram of this experiment is shown below.

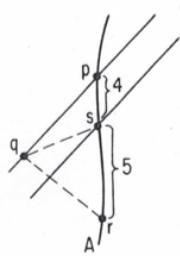


Fig. 67
One method for *A* to determine the "length" of the pole.
A computes the apparent speed of the pole and the apparent time taken by the pole to pass him.

Event *q* on the rear end of the pole is chosen so that a light-pulse from *q* just reaches A as the front end of the pole does (event *s*). Event *r* is chosen so that a light-pulse from *r* passes through *q*. We also choose some numbers for elapsed times as shown (equivalent to choosing relative speed of observers). Elapsed time, according to A, between *s* and *p* is 4 seconds and between *r* and *p* is 9 seconds so that the elapsed time between *r* and *s* is 5 seconds. What is the length of the pole?

We first compute the speed of the pole as discussed earlier. Since the numbers

are identical to the earlier case we have speed of pole = $5c/13$. Thus, the apparent length of the pole is (speed)(elapsed time between front and rear passing) = $(5c/13)(4) = 4.6 \times 10^{10} \text{ cm}$ (a very long pole!). That is the apparent length of the pole calculated using experimental measurements.

Another possible technique for measuring the length is shown below.

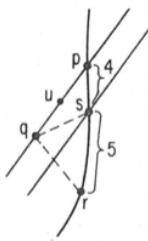


Fig. 68
A second notion of "length" of the moving pole. A determines an event u which he regards as simultaneous with event s , and then computes the apparent spatial distance between events u and s .

Here A identifies some event u on the rear end of the pole which A thinks is simultaneous with s (remember she has an Aristotelian setup with time-surfaces). u and s are at opposite ends of the pole and are simultaneous (so that pole has not moved between u and s). A then computes the spatial distance = apparent length. According to A, the elapsed time between s and p is 4 seconds. Thus, if u is simultaneous with s , then according to A, the elapsed time from u to p must be 4 seconds also. But A thinks the speed of the rear end is $5c/13$. Thus, A says that the apparent spatial distance between u and p is $4(5c/13) = 20c/13$. So we now have two events, which A thinks are simultaneous and we have A's apparent spatial distance between them, namely, $20c/13$. Thus, A says the length of the pole is $20c/13 = 4.6 \times 10^{10} \text{ cm}$. The same as before!

Let us now consider a third experiment as shown below.

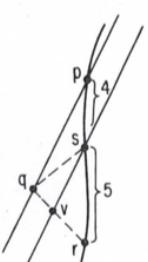


Fig. 69
A third notion of "length" of the moving pole. A effectively times the passage from one end of the pole (event v) to the other end (event q) and back (to event s).

A identifies the event v at which the light she sent to the rear end of the pole meets the front end of the pole. The argument now is - light, beginning at v (front end) travels to q (rear end) and then bounces back traveling the length of the pole again and returning to s (front end). If we can determine the total elapsed time (Δt) between v and s , then the spatial distance (Δx) or apparent length is $c(\text{total elapsed time})/2$. Now $\Delta x = (\text{apparent speed of front of pole})(\text{apparent elapsed time from } v \text{ to } s) = 5c/13 \times \Delta t$. We also have, according to A, that r and s occur at the same spatial position (on A's worldline). Thus $\Delta x = \text{apparent spatial distance between } v \text{ and } s$ also. In addition, the elapsed time between v and s is also Δt and since the apparent elapsed time from r to s is 5 seconds, the apparent elapsed time from r to v must be $(5 - \Delta t)$ seconds. Events r and v are light like related - therefore, $\Delta x = (5 - \Delta t)c$.

We now have two expressions for Δx and we thus obtain

$$\Delta x = (5 - \Delta t)c = 5c/13 \times \Delta t$$

or $\Delta t = 65/18 \text{ sec}$ = according to A, the elapsed time between events v and s . If we now compute the apparent length of the pole $c(\text{total elapsed time})/2$ we get $5.4 \times 10^{10} \text{ cm}$ which is different than the earlier results (17% larger). There is nothing wrong with the experiment in this case. The length calculated was the appropriate one for that experiment as were the others. Length has many different meaning in relativity. All methods would give the same answer if the pole was moving slowly; they give different answers if the pole is moving fast. Our everyday ideas and language are not appropriate for relativistic speeds!

We note that the *defined length or proper length* of an objects is the length measured by an observer at rest relative to the object and this corresponds to measuring the spatial separation between the ends of the object simultaneously(see Boccio or Mermin notes). The Geroch text does even more examples, but that is enough for us. It should be clear by now we can answer any question asked from the viewpoint of any observer.

3.7 Einstein's Equation: The Final Theory

We now have almost all the ideas needed to develop Einstein's general theory of relativity. We need only add one more ingredient (Einstein's equation) and rearrange all the ideas into a coherent picture. So let us first summarize what we have already done and put into the form of a physical theory along the way. The starting point will be the two books - one listing all possible events and the other listing, for every pair of nearby events, the interval between them. Instead of regarding these books, as earlier, as a summary of what the world is like (remember they were constructed by observers wandering around making measurements), imagine now a factory which constantly turns out 2-volume sets - one after another, all different from each other. The first book of each set lists a collection of points (labeled by letters). The second book of each set

lists pairs of points and for each pair, a number of seconds squared. The total output of this factory, over a long period of time, will be a large pile of 2-volume sets. Each set is called a space-time geometry. We make no commitment at this point as to any connection of the sets to our world, any other world, or anything physical. They are just a bunch of sets!

We now select, at random, one of the sets and give it to our friend in the closet who understands relativity. He then uses the 2 books to decide define the worldlines corresponding to light-pulses and clocks. He can then predict spatial and temporal intervals between any pair of events and so on. For every physical experiment he can think of he is able to make predictions about what he thinks happens (he assumes that is the way the outside world actually works).

Of course, this is silly. This was a random set of two books and thus, does not necessarily describe the real world. Each set is a model of a world but not necessarily the actual world. We need something else to find the correct set.

Now, based on our discussions so far, any physical theory seems to consist of three things:

- (1) the statement of a certain class of models
- (2) a collection of techniques by which, from a given model, one can make detailed physical predictions about the world
- (3) all physical predictions which flow via (2) from one particular model in the class (1) in fact agree completely with the actual physical experiment carried out in our world

Now the *beauty* or *aesthetic appeal* of a theory refers to how simple the statement in (1) is. There also may be some murmurings about as to how *natural looking* (2) is. The *technical simplicity* of a theory refers to how simple and straightforward the techniques in (2) are. The *generality* of a theory refers to how large the variety of physical predictions in (2) is. As more and more experiments are performed and the statement in (3) remains valid, we become more convinced of the validity of the theory.

Our present theory fits in as follows:

- (1) the models are the space-time geometries (output from factory)
- (2) the techniques for making predictions is everything we have been talking about
- (3) all the predictions which flow via (2) from one particular model in class (1) in fact agree completely with all outside experiments. This is the crux of the theory!

Unfortunately, there are a number of unpleasant feature of this theory. In particular, we have these:

- (1) There are more models in class (1) than we would like. Generally, the fewer the models the better the theory(remember (3) says there was only one that works). The worst case scenario would be if each experiment needed its own model!
- (2) The predictions from (2) are complete as far as spatial and temporal things(kinematic things) are concerned, but are very incomplete as far as dynamical things are concerned. For example, fix event p on the world-line of A. A decides that at p she will throw off a particle with a certain speed in a certain direction - this determines the direction of the world-line of the particle as it emerges from p . Now A issues further instructions that the particle, after release, must not be disturbed. It is to move freely however it wishes. All of these conditions will uniquely determine how the particle will move(that is what we observe in the actual world). We could observe the particle and then plot the events it experiences and thus obtain its subsequent (after release) worldline. This is not enough, however. We do not want to just represent what has happened, we want to be able to predict what will happen. We are unable to do this as yet, i.e., this cannot be done with only the two books. All that he could do is tell you all the possible timelike worldlines passing through p - he cannot tell you which will actually occur. Note that we can make some dynamical predictions, i.e., if the particle were a photon, then the two books are sufficient to predict its worldline.
- (3) The theory so far has nothing to do with gravitation. How is general relativity supposed to be a theory of gravitation?

So we have a theory with three serious objections. Normally, one would attempt to deal with each objection one at a time. In this case, however, Einstein was able to take care of all three at once! He found a certain equation which we call Einstein's equation that can be written symbolically(because the mathematics really necessary to write it down is very difficult) as shown below.

$$\left(\begin{array}{c} \text{Curvature of} \\ \text{space-time} \\ \text{geometry} \end{array} \right) = G \left(\begin{array}{c} \text{Mass density} \\ \text{of matter} \\ \text{in space-time} \end{array} \right)$$

We now explain the meaning of all parts of the symbolic equation.

First, consider the term *curvature of space-time geometry*. To set the stage we ask the question: How does one know that the surface of the earth is curved? We will only allow methods of explanation or measurements that operate entirely within the surface itself,i.e., we obviously could tell by getting *outside* the earth's surface - just go to the moon and look at the earth, etc, but in the case of space time - we cannot get outside space-time to look at it and see what is happening.

We might imagine the earth inhabited by small, very flat, 2-dimensional ants, which crawl about the surface. These ants cannot move off the surface or be influenced by anything off the surface. Can such ants determine whether or not the surface of the earth is curved? To see how to proceed, we ask a more general question. What is the sum total of all the information (of a geometrical character) that these ants could accumulate about the surface of the earth?

They can locate points on the surface of the earth and they could measure distances between pairs of points and nothing else. Thus, they can acquire information which is already in the two books. Thus, we ask again - can the ants determine (or can we determine from only the two books) whether or not the surface of the earth is curved? The answer is yes.

Consider the triangle on the earth as in the figure below.

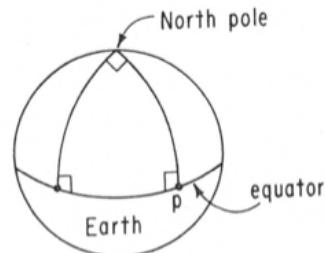


Fig. 74

A "triangle" on the surface of the earth. The "sides" are a portion of the equator, and two longitudes. All three angles are right angles.

The ants could draw such a triangle and they could determine the angles of the triangle. In this example all the angles are 90° so that the sum of the angles is 270° . If you remember your Euclidean geometry, the sum of the angles of a triangle drawn on a flat plane is 180° . Clearly the surface is not a plane! More geometry says that the angle difference is proportional to the area of the triangle with the proportional constant being the square of the radius of the sphere. Thus, the ants can measure the radius without ever leaving the surface(they need to learn some geometry).

Like ants on the surface of the earth, we cannot leave space-time. We have access only to events in space-time and intervals between them if they are nearby. Thus, as is the case with the ants, we are able to use information in the two books to determine the curvature of space-time. This says, in effect, that the information in the two books can be used to introduce a quantity which is the *curvature of the space-time geometry* itself. The mathematics involved is very complicated but that is not important. We only need to know it is possible. This quantity representing curvature, which can be calculated using the two books, is the quantity which appears on the left-hand side of Einstein's equation.

It is important to not try to visualize *curvature* in 4 dimensions as any real analog of curvature of surface in 3-dimensions. Our curvature is a mathematical quantity which is zero for flat surfaces and not zero for non-flat surfaces. That is all that is important. The curvature of space-time is not a constant (like the curvature of a sphere). It varies with space and time - space-time can have regions of positive, zero and negative curvature.

On the right-hand side is the constant G , which is Newton's gravitational constant and $G = 6.7^{-8} \text{ cm}^3/\text{gm} - \text{sec}^2$. We are left now with the quantity we called *mass density of space-time*. This quantity is essentially what its name suggests. It expresses how much matter exists in any given region of space-time. It also varies over space-time.

Einstein's equation requires that the curvature of the space-time geometry equals the constant G times the mass density of matter in space-time. This is different equation for each event in space-time (since all quantities vary with space and time). The equation must hold for every event in space-time.

Let us now see how we solved the three problems mentioned earlier.

First, the imposition of Einstein's equation will cut down on the number of two book sets they are valid (they must satisfy the equation at every event). and thus cut down on the number of potentially valid models. For each we could construct a third book as follows. Look up an event p in the first book. Use the second book to compute the curvature of the space-time geometry at p . Divide the curvature by G to obtain the *mass density of the matter* at p and record in the third book. Thus, Einstein's equation can be used to predict the mass

density from the curvature! It is also possible to use Einstein's equation to go the other way - to calculate the curvature from the mass density! Einstein's equation is the fundamental thing all by itself. The equation asserts that a distribution of matter requires a certain curvature in the space-time geometry - matter causes curvature in the space-time geometry.

An analogy, which should not be pushed too far, is shown below where a rubber sheet represents the geometry of space-time and its shape the curvature of space-time. Introducing matter as shown distorts the surface and changes the curvature and changes the path of motion of a steel ball released on the sheet.

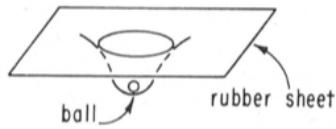


Fig. 76
An analogy to Einstein's equation. Matter, placed on the rubber sheet, causes curvature of that sheet.

Second, Einstein's equation allows us to make dynamical predictions in a subtle and elegant way. Remember, that we allowed observer A at event p to emit a particle which was free to move as it wished thereafter. We observed, however, that the particular followed a unique worldline determined by the initial conditions at emission. We did not earlier have any means to make the prediction of a unique worldline. Einstein's equations fixes this problem. Let us fix the space-time geometry and consider a number of possibilities for how the emitted particle might move as shown below.

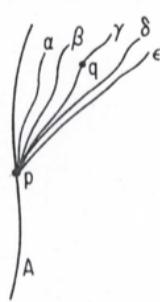


Fig. 77
Given event p , and the initial direction of a world-line from p , there are many world-lines having that direction at p . Which of all these possibilities will actually be realized by a thrown particle?

For each of the five possibilities shown we can construct a three volume set of books. The first two books will always be the same since we are in every case dealing with the same events and the same intervals(remember we fixed the space-time geometry). The third book in each of the five sets will be different, however. A particle has mass and thus its introduction into space-time represents some mass density. Consider for example, the event q . The sets to represent situations α , β , δ , and ϵ will all say, in their third books, *event q - mass density zero* for in these situations there is no particle experiencing event q . The set representing γ , however, will say, in the third book, *event q - mass density (not zero)* for in situation γ there is a particle experiencing event q . Only one set of books can be labeled *satisfies Einstein's equation*. In this way, Einstein's equation restricts what matter is allowed to do - it thus offers us the possibility for some dynamical laws.

What if we use the reverse interpretation of Einstein's equation - above we used space-time geometry determines distribution of matter; now we use the matter causes curvature interpretation. In this case, we think of all the space-time events as having been given a priori, and the matter distribution as having been given also. If we look in the third book now we can determine the worldline of the emitted particle by finding those events that have a non-zero mass density associated with them and collect them together into the worldline. If we now impose Einstein's equation, it says that the intervals cannot be chosen arbitrarily, but rather must be chosen so that the resulting curvature of the resulting space-time geometry satisfies Einstein's equation, i.e., is equal to the (known, in this case) right side of the equation. So now Einstein's equation no longer acts to restrict the actual worldline of the particle, rather it acts to restrict the intervals, given that worldline.

We now require that A make physical statements about what the particle is doing, in particular, the particle worldline and its corresponding intervals(which determine how clocks tick, how light-pulses move and so on - that is, how A's instruments, that she will use to make measurements, actually work). Here is how Einstein's equation now works. First, it cannot do anything about the worldline, but it cleverly adjusts what the intervals will be so that A ends up making the *correct* (physically realized) statements about what the particle is doing. Once again Einstein's equation provides dynamical information about how the particle moves - more precisely, information about what A will have to say about its motion.

From either approach we get the same result. On imposing Einstein's equation, one loses the right to specify both the space-time geometry and the the worldline of the particle(or equivalently the mass density associated with the particle) at will. The two things are connected via Einstein's equation. The result of this connection is that the dynamical behavior of physical phenomena is now tied down to space-time itself.

Let us again make a physical analogy. Consider the rubber sheet and let it be in some curved configuration as shown below.

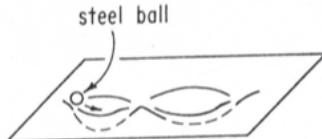


Fig. 78
An analogy to Einstein's equation. Curvature (of the rubber sheet) causes dynamical effects of matter.

The configuration was created by some distribution of matter placed on the sheet. Let us now place a small steel ball on the sheet. The particular curvature dictates how the ball will subsequently move on the sheet. This reaction to curvature is what, within the theory, dictates how matter moves about. Remember, do not push the analogy too far!

What about the third difficulty? Where is gravitation in all of this? Now, Einstein's equation can be interpreted as requiring that *matter causes curvature in space-time* and that it can also be interpreted as requiring that *matter move in certain ways in response to curvature in space-time*. Gravitation thus arises as follows. Let there be two massive bodies in the world. Then, according to Einstein's equation, each will cause a certain amount of curvature in space-time. Further, according to Einstein's equation, each body will be forced to move in a certain way in response to the curvature caused by the other. The net result, then, is that each body influences the other. This influence is what we call gravitation. The curvature acts as an intermediary between the two bodies - it is the gravitational field.

We can now state what the final theory is. First, the models, which consist of the three volume sets which satisfy Einstein's equation. The physical interpretation of these models follow the lines we have been discussing - techniques, computations, descriptions, interpretive remarks, etc. The physical interpretations associated with a model correspond to the physical phenomena observed in real experiments. Gravitation enters the theory because of the limitations on valid models. This, then, is the general theory of relativity.

Some thoughts. Theories consist of an enormous number of ideas, arguments, hunches, vague feelings, value judgments, and so on, all arranged in a maze. The various ingredients are connected in a complicated way. All of this (the nebulous mass) together is the theory. When presented however, the theory must be laid out in some linear fashion, consisting of one point after another, each connected in some more or less direct way to its predecessor. One learns the theory in the linear way and one proceeds to form one's own view of the totality involved - one creates one's own nebulous mass taken as a whole. You rearrange the

points, make new connections, use hunches and vague feelings and so on. One usually isolates certain points and calls them *postulates*. There is no proof of the theory attempted in any manner!

After that, what do relativists do? Some spend time trying to figure out which of the models corresponds to the real world by comparing prediction for small systems to experiments. Others attempt to find better techniques for extracting information from models that can be compared to experiments. Still others are always looking for new theories - to replace Einstein's equation - because in the end all theories fail at some point and must be replaced.

What are possible objections to the theory? In the small region of space-time nearby any observer, we do not see any evidence that it is curved. Where is the curvature? It is a matter of scale or a question of how curved. The surface of the earth is curved but you do not notice it or need to take it into account when you throw a ball, etc. If you fired a rocket and ignored the curvature, you would miss however. In the entire solar system the effects of curvature are still very small but are now observable as we mentioned earlier. In the galaxy the effects are even more dramatic and on a cosmological scale the curvature effects dominate everything. In regions of high mass density, curvature effects also dominate everything. This case is the so-called black hole, which we will now consider.

3.8 An Example: Black Holes

One of our models - one of the three volume sets - one that satisfies Einstein's equation - is the so-called *black hole*. Let us look at what the books say about this solution and then look at whether this corresponds to anything in the real world.

The first task, it is to describe what space-time is, i.e., express the contents of the first book (listing the events) and the second book (listing the intervals). For the events we draw a picture in 3-dimensional space as shown below.

We are not saying that the events are in ordinary 3-dimensional space, but only that we can characterize them using the diagram. In the picture we have a vertical cylinder and its axis. At this point the cylinder and the axis are just guideposts to help us locate things and describe what is going on geometrically. Any physical interpretations will come later. We label different regions of the diagram as follows:

1. The region outside the cylinder is called the *external region*
2. The cylinder itself is called the *horizon*
3. The region inside the cylinder (excluding the axis) is called the *internal region*
4. The axis is called the *singularity*

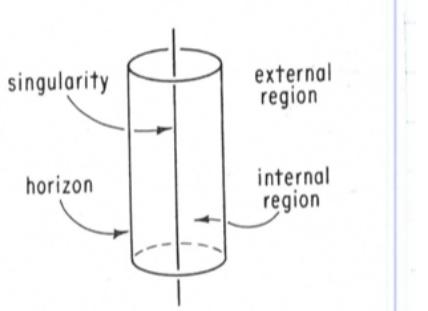


Fig. 80
The set of events for a black-hole space-time. We divide the events into three classes: those in the external region, on the horizon, and in the internal region.

These names are only for describing where a point lies at this time and we should not attach any other meanings. The actual names will be justified eventually.

The figure summarizes the contents of the first book. Now to deal with the second book. We need to be able to concisely summarize the content of the second book without listing every interval for every pair of nearby points. This is done as follows. Fix some point p in this space-time. The second book then lists every point q near p , the interval between p and q . We use this information to determine the light-cone of point p - the locus of all points q which have zero interval from p . Thus, nearby intervals imply *local* (=near p) light-cones. We then attach to each point in the diagram a small cone to represent the local light-cone for that point. These cones represent some but not all of the information in the second book.

What can we tell about intervals if we only know these local light-cones?

1. If q has zero interval from p , i.e., p and q are lightlike related. In this case q will lie on p 's light-cone or p will lie on q 's light-cone.
2. If q and p are timelike related, then q lies inside p 's light-cone or p lies inside q 's light-cone.
3. If q and p are spacelikelike related, then q lies neither inside nor on p 's light-cone or p lies neither inside nor on q 's light-cone.

Thus, we can determine, using just light-cones, the sign of the interval between two nearby events - whether they are spacelike, timelike or lightlike related. The interval gives more information than just its sign, but the rest of the information is difficult to represent on the diagram. Leaving out this information will not limit our ability to make physical interpretations of the resulting space-time geometry.

Now we attach to each point in the figure a small light-cone to represent the information from the second book as shown below.

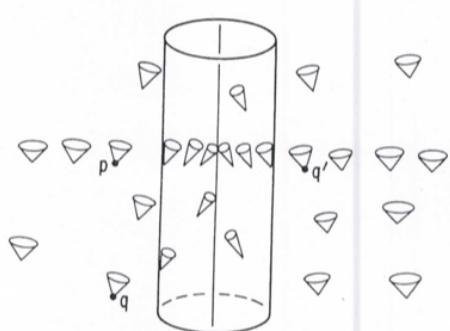


Fig. 81
The light-cones for the black-hole space-time. The cones are "vertical" in the distant external region, become tangent to the horizon, and then, in the internal region, lean toward the singularity. It is intended that the figure have the symmetries of translation up and down, and rotation about the axis.

Note, we are not in any way justifying the choices at this point. We are just summarizing the contents of the second book, which is supposed to represent the *black hole* solution for Einstein's equation. In the diagram, the light-cones in the external region far from the horizon are standard light-cones (like we would have drawn in special relativity or flat space). They sit up straight. As one moves in towards the horizon, however, the light-cones begin to do two things: they become narrower and they begin to tip in toward the horizon. This *narrowing and tipping* continues all the way to the horizon. The light cones of points on the horizon are so tipped that the cones are tangent to the cylinder. As one proceeds into the internal region, the light cones continue to become narrower and continue to tip. As one approaches the singularity, finally, the cones have become very narrow and have tipped 45° in the diagram. The figure only has light-cones for a few representative points. The entire picture remains the same on vertical translation up or down. Thus, light cones at p and q are the same (they are on the same vertical line). Thus, we really only need to know the light-cones on one horizontal plane. The entire picture remains the same under rotation of the diagram about the axis - the light-cones at p and q' are related by a rotation about the axis. Thus we only need to know the light-cones on a horizontal line from the axis to infinitely far away.

This is our space-time geometry or part of it since we have only shown the points and signs of the intervals.

we need to comment about points on the axis. Although these points exist in the 3-dimensional space we are using to represent space-time, they are not real events in this particular space-time. These points are not in the first book. The reason is the following. Any point p must have a *unique* light-cone. The light-cones from all directions tip over to 45° as one approaches the axis. Since we can approach a point p on the axis from any direction, the limits of the light-cones must be the same on the axis no matter what direction we approach from - this is not the case - thus the axis is not in space-time. This is shown in the figure below.

Thus, as we approach the axis, the light-cones have a *singular behavior* - hence the name *singularity* for the axis.

We now need to apply the great variety of techniques we have discussed for extracting physical prediction from a given space-time geometry to the particular geometry called the *black hole* - this will generate physical predictions about black holes. Everything is now fixed by the space-time geometry. We can no longer add any new properties about black holes - we can only generate allowed consequences of the model. If the predictions do not agree with experiment, we discard the model and start again with a different model.

Now worldlines of observers are timelike lines ($v < c \rightarrow$ inside light-cones) Three typical observer worldlines are shown in the figure below.

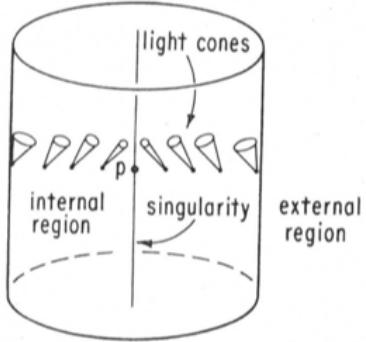


Fig. 82
More detailed view of the light-cones near the singularity.

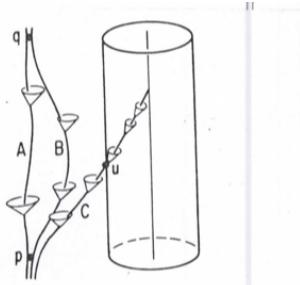


Fig. 83
Three observers in our black-hole space-time. Observer *A* remains in the distant external region, *B* ventures somewhat closer to the hole, and *C* passes through the horizon to the interior of the hole.

We have superimposed these observer worldlines into the space-time of the black hole. The worldline A represents an observer who keeps well away from the black hole. If A were orbiting around the black hole the worldline would be a spiral as in part (A) of the figure below

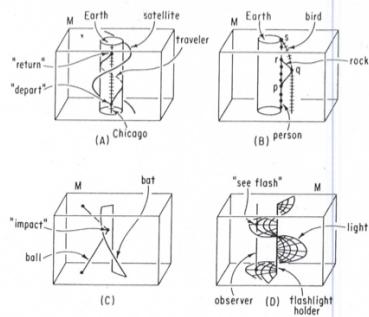


Fig. 14
Four examples of the interpretation, in the Aristotelian view, of more complicated space-time diagrams.

Observer B has somewhat more curiosity about the black hole. She begins with A. At event p , however, she decides to take a closer look for herself. She thus goes in (closer to the horizon) while remaining in the external region. Eventually her curiosity wanes and she goes back to A joining him at event q . Observer C is overwhelmed by curiosity about this black hole. He goes directly toward it and then, at event u , crosses the horizon. He then remains in the internal region for an interval and finally hits the singularity(his worldline seems to end at the singularity). Note that all the worldlines are time like since they remain inside the local light-cones.

Remember there is no event available to C at the singularity - there is no possibility for further extension of C's worldline after it arrives at the singularity. What then does it mean physically to say C's worldlines hits the singularity? Mathematically, C's worldline just ends. Physically, this means that C is *snuffed out of existence* - after some finite time according to himself, he ceases to exist in space-time.

From C's point of view things are easy to describe -he exists at his clock reading 1, at his clock reading 1-1/2, at his clock reading 1-3/4, at his clock reading 1-7/8 and so on, but simply does not exist at clock readings of 2 or greater. C will never say I am experiencing an event and my clock reads 2. Does C disappear before one's(other observers) eyes? Not necessarily. This last question corresponds to an actual experiment that might be performed on C (where a second observer receives light-pulses from C and interprets things about C in terms of those received pulses). In order to predict what would happen in such an experiment we need to draw a diagram that contains these light-pulses and let the diagram tell us where the pulses go and thus what we will see. We will investigate such visual experience experiments later - we cannot say anything or answer such questions without doing an experiment - that is the way physics works.

Let us suppose that C, at some point on his worldline, was handed the space-time diagram for this black hole. He can tell that if he does not change what he is doing, then he will be *snuffed out*. Clearly, that is not what he wants so he decides to stifle his curiosity and return to the safety of A. Will he be able to do so? It depends on when he makes the decision to try to return to A. Three possibilities are shown below.

If C makes the decision at event v (in the external region), then, even though the light-cones have started to tip, the back side (side away from horizon) is not yet vertical (this occur first at the horizon). At event v C has the choice(by accelerating) of having his worldline go in any direction in the local light-cone of event v . Clearly, he can choose a new worldline they is going outward toward A (as shown). As C moving along this new worldline, the light-cones straighten up and it becomes even easier to get to A. So if the decision is made at event v all is OK.

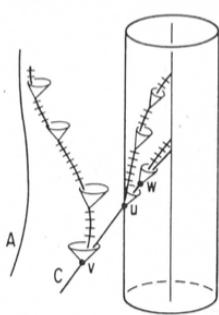


Fig. 84
The result of C's deciding, at various events along his world-line, that he wishes to return to the external region.

Suppose, however, C waits until event u to try to adjust his worldline. The new worldline must, however lie inside the local light-cone of event u . But at event u the light-cone is already tangent to the horizon. Thus, any direction inside the light-cone is going to take C into the internal region and closer to the singularity and farther from A. So one moment later, no matter what he does, C is going to be in the internal region. In the internal region all light-cones are tipped event more toward the singularity and thus, C has no choice but to continue moving toward the singularity. So a decision made at u is already too late - he will be snuffed out no matter what he does. A decision at event w is even worse for C.

We summarize as follows. If you are in the external region, then you always have the option of remaining in the external region (no matter how close to the horizon your worldline may go). Crossing the horizon, however, is a disaster. Once the horizon is crossed, no matter what you do, you will go further into the internal region and eventually your worldline ends on the singularity. Crossing the horizon can be experienced only once!

Let look in more detail at what is happening as C approaches and crosses the horizon. The figure below

shows C's worldline if he did nothing to change it. At event v C remembers he has a dentist appointment at event q in the external region. C does not do anything at v to change his worldline however. How long can C wait before deciding to do something and still make the appointment? Let us assume that the elapsed time along C's worldline between v and u (where C would cross the horizon if he did nothing) is 10 seconds. So if C waits 10 seconds, then all is lost - he could not make any appointments in the external region after reaching u . Suppose then he waits 9.5 seconds after v (event s) and at that point decides to go back to A. C can change his worldline to any worldline within the local light-cone at s . s is so close to the horizon, however, that the back side of the tipped light-cone is almost vertical. Thus, beginning at s , C cannot make very much headway towards A. He can, by accelerating very strongly, slow but

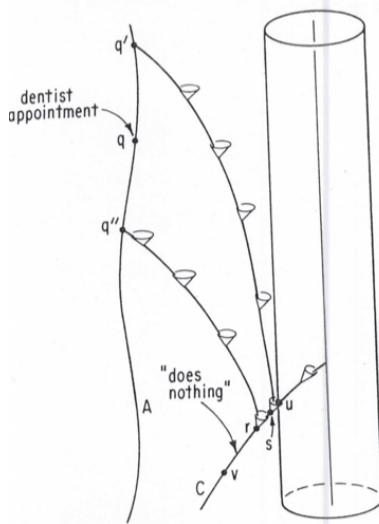


Fig. 85
The nearer to event u that C arrives at his decision to return to the external region, the later C arrives in this region.

surely move away from the horizon and into regions where the light-cones are closer to straight up and he can make more headway toward A. Unfortunately, C wasted a lot of time in almost pure vertical motion on the diagram and can at best only make it back to A at event q' - missing his appointment. If he made his decision earlier say at 8 seconds after v (event r) where the light cones are not very tipped over, he can make lots of headway toward A right from the beginning(while remaining inside the local light-cones) and arrive at A at event q'' and thus make his appointment. Thus, the last few seconds before u are crucial. The closer to u where C makes his decision, the longer it will take him to get back to A - it might take years or centuries if C was very close to u .

Let us now consider another experiment as shown below.

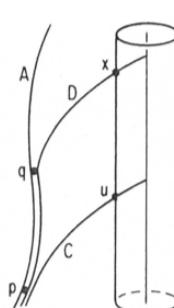


Fig. 86
A space-time diagram representing two individuals, C and D , who enter the black hole, D later than C .

We have A remaining in the external region far from the horizon. C starting trip at p and going through the horizon and getting *snuffed out*. and C's friend D who starts her trip to cross the horizon later than C at q . Being a friend and knowing that C will soon be *snuffed out*, D wants to join C for his last few moments(commit suicide together - experimentalists are weird people). Is it actually possible for D to join C? From the diagram, the clear answer is no. This is because D's worldline must always remain inside the local light-cones along her worldline.

We now modify the experiment again as shown below.

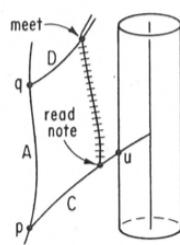


Fig. 87
The result of C's deciding, before event u , to return to D.

A's worldline is the same. D's worldline is the same. D, however, hand C a note at p indicating that she would not begin the trip across the horizon until event q . If C reads the note before u , then he could change his world line and meet D (as shown) where they could decide together whether to stay in the external region or plunge together across the horizon to the singularity. More interesting, however, is if C does not open the note until event u as in the figure below.

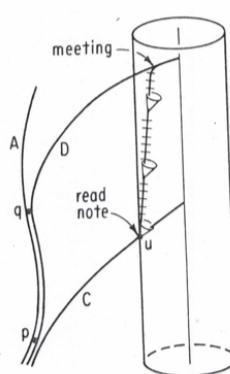


Fig. 88
The result of C's deciding, at event u , to return to D. Now,
C and D are only able to meet within the internal region.

C can, in fact, still meet with D before the end. The changed worldline shown

is perfectly valid, i.e., C's worldline is always inside the local light-cone. He can travel a very large vertical distance (toward D) on the diagram for a very small horizontal distance (toward singularity) due to the nature of the light-cones.

Finally, we consider a last case where C opens the note at event v as shown below.

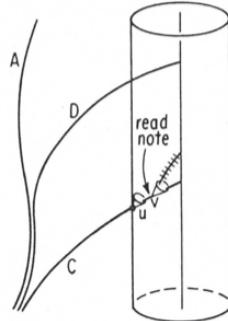


Fig. 89
The result of C's deciding, after event u , to return to D.
In this case, C is unable to do so.

At this event, the light cones have tipped so much that C no longer can reach D before the end (remember he needs to stay inside the local light-cones)

. Note that, given a space-time geometry (the black hole in this case) we have had all of these discussions and made all of these physical interpretations based only on the idea that observer worldlines have to be inside the local light-cone. We now use another idea to make further interpretations. In particular, we use the fact that the worldline of a light-pulse is a lightlike line. Using this we can draw within the diagrams the worldlines of various light-pulses. Since we see things visually with light-pulses, we can then discuss what physical phenomena will look like visually to different observers.

We consider the situation shown below.

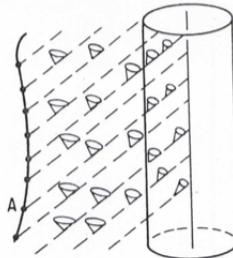


Fig. 90
Space-time diagram representing observer A's emission of light, which enters the black hole.

Here A again remains in the external region. Now A continually sends out light-pulses inward toward the black hole. These worldlines are tangent to the local light-cones as shown. They all go through the horizon to the singularity. To make the experiment interesting we now introduce observer C as shown below. We consider the situation shown below.

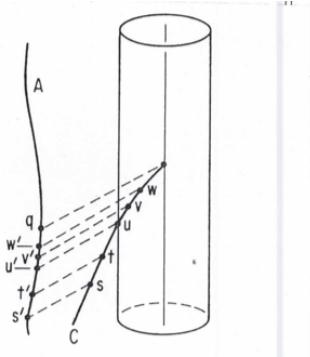


Fig. 91
Light from observer A as seen by observer C, who enters the black hole. C is able to maintain surveillance of A during his journey.

We consider events s , t , u , v and w on C's worldline and the corresponding events s' , t' , u' , v' and w' on A's worldline. The light-pulse emitted by A at s' reaches C at s , the light-pulse emitted by A at t' reaches C at t , and so on. What will C see visually if he looks at A during his trip into the hole? As indicated, at successive events C will experience the light emitted from successive events on A's worldline. This is what happens everyday when you look at others moving around. Nothing strange happens here either even as C crosses the horizon. Nothing dramatic happens even as C approaches the singularity. He no longer sees anything after he is snuffed out - what did you expect?

C could watch A for a longer period by altering his worldline inside the internal region as shown below.

But still nothing dramatic occurs. So nothing strange for the observer falling into the black hole *seeing* other observer outside the black hole who are emitting light-pulses. We now consider a different situation. The light-pulses will be emitted by C as he goes into the black hole. We draw these light-pulses as shown below.

The behavior of the light-pulses emitted by C is more interesting. Those emitted from an event like v , in the external region, just go out into the external region. The closer the point of emission on C's worldline is to the event u at which C crosses the horizon, the *farther up* the diagram the light-pulse goes before it eventually gets away from the hole (the dashed lines are the lightlike worldlines of the light-pulses).

What happens to the light-pulse emitted at u ? the worldline of this pulse is

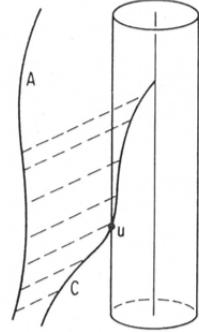


Fig. 92
By accelerating away from the singularity just after event u , C is able to watch A 's activities for a longer time.

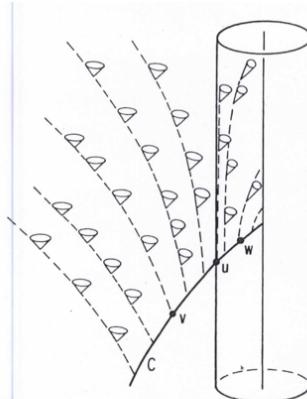


Fig. 93
Light-rays as emitted by an observer C who enters the black hole.

just the vertical line tangent to the horizon (it is a lightlike line since the local light-cones are also tangent to the horizon). Finally, we indicate the worldlines of the light-pulses emitted by C after he crosses the horizon at u . These pulses just go into the singularity. The longer after u the more directly the pulse goes into the singularity. Some, as indicated might take a long time to get into the singularity (if emitted just after crossing the horizon).

We can now determine what someone (A) in the external region watching C visually would see (see figure below).

A receives the signals at a succession of events on his worldline from a succession of events on C 's worldline as shown. No light from C reaches A from event u and after on C 's worldline. Thus pulses either remain on the horizon or go to the singularity. Thus, A is only able to see C during C 's stay in the external region. Does A see C disappear somehow as C reaches u ? Not really.

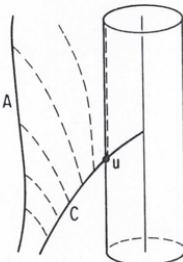


Fig. 94
A watches C by means of the light received from C. In this case, A remains outside the hole, while C goes in. The result is that A, at any of his moments, can still see C, but that C's activities after entering the hole are forever hidden from A.

The point is that, although A may watch C forever (according to A), all A will see visually is C getting closer and closer to the horizon, i.e., at some event A might receive the light C sent out 1 second before u (according to C). An hour later A will be receiving light C sent out at say 0.3 seconds before u . A day later A will be receiving light C sent out say 0.02 seconds before u and so on. Even after hundreds of years, A will still be receiving light sent by C just before u (say 0.0000007 second according to C before u). A never sees C crossing the horizon, never sees C in the internal region and certainly never see C reaching the singularity. According to A, C (when watched visually) seems to slow down as time goes on. This is clear from the diagram since it takes forever for A to see C's last second before C experiences u . Thus for A, years and years seem to pass while A only sees visually C going through a small fraction of a second of his life. Eventually, C appears to be frozen according to A.

Let us now introduce another observer D who stays with A until event q also watching C as shown below.

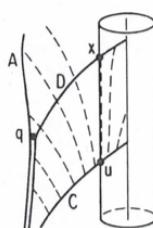


Fig. 95
Observer D, by deciding to enter the black hole, is even able to watch C's entry into the black hole.

Eventually D leaves A and go toward the horizon and singularity as shown. D wants to see C crossing the horizon and entering the internal region. D understands that he must enter the internal region to accomplish this. As D crosses the horizon at event x he sees the light emitted by C at u (where C crossed

the horizon). As D proceeds into the internal region he receives light from C in the internal region as shown. D is unable to see C actually hit the singularity. There is a critical point on C's worldline between u and the singularity from which a light-pulse just manages to make it to D as D reaches the singularity (so it never actually reaches D).

So D sees the following. As D begins to race toward the black hole, D sees C speeding up again (before q D thought C was slowing down). As D crosses the horizon, he sees C crossing the horizon. Thereafter, D sees C for a while in the internal region and finally D gets snuffed out at the singularity without ever seeing C get snuffed out.

Thus, nobody ever sees the final moments of C's life except someone accompanying C (on the same worldline); all light emitted just before reaching the singularity just goes directly into the singularity and thus the light itself is snuffed out before anyone has the opportunity to experience it.

As a final example of the visual effects, we consider the situation where A (remains in external region) simply looks into the black hole. What will A see? A will receive all light-pulses that reach his worldline as shown in the figure below.

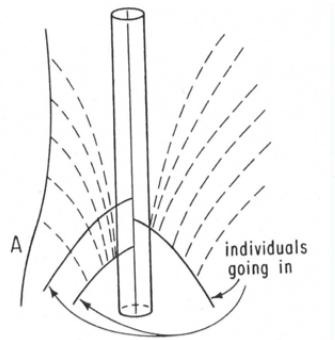


Fig. 96
A's visual impressions on looking at the black hole from the external region. A is able, at all times, to see every individual who has ever entered the hole.

The pulses come from nearer and nearer the horizon the further back along the pulse one goes - thus A sees visually everybody that has gone into the black hole (on A's side). Those having gone in most recently will be seen farthest from the horizon and will appear to be moving the fastest. Those who went in a longer time ago will appear to be closer to the horizon and will appear to be slowed down more. Nobody who ever went into the hole will be totally invisible to A. The black hole thus gives a visual record of every observer who ever plunged through the horizon. No information as to who did and did not go into the black hole is ever lost to A.

We now know why a black hole is so named. Black refers to the fact that no light ever comes out of the hole itself - no light-pulse ever passes from the internal region to the external region - the hole sends out no light - it appears black. Hole refers to the fact that observers (such as C) can go into the object (through the horizon and into the internal region) but no observer can come back out (internal to external regions). The horizon itself is a *one-way membrane* - in but not out.

Some other conclusions that could be drawn from these diagrams if we add to them numerical information about the interval on the worldlines. They are as follows. In the figure below

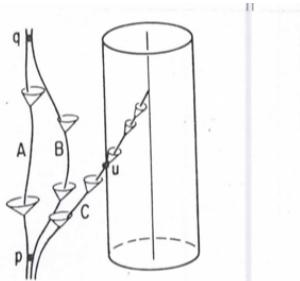


Fig. 83
Three observers in our black-hole space-time. Observer A remains in the distant external region, B ventures somewhat closer to the hole, and C passes through the horizon to the interior of the hole.

B would experience a smaller elapsed time between p and q than A would. In the figure below

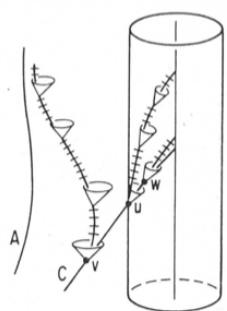


Fig. 84
The result of C's deciding, at various events along his world-line, that he wishes to return to the external region.

C would experience a smaller elapsed time between u and the singularity by taking the railroad-track worldline than by taking the direct route. In the figure below,

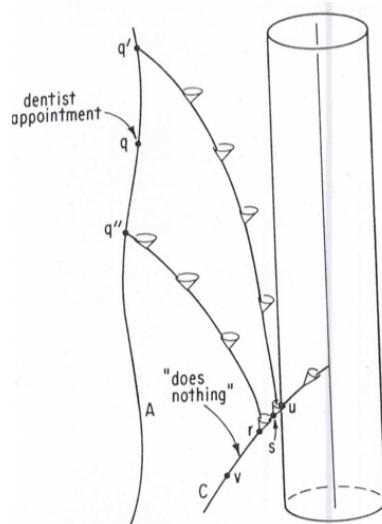


Fig. 85
The nearer to event u that C arrives at his decision to re-turn to the external region, the later C arrives in this region.

C, by delaying the start of his return journey to A until he is very near u can cause his apparent elapsed time in the return journey to A to be as small as he wishes, even though A may say that C was away for a very long time.

As a final example we consider the following. The world-surface of a rope is a two-dimensional surface. We again let A remain outside the hole while C ventures in. This time, however, C ties one end of a long rope around his waist and A is instructed to feed out rope as C goes into the hole. This is shown below.

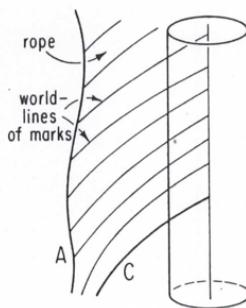


Fig. 97
Observer A lowers C into the hole by means of a rope, and continues to feed out the rope at all later times. Included in the figure are the world-lines of "marks" made on the rope. The rope, then, is continually being fed into the singularity.

We assume that marks have been made at regular interval along the rope and the worldlines on the diagram are those of the marks. Note how new marks (new worldlines) continually appear from A's worldline as the rope is fed out. The worldlines of the marks are timelike and after they cross the horizon they proceed to the singularity (C reaches the singularity first with his end of the rope and the rest of the rope is continually pulled into the singularity snuffing out successive marks).

Now, at event q in the figure below

A becomes worried about C and decides to try to pull him out of the hole using the rope. So, A pulls on the rope at q . After q , worldlines associated with rope marks no longer emerge from A - they instead come back to A as A pulls in the rope. Now any marks worldlines that have already crossed the horizon (are in the internal region). Thus, there will be fewer marks available between A and the hole. Physically, this means that the rope is being stretched. Thus, A will have to pull harder and harder on the rope in order to continue reeling it in, i.e., in order to continue to receive on his worldline additional marks. Eventually, presumably, the rope will break as shown in the diagram. One end will just continue falling into the black hole, eventually being absorbed by the singularity, while A will be able to recover the rest as illustrated in the diagram. Thus, A will not be successful in pulling A out of the black hole using the rope.

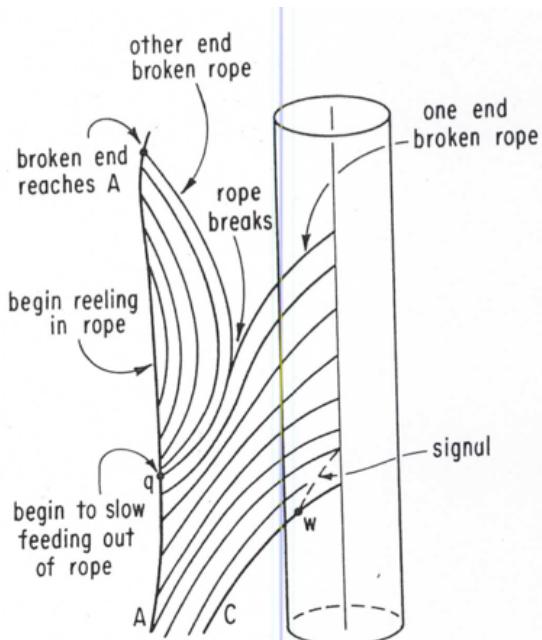


Fig. 98

A space-time diagram of what happens if *A* decides to retrieve *C* by pulling on the rope. Rather than *C*'s returning, the rope breaks.

That completes the general discussion of the black hole space-time geometry. Experimentally, black holes are now known to exist - both isolated, in binary systems with other stars, and very large ones at the center of galaxies.

We have now presented Einstein's theory of relativity. Along the way we have seen how physics, especially theoretical physics works.

Chapter 4

Notes on Weinberg

The First Three Minutes

4.1 Introduction: The Giant and the Cow

There exist endless numbers of ancient myths and stories about the origin of the universe. Most present the reader with new problems to answer equal in number to the answers they purport to give. The complications that are raised can be summarized in the words - complicated initial conditions or why or how did the conditions necessary for the story to work occur in nature - do we always need a god to intervene to make things work, and so on.

Science and scientists have worried about the question from the beginnings of science. No progress could be made in any scientific explanations because the experimental data was non-existent and there were no theoretical foundations that could be applied. This all changed in the latter half of the 20th century. A theory of the early universe now exists - the so-called big-bang theory (modified by inflation) along with specific recipes for the contents of the universe. This is the theory we will be discussing.

We start with a summary of the history of the early universe as it is presently understood by modern theory. We will explain all the details during our discussions.

It all started with an *explosion*. This is not any ordinary explosion as might occur today, which would have a point of origin(center) and would spread out from that point. The *explosion* we are proposing occurred simultaneously everywhere, filling all space(infinite in extent or cyclic) from the beginning, with every particle of matter rushing apart from every other particle.

At about 1/100 of a second, the earliest time that could be talked about in 1977(we will push that number closer to zero later with modern developments),

the temperature of the universe was about 10^{11} degrees Centigrade. This is much hotter than the center of even the hottest star - no ordinary components of matter as we know them - molecules, atoms, nuclei - could hold together at this temperature. At this time, the matter that was rushing apart consisted of electrons(mass and negative charge), positrons(anti-electrons)(mass and positive charge), neutrinos(3 kinds and almost massless; no charge) and photons(massless; no charge). The number and average energy of the photons in the early universe was about the same as for electrons, positrons and neutrinos. These particle - electrons, positrons, neutrinos and photons - were continually being created out of pure energy and then after short lives being annihilated again. The numbers that existed were not set initially but fixed by a balancer between processes of creation and annihilation. A calculation then shows that the density of this cosmic soup at a temperature of 10^{11} degrees Centigrade was about 4×10^9 times the density of water!. There also existed a small contamination of heavier particles (protons and neutrons) - about one proton and one neutron for every 10^9 photons or electrons or positrons or neutrinos. As we will see later this ratio is measurable in the so-called cosmic radiation background we will discuss later.

As the *explosion* continued, the temperature dropped reaching 3×10^{10} degrees Centigrade after about 1/10 of a second; 10^{10} degrees Centigrade after about one second; and 3×10^9 degrees Centigrade after about 14 seconds. At this point the temperature was low enough that electrons and positrons began to annihilate faster than they could be recreated from photons and neutrinos. The extra energy released in the annihilation processes temporarily slowed the cooling process but the temperature continued to drop reaching 10^9 degrees Centigrade after about 3 minutes. At this point the temperature was low enough for protons and neutrons to begin to form complex nuclei such as deuterium and then to form the most stable of light nuclei, helium.

At the end of three minutes the contents of the universe were mostly in the form of photons, neutrinos and antineutrinos. There was a small amount of nuclear material - 73% hydrogen(protons) and 27% helium(alpha particles) along with an equally small number of electrons left over from the electron-positron annihilation processes. The matter continued to rush apart becoming steadily cooler and less dense, Much later, after a few hundred thousand years, it became cool enough for electrons to join with nuclei to form atoms of hydrogen and helium. The resulting gas would begin to form clumps under the influence of gravitation(where did that come from?) which would ultimately condense to form the galaxies and stars of the present universe. All the ingredients for these processes was created in the first three minutes!

There is much vagueness in this theory about the details of the beginning (less than 1/100 of a second). Also some initial conditions seem to need precise values without any explanation of how and why. We will see have modern developments straighten out these difficulties later. The reason the big-bang theory is

preferred over all others is because of experimental evidence supporting it - in the end that is the most important thing! We embark on a process where we will fill in the details - experimental evidence and theoretical considerations. After understanding the period beyond 1/100 of a second will will endeavor to update Weinberg's work to include *inflation theory* so that we can talk about the period before 1/100 of a second. In the end experimental observations will have be our main guide for how to proceed - that is the way of theoretical physics. Finally, we will look at the future of the universe - what will happen!

4.2 The Expansion of the Universe

Look at the night sky. On the short time scale of our lives it looks like a unchanging universe! Other than local movements within our local solar system, the stars seem motionless. The stars, however, are moving, with speeds as high as a few hundred kilometers per second. Thus, in a year a fast star might travel on the order of 10^{10} km . The distance to the nearest star is 4.2 ly or

$$4.2 \text{ ly} \frac{c \cdot 1 \text{ year}}{\text{ly}} \frac{365 \times 24 \times 3600 \text{ seconds}}{1 \text{ year}} = 1.19 \times 10^{17} \text{ km}$$

So this so-called *proper motion* of a star is a negligible fraction of the distance between stars(0.00001%). That is why it looks like things are not changing! A relatively fast star is Barnard's star ($56 \times 10^{12} \text{ km}$ away moving at 89 km/sec or $2.8 \times 10^9 \text{ km}$ per year so that its apparent position shifts in one year by an angle of 0.0029 degrees as shown below.



Proper motion of Barnard's Star: The position of Barnard's star (indicated by white arrow) is shown in the upper photograph. The lower photograph shows the position of Barnard's star relative to the highest background star is readily apparent. In the 122 years between the two photographs, the star has moved less than the "proper motion" of 0.17 minutes of arc per year. (Yerkes Observatory Photograph)

Even this extreme case is just detectable! Other images of the seemingly unchanging universe are shown below.



Figure 4.1: The Milky Way in Sagittarius



Figure 4.2: The Spiral Galaxy: M104



Figure 4.3: The Andromeda Galaxy: M31

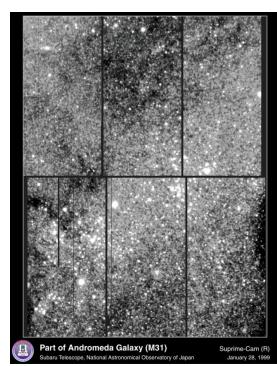


Figure 4.4: Details of the Andromeda Galaxy: M31

This appearance of an unchanging universe is an illusion! Observational evidence(as we will see) indicates that the universe is in a state of violent explosion where galaxies are rushing apart at speeds approach the speed of light. If we extrapolate their motion back in time, then in the past everything was much closer - in fact so close, they could not have had separate existence - this is the so-called *early universe* that we will be discussing to start with.

Observational evidence for the expansion of the universe comes from the fact that it is possible to determine the motion of a luminous body in a direction along a line of sight more accurately than motion at right angles to the line of sight. This ability relies on the Doppler effect that we have discussed earlier. We found these results earlier. If a source of light waves has a period of T seconds and the source is moving away from the observer at velocity V , then the period T' seen by the observer is given by

$$\frac{T'}{T} = \sqrt{\frac{c+V}{c-V}}$$

Since wavelength and period are related by $\lambda = cT$ and $\lambda' = cT'$ we then have

$$\frac{\lambda'}{\lambda} = \sqrt{\frac{c+V}{c-V}}$$

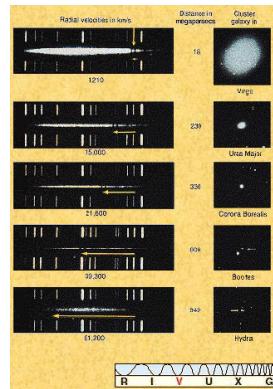
For $V \ll c$ we then have

$$\frac{\lambda'}{\lambda} \approx 1 + \frac{V}{c}$$

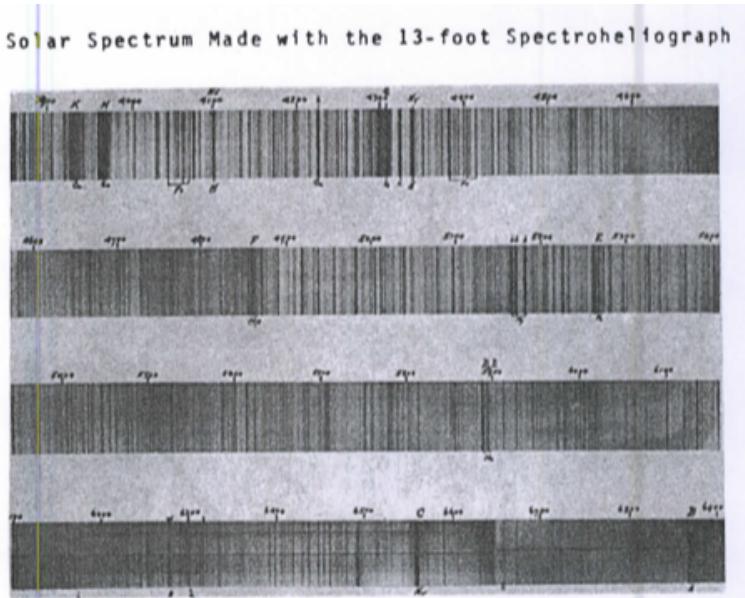
which corresponds to a shift of wavelengths toward the red end of the spectrum. If the source were moving toward the observer instead we have

$$\frac{\lambda'}{\lambda} \approx 1 - \frac{V}{c}$$

which corresponds to a shift of wavelengths toward the blue end of the spectrum. Some examples are shown below.



If one looks at the spectrum of a star(the sun) we have as shown below



The Spectrum of the Sun: This photograph shows light from the sun, broken up into its various wavelengths by a 13-foot focus spectrograph. On the average, the intensity at different wavelengths is about the same as would be emitted by any totally opaque (or "black") body at a temperature of 5800° K. However, the vertical dark "Fraunhofer" lines in the spectrum indicate that light from the sun's surface is being absorbed by a relatively cool and partly transparent outer region, known as the reversing layer. Each dark line arises from the selective absorption of light at one definite wavelength; the darker the line, the more intense the absorption. Wavelengths are indicated above the spectrum in Angstrom units (10^{-8} cm). Many of these lines are identified as due to absorption of light by specific elements, such as calcium (Ca), iron (Fe), hydrogen (H), magnesium (Mg), sodium (Na). It is partly through the study of such absorption lines that we can estimate cosmic abundances of the various chemical elements. Corresponding absorption lines in the spectra of distant galaxies are observed to be shifted from their normal positions toward longer wavelengths; it is from this red shift that we infer the expansion of the universe. (Hale Observatories Photograph)

The Doppler effect is dramatically important for astrophysical observations when it is applied to the study of individual spectral lines. Fraunhofer had already discovered that when light is passed through a slit and then a glass prism, the resulting spectrum of colors is crossed by hundreds of dark line each one an image of the slit (see above). The dark lines each correspond to a definite wavelength. They are identical for many different sources of light and are produced by selective absorption of light of definite wavelengths as the light passes through the stellar atmosphere. Each line corresponds to the existence of a specific chemical element in the atmosphere. The actual shift in these dark lines due the movement of the source relative to the observer together with a known spectrum for the same source at rest allows one to determine the speed of the source.

When you look at the night sky two objects of great cosmological importance can been seen. As seen in Fig 2.1 above, we have the band of light called the Milky

Way galaxy stretching across the sky. It represents a flat disk of stars (we are in the disk). Seen from outside we have a similar galaxy in Fig 2.2 (M104) seen edge-on. Finally in Figs 2.3 and 2.4 we have our nearest neighbor galaxy - the Andromeda galaxy (M31) and a close-up showing stellar details. Such galaxies are of the order of 100,000 light-years in diameter, of order 10,000 light-years in thickness, have typical masses of order 10^{11} solar masses (even more when we include dark matter, if it exists). Massive telescope were eventually able to distinguish spiral and elliptical galaxies. Within these galaxies are variable stars (Cepheid variables) which have a definite relationship between their period and their absolute luminosity, which is the total radiant power emitted by an astronomical object in all directions. Absolute luminosity is observable only at the star. Away from the star, we only measure apparent luminosity, which is the radiant power received at the distant detector (reduced by distance - $1/r^2$ fall off - the so-called inverse square law. With Cepheid variables, we measure the period T , which tells us the absolute luminosity $L_0 = f(T)$. We then measure the apparent luminosity $L = L_0/d^2$, where d = the distance between us and the star and thus we can determine d .

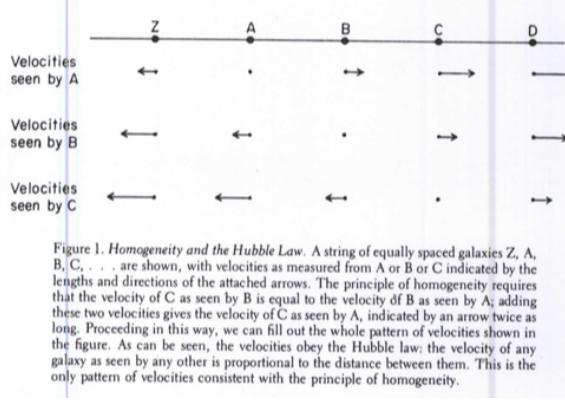
$$L = \frac{L_0}{d^2} = \frac{f(T)}{d^2} \rightarrow d = \sqrt{\frac{f(T)}{L}}$$

This type of calculation showed that most observed galaxies were outside our own galaxy (it was thought early on that they were just objects within our galaxy). Thus, it was found that an enormous number of galaxies like our own fill the universe to great distances in all directions.

At the same time the light from these galaxies was also measured to have shifted spectra indicating the galaxies were moving away or toward the earth. Some examples are shown in a figure above. At first it was thought that these might be merely relative velocities, reflecting a motion of our own solar system toward some galaxies and away from others. However, this explanation quickly became untenable as more and more of the larger spectral shifts were discovered - all toward the red end of the spectrum(moving away). It seemed that, except for a few galaxies in our local neighborhood, all other galaxies are generally rushing away from our own. This does not place our galaxy at some special place of central importance - it seems to say that the universe has undergone some sort of explosion in which every galaxy is rushing away from every other galaxy.

In 1929, Hubble discovered that the red shift of galaxies increases roughly in proportion to their distance from us. It does not matter who the *us* is - the universe looks the same from the viewpoint of any observer - all the galaxies are rushing away with speeds proportional to distance. - this is the so-called Cosmological Principle(valid in the universe on the large scale - about equal to the distance between clusters of galaxies - $10^8 ly$). In fact, one can use the principle to show that the relative speed of any two galaxies must be proportional to the distance between just as Hubble discovered from observations.

To see this consider the figure below



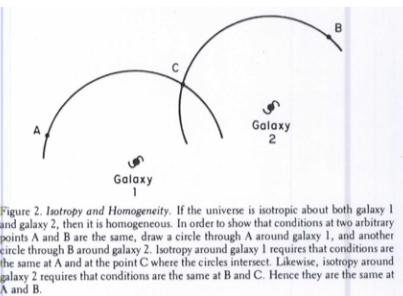
Look at the three typical galaxies A, B, and C strung out along a straight line as in the figure. Suppose that the distance between A and B is the same as the distance between B and C. Whatever the speed of B as seen from A, the Cosmological Principle requires that C should have the same relative speed relative to B. C, however, is twice as far away from A as is B, is also moving twice as fast relative to A as is B and so on for any other galaxies in the chain. Alternatively we have

$$\frac{\lambda'}{\lambda} = 1 - \frac{V}{c}$$

$$z = \text{redshift} = \frac{\lambda' - \lambda}{\lambda} = \frac{\lambda'}{\lambda} - 1 = \frac{V}{c}$$

so that $z = \frac{V}{c} \propto \text{distance}$. Philosophers love this result - they would ask - why should any part of the universe or any direction be any different from any other? The answer, observationally, seems to be - they are not!

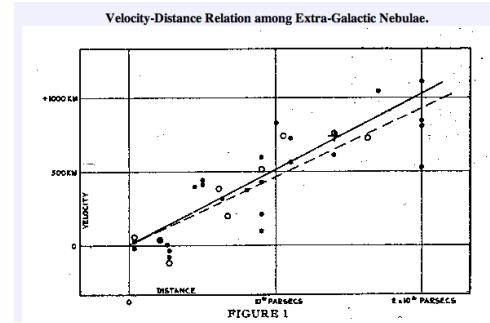
So there is nothing special about mankind's location in the universe. The universe is *isotropic* around us(our galaxy) and it must be isotropic around any other galaxy. Now, any point in the universe can be carried to any other point by a series of rotations around fixed centers as shown below.



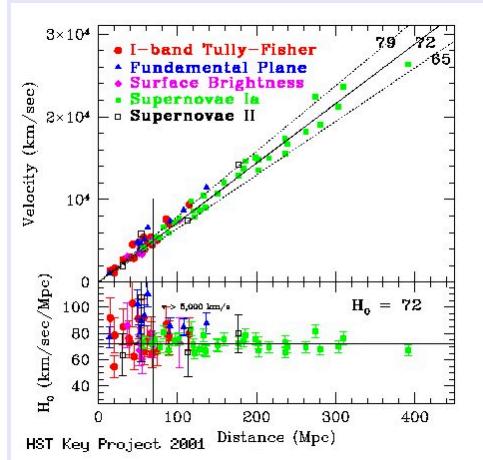
so if the universe is isotropic around every point, it is necessarily also homogeneous!

We also note that we used the Newtonian velocity addition rule in our discussion so the result is only valid for $V \ll c$. This was OK for Hubble, because none of the galaxies he observed were moving very fast. At some point in our discussions we will have to incorporate the relativistic velocity addition rules we have derived so that velocities do not exceed the speed of light, i.e., we will have use general relativity. Long before that the entire concept of distance will begin to fail and we will have deal with that also.

Hubble found distances using apparent luminosity of Cepheid variable stars and velocities using Doppler shifts and concluded (1929) that there is a *roughly linear relation between velocities and distances*. His actual data as shown below



would not, however, allow you to draw such conclusions - he knew what he wanted to be true and drew the corresponding conclusions. Modern data does not have any problem as shown below.



So Hubble drew the correct conclusions, but using a very dangerous and almost

dishonest process, which in many other similar cases has led science down blind allies and disastrous paths. The proportionality constant is called the *Hubble constant* H .

What does all this say about the origin of the universe? If everything is rushing apart, then at some earlier time they must have been closer together. If the velocities had been constant throughout this time, then the time taken for any pair of galaxies to reach their present separation is just the present distance divided by their relative velocity. With velocity proportional to present separation this time would then be the same for all pairs of galaxies - they must have all been close together at some time in the past! Using the Hubble constant of $H = 15 \times 10^6 \text{ km/sec}/10^6 \text{ ly}$ the time since the galaxies began to move apart would be $1/H = 20 \times 10^9 \text{ years}$. This number(age) calculated from the Hubble constant is called the *characteristic expansion time*. Since, as we will see, the galaxy velocities have not been constant for all time(they have been slowing down), the true age of the universe must be less than the *characteristic expansion time* for this value of the Hubble constant.

This result is usually taken to mean that the universe size is increasing. All this means is that the separation distance between any pair of distant galaxies is increasing. We want to avoid thinking of the universe as some volume in 3-dimensional space and thus being mislead about what is really happening. We note at this point, that we must be careful about drawing conclusions from one set of experiments. It is difficult to measure velocity versus distance(uncertainties about extragalactic distance scales or even what they mean are the main concern), redshifts may come other sources (gravity) than just motion, etc. The calculated age need to be confirmed in other unrelated experiments and so on. One example is the age of stuff in our galaxy is estimated to be about $10 - 15 \times 10^9 \text{ years}$ from measurements of the relative abundances of radioactive isotopes and from calculations concerning stellar evolution. There is no connection between these new results and redshifts and since the calculated ages are very similar, the presumption is strong that the age of the universe deduced from the Hubble constant really does represent a true beginning. These ideas then are the beginning of the so-called big-bang cosmology as a theory of the universe.

Our picture of the universe so far is one of an expanding swarm of galaxies. Light has only played the role of a messenger telling us about the distances and velocities of the galaxies. It turns out, however, that conditions in the early universe were very different - light was the dominant constituent of the universe - ordinary matter was a negligible contamination. Let us therefore look more closely at the redshift in terms of the behavior of light waves in an expanding universe. This will be very important later on.

Consider a light wave traveling between two galaxies. The separation between the galaxies equals the light travel time times the speed of light and the increase

in the separation during the light transit time equals the light transit time times the relative velocity of the galaxies. If we calculate the fractional increase in separation we have

$$f = \frac{\text{increase in separation}}{\text{mean separation during increase}} = \frac{v_{rel} \times t_{transit}}{c \times t_{transit}} = \frac{v_{rel}}{c}$$

that is, the transit time cancels out!. This same ratio, as we saw earlier, gives the fractional increase in wavelength of light during the journey. Therefore, *the wavelength of any light ray increases in proportion to the separation between the galaxies as the universe expands*. If the wavelengths of light appear to stretch (confirmed by observation), then we conclude that the universe itself is also large by the same amount, i.e., if the wavelength of light increases by 10% during its transit from some galaxy to earth, then during the same time the universe has gotten 10% larger.

So far we have only been concerned with *kinematic* effects - the description of the motion without any consideration of the forces that produce the motion. Physical theory also wants to understand the dynamics of the universe - in this case that is a study of the cosmological role of the gravitational force between astronomical bodies. Newton was the first to tackle this problem. He thought that matter had to be distributed throughout an infinite space, which is so difficult to work with that it stifled all progress until Einstein developed general relativity as we have discussed. Einstein's equations admitted solutions that predicted the existence of a redshift proportional to distance - the so-called *de Sitter model*. It was probably this theory that pushed Hubble to say his data had the same property. The problem was that solutions of this type were *static* - unchanging and that is not the universe we now know from observations (they did not know back then). We need solutions which are not static, but are homogeneous and isotropic. Finally in 1922 Friedmann found such a solution to Einstein's equations - it provided the mathematical background for most modern cosmological theories.

There are two types of Friedmann models. If the average density of the matter of the universe is less than or equal to a certain critical value, then the universe must be spatially infinite. In that case, the present expansion of the universe will go on forever. On the other hand, if the density of the matter of the universe is greater than this critical value, then the gravitational field produced by matter curves the universe back on itself - it is finite though unbounded, like the surface of a sphere. In this case, the gravitational fields are strong enough eventually to stop the expansion of the universe - the universe will ultimately implode back to indefinitely large density. This critical density turns out to be proportional to the square of the Hubble constant which is about $5 \times 10^{-30} \text{ gm/cm}^3$ or about 3 hydrogen atoms per cubic meter of space.

The motion of a typical galaxy in the Friedmann models is just like that of a stone thrown upward from the surface of the earth. If the stone is thrown fast

enough or equivalently if the mass of the earth is small enough, then the stone will gradually slow down but will nevertheless escape to infinity. This corresponds to the case of a universe density less than the critical density. On the other hand, if the stone is thrown with too little speed, then it will rise to a maximum height and then plunge back downward, which correspond to the universe density above the critical value. In these models the galaxies are not rushing apart because some mysterious force is pushing them apart (stone is not being repelled by the earth) - the galaxies are moving apart because they were thrown apart by some sort of *explosion* in the past!

Many of the detailed properties of the Friedmann models can be calculated without using the full blown general relativity. In order to calculate the motion of a galaxy relative to our own, draw a sphere with our galaxy at the center and the other galaxy on the surface. The motion of the galaxy on the surface of the sphere depends only on the mass of the matter inside the sphere - the outside matter has no effect! This result was known to Newton and still hold in Einstein's theory (Birkhoff's theorem - see below). The only requirement for the theorem to hold is homogeneous and isotropic matter distribution.

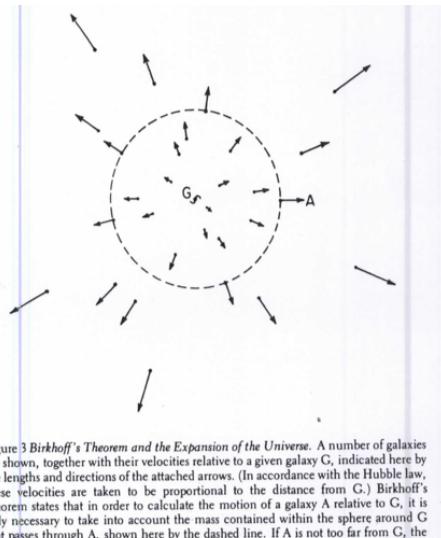


Figure 3 Birkhoff's Theorem and the Expansion of the Universe. A number of galaxies are shown, together with their velocities relative to a given galaxy G , indicated here by the lengths and directions of the attached arrows. (In accordance with the Hubble law, these velocities are taken to be proportional to the distance from G .) Birkhoff's theorem states that in order to calculate the motion of a galaxy A relative to G , it is only necessary to take into account the mass contained within the sphere around G that passes through A , shown here by the dashed line. If A is not too far from G , the gravitational field of the matter within the sphere will be moderate, and the motion of A can be calculated by the rules of Newtonian mechanics.

We can use the theorem to calculate the critical density of the Friedmann models. As in the figure above, we can use the mass within the sphere to calculate the velocity that the galaxy on the surface would have to have to escape to infinity. This escape velocity is proportional to the radius of the sphere (larger radius means larger mass inside the sphere so that escape velocity is larger). On the other hand, Hubble's law says that the actual velocity of the galaxy is proportional the radius of the sphere (distance from us) also. Therefore the ratio of the actual velocity to the escape velocity does not depend on the radius of

the sphere. Therefore, depending on the values of the Hubble constant and the cosmic density every galaxy which moves according to Hubble's law will either exceed escape velocity (go to infinity) or be less than escape velocity(fall back toward us). The critical density is the value of the cosmic density at which the escape velocity equals the Hubble law velocity. Let us derive the result now.

For a sphere of galaxies of radius R we have a total mass (using cosmic density ρ)

$$M = \frac{4\pi R^3}{3} \rho$$

The gravitational potential energy of any galaxy (mass m) at the surface of the sphere is

$$U = -\frac{mMG}{R} = -\frac{4\pi mR^3\rho G}{3} , \quad G = 6.67 \times 10^{-8} \text{ cm}^3/\text{gm sec}^2$$

The velocity of this galaxy is given by Hubble's law as $V = HR$. Therefore the kinetic energy is

$$K = \frac{1}{2}mV^2 = \frac{1}{2}mH^2R^2$$

The total energy is then

$$E = K + U = mR^2 \left[\frac{1}{2}H^2 - \frac{4}{3}\pi\rho G \right]$$

which is a constant during the motion of the galaxy. The condition $E = 0$ correspond to just enough energy to escape to infinity. Thus the critical density condition is

$$\frac{1}{2}H^2 - \frac{4}{3}\pi\rho_{crit}G = 0 \rightarrow \rho_{crit} = \frac{3H^2}{8\pi G}$$

This result, although calculated without general relativity, is still valid in a relativistic universe if the density ρ is interpreted as the total energy density divided by c^2 instead of just the mass density. Putting in numbers we find $\rho_{crit} = 4.5 \times 10^{-30} \text{ gm/cm}^3$.

The detailed time dependence of the size of the universe can also be worked out - very mathematical. For once let me show you the equations from GR:

$$R\dot{\rho}c^2 = -3(\rho c^2 + p)\dot{R} , \quad p = \text{pressure}$$

The pressure is given by $w\rho c^2$ where $w = 0$ for ordinary matter and $w = 1/3$ for radiation. The results of solving these equations is shown below.

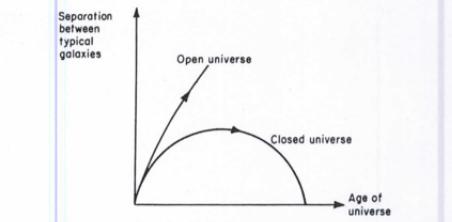
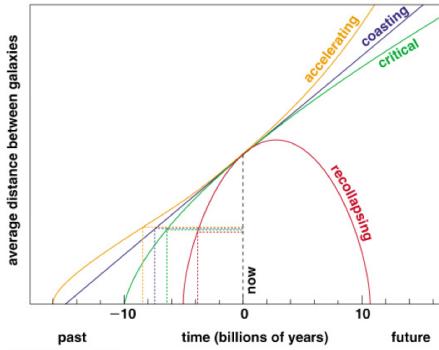


Figure 4. *Expansion and Contraction of the Universe.* The separation between typical galaxies is shown (in arbitrary units) as a function of time, for two possible cosmological models. In the case of an “open universe,” the universe is infinite; the density is less than the critical velocity; and the expansion, though slowing down, will continue forever. In the case of a “closed universe,” the universe is finite; the density is greater than the critical density; and the expansion will eventually cease and be followed by a contraction. These curves are calculated using Einstein’s field equations without a cosmological constant, for a matter-dominated universe.

and even fancier



Note the three solutions - open, closed and flat universes. There is one result that will be very important to us later which we will derive now.

If we go back to the total energy result we derived earlier and now put in the time dependence we have

$$E = mR^2(t) \left[\frac{1}{2}H^2(t) - \frac{4}{3}\pi\rho(t)G \right]$$

where now all the variables take on their values at the particular time t . The energy, however, must be constant. As we will show shortly, as $R(T) \rightarrow 0$, $\rho(t)$ increase at least as fast as $1/R^3(t)$, so that $\rho(t)R^2(t)$ grows at least as fast as $1/R(t)$ as $R(T) \rightarrow 0$. In order to keep the energy constant, the two terms in the bracket must nearly cancel. Therefore for $R(T) \rightarrow 0$ we have

$$\frac{1}{2}H^2(t) \rightarrow \frac{4}{3}\pi\rho(t)G$$

The characteristic expansion time is the reciprocal of the Hubble constant or

$$t_{exp}(t) = \frac{1}{H(t)} = \sqrt{\frac{3}{8\pi\rho(t)G}}$$

Now let us return to see how the density ρ varies with distance R .

If the mass density is dominated by the mass of nuclear particles (the matter-dominated era), then the total mass within a comoving sphere of radius $R(t)$ is proportional to the number of such particles within the sphere - this must remain constant.

$$\frac{4}{3}\rho(t)R^3(t) = \text{constant}$$

which says that

$$\rho(t) \propto \frac{1}{R^3(t)}$$

If, instead, the mass density is dominated by the mass equivalent to the energy of radiation (radiation-dominated era), then $\rho(t)$ is proportional to the fourth power of the temperature (Stefan-Boltzmann law). The temperature, however, varies like $1/R(t)$, so then

$$\rho(t) \propto \frac{1}{R^4(t)}$$

In order to be able to deal with both cases in our discussions we will write

$$\rho(t) \propto \frac{1}{R^n(t)}$$

with

$$n = \begin{cases} 3 & \text{matter-dominated era} \\ 4 & \text{radiation-dominated era} \end{cases}$$

Our earlier conclusion that $\rho(t)$ increases at least as fast as $1/R^3(t)$ is now clear.

Now the Hubble constant is proportional to $\sqrt{\rho}$ and therefore

$$H(t) \propto \frac{1}{R^{n/2}(t)}$$

The velocity of a typical galaxy is then

$$V(t) = H(t)R(t) \propto \frac{1}{R^{1-n/2}(t)}$$

Since the velocity is the derivative of the distance, the mathematics of calculus says that if the velocity is proportional to some power of the distance as it is above, then the time it takes to go from one point to another is given by

$$t_1 - t_2 = \frac{2}{n} \left[\frac{R(t_1)}{V(t_1)} - \frac{R(t_2)}{V(t_2)} \right] = \frac{2}{n} \left[\frac{1}{H(t_1)} - \frac{1}{H(t_2)} \right]$$

Finally we have

$$t_1 - t_2 = \frac{2}{n} \sqrt{\frac{3}{8\pi G}} \left[\frac{1}{\sqrt{\rho(t_1)}} - \frac{1}{\sqrt{\rho(t_2)}} \right]$$

Therefore, whatever the value of n , the time elapsed is proportional to the change in the inverse square root of the density. For example, during the who of the radiation-dominated era after the annihilation of electrons and positrons, the energy density was given by

$$\rho = 1.22 \times 10^{-35} [T(\text{°K})]^4 \text{ gm/cm}^3$$

with $n=4$. Therefore the time required to cool from 10^8 degrees to 10^7 degrees was $t = t_1 - t_2 = 1.90 \times 10^6 \text{ sec} = 0.06 \text{ years}$.

We can derive a more general result. The time required for the density to drop to a value ρ from a value very much greater than ρ is

$$t = \frac{2}{n} \sqrt{\frac{3}{8\pi\rho G}} \begin{cases} t_{exp}/2 & \text{radiation-dominated era} \\ 2t_{exp}/3 & \text{matter-dominated era} \end{cases}$$

Thus, for example, at $3000^\circ K$ the mass density of photons and neutrinos was

$$\rho = 1.22 \times 10^{-35} [3000]^4 \text{ gm/cm}^3 = 9.9 \times 10^{-22} \text{ gm/cm}^3$$

This is much less than the density at $10^8^\circ K$ that the time require for the universe to cool from very high early temperature to $3000^\circ K$ is given by ($n = 4$), $t = 2.1 \times 10^{13} \text{ sec} = 680000 \text{ years}$.

Finally, since the density is proportional to $1/R^n$, the time is proportional to $R^{n/2}$ or

$$R \propto t^{2/n} = \begin{cases} t^{1/2} & \text{radiation-dominated era} \\ t^{2/3} & \text{matter-dominated era} \end{cases}$$

Now one way to tell whether or not galactic velocities exceed the escape velocity is to measure the rate at which they are slowing down. If this deceleration is less(or greater) than a certain amount, then the escape velocity is (or is not) exceeded. In practice, this means that one must measure the curvature of the graph of redshift versus distance for very distant galaxies as shown below.

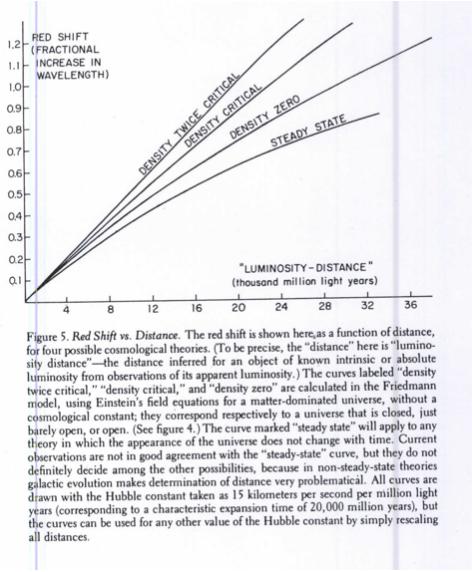


Figure 5. Red Shift vs. Distance. The red shift is shown here as a function of distance, for four possible cosmological theories. (To be precise, the "distance" here is "luminosity distance," or distance inferred for an object of known intrinsic or absolute luminosity from observations of its apparent luminosity.) The curves labeled "density luminically" "density critical," and "density zero" are calculated in the Friedmann model, using Einstein's field equations for a matter-dominated universe, without a cosmological constant; they correspond respectively to a universe that is closed, just barely open, or open. (See figure 4.) The curve marked "steady state" will apply to any theory in which the appearance of the universe does not change with time. Current observations are not in good agreement with the "steady-state" curve, but they do not definitely decide among the other possibilities, because in non-steady-state theories galactic evolution makes determination of distance very problematical. All curves are drawn with the Hubble constant taken as 15 kilometers per second per million light years (corresponding to a characteristic expansion time of 20,000 million years), but the curves can be used for any other value of the Hubble constant by simply rescaling all distances.

As one proceeds from a more dense finite universe to a less dense infinite universe, the curve of redshift versus distance flattens out at very large distances. By the 1970's the program to determine this graph had not produced conclusive results (The situation is very different now as we will see later).

The problem is in estimating the distances. What is distance when we cannot directly use a *ruler*? Close to us we can use geometry - measure angles and determine distance directly - this only works, however, for small distances - difficult to measure angles precisely. We then compare the direct distances to the distance calculated with Cepheid variables in some overlap region (this calibrates the Cepheid variable distance) and then use Cepheid variables to go further out in distance. Eventually, however, it becomes impossible to pick out any Cepheid variable star from within a galaxy. We could use the same technique substituting entire galaxies but we don't know the absolute luminosity of a typical galaxy. In fact, a distant (and older) galaxy may not have the same properties as closer (younger) galaxies due to galactic evolution. So distance measurement is very difficult to do and is a major limitation.

IN the 1970's the best inference from the data was that the deceleration of distant galaxies seems very small, which would mean that they are moving at more than the escape velocity (I note, as we will see later, that they actually seem to be accelerating according to modern data). Generally, the uncertainties in astrophysical type measurement are large and make drawing conclusions difficult. Luckily for us, we do not have to pin down the large-scale geometry of the universe to draw conclusions about its beginning. The reason is that the universe has a *horizon* and this horizon shrinks rapidly as we look back toward

the beginning. No signal can travel faster than light so at any time we can only be affected by events occurring close enough so that a ray of light would have had time to reach us since the beginning of the universe - the events that could affect us at any time are those in our past - in our backward light-cone - the set of past events that are timelike related to our present event. Any event that occurred beyond this distance could have no effect on us - it is beyond the *horizon*, i.e., there is at any time t after the beginning a horizon at a distance of order ct , from beyond which no information could yet have reached us. Since $R(t)$ vanishes less rapidly as $t \rightarrow 0$ than the distance to the horizon, at a sufficiently early time any given *typical* particle is beyond the horizon. If the universe is now 10^{10} years old, the horizon distance is now about 3×10^{10} light-years. But when the universe was a few minutes old, the horizon was only at a distance of a few light-minutes - less than the present distance from the earth to the sun. The entire universe, of course, was smaller than also (the separation between bodies decreases). However, as we look back toward the beginning, the distance to the horizon shrinks faster than the size of the universe, i.e as we found earlier, the size of the universe is proportional to $t^{1/2}$ or $t^{2/3}$ while the distance to the horizon is proportional to the time - thus, for earlier and earlier times, the horizon encloses a smaller and smaller portion of the universe as shown below.

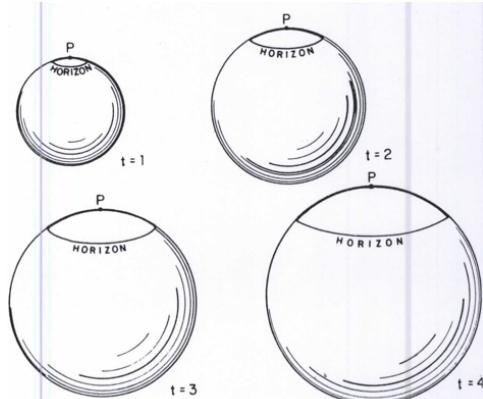


Figure 6. *Horizons in an Expanding Universe*. The universe is symbolized here as a sphere, at four moments separated by equal time intervals. The "horizon" of a given point P is the distance from beyond which light signals would not have had time to reach P. The part of the universe within the horizon is indicated here by the unshaded cap on the sphere. The distance from P to the horizon grows in direct proportion to the time. On the other hand, the "radius" of the universe grows like the square root of the time, corresponding to the case of a radiation-dominated universe. In consequence, at earlier and earlier times, the horizon encloses a smaller and smaller portion of the universe.

As a consequence of this closing in of horizons in the early universe, the curvature of the universe as whole makes less and less difference as we look back to earlier and earlier times. Thus, we can still deal with the beginning without knowing the large-scale geometry!

We now have a view of the universe that is as simple as it is grand. The universe is expanding uniformly and isotropically - all galaxies see the same things in all directions. As the universe expands, the wavelengths of light rays are stretched out in proportion to the distance between the galaxies. The expansion was not thought to be due to any cosmic repulsion in the 1970s but completely due to an initial explosion (modern data is not so sure). The expansion velocities in the 1970s seemed to be slowing down(slowly) under the influence of gravity suggesting a low matter density and a gravitational field too weak to ever stop the expansion. Extrapolation backwards suggest an age of the universe on the order of $(10 - 20) \times 10^9$ years.

4.3 The Cosmic Microwave Background

We now look at a very different set of observations. We stop looking at light emitted in the last few hundred million years from galaxies like our own, but instead look at the diffuse background of radio static noise left over from near the beginning of the universe. It is an interesting story. In 1964 Bell Labs had an antenna built for radio communication with the Echo satellite - it was a 20 foot horn reflector with ultra-low noise, which attracted radio astronomers to the instrument. Penzias and Wilson attempted to use the antenna to measure the intensity of radio waves emitted from our galaxy at high galactic latitudes (out of the plane of the Milky Way). A very difficult measurement. The radio waves from our galaxy are more like noise - much like the static on a radio during a thunderstorm. This radio noise is hard to distinguish from always present electrical noise produced by random electron motion in the antenna and other connected electrical circuits or from radio noise picked up by the antenna from the earth's atmosphere.

If one is studying a small (localized) source like a star or a distant galaxy, then one can point the antenna at the star and receive signals, then point the antenna away from the star and receive background signals (noise, etc) only and then subtract them out of the star signal (since they are the same no matter where we point the antenna). Penzias and Wilson, however, were trying to measure the signal coming from our own galaxy - in effect, from the sky itself - there is no way to subtract off the spurious signals. They therefore set out to identify and minimize all electrical noise produced in the antenna receiving system. To study the noise from the antenna structure they set the antenna to look at radio waves of a short wavelength 7.35 cm (these are called microwaves). They thought that radio noise from the galaxy would be negligible at this wavelength and therefore all the signal would be from the antenna structure. Noise from

the earth's atmosphere would be easy to measure and subtract off since it could be easily identified by the dependence of the signal on direction (it depends on the thickness of the atmosphere along the antenna direction). Once having understood the noise this way, they could then proceed to study the radio waves from the galaxy which were expected to be at wavelength 21 cm (characteristic of hydrogen which is the dominant stuff in the universe).

To their surprise, they found a very large signal of microwave noise at 7.35 cm that was independent of direction. They also found that this *static* did not vary with time of day or, as the year went on, with the season. It could not be coming from our galaxy since they would have already seen it in measurement taken on the Andromeda galaxy (almost the same as the Milky Way galaxy). They lack of any dependence on direction suggested that the radio noise was coming from a much larger volume of the universe than just our own galaxy.

Still thinking it was somehow antenna noise, they took the apparatus apart, cleaned it, and rebuilt it - no effect - the mysterious noise was still there undiminished.

Now any body at any temperature above absolute zero always emits radio noise (due to the thermal motions of the electrons within the body) The radio noise at any wavelength depends only on the temperature - the higher the temperature, the more intense the static. Thus, one can state an equivalent temperature for any radio source. Penzias and Wilson found an equivalent temperature for their observed noise between 2.5 and 4.5 degrees above absolute zero or about 3.5 degrees Kelvin or $3.5^\circ K$. This was much greater than expected, but very low in absolute terms. They hesitated publishing their results. They certainly did not think that this was the most important cosmological advance since the discovery of red shifts!

Several theories existed at the time of the measurements that, although not completely correct, had the right idea about what was happening. Most of the theories realized that if there had not existed an intense background of radiation during the first few minutes of the universe, then nuclear reactions would have proceeded so rapidly that a large fraction of the hydrogen present would have been *cooked* into heavier elements - this could not have happened because about $3/4$ of the present universe is hydrogen. The rapid nuclear cooking could have been prevented only if the early universe was filled with intense radiation having an enormous temperature at very short wavelengths - this would blast apart any nuclei that formed as fast as they formed.

As we will see later, this radiation has survived the expansion of the universe but its equivalent temperature decreased as the universe expanded in inverse proportion to the size of the universe. Therefore, the present universe should be filled with radiation, but with an equivalent temperature significantly less than its value in the first few minutes. The first theoretical estimates showed the

expected temperature to be in the same range as that measured by Penzias and Wilson. Thus, the temperature recorded by the antenna is not the temperature of the present universe, but seems to be the temperature that the universe had near its beginning - greatly reduced by expansion.

Is this microwave radiation actually left over from the beginning of the universe? To answer this question we need to discuss what we expect theoretically: What are the general properties of the radiation that *should* be filling the universe if current cosmological ideas are correct? We will need to consider what happens to radiation as the universe expands - not only at the time of nucleosynthesis(at the end of three minutes) but also in all the time that has elapsed since then.

We need to give up the classical picture of radiation in terms of electromagnetic waves and switch to the quantum view that radiation consists of particles known as photons. An everyday light wave contains an enormous number of photons traveling along together. Classically, the energy is assumed to be a continuous quantity. If, however, we were to measure the enrgy of a light wave with extreme precision, we would find that it always comes in multiples of a definitie quantity, which is the energy of a single photons in the wave, i.e., $E_{wave} = N E_{photon}$ where $E_{photon} = \hbar\nu = \hbar c/\lambda$ where N is a very large number (10^{15} per cm^3), ν = the frequency of the wave, λ = the wavelength of the wave and $c = \lambda\nu$. We note that a typical frequency is $\nu = 10^{14} sec^{-1}$ and therefore $E_{photon} = \hbar\nu \approx 10^{-20} Joules$, which is extraordinarily small. We know there are photons from the interaction of light with atoms which usually take place one photon at a time and cannot be explained except via the photon idea. Photons have zero mass and zero charge, carry a definite energy $E_{photon} = \hbar\nu$ and momentum $p = E_{photon}/c$.

What happens to an individual photon as it travels through the universe? Not much as it turns out. The light from objects as much as 10^{10} light-years away seems to reach us easily. This indicates that whatever matter may be present in the universe it is sufficiently transparent so that photons can travel for very large distances(for a large fraction of the of the universe) without being scattered or absorbed.

The redshift data tells us that the universe is expanding - the universe must have been more compressed in the past than now. In general, if you compress matter its temperature will rise As we will find later, there was a period which probably lasted for the first 700000 years of the universe when the contents were so hot and dense that the formation of stars and galaxies was impossible - matter remained as nuclei and electrons. Under these extreme conditions a photon could not travel very far with out being scattered or absorbed(as it can now). The mean free time between photon interactions with matter was much shorter than the characteristic time for the universe expansion. Therefore, even though the universe was expanding very rapidly at first, to an individual photon or electron or nucleus the expansion was taking a large amount of time - time

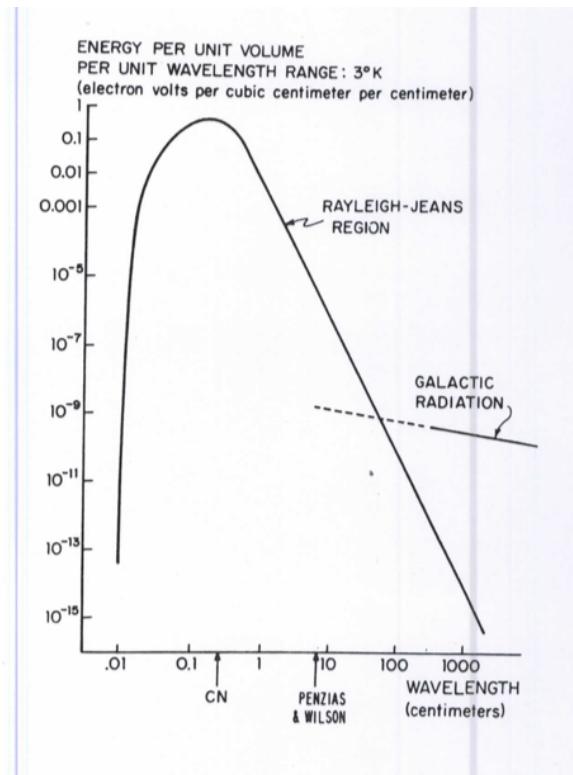
enough for each particle to be scattered or absorbed and reemitted many times as the universe expanded. Any system like this where individual constituents have time for many interactions usually comes to a state of equilibrium. The properties of such a state are not determined by the initial conditions but instead by the requirements needed to maintain the equilibrium. Each of the particles is constantly changing in this state, but all the statistical variables such as distributions(probability) of energy or position are remaining essentially constant. This type of equilibrium is called *thermal equilibrium*. because it is characterized by a definite temperature which is essentially uniform throughout the system.

Most physical systems in the world are far from thermal equilibrium - the surface of the earth certainly is not a state of equilibrium. However, at the center of a star, there is nearly perfect thermal equilibrium and thus we can estimate the conditions are like with great confidence. The universe is certainly not in a state of perfect thermal equilibrium - it is expanding! However, as we mentioned the times scale of interaction in the early universe compared to the time scale of expansion - interactions taking place on a much faster time scale than the expansion - was such that the universe could be regarded as evolving *slowly* from one state of nearly perfect thermal equilibrium to another. It turns out that the properties of any system in thermal equilibrium are entirely determined once we specify the temperature of the system and the densities of a few conserved quantities. Thus, the universe preserves only a very limited memory of its initial state. This is not good if we want to know what the actual big-bang was like, but is great if we want to infer the course of events after the beginning without having to make too many arbitrary assumptions.

We now ask - What are the general properties of radiation in thermal equilibrium with matter?

By the 1890s it was known that the properties of radiation in a state of thermal equilibrium with matter depend only on the temperature - the amount of radiation energy per unit volume in a small range of wavelengths is given by a universal formula which involves only the wavelength and the temperature. This same formula, as it turns out, gives the intensity of the radio noise in terms of the equivalent temperature. For various obscure reasons this radiation became known as *black-body radiation*.

A full theoretical understanding of black-body radiation require quantum mechanics, which we will not go into here. Weinberg gives some ideas about it in the text. The characteristic black-body radiation distribution called the Planck distribution takes the form shown below.



This is a plot of energy density per unit wavelength range as a function of wavelength (log-log scales) for black-body radiation with a temperature of $3^{\circ} K$. For a temperature which is greater than $3^{\circ} K$ by a factor f , it is only necessary to reduce the wavelength scale by a factor $1/f$ and increase the energy density scale by a factor f^5 and the curve is exactly the same! The arrows indicate the Penzias and Wilson measurement.

Returning to our discussion of the observed microwave radiation, the universe must have been so hot and dense there were only free nuclei and electrons and the scattering of photons by free electrons maintained a thermal equilibrium between matter and radiation. As time passed, the universe expanded and cooled eventually reaching a temperature ($3000^{\circ} K$) cool enough to allow the combination of nuclei and electrons into atoms. The sudden disappearance of the free electrons broke the thermal contact between radiation and matter and the radiation continued to expand freely. At that moment the energy of the radiation at all wavelengths was determined by the conditions of thermal equilibrium - the Planck distribution for a temperature of $3000^{\circ} K$. the typical photon wavelength would have been 0.0001 cm or 10000 \AA and the average distance between photons was roughly the typical wavelength.

What has happened to the photons since then? Individual photons would not be created or destroyed - the average distance between them would increase in proportion to the size of the universe - in proportion to the average distance between typical galaxies. The associated cosmological redshift cause the wavelength of any individual photon to increase in proportion to the size of the universe. The photons would thus remain one typical wavelength apart just as for black-body radiation - the radiation filling the universe would continue to be black-body radiation as the universe expanded even though it is no longer in thermal equilibrium with matter. We give the mathematical proof below.

The Planck distribution gives the energy du of black-body radiation per unit volume, in a narrow range of wavelengths from λ to $\lambda + d\lambda$, as

$$du = \frac{\frac{8\pi hc}{\lambda^5} d\lambda}{e^{hc/k\lambda T} - 1}$$

where T is the temperature; k is Boltzmann's constant ($1.38 \times 10^{-16} \text{ erg}/\text{K}$); c is the speed of light; e is a numerical constant (3.718); and h is Planck's constant ($6.625 \times 10^{-27} \text{ erg-sec}$). The Planck formula for du reaches a maximum at a wavelength $\lambda = 0.2014052hc/kT$ and then fall steeply off for decreasing wavelengths. The total energy density for all wavelengths is

$$u = \int_0^\infty \frac{\frac{8\pi hc}{\lambda^5} d\lambda}{e^{hc/k\lambda T} - 1} = \frac{8\pi^5 (kT)^4}{15(hc)^3} = 7.56464 \times 10^{-15} [T(\text{K})]^4 \text{ erg/cm}^3$$

which is the Stefan-Boltzmann law we mentioned earlier.

The Planck distribution can easily be interpreted in terms of photons. Each photon has an energy $E = hc/\lambda$. Hence the number dN of photons per unit volume in black-body radiation in a narrow range of wavelengths from λ to $\lambda + d\lambda$ is

$$dN = \frac{du}{hc/\lambda} = \frac{\frac{8\pi}{\lambda^4} d\lambda}{e^{hc/k\lambda T} - 1}$$

and the total number of photons is

$$N = \int_0^\infty dN = 60.42198 \left(\frac{kT}{hc} \right)^3 = 20.28 [T(\text{K})]^3 \text{ photons/cm}^3$$

and the average photon energy is

$$E_{\text{photon}} = \frac{u}{N} = 3.73 \times 10^{-16} [T(\text{K})] \text{ ergs}$$

Now let us consider what happens to black-body radiation in an expanding universe. Suppose the size of the universe changes by a factor f . As we saw earlier, wavelengths will change in proportion to the size of the universe, $\lambda' = f\lambda$. After the expansion, the energy density du' in the new wavelength range λ' to $\lambda' + d\lambda'$ is less than the original energy density du in the old wavelength range λ to $\lambda + d\lambda$ for two different reasons:

1. Since the volume of the universe has increased by a factor f^3 , as long as no photons have been created or destroyed, the number of photons per unit volume has decreased by a factor $1/f^3$.
2. The energy of each photon is inversely proportional to its wavelength and is therefore decreased by a factor $1/f$.

It then follows that the energy density is decreased by an overall factor $1/f^3$ times $1/f$ or $1/f^4$. Thus,

$$du' = \frac{1}{f^4} du = \frac{\frac{8\pi hc}{\lambda^5 f^4} d\lambda}{e^{hc/k\lambda T} - 1}$$

If we rewrite this formula in terms of the new wavelength λ' , it becomes

$$du' = \frac{\frac{8\pi hc}{\lambda'^5} d\lambda'}{e^{hc/f/k\lambda' T} - 1}$$

but this is exactly the same as the old formula for du in terms of λ and $d\lambda$, except that T has been replaced by a new temperature $T' = T/f$. Thus, we conclude that freely expanding black-body radiation remains described by the Planck formula, but with a temperature that decreases in inverse proportion to the scale of the expansion.

Penzias and Wilson found a temperature of about $3^\circ K$, which would be expected if the universe had expanded by a factor of 1000 since the time when the temperature was high enough ($3000^\circ K$) to keep matter and radiation in thermal equilibrium. If this interpretation is correct, then the $3^\circ K$ radio static is by far the most ancient signal received by astronomers - it was emitted long before the light from the most distant galaxies that we can see!

Subsequent measurements have shown that this temperature is characteristic of all wavelengths as expected. However, all the measurements were on the long wavelength end of the Planck distribution for $3^\circ K$, i.e., above the maximum which occurs for at 0.1 cm . In order to confirm that we really have black-body radiation, we need to check at all wavelengths- not just radio/microwave radiation but also infrared radiation. Unfortunately, the earth's atmosphere which is nearly transparent at wavelengths above 0.3 cm becomes increasingly opaque at shorter wavelengths so ground-based astronomy will not work. Continued experimentation required getting above the atmosphere. This was done using rockets and balloons and black-body radiation at about $3^\circ K$ was confirmed to wavelengths as low as 0.06 cm . To use a satellite, was a very difficult undertaking since the detectors had to be cooled to liquid helium temperature (comparable to the black-body temperature itself). This was easier to do on earth and in balloons and rockets. Another reason to go to satellites was to get above the atmosphere. We have using the idea that all the radiation was isotropic - same in all directions up to now. This was needed to be consistent with the Cosmological Principle. One, however, wants to investigate the possible directionality

of the radiation. With the atmosphere in the way this is almost impossible since the atmosphere has its own directional effects that would mask it.

Why do we want to look for direction dependence?

There might be fluctuations in the intensity with small changes in direction caused by the actual lumpiness of the universe either at the time the radiation was emitted or since then. Galaxies in early stages might show up as hot spots. Also there is a small directionality due to the earth's motion through the universe. If the earth were moving at about 300 km/sec with respect to the average matter in the universe (around the sun + sun around center of galaxy + galaxy motion), then we would have a Doppler effect where the wavelength from ahead or astern would be decreased or increased respectively by about 0.1%. Non-satellite instruments do not have this kind of accuracy. Since this book was written, a Swarthmore graduate John Mather has led a project (COBE) to measure both the background radiation at all wavelengths and the fluctuations in the background. We present below the COBE results which we will talk about later. Also included are later WMAP and PLANCK data.

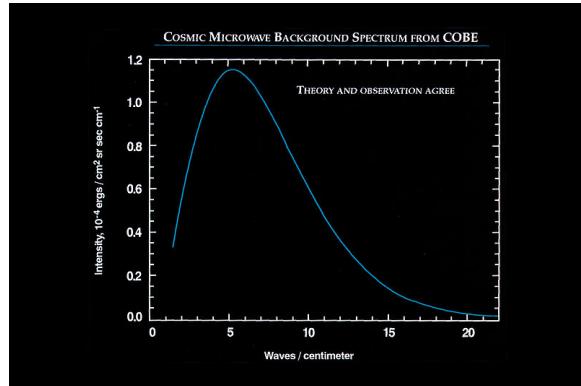


Figure 4.5: COBE Black-Body Data

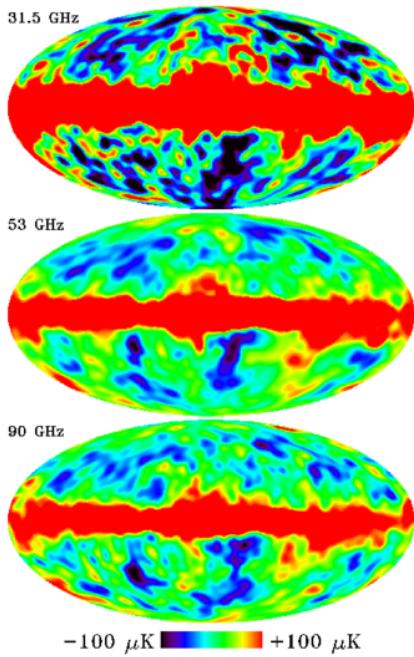


Figure 4.6: COBE Full Sky - Different Wavelengths - Data

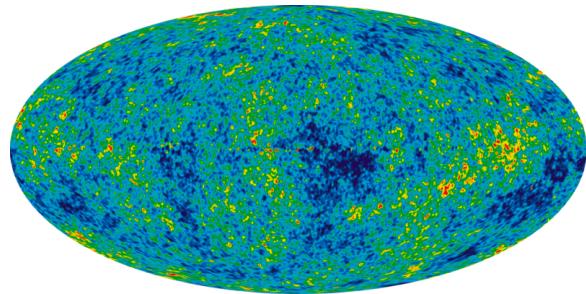


Figure 4.7: WMAP Full Sky Data

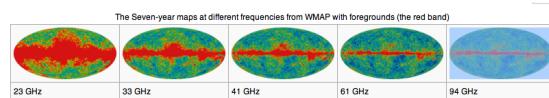


Figure 4.8: WMAP Full Sky - Different Wavelengths - Data

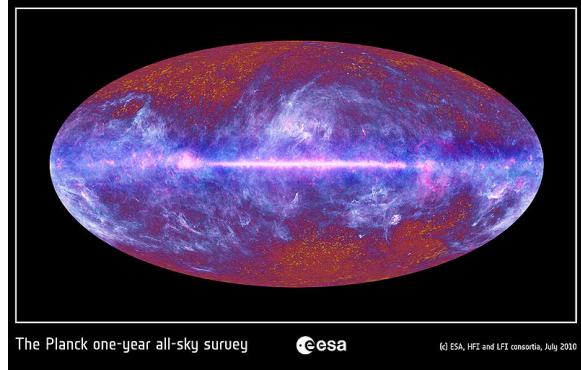


Figure 4.9: PLANCK Full Sky Data

Clearly it is exactly Black-Body radiation and there are fluctuations to be understood.

What cosmological insight can we draw from the particular temperature value $3^{\circ} K$? It will allow us to determine one crucial number that we will need to follow the history of the first three minutes.

At $3^{\circ} K$ there are 550000 photons per liter. The density of nuclear particles (neutrons and protons) in the present universe is between 6 and 0.03 particles per 1000 liters. Thus there are between 10^{10} and 2×10^{10} photons for every nuclear particle in the universe today. This number has been roughly constant for a long time. No creation of either has taken place during the expansion after $3000^{\circ} K$. This is the most important quantitative conclusion we can draw from the early data. For practical purposes we will assume the number is 10^9 photons per nuclear particle.

The differentiation of matter into galaxies and stars could not have begun until the time when the cosmic temperature became low enough for electrons to be captured into atoms. In order for gravity to produce the clumping of matter into isolated fragment it is necessary for gravity to overcome the pressure of matter and the associated radiation. The gravitational force within a clump increases with the size of the clump while the pressure does not depend on the size; hence at any given density and pressure there is a minimum mass which is susceptible to gravitational clumping. This is called the *Jeans mass*. It turns out that the Jeans mass is proportional the $(\text{pressure})^{3/2}$ (see derivation below).

In order for a clump of matter to form a gravitationally bound system, it is necessary for its gravitational potential energy to exceed its internal thermal energy. The gravitational potential energy of a clump of radius r and mass M

is of order

$$U_{grav} \approx -\frac{GM^2}{r}$$

The internal energy per unit volume is proportional to the pressure p , so the total internal energy is of order

$$E_{int} \approx pr^3$$

The gravitational clumping should be favored if

$$\frac{GM^2}{r} \gg pr^3$$

But for a given density ρ we can express r in terms of M through the relation

$$M = \frac{4\pi}{3}\rho r^3$$

The condition for gravitational clumping can therefore be written

$$GM^2 \gg p \left(\frac{M}{\rho} \right)^{4/3}$$

or $M \geq M_J$ where M_J is the quantity known as the *Jeans mass*:

$$M_J = \frac{p^{3/2}}{G^{3/2}\rho^2}$$

For example, just before the free electrons and the free nuclei combined into hydrogen(recombination), the mass density was $9.9 \times 10^{-22} \text{ gm/cm}^3$ (calculated earlier) and the pressure was $p \approx \rho c^2/3 = 0.3 \text{ gm/cm sec}^2$ so that the Jeans mass is $M_J = 9.7 \times 10^{51} \text{ gm} = 5 \times 10^{18} M_\odot$, where M_\odot is one solar mass. Note that our galaxy mass is about $10^{11} M_\odot$. Galaxies are not massive enough to have formed at this time or before. After recombination, the pressure dropped by a factor of 10^9 , so the Jeans mass dropped to $M_J = 1.6 \times 10^5 M_\odot$ and galaxies were able to form. We know they can form but we still do not know how they form.

So before recombination, there were no stars or galaxies in the universe, only an ionized and undifferentiated soup of matter and radiation.

Another remarkable consequence of the large ratio of photons to nuclear particles is that there must have been a time, not too far in the past, when the energy of radiation was greater than the energy contained in the matter of the universe. The energy in the mass of a nuclear particle $mc^2 \approx 9.4 \times 10^8 \text{ eV}$. The average energy of a 3° K black-body radiation photon is about 0.0007 eV so that even with 10^9 photons per neutron or proton most of the energy of the present universe is in the form of matter, not radiation. However, at earlier times, the

temperature was higher, so the energy of each photon was higher, while the energy of the nuclear particles is unchanged. With 10^9 photons per neutron or proton, in order for the radiation energy to exceed the energy of matter we must have a temperature of about $4000^\circ K$. This is the temperature that marks the transition between a *radiation-dominated* era in which most of the energy in the universe was in the form of radiation, and the present *matter-dominated* era in which most of the energy is in the masses of the nuclear particles.

It is striking that this transition $4000^\circ K$ occurred about the same time as the content of the universe were becoming transparent to radiation $3000^\circ K$. We do not know why. It is also not clear which change occurred first (the numbers used to calculate the transition temperature are fairly uncertain).

None of these uncertainties will affect our study. We only need to know the early era was radiation-dominated with only a small contamination of matter.

4.4 Recipe for a Hot Universe

We now know the universe is expanding and that it is filled with a universal background of radiation, now at a temperature of about $3^\circ K$. This radiation was left over from a time when the universe was effectively opaque, when it was 1000 times smaller (the average distance between a typical pair of particles was a 1000 times smaller) and hotter than at present. In order to account for the first three minutes, we must now use theory to look back even earlier when the universe was even smaller and hotter, to discover the physical conditions that prevailed.

During the radiation-dominated era, the numbers of photons are so large and the energy so large, that we can consider the universe as if it were filled only with radiation (assume essentially no matter). We note that the radiation-dominated era began at the end of three minutes when the temperature dropped below 10^9 degrees Kelvin. Prior to this period, matter was important, but not the kind of ordinary matter we usually talk about.

We need to find out how the temperature was related to the size of the universe so that we can figure out how hot things were at any given moment. We know that if the radiation were expanding freely (no interactions), then the wavelength of each photon is stretched out (redshift) in proportion to the size of the universe as the universe expands. We also know that the average wavelength of black-body radiation is inversely proportional to its temperature. Thus the temperature would have decreased in inverse proportion to the size of the universe, just as it is doing now. However, the expansion was not free during this era - the photons were experiencing rapid collisions with the small number of electrons and nuclear particles - remember the universe was essentially opaque during this era. Luckily for us, however, most of the time the photons were freely

traveling between collision (such a small amount of matter), and therefore they essentially behaved as if they were free as far as the temperature change due to the expansion of the universe was concerned and the earlier argument still works!

As we look earlier in the evolution of the universe, we will come to a time when the temperature was so high that photon-photon collisions began to produce material particles out of pure energy. The particle produced this way will turn out to be just as important as the radiation during the first three minutes - both for determining the rates of nuclear reactions and in determining the rate of expansion of the universe itself. What is the temperature threshold for this particle production to occur? The process is understood in terms of the quantum theory of light. Two photons can collide, disappear and all their energy and momentum reappear in the production of material particles($E = mc^2$). Ordinary nuclear reactions also use this form of energy conversion but generally only a small fraction of the mass is converted. In order for two photons to produce two material particles of mass m the energy of each photon must be greater than or equal to mc^2 .

Can photons have this energy ? or What is the characteristic energy of individual photons in the radiation field? We can estimate by a rule from Statistical Mechanics, namely $E_{char} \approx kT$ where k is Boltzmann's constant and $k = 0.00008617\text{ eV/}^\circ\text{K}$ and T is the temperature in degrees Kelvin. Therefore, at 3000°K when the contents of the universe were just becoming transparent, each photon had a characteristic energy of 0.26 eV . This is the characteristic energy of reactions involving atoms - that is why the radiation was able to prevent the recombination of nuclei and electrons into atoms.

To produce particles, however, the characteristic energy needs to be larger than mc^2 and the corresponding threshold temperature would be mc^2/k . The electron e^- and the positron e^+ (antiparticle) are the particles with the smallest mass. The rest energy mc^2 of the electron(positron) is 0.511002 MeV . Therefore the threshold temperature at which photon can have this much energy is $mc^2/k = 6 \times 10^9\text{ }^\circ\text{K}$ (NOTE: $T_{center\ of\ sun} = 15 \times 10^6\text{ }^\circ\text{K}$).

At this temperature or higher, electrons and positrons would have been freely created in photon collisions and would be present in very large numbers. That is why we do not see electrons and positrons popping out of empty space whenever the sunlight is bright! Similar rules apply for all other particles and their associated antiparticles.

The table below summarizes some facts for several particles.

Class	Name	Symbol	Energy(MeV)	$T_{thresh}(10^9 \text{ } ^\circ K)$	Species	Mean Life(sec)
	Photon	γ	0	0	1	stable
Lepton	Neutrinos	$\nu_e, \bar{\nu}_e, \nu_\mu, \bar{\nu}_\mu$	0	0	2	stable
Lepton	Electron	e^-, e^+	0.5110	5.930	7/4	stable
Lepton	Muon	μ^-, μ^+	105.6	1226.2	7/2	2.197×10^{-6}
Hadron	π mesons	π^0	135.0	1566.2	1	0.8×10^{-16}
Hadron	π mesons	π^+, π^-	139.6	1566.2	2	2.6×10^{-8}
Hadron	Proton	p, \bar{p}	938.3	10888	7/2	stable
Hadron	Neutron	n, \bar{n}	939.6	10903	7/2	920

Table 4.1: Properties of Some Elementary Particles

The particles that are present in large numbers at different times in the history of the universe can be read off the table by looking at the threshold temperature column.

The number of particles present at temperatures above the threshold temperature is governed by the thermal equilibrium - the number must have been just high enough so that precisely as many were being destroyed each second as were being created. Since the rate of 2 photons into 2 particles is the same as the rate of 2 particles into 2 photons, the condition of thermal equilibrium requires that the number of particles of each type, whose threshold temperature is below the actual temperature should be about equal to the number of photons. If number less than photons, will be created faster and number rise to balance and if number smaller, will be destroyed faster and numbers decrease! Thus, above electron threshold temperature, say at $6000 \text{ } ^\circ K$ the number of positrons and electrons is the same as the number of photons and the universe is a soup of all three!

On the other hand, high above the threshold temperature, material particles basically have energy kT , which is much larger than mc^2 and the mass can be neglected so that material particle behave like photons. In this case, the pressure and energy density contributed by each type of material particle is proportional to T^4 in the same way as photons. The universe is then composed of a variety of types of *radiation* - one for each species - the effective species number is listed in the table above. The energy density of the universe is then proportional to T^4 and the effective number of species with threshold temperature below the actual temperature.

When was the universe at these very high temperatures? The balance between the gravitational field (attractive) and the outward momentum pressure (repulsive) of the contents of the universe determines the rate of expansion of the universe. It is the total energy density that is the source of the gravitational field at early times in the universe. The energy density depends essentially only

on the temperature - thus the cosmic temperature can be used as a kind of clock - cooling instead of ticking as the universe expands. We showed earlier that the time required for the energy density of the universe to fall from one value to another is proportional to the difference of the reciprocals of the square roots of the energy density (proportional to number of species and T^4). Thus, as long as the temperature does not cross a threshold, the *time required for the universe to cool from one temperature to another is proportional to the difference of the inverse squares of these temperatures*. Therefore, if we start at a temperature of $10^8 \text{ }^\circ\text{K}$ then after about 700000 years the temperature would reach $3000 \text{ }^\circ\text{K}$ as shown below.

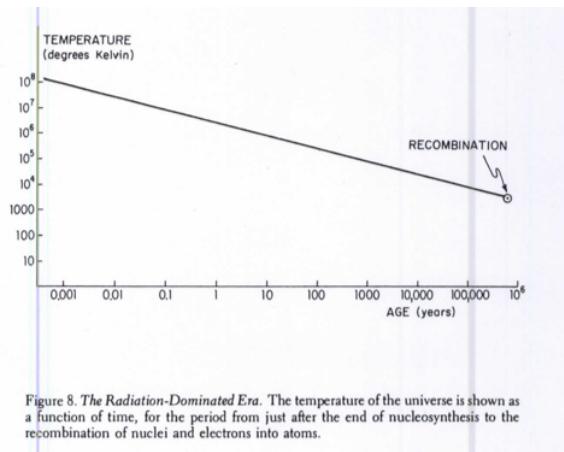


Figure 8. *The Radiation-Dominated Era*. The temperature of the universe is shown as a function of time, for the period from just after the end of nucleosynthesis to the recombination of nuclei and electrons into atoms.

During the first few minutes the number of particles and antiparticles cannot have been precisely equal. Since they would then have all annihilated when the temperature dropped below $10^9 \text{ }^\circ\text{K}$ and nothing would be left except radiation - we have evidence against this - us! There must have been an excess of particles over antiparticles that led to the matter in the present universe. What exactly were the constituents making up this excess? The list of possibilities is endless! In order to figure out what was there, we need to look in more detail at the thermal equilibrium. In a state of thermal equilibrium, there are quantities that do not change - *conserved quantities*. For example total energy - even though collisions are constantly redistributing the energy, the total remains constant - this is called a *conservation law*. For each such conservation law there exists a value that must be specified in advance in order to work out the properties of a system in thermal equilibrium - the values cannot be determined from the conditions for equilibrium. once these constant values have been specified for a system in thermal equilibrium then all of its properties are uniquely determined. Since the universe has passed through a state of thermal equilibrium, all we need to do to give a complete recipe for the contents of the universe at early times is to know what physical quantities are conserved as the universe expands and what were the values of these quantities.

It turns out that there are just three conserved quantities whose densities must be specified in our recipe for the early universe:

1. Electric Charge - Particles and antiparticles can annihilate or be created, but total charge never changes.
2. Baryon Number - The number of protons, neutrons, mesons minus the number of their antiparticles. Baryons and antibaryons can annihilate or be created, but total baryon number never changes.
3. Lepton Number - The number of electrons, muons, neutrinos minus the number of their antiparticles. Leptons and antileptons can annihilate or be created, but total lepton number never changes.

To complete the recipe for the contents of the universe at any given time, we must specify the charge, baryon number, and lepton number densities as well as the temperature at that time. The conservation laws then say that as the universe expands the three total quantities remain fixed. Thus, charge, baryon number, and lepton number densities vary with the inverse cube of the size of the universe. On the other hand, the photon density also varies with the inverse cube of the size of the universe - the photon density is proportional to T^3 and the temperature varies with inverse size of the universe. Thus, the totals for all these numbers charge, baryon, lepton, and photon which equal their respective density times the volume remain fixed and the recipe can be specified by giving the charge, baryon number, and lepton number ratios to the photon number.

The net charge of the universe must essentially be zero or the cosmic electric charge per photon is negligible. We can see this from the numerical considerations below. If the earth and the sun had an excess of positive over negative charges of about one part in 10^{36} electric force between them would be greater than the gravitational force between the earth and the sun. Gravity dominates on all scales in the universe - hence our conclusion.

The only stable baryons are the nuclear particles - the proton and the neutron and their antiparticles. As far observation is concerned there is no appreciable amount of antimatter in the universe. Thus, the baryon number is essentially the number of nuclear particles. At the present time there is one nuclear particle for every 10^9 photons in the microwave radiation background. Thus, the baryon number per photon is about 10^{-9} . This is a remarkable result. Let us see why. Consider a time in the past when the temperature was above $10^{13} \text{ }^\circ\text{K}$, which is the threshold temperature for protons and neutrons. At that time, the number of nuclear particles and antiparticles was about equal to the number of photons. But baryon number is the difference between particle and antiparticle numbers. If this difference was 10^9 smaller than the photon number(as above) and hence also about 10^9 smaller then the total number of nuclear particles, then the number of nuclear particles would have exceeded the number of antiparticles by only one part in 10^9 . In this view, when the universe cooled below

the threshold temperature for nuclear particles, the antiparticles all annihilated with corresponding particles, leaving behind the tiny excess of particles over antiparticles as a residue which eventually turns into the world we know now.

Since the net charge in the universe is zero, there exist exactly one electron for each proton. Proton make up about 87% of all nuclear particles in the present universe. If electrons were the only leptons in the present universe, then the lepton number per photon would be about the same as the baryon number per photon. However, there is another stable lepton, namely the neutrino. It is massless, has zero charge. Thus, we need to figure out the numbers of neutrinos and antineutrinos. Not easy to do! Neutrinos only interact very weakly with matter and radiation (no charge, no mass). A good number to remember - it would take 46 light-years of lead to stop a neutrino!

So the universe might be filled with neutrinos and we could not easily tell. It would easily be possible for the number of neutrino and their corresponding energies to be comparable to photons. However, once again lepton number is a difference and we still assume it is very small compared to the photon number while the number of electrons or neutrinos is about the same as the number of photons above threshold temperatures.

We now have the recipe for the contents of the early universe. Take a charge per photon equal to zero, a baryon number per photon equal to one part in 10^9 . Take the lepton number per photon to be uncertain but small. Take the temperature at any given time to be greater than the temperature $3\text{ }^\circ K$ of the present radiation background by the ratio of the present size of the universe to the size at that time. Stir well, so that the detailed distribution of particles of various types are determined by the requirements of thermal equilibrium. Place in an expanding universe, with a rate of expansion governed by the gravitational field produced by this medium. After a long enough wait, this concoction should evolve into our present universe.

4.5 The First Three Minutes

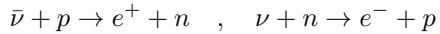
We can now follow the course of cosmic evolution through its first three minutes. The time scale used will be temperature-based, i.e., we will take a picture each time the temperature drops by a factor of about 3. We start the story at $1/100$ second (we will look at the earliest period later since this is the period where most of the physics was discovered after Weinberg wrote his book).

First Frame. The temperature of the universe is $10^{11}\text{ }^\circ K$. The universe is in a state that is easy to describe. It is filled with an uniform soup of matter and radiation, each particle of which has rapid collisions with other particles. Even though it is rapidly expanding, the universe is approximately in perfect thermal equilibrium. Statistical mechanics rules - means that nothing depends

on what happened before the first frame (except that the temperature value and the size value - how they got those values is a story for later). We also know the conserved quantities - charge, baryon and lepton number are all very small or zero. The most abundant particles are those with a threshold temperature below $10^{11} \text{ }^{\circ}\text{K}$ - electrons, positrons, photons, neutrinos and antineutrinos. The universe is so dense that even neutrinos (remember 46 light-years of lead to stop them!) are kept in thermal equilibrium with the other particles by rapid collisions. Since the threshold temperature for electrons and positrons is well below $10^{11} \text{ }^{\circ}\text{K}$ they act like pure radiation (neglect their mass) - like photons and neutrinos.

What is the energy density of the different kinds of radiation? From the table listing of effective species number we see that it turns out that the total energy density of the soup is $9/2$ greater than it would be if we only had photons. The Stefan-Boltzmann law says the energy density of electromagnetic radiation at $10^{11} \text{ }^{\circ}\text{K}$ is $4.72 \times 10^{44} \text{ eV per liter}$. Thus the total energy density of the universe-soup is $21 \times 10^{44} \text{ eV per liter}$. This is equivalent to a mass density of $3.8^9 \text{ kg per liter}$ or 3.8^9 times the density of water.

The universe is rapidly expanding and cooling at this time. The rate of expansion is given by the condition that all parts of the universe are moving at the escape velocity away from some arbitrary center. Since the universe is so dense, the escape velocity is very high and the characteristic time for expansion of the universe is about 0.02 seconds(see earlier calculations - page 16). This *characteristic expansion time* is approximately 100 times the time interval during which the size of the universe would increase by 1%. It is equal to the reciprocal of the Hubble constant at that time. During this period the small number of existing protons and neutrons will be experiencing rapid reactions with the dense radiation. The two dominant reactions are



which occur at equal rates. Thus, the numbers of protons and neutrons do not change. Equilibrium requires these two numbers to be about the same at this time.

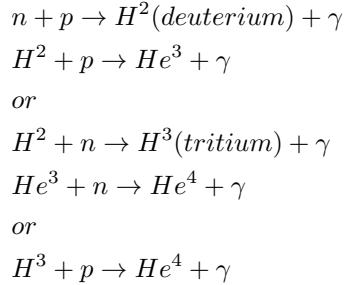
Second Frame. The temperature of the universe is $3 \times 10^{10} \text{ }^{\circ}\text{K}$. Since the first frame 0.11 seconds have elapsed. Nothing has changed qualitatively - the contents of the universe are still the same and all are still in thermal equilibrium and high above their threshold temperatures. Thus, the energy density has simply decreased like T^4 to about 3×10^7 that of the rest energy density of water. The rate of expansion has decreased like T^2 so the characteristic expansion time of the universe has now increased to 0.2 seconds. No nuclei form as yet, but the rate of neutrons to protons is now greater than for protons into neutrons and the nuclear balance has shifted from 50-50 to 38-62 neutrons-protons.

Third Frame. The temperature of the universe is $10^{10} \text{ }^{\circ}\text{K}$. Since the first

frame 1.09 seconds have elapsed. At about this time, the decreasing density and temperature have increased the mean free time (between collisions) of neutrinos so much that they are beginning to act like free particles (no interactions), no longer in thermal equilibrium with electrons, positrons or photons. They no longer play any role in the evolution of the universe except that their energy will still act as part of the source of the gravitational field. The neutrino wavelength will continue to stretch in direct proportion to the size of the universe. The total energy density has continued to decrease as T^4 to about 4×10^5 that of the rest energy density of water. The rate of expansion has decreased like T^2 so the characteristic expansion time of the universe has now increased to 2.0 seconds. The temperature is now only 2 times the electron threshold temperature - this means they are just beginning to annihilate more rapidly than they are recreated out of radiation. It is still too hot for nuclei to form. The nuclear balance has shifted to 24-76 neutrons-protons.

Fourth Frame. The temperature of the universe is $3 \times 10^9 \text{ }^\circ K$. Since the first frame 13.82 seconds have elapsed. We have dropped below the threshold temperature for electrons and positrons - they are rapidly disappearing as major constituents of the universe. The energy released in their annihilation slows down the cooling of the universe - neutrinos do not pick up any of this energy so they are 8% cooler than the electrons, positrons and photons. Since electrons and positrons are disappearing the energy density is slightly smaller than just the T^4 drop off.

The temperature is now low enough for stable nuclei like helium to form but it does not happen immediately because the expansion rate is still too rapid and the formation of helium requires a complex series of fast two-particle reactions.



To work the first step, the production of deuterium, must take place. Deuterium is a weakly bound structure compared to helium. It takes only 1/9 as much energy to break apart deuterium as it takes to break apart helium. At $3 \times 10^9 \text{ }^\circ K$ nuclei of deuterium are broken apart as soon as they are formed and thus the process to produce helium cannot take place. The nuclear balance has now shifted to 17-83 neutrons-protons.

Fifth Frame. The temperature of the universe is $10^9 \text{ }^\circ K$, which is only 70 times hotter than the center of the sun. Since the first frame 3 minutes and 2

seconds have elapsed. Electrons and positrons have mostly disappeared leaving photons and neutrinos. The energy released by electron-positron annihilations has given the photons a temperature 35% higher than that of the neutrinos. The universe is now cool enough for H^3 and He^3 and He^4 to hold together - the deuterium bottleneck, however is still effective and the heavier nuclei do not form as yet. Most particle collisions have ceased. Now however, the decay of the neutron (lifetime = 15 minutes) now starts to become important. In each 100 seconds, 10% of the remaining neutrons decay into protons. The nuclear balance has now shifted to 14-86 neutrons-protons.

A Little Later. Shortly after the fifth frame a dramatic event occurs - the temperature drops to a point at which deuterium nuclei hold together thus eliminating the bottleneck to the production of heavier nuclei. The production stop at helium however because of other bottleneck - there are no stable nuclei with 5 or 8 nuclear particles! At this point all the remaining neutrons are cooked into helium nuclei. For a density of 10^9 photons per nuclear particle, this nucleosynthesis will begin at a temperature of $9 \times 10^8 \text{ }^\circ K$. Approximately 3 minutes and 46 seconds have passed. Neutron decay had shifted the nuclear balance to 13-87 neutrons-protons as nucleosynthesis begins. After nucleosynthesis, the fraction by weight of helium is just equal to the fraction of all nuclear particles that are bound into helium; half are neutrons and essentially all neutrons are bound into helium so that the fraction by weight of helium is twice the fraction of neutrons among nuclear particles or about 26%. If the nuclear particle density were a little higher, then nucleosynthesis would start earlier when not so many neutrons have decayed and thus more helium would have been produced (not more than 28% by weight). See the figure below.

What now happens in the next frame?

Sixth Frame. The temperature of the universe is $3 \times 10^8 \text{ }^\circ K$. Since the first frame 34 minutes and 40 seconds have elapsed. Electrons and positrons are almost completely annihilated except for a small (one part in 10^9) excess of electrons needed to balance the charge of the protons. The energy released by electron-positron annihilations has given the photons a temperature 40.1% higher than that of the neutrinos. The total energy density has continued to decrease to 9.9% that of water. Of this 31% is in the form of neutrinos and 69% in form of photons. The characteristic expansion time is now 1 – 1/4 hours. Nuclear processes have stopped - the nuclear particles are now either bound in helium nuclei or are free protons (hydrogen nuclei) with about 22 – 28% by weight. The universe is still too hot for stable atoms to form.

The universe continues to expand and cool without any significant events for about 700000 years. At that time the temperature will have dropped enough that electrons and nuclei can form stable atoms. The disappearance of free electrons makes the universe transparent to radiation. The decoupling or radiation and matter now allows matter to begin to form stars and galaxies under the influence of gravity.

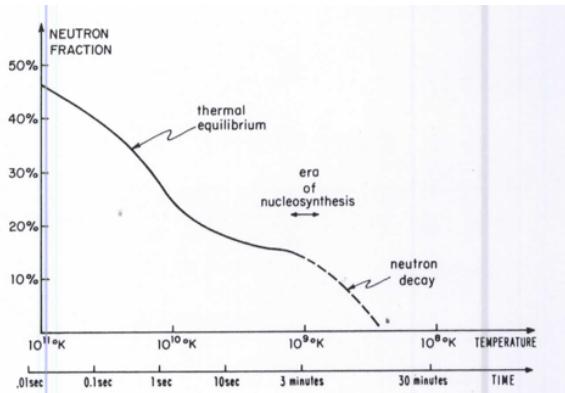


Figure 9. *The Shifting Neutron-Proton Balance.* The fraction of neutrons to all nuclear particles is shown as a function both of temperature and of time. The part of the curve marked "thermal equilibrium" describes the period in which densities and temperature are so high that thermal equilibrium is maintained among all particles; the neutron fraction here can be calculated from the neutron-proton mass difference, using the rules of statistical mechanics. The part of the curve marked "neutron decay" describes the period in which all neutron-proton conversion processes have ceased, except for the radioactive decay of the free neutron. The intervening part of the curve depends on detailed calculations of weak-interaction transition rates. The dashed part of the curve shows what would happen if nuclei were somehow prevented from forming. Actually, at a time somewhere within the period indicated by the arrow marked "era of nucleosynthesis," neutrons are rapidly assembled into helium nuclei, and the neutron-proton ratio is frozen at the value it has at that time. This curve can also be used to estimate the fraction (by weight) of cosmologically produced helium: for any given value of the temperature or the time of nucleosynthesis, it is just twice the neutron fraction at that time.

That is the theory - any theory must be testable by experimental observations. The 22 – 28% by weight of helium has been confirmed not only in stars but in the universe as a whole. On the other hand, the abundance of heavier elements does vary dramatically from place to place. This is what is expected if the heavy elements were produced in stars, but helium produced in early universe before any stars existed. Small traces of deuterium have also been found. Deuterium abundance is very sensitive to the density of nuclear particles at the time of nucleosynthesis - higher densities make nuclear reactions proceed faster so all deuterium would have been cooked into helium. Calculation show that the abundance of deuterium (by weight) produced in the early universe is as shown in the table below.

Photons/nuclear particle	Deuterium abundance(parts per million)
10^8	0.00008
10^9	16
10^{10}	600

If we could determine the primordial deuterium abundance that existed before stellar cooking began we could make a precise determination of the photon-to-nuclear particle ratio. Knowing the present radiation temperature of $3^{\circ}K$, we could determine a precise value for the present nuclear mass density of the universe and decide whether it is open or closed.

Satellite observations indicates deuterium abundance of 20 parts per million and that leads to a present density value of 500 nuclear particles per million liters. This is very much less than the critical density for a closed universe. - therefore the universe is open and will expand forever. We will have more to say about this matter latter on - beyond what Weinberg knew about!

Some of the conclusions that Weinberg has drawn in his account of the first three minutes must seem very tenuous and based on flimsy evidence at best. But that is the way physics progresses - keep an open mind - try all kinds of theories based on one's best assumptions - test the consequences - continue until get a better and better theory. One must put forward all kinds of ideas in order to find the right ones in the end. As we will see shortly the model has its flaws and new experiments will lead to changes that make a better theory based on the old one.

4.6 A Historical Diversion

Just read this historical material on your own.

4.7 The First One-Hundredth Second

Here we now substitute a more modern approach due to Alan Guth.

4.8 What Lies Ahead

Here we now substitute a more modern approach due to Alan Guth.

Chapter 5

Notes on Inflationary Universe

5.1 The Standard Model of Particle Physics: 1970's

In the early days of particle physics it was thought that there existed four fundamental interactions - collisions, decays, annihilation and creation processes, and so on. They were (from weakest to strongest):

1. Gravitation is the interaction we have been talking about during this class. It is the weakest of the known forces on a small scale, but dominates the others on the cosmological scale. The force of gravity between two elementary particles, supposedly mediated by the exchange of a particle called a *graviton*, is so weak that it has never been detected in experiments. For example, the gravitational force between a proton and an electron is 2×10^{39} times weaker than the electrical attraction.
2. Weak interactions account for the radioactive decay of nuclei. The decay of the neutron into a proton + electron + antineutrino is an example of a weak interaction. It is very short range (1/100 of the size of an atomic nucleus) so has no effect in everyday events.
3. Electromagnetism like gravity, plays a major role in everyday life. It includes both electric forces and magnetic forces. Special relativity is important since magnetic fields are just electric fields observed in a moving frame. Changing electric fields produce magnetic fields and vice versa. Electric forces bind the electrons of an atom to its nucleus. Light is composed of electromagnetic waves, as are microwave, radio waves, gamma rays and x-rays. Since ordinary matter is neutral (no net charge) all attractive and repulsive electric forces essentially cancel in the everyday world. Good thing - if we removed all the negative charge from two one-pound iron balls, then at one foot apart, they would repel each other with a force of 7×10^{18} tons.
4. The strong interactions, which have a range roughly the size of a nucleus (10^{-13} cm), bind protons and neutrons inside a nucleus. The protons and

neutrons are each believed (as we will see) to be composed of three quarks also held together by the strong interactions.

Before 19070, one of these interactions - electromagnetism - was described by a successful theory due to Maxwell. It was completely consistent with special relativity. When quantum theory came on the scene the view of electromagnetism had to be modified - waves were replaced by photons. Quantum theory generally predicts no unique outcome for an experiment, but instead predicts the probabilities of alternative outcomes. The behavior of any one photon is unpredictable, but the average behavior of large numbers of photons closely mimics the continuous waves of the classical Maxwell equations.

The full theory of relativistic electrons and positrons interacting with photons - quantum electrodynamics(QED) - was formulated in the 1930s and made fully successful(elimination of mathematical inconsistencies by a technique called *renormalization*) in the 1940s. QED is the most precise physical theory ever devised - its predictions agree with experiment to 11 or more decimal places! The interaction was due to the exchange of photons between particle with charge.

The other three interactions were only partly understood at this time. Gravitation, which had been described by Newton was superseded by Einstein's General Relativity. General Relativity seems to work very accurately for classical experiments such as planetary motions and so on. But it is a purely classical theory and therefore cannot be the final answer. As formulated by Einstein, it is inconsistent with quantum theory. Attempts were made to along the same lines as QED to construct a quantum gravity theory - they failed because the renormalization tricks would not work in this case. There existed a successful weak interaction theory in the 1970s. It was similar in form to QED. But had the same problems as quantum gravity with renormalization although a point-interaction(extreme short range) model seemed to have some success. For strong interactions the level of ignorance was very high. Patterns of behavior had been noticed and the quark idea had been proposed in 1964. Nobody understood how quarks interact or why no experiment had shown any direct evidence of their existence. Confusion reigned!

The 1970s saw the creation of the theory called the *standard model of particle physics*. The weak, electromagnetic and strong interactions are all described with sufficient accuracy to agree with every experiment. The model proposes that nature can be described by a set of fundamental building blocks and a set of equations describing how they interact. The fundamental particles are shown below.

	1st Generation	2nd Generation	3rd Generation
Quarks:	Up Down	Charm Strange	Top Bottom
Leptons:	Electron Electron-Neutrino	Muon Muon-Neutrino	Tau Tau-Neutrino
Force Carriers:	Weak Interactions: W^+, W^-, Z^0		
	Electromagnetism: Photon		
	Strong Interactions: 8 Gluons		
	1 Higgs Particle		

Figure 5.1: The Content of the Standard Model

The first part of the puzzle was solved by Weinberg and Salam in 1968 in a unified theory of weak and electromagnetic interactions called the *electroweak* (EW) theory. The transmitters of the weak interactions were the neutral Z^0 and a pair of oppositely charged particles - the W^\pm . While the photon is massless and always travels at the speed of light, the weak transmitters are very massive and move slowly. The theory was renormalized (all mathematical inconsistencies removed) in 1971 by t'Hooft. The theory worked in that it agreed with experiment.

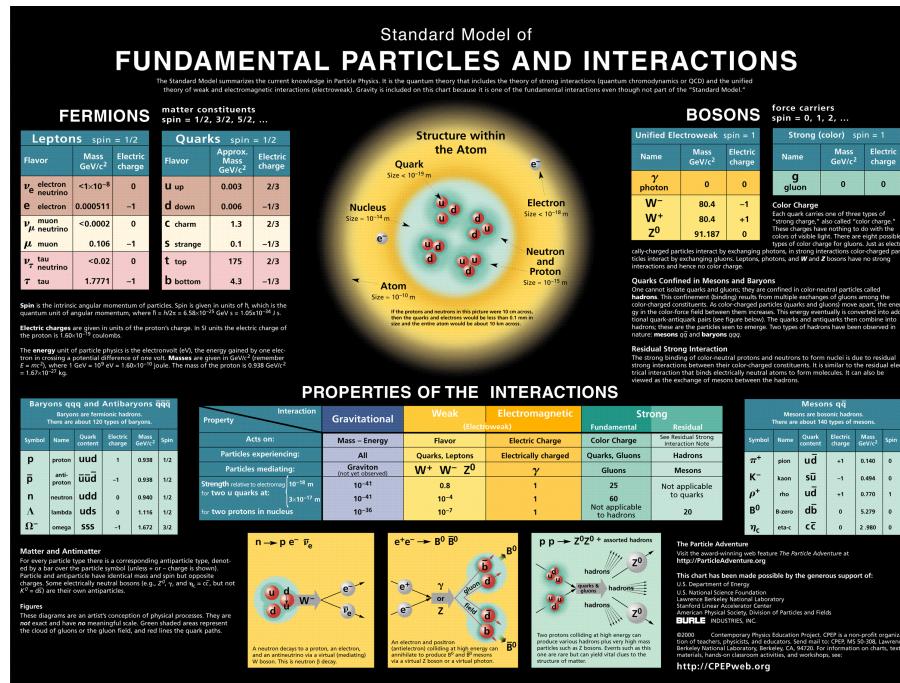
At about the same time, the quark idea was used to explain all of the observed particles that were appearing experimentally due to the strong interaction. The quarks were very strange objects - they had never been observed in the laboratory, they had fractional charge, and so on. A summary of the original quark model is shown below.

QUARKS		ANTIQUARKS	
Flavor of Quark	Charge	Flavor of Antiquark	Charge
Up	$+\frac{2}{3}e$	\bar{U}	$-\frac{2}{3}e$
Down	$-\frac{1}{3}e$	\bar{D}	$+\frac{1}{3}e$
Strange	$-\frac{1}{3}e$	\bar{S}	$+\frac{1}{3}e$

Proton	Neutron	Δ^{++}	π^+

Several physicists working with a theoretical ideas first formulated in 1954 by Yang and Mills - so-called *gauge theories* (QED is an example) invented quantum chromodynamics(QCD) - a theory of the strong interaction. While other theory describe interactions which get stronger as the energy increase, in QCD the interactions become weaker - called *asymptotic freedom*. QCD proposed the three flavors of quarks - up, down and strange - each exist in three variations called *colors* - nothing to do with real color! The colors are the equivalent of charge for QED. The Yang-Mills interactions between the nine quarks are transmitted by 8 massless gluons (the role of the photon in QED and The Z and W particles in the electroweak theory - both of which are also Yang-Mills type theories!). Asymptotic freedom means that quarks interact very weakly in high energy processes such as collisions, but within the proton, which is a low energy system, the quarks are bound strongly to each other. The theory leads to quark confinement - they cannot be observed directly - the only observable particle are *colorless* - in baryons we have three quarks - one of each color to produce a colorless particle and in meson we have a quark and an antiquark to produce a colorless particle.

So now there were two working Yang-Mills theories - EW and QCD. Developments, both experimental and theoretical in the 1970s led to the standard model - a combination of EW and QCD. The particles involved are shown below.



The standard model works! The only interaction now left out was gravity.

5.2 Grand Unified Theories (GUTs)

The basic goal of a GUT are theories of weak, electromagnetic and strong interactions in terms of a single unified interaction (not three distinct interaction as in the standard model). GUT are attractive for these two reasons.

1. GUTs are the only known theories that predict that the proton charge equals the electron charge in magnitude. We know they are equal experimentally, but no one knew why! In the standard model they could have any value - they must be restricted a priori to be equal! GUTs contain a fundamental symmetry (like invariance under rotations) that relates the behavior of electron (actually all 6 leptons) to the behavior of the 6 quarks, two of which make up a proton. The symmetry guarantees that the charges are equal - if not - if the symmetry was violated - if the symmetry were broken - by even the smallest amount, then the theory would no longer be well-defined mathematically - GUTs would have to be abandoned even if the charges were different in the 22nd decimal place! The continued experimental equality confirms that we should be looking for a GUT to go beyond the standard model.
2. The GUT theory makes an interesting prediction about the relative strengths of the three interactions. The standard model describe the observed properties of the weak, electromagnetic and strong interactions in terms of three fundamental interactions. One is the color interactions of QCD describing the strong interaction and labeled by the symbol SU(3) (represents the group of symmetries associated with the strong interaction - what is conserved and what changes under the exchange of gluons). The weak and electromagnetic interactions, although unified in the EW theory, are nonetheless described in the theory by two fundamental interactions, SU(2) and U(1) - the EW theory is unified in the sense that the SU(2) and U(1) interactions are twisted together to describe both the weak and electromagnetic interactions. Each of the three interactions SU(3), SU(2) and U(1) has a different strength. The value of the strength for each is not predicted by the standard model theory, but must be input from experiment.

Calculations in the standard model seem to show, however, that at higher energies the strength vary and start to approach each other in value as shown below.

We note that the left of the curve correspond to the typical energies in present day experiments. All three strength line seem to meet at one point at an energy above 10^{16} GeV (note that the LHC has an energy of $10 \text{ TeV} = 10^4 \text{ GeV}$ which is significantly smaller). While two curves will typically cross at a single point, there is no a priori reason for all three curves to meet at a single point unless there is some underlying reason. GUTs provide a reason (the standard model does not). Based on the idea that the three fundamental interactions of the standard model actually arise from a single fundamental interaction, GUTs

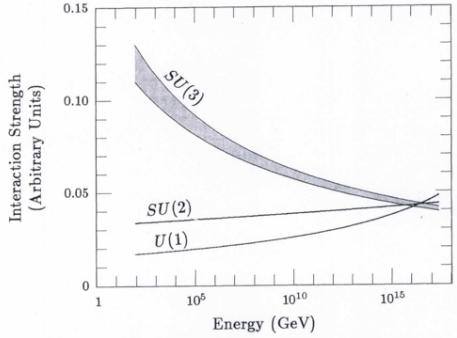


Figure 5.2: Dependence of Interaction Strength on Energy

imply that all three curves must meet. Thus if any two are measured the third can be predicted. The GUT predictions agree with experiment.

How, at energies below 10^{16} GeV , does the single *unified* interaction simulate the existence of three very distinct types of interactions - the strong, weak and electromagnetic? The secret behind this masquerade is a phenomenon called *spontaneous symmetry breaking*. Spontaneous symmetry breaking is not a new idea associated with GUTs. The concept was also used to construct the electroweak theory and is closely connected to the Higgs particle that was in an earlier table but we have not yet said anything about. According to the general definition, a spontaneously broken symmetry is a symmetry which is present in the underlying theory describing a system, but which is hidden when the system is in its equilibrium state. It also occurs in much simpler systems (than GUTs or EW theory) such as the formation of a crystal. Let us use this latter example to see how spontaneous symmetry breaking works.

To make the analogy to GUTs as clear as possible, we consider a very simple type of crystal called *orthorhombic* (the mineral topaz). These crystals have a rectangular structure as shown below

so all the angles are right angles. Unlike a cubic crystal, the three principal lengths of the orthorhombic crystal are different (important for comparison to GUTs). An outline of the crystal/GUT analogy is shown below.

Starting at the top of the table, the first row indicates the symmetry that is involved. For the case of a crystal, the relevant symmetry is rotational invariance. The physical laws that describe the system are rotationally invariant - they make no distinction between one direction in space and another. In the case of the GUT, the symmetry is more abstract, having nothing to do with orientation in physical space. Instead, the symmetry is what is sometimes called an *internal* symmetry - one that relates the behavior of one type of particle to the behavior of another. In this case, the underlying symmetry of the GUT has two manifestations.

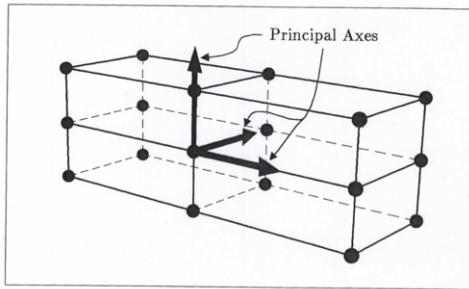


Figure 5.3: The Structure of an Orthorhombic Crystal

	CRYSTAL	GUTs
SYMMETRY	Rotational Invariance	Three interactions indistinguishable Electron, neutrino, and quark indistinguishable
SPONTANEOUS SYMMETRY BREAKING	Crystal axes pick out three distinguishable directions	Higgs fields pick out three distinguishable particles — electron, neutrino, and quark — and also three distinguishable interactions
LOW ENERGY PHYSICS	Three fundamental axes of space Three fundamental speeds of light	Three distinguishable particles Three distinguishable interactions
HIGH TEMPERATURE PHYSICS	Crystal melts—rotational invariance restored	Phase transition at $T \approx 10^{29} \text{ }^{\circ}\text{K}$ — Symmetry restored

1. The symmetry implies that the three interactions of the standard model SU(3), SU(2) and U(1) are really one interaction and hence indistinguishable.
2. The GUT symmetry implies that the underlying laws of physics make no distinction between an electron, a neutrino, or a quark (the 12 constituents of all matter).

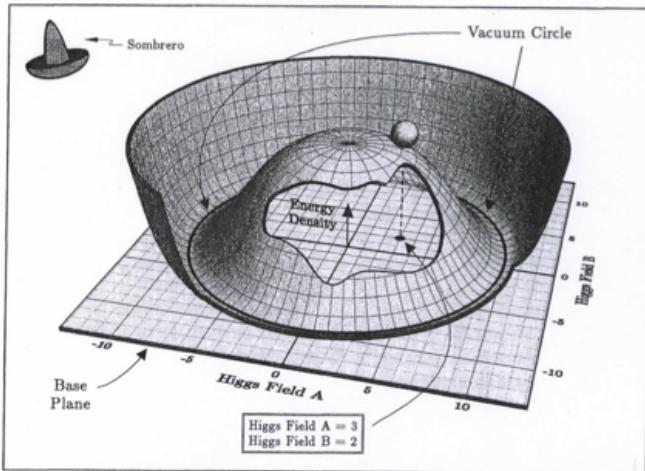
Both manifestations of the GUT symmetry are analogous to the indistinguishability of different directions of space for the crystal.

The second row of the table describes the mechanism of spontaneous symmetry breaking - what is it that breaks the symmetry? In the case of the crystal, the atoms arrange themselves along crystal axes which are picked out at random by the first few atoms as the crystal starts to grow. Thus the three directions of the principal axes (see figure) become distinguishable from each other and from other directions. In the construction of GUTs, theorists include a set of fields(particles) specifically for the purpose of spontaneously breaking the symmetry. These fields are known as *Higgs fields* and the spontaneous symmetry breaking mechanism is called the *Higgs mechanism*. Such Higgs fields were introduced to make EW theory work and after that were a much used part of the theoretical physicist's toolkit. In the standard model of particle physics, for example, a field is introduced for each of the fundamental particles. The electromagnetic field associated with the photon is familiar - this photon picture was the prototype on which the rest of the theories are based. The standard model includes a field associated with the electron, a field associated with the green-bottom quark, and so on. All these fields, including the electromagnetic and Higgs fields, are treated on an equal footing. Each field is postulated to exist, and to evolve according to a specified set of equations. Quantum theory implies that the energy of the electromagnetic field is concentrated in bundles, called photons, which can be interpreted as particles of light. Similarly, quantum theory implies that the energy of the electron field is concentrated in bundles which we interpret as particles called electrons. In modern particle theories every fundamental particle is described as a bundle of energy of some field. The energy of the Higgs fields are concentrated into Higgs particles. The Higgs particles associated with the breaking of the GUT symmetry are expected to have masses corresponding to energies in the vicinity of 10^{16} GeV , which means that they are far too massive to be observed experimentally! The Higgs particle of the standard model, however, is expected to have a mass in the vicinity of 10^3 GeV and it is being looked for in the LHC.

Although ideas from quantum theory were needed above to explain how particles arise from fields, the mechanism of spontaneous symmetry breaking itself can be described classically. Physicists designed GUTs so that so that they will lead to spontaneous symmetry breaking, by formulating them so that the energy density of the Higgs fields behaves in a peculiar way. For most fields, such as electric and magnetic fields, the energy density of the field has its lowest possible value - zero - when the field vanishes. For the Higgs fields, however, the theories are constructed so that the energy density is lowest when the Higgs fields have nonzero values. The Higgs fields in empty space - called the *vacuum* - will therefore have nonzero values, since they will settle into the state of lowest possible energy density.

The figure below illustrates a sample energy density diagram for a set of two

Higgs fields.



The simplest GUT actually requires 24 Higgs fields, but two will be enough to see how the mechanism works. Call the two Higgs fields A and B. Since the Higgs fields interact strongly with each other, the total energy density depends on both of the Higgs fields and cannot be expressed as a simple sum of the individual energy densities. The 3D diagram show the energy density of the Higgs fields, for any specified values for the two Higgs fields.

To understand the implications of this diagram, imagine a small ball located on the surface, as shown, directly above the point in the base plane corresponding to the values of the Higgs fields. Since energy is required to lift the ball, the gravitational energy increases with height. Thus, the gravitational energy of the ball is proportional to the energy density of the Higgs fields, which on the diagram is the height of the surface. Although the properties of gravitational energy may not be familiar, the *effects* of gravity are easy to visualize. The ball will be pulled downward toward the curve labeled *vacuum circle*. The evolution of the Higgs field is very similar to the motion of the ball.

If both Higgs fields vanish, then the imaginary ball is sitting at the peak of the mountain in the middle so that the energy is relatively high. We note that this same mechanism works successfully in exactly this way in the EW theory so nature already behaves this way in a known case. Now, clearly, the state of lowest possible energy density is not unique - any of the points on the vacuum circle correspond to zero energy density. The imaginary rolling ball can come to rest eventually at any point on this circle. Thus, the values of the Higgs fields are not determined by energy considerations alone. Just as the atoms in the crystal can align equally well along an infinite number of possible orientations, the set of Higgs fields in the vacuum can settle equally well at any point on

the vacuum circle. Some particular point on the vacuum circle would be chosen randomly in the early history of the universe, just as the directions of the crystalline axes are chosen randomly as the crystal begins to form. This random choice of nonzero Higgs field values breaks the GUT symmetry, just as the particular orientation of the crystal breaks the rotational symmetry. In both cases, the underlying laws of physics remain exactly symmetric - the symmetry is broken *spontaneously* in the sense that it is only an accident of history that chooses the orientation of the crystal or the point on the vacuum circle for the Higgs fields.

The other particles in the theory interact with the Higgs fields so they are affected by the random choice of Higgs fields values. Since different particles interact with different Higgs fields, distinctions arise between particles that would otherwise be indistinguishable. For example, suppose that the fields in the vacuum settle at the point for which Higgs field A has a value 10 and Higgs field B has the value 0. Then, one might guess, the particles that interact with Higgs field A will behave very differently from those particles that interact with Higgs field B. Just as a finger pressing against a violin string can radically alter the pattern of its vibrations, the large value of Higgs field A in this example can radically change the patterns of vibrations of the other fields with which it interacts. Since particles are the bundles of energy of the vibrating fields, the properties of the particles are dramatically affected. In particular, the mass of a particle is determined by its interaction with the Higgs fields, so the masses of the particles that interact with Higgs field A will become different from the masses of the particles that interact with Higgs field B.

In a full GUT with a larger number of Higgs fields, some of the particles are caused to act like electrons, some like neutrinos and others like quarks. Similarly, some force-carrying particles will be caused to act like the gluons of the strong interactions, others will be caused to act like W^\pm and Z^0 of the weak interactions and one will be caused to act like the photon of the electromagnetic interaction. The distinctions are a direct cause of the way the force-carrying particles interact with the Higgs fields. Since the masses of the force-carrying particles is caused by the Higgs mechanism, the large masses of the W^\pm and Z^0 (order of 90 proton masses) are attributed to the Higgs mechanism. Quantum theory says that the range of a force is inversely proportional to the mass of the exchanged force-carrying particle - thus the very short range of the weak interaction is understood as a consequence of the Higgs mechanism.

The third row in the table describes the behavior of low energy physics in the two systems. Let us imagine intelligent creatures - orthorhombons - living in the crystal. They can move about and do experiments but cannot perturb the crystal in which they live. They would consider the crystal as a fixed property of space. Their texts would not mention rotational symmetry but would instead discuss space and its three primary axes. The structure would affect the speed of light in different direction and their table of constants would list the three

different speed values.

If GUTs are correct, then our universe is similar to this crystal world. Orthorhombons live inside a crystal, the effects of which they mistakenly view as fixed attributes of space. We live in a region full of Higgs fields the effects of which we mistakenly view as fixed attributes of the laws of physics. Our tabulation of properties of three distinct interactions is the same as their three speeds of light. Similarly, the distinct properties that we observe for electrons, neutrinos and quarks are not fundamental - they represent the different ways the particles interact with the fixed Higgs fields that exist everywhere.

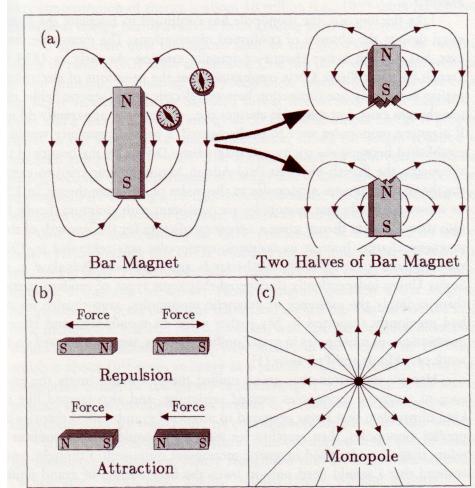
Finally, the last row of the table describes the high temperature behavior of the two systems. If the crystal is heated sufficiently, it will melt and become a liquid - this is a phase transition - a sudden change in the behavior of the system as the temperature is varied. Rotational symmetry is thus restored at high temperatures. In GUTs, there also exists a symmetry-restoring phase transition that occurs at very high temperatures. To visualize this transition, remember that at zero temperature the two Higgs fields assume a random pair of values on the vacuum circle. As the temperature is raised, the Higgs fields acquire thermal energy and begin to oscillate - as if the surface were connected to a vibrator which cause the ball to move and jiggle about. When the temperature is low, we only have small oscillation which remain centered on the vacuum circle. Since, on the average, the values of the Higgs fields are still nonzero, the symmetry remains spontaneously broken. Once the temperature exceeds a certain value, however, the ball begins to move about wildly, sometimes crossing over the central peak. Its average position becomes the center of the peak (the highest point) and all evidence of the initial zero-temperature values of the Higgs fields is lost. Each Higgs field then has an average value of zero so the GUT symmetry is restored. The SU(3), SU(2) and U(1) interactions all merge into a single interaction and there is no distinction between electrons, neutrinos and quarks.

For a typical GUT, this transition occurs at $10^{29} \text{ }^{\circ}\text{K}$ or an average thermal energy of 10^{16} GeV ! The diagram showing the three strengths meeting at a point now tells this story. The unification scale, where the three lines meet, is identified with the typical energies of the Higgs fields. Above that point we have exact GUT symmetry(the initial Higgs fields values have no effect) and all particles are identical and there exists only a single interaction. Below that point, the effect of the initial Higgs fields is substantial. We have three different interaction and the standard model!

As we will see, GUTs not only are a good model of the world of elementary particles, but through the inflationary universe theory we will soon describe, they also can help explain the origin and structure of the universe.

5.3 The Magnetic Monopole Problem

Magnetic monopoles are hypothetical particles that produce a special kind of magnetic field. The magnetic field of ordinary magnets is produced by the motion of electrons in the material and all such magnets have a N and S pole of equal strength as shown below.



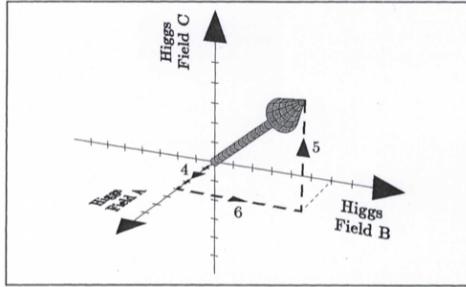
The magnetic field lines which can be followed by placing a compass near the magnet extend from the N pole to the S pole. One can verify that two N poles(or two S poles) repel while a N pole and a S pole attract. If the magnet is broken in two, one does not obtain separate N and S poles, each piece is just a smaller magnet with both a N and S pole. A monopole is an isolate pole, either N or S. The magnetic field of a monopole points radially outward just as the electric field for an electric charge or the gravitational field of a spherical mass. So far magnetic monopoles have not been observed.

There was a potential problem for GUTs however. It seems that GUTs contain magnetic monopoles - they are extraordinarily massive on the order of 10^{17} GeV . Thus, there was not any expectation of see any monopoles in any high-energy experiments that could be done. On the other hand, the question arises - what about the big bang? How many monopoles would have been produced in the big bang?

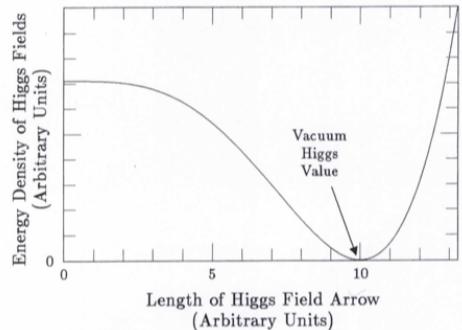
The calculations are very difficult, but the conclusions were inescapable - GUTs lead to a fantastic overproduction of magnetic monopoles in the early universe - hence the magnetic monopole problem.

The simplest GUT that contain magnetic monopoles has three Higgs fields A, B, and C. Remember each Higgs filed is specified by one number - its value. The

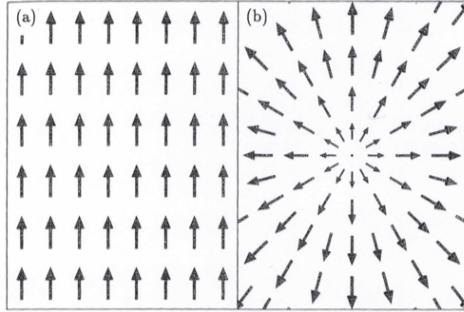
figure below show a graphical way of describing the value of the three Higgs fields at any point.



We then need to draw such an arrow at every point to completely describe the Higgs fields, although we only need to draw a representative sample to understand what is going on. The energy density depends on the length of the arrow. It is high when the arrow has zero length and has a minimum value when the arrow has a specific nonzero length - on the vacuum circle. Same as earlier discuss, except three Higgs field axes instead of two (surface is in 4 dimensions!). In the vacuum, energy density is at minimum and Higgs field arrow length equals the vacuum Higgs value, but its direction is arbitrary. Once again the Higgs filed is not fixed by energy considerations alone but determined by random processes in the early universe. In addition to energy density of the Higgs field as shown below



the theory allows that any variation of the Higgs field from point to point in space also contributes to the energy. Since the vacuum has the least possible energy, the Higgs fields in the vacuum cannot vary from place to place. The figure below(a) shows one possible picture of the Higgs fields in the vacuum.



In this case, only Higgs field C is nonzero, but that is an arbitrary choice(only one will be nonzero - we do not know which one).

The Higgs fields of the monopole are more difficult to visualize since the monopole is a 3D object. The part (b) of the above figure shows a 2D slice of the monopole field (a plane that cuts through the center). All three Higgs fields vanish at the center point of the monopole - the Higgs arrow has zero length - this creates a large energy density which is the reason for the large mass of the monopole. Everywhere else the Higgs arrow points radially outward with its length increasing to a maximum of the vacuum circle value. Clearly the picture resemble the magnetic field of a monopole - hence the name magnetic monopole. The detailed calculations of the theory show that the Higgs fields picture does lead to the appropriate magnetic field picture for the magnetic monopole. The magnetic strength is 68.52 times stronger than the electric charge and hence the force between two monopoles is 4695 time stronger than the force between two electrons at the same separation.

If we consider the conditions of the early universe when the temperature was above $10^{29} \text{ }^{\circ}\text{K}$, the Higgs fields were oscillating wildly, but the average value at any point averaged over a small interval of time was equal to zero. As the universe cooled below about $10^{29} \text{ }^{\circ}\text{K}$, the Higgs fields underwent a phase transition in which the oscillations diminished and a nonzero average value was establish.

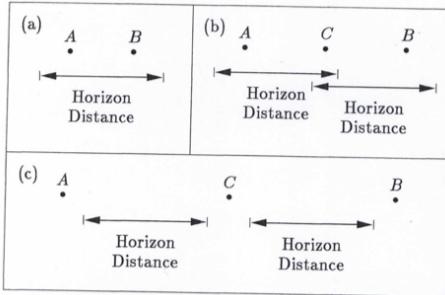
The cooling was taking place throughout the universe, but the Higgs fields at a particular location could interact only with fields in nearby regions. Thus, the random determination of the direction in which the Higgs arrow began to point was made independently in many different small regions of space. Thus, the Higgs field shortly after the phase transition could not have looked anything like the highly organized pattern that describes the vacuum as in part (a) of the above figure, but instead must have been a jumble in which the Higgs field arrow

varied randomly from one small region to another. The initial jumble would not last long however. The variation of the Higgs field from place to place requires energy and the energy density was decreasing as the universe expanded and cooled. Therefore, the Higgs fields smoothed out as the universe cooled with the Higgs arrow in one region tending to align with the Higgs arrow in neighboring regions. The tendency of the Higgs field arrows to align, however, could proceed only in regions for which the neighboring regions showed some degree of consistency. For some regions, however, this consistency would be lacking - for example - consider the point at the center of the monopole in part(b) of the above figure at which the Higgs fields are all zero. Surrounding this central point, the Higgs arrows point in every possible direction. If the Higgs field at the center point became nonzero, then the Higgs arrow would have to point in some direction. If the arrow began to align with fields in the region above the central point, the arrow would become further from alignment with the Higgs arrow below the central point. The net effect would be to increase the energy - this means that once a Higgs field pattern as in part(b) is encountered, the alignment process must stop!

The monopoles, therefore, are the surviving remnants of the chaos in the Higgs fields immediately after the phase transition.

How do we estimate the amount of monopole production?

We must understand the degree of chaos in the Higgs fields that would result from the phase transition. Now nothing travels faster than light and the phase transition happened so early in the history of the universe that even signals traveling at the speed of light could not have gotten very far. Specifically, we can invoke the *horizon distance* that we talked about earlier in Geroch - it is the total distance that a light pulse could have traveled from the instant of the big bang until the time under consideration. Suppose we imagine two points A and B immediately after the phase transition, as shown below.



If these two points are separated by more than two horizon distances, then it is not even possible for an event at some third point C to have a common influence on both points A and B. If the separation is more than tow horizon distances, the Higgs arrow at A can have no tendency whatever to align with the Higgs arrow at B. This is a measure of the chaos of the Higgs fields. They could not have been less chaotic than this! Since monopoles are remnant of the chaos, this statement about the degree of chaos can be converted into a statement about the number of monopoles produced.

At this point, the important question was whether the monopole problem could be avoided. Can we find some set of circumstances under which GUTs would be compatible with the hot big bang cosmology?

The monopole estimate is large because the horizon distance is short implying a large amount of chaos in the Higgs fields. Suppose however that the phase transition were delayed somehow. Then the horizon distance would have time to grow. The Higgs fields would have time to align over longer stretches of distance so the degree of chaos would decrease. The horizon bound on the number of monopoles would also decrease - in fact it might be possible to melt away the monopole problem.

It turns out that when the GUT phase transitions are looked at more carefully (along with horrendous calculations) that there are actually two early universe phase transitions instead of one. The monopoles are produced at the second phase transition, the one that occurs at the lower temperature. In fact it was possible for the second phase transition to occur at a low temperature - this means much later - more time - which is exactly what is needed to suppress the monopole production.

During the calculations it turned out that the range of allowed values for a set of unknown parameters was incredibly narrow! It was so narrow that the parameters would have to be specified to 12 decimal places for the monopole suppression to work. While it is true that the true values might lie in this narrow range, no reason was known as to why. Was the second phase transition real?

There was an alternative possibility however. The natural temperature of the second transition might be high but the phase transition might occur only after a large amount of *supercooling*. Supercooling is a situation in which the a phase transition is delayed so that the temperature falls well below the normal temperature of the phase transition before the transition takes place. Water, for example, can be supercooled by more than 20°C below its freezing point without turning it to ice. If the GYT phase transition were delayed by supercooling, then the monopole production would be strongly suppressed.

A phase transition with a large amount of supercooling is called a *first-order*

transition. Boiling water is an example of a first-order transition. When water boils, the temperature of the water rises slightly above the boiling point. Bubbles of steam then form randomly in the hot liquid. As each bubble grows, it absorbs energy from its surroundings, preventing the temperature from rising much above the boiling point until the water boils away. The heat energy absorbed by the growing bubbles is used to convert water to steam since steam is the higher energy phase. The description of the first-order phase transition in the early universe is very similar, except that the temperature is falling rather than rising. Thus, the temperature of the early universe would fall below the normal temperature of the phase transition. To suppress monopoles, it must fall very far below the normal temperature. Bubbles of this new phase would then begin to form randomly and grow, just like bubbles of steam. One obvious difference is that bubbles of steam rise to the surface since gravity pulls downward more strongly on the water. In the early universe, however, the bubbles would just grow spherically, expanding at nearly the speed of light until they collide with other bubbles. Eventually, one might guess, these bubbles would merge to fill all of space completing the phase transition (not exactly true as we will see).

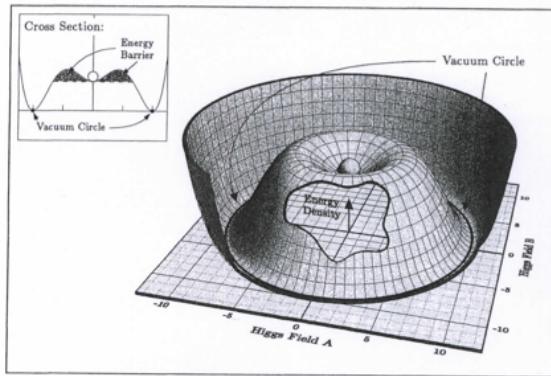
For GUT phase transitions, the phases are described in terms of Higgs fields. IN the high temperature phase before the transition, the Higgs fields are undergoing large oscillations with an average value of zero. In the new lower temperature phase, inside the growing bubbles, the decrease in available energy causes the Higgs fields to settle into a low energy state. The Higgs arrow continues to oscillate due to the thermal energy but the oscillations are now smaller than in the high temperature phase and are centered about an average value for which the length of the arrow is near the vacuum Higgs value. Within each bubble, however, the direction of the Higgs arrow is random. The Higgs arrow is nearly constant within each bubble, but the arrow of one bubble will have no tendency to align with the arrow of any other bubble. So for a first-order transitions, the degree of chaos is determined by the distribution of bubbles. Each bubble expands at nearly the speed of light after it materializes, but the rate at which bubbles materialize depends sensitively on the details of the underlying particle theory. Rapid bubble materialization would lead to a dense foam of small bubbles with a high degrees of chaos. If the bubbles materialized at a slow rate, however, then a small number of bubbles would have time to fill space, minimizing the degree of chaos. Thus slow bubble formation is needed to suppress monopole production.

5.4 The Inflationary Universe

The next question to be faced was the following - Would the gravitational field due to supercooled matter change the expansion of the universe - the assumption had been that it would not up to that point. Up to now we have been considering an energy density of the Higgs fields that was represented by the

sombrero shaped surface shown in an earlier figure. For such a shape, one does not get a large amount of supercooling. But remember, that shape was only an illustration of the properties. The actual shape of the Higgs field energy graph depends on the details of the underlying GUT - they are not known. We are only interested at this point in finding any version of a GUT that avoids the monopole problem,i.e., we just want to know it is possible.

In order to have a large amount of supercooling, the Higgs energy density must resemble the shape shown below.



As earlier, the state of lowest energy is achieved when the Higgs fields have values on the vacuum circle for which the energy density is zero. The new feature is the dip in the center of the peak. To understand how this leads to supercooling, we again visualize the values of the two Higgs fields as a ball rolling on the surface. The effects of high temperature can again be simulated by vibrations. If the temperature is high enough, then the ball is tossed about so violently by random thermal agitation that the central peak becomes insignificant and the average position of the ball is at the center. At zero temperature, however, there are no vibrations. If the ball were placed anywhere outside the dent, then it would settle into a randomly determined point on the vacuum circle. At low temperatures, the average position of the ball would lie near the vacuum circle and the ball would have small oscillations about this average position. As the system cools from high temperatures, the phase transition is marked by the average position of the ball moving away from the center.

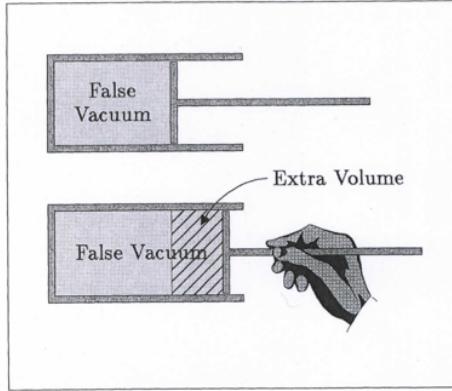
If the ball starts at high temperature and cools, however, there is a chance that it can become stuck in the dip as shown in the diagram. This is the case that corresponds to extreme supercooling since it is difficult for the ball to jump out of the dip. If the temperature fell all the way to absolute zero while the ball was in the dip, then all vibrations would cease and the ball would have no energy to jump over the energy barrier that separates it from the vacuum circle. Classical physics says that the ball will stay in the dip forever in this situation. This implies that the region of space can supercool if the Higgs fields in the region have

values near zero corresponding to the ball in the dip. The state, the product of extreme supercooling is called the *false vacuum*.

Even though classical physics would imply that the state is absolutely stable, quantum theory implies that it can decay by a process called *tunneling*. The *false vacuum* is not the state of lowest possible energy - the energy is lower on the vacuum circle. But the false vacuum can only lower its energy by quantum tunneling which is a slow process. In a time interval too short for tunneling to occur, the energy density of the false vacuum cannot be lowered. The false vacuum act *temporarily* as a vacuum so the word *false* really indicates *temporary*.

It turns out that the false vacuum has a peculiar property that makes it very different from any ordinary material. For ordinary materials, of any type, the energy density is dominated by the mass of the particles ($E = mc^2$). If the volume is increased, then the density of particles and hence the energy density decreases. The energy density of the false vacuum is not attributed to the particles but to the Higgs fields. Even as the universe expands the energy density of the false vacuum remains constant (as long as it does not decay).

The idea that a material can expand at constant energy density does not seem to make sense. Where does the energy come from? Consider the situation shown below



which shows a piston chamber filled with false vacuum surrounded by a region of ordinary vacuum. We assume that the false vacuum does not have time to decay and system is so small that the gravitational field created by the false vacuum can be ignored. What happens if the piston was moved so that the volume of false vacuum was increased as shown. The total energy inside the piston has to increase. The only possible source of this energy is the hypothetical hand that moves the piston outward. It must supply the huge amount of energy - this says that it is pulling against a very large force. The only other force acting on the piston is due to the pressure of the false vacuum. If this pressure were positive, as with normal pressure, then it would push on the piston assisting the hand (not resisting). To resist the motion, the pressure of the false vacuum must be large and *negative*!

Thus, a material with a constant energy density leads to a negative pressure. The false vacuum creates a suction, even when no pressure is applied from the outside (suction through a drinking straw is really just a pressure below that of the surrounding air - not a negative pressure). You might think that that this suction due to the false vacuum would slow the expansion (maybe even reverse it) but exactly the opposite happens! The pressure does not slow the expansion because pressure results in a force only if it nonuniform. A bottle with vacuum inside implodes - an open bottle does not even though pressure exists - it is the same on all surfaces! The false vacuum in a supercooled universe would fill space uniformly so that the forces created by the negative pressure would cancel.

Nevertheless, the negative pressure of the false vacuum leads to very peculiar gravitational effects. In general relativity, all forms of energy create gravitational fields. In addition, GR says that pressure can create a gravitational field (remember the Friedmann equations). Usually, this contribution is negligible. In this room the pressure gravitational field ratio to the mass gravitational field is 10^{-11} ! In the early universe, however, the pressure were so large that the resulting gravitational fields were important. GR says that a positive pressure creates an attractive gravitational field and a negative pressure creates a repulsive gravitational field. For the false vacuum, the repulsive component is three time larger than the attractive component.

The false vacuum actually leads to a strong gravitational repulsion.

This is the same effect as inserting a term called the cosmological constant into Einstein's equations of GR. Einstein rejected such a term because it would lead to an accelerating expansion which was not experimentally observed (it is now however and the extra term go by the exotic name of *dark energy* - more later). There is an important difference however between the two repulsive effects. The cosmological constant term is a permanent term in the equations of gravity while the false vacuum is an ephemeral state that exerts its influence for only a brief moment in the early history of the universe. Calculations show

that the gravitational repulsion causes the universe to expand exponentially. This means that the expansion is characterized by a *doubling time* which for GUTs is 10^{-37} seconds. In this interval of time all distances in the universe are stretched to double their original size and so on for each doubling time. In 100 doubling times, which is only 10^{-35} seconds, the universe would be 10^{30} times its original size - exponential growth is fast! For comparison, in standard cosmology, the universe would grow during this same time interval by only a factor of 10! An exponentially expanding space like this was actually found as a solution to Einstein's equation in the 1920s - called the de Sitter solution

Since the supercooled false vacuum state was not stable, the exponential expansion would not continue forever. Eventually the false vacuum would decay by quantum tunneling. The field does not tunnel everywhere in space at once, but instead follows the pattern of first order transitions, like the way water boils - bubbles of new phase form randomly in space, just as bubbles of steam in a pot. Each bubble begins small, but the bubbles in false vacuum grow at nearly the speed of light until the bubbles merge to fill the space. Inside each bubble is essentially the ordinary vacuum where the Higgs fields have values near the vacuum circle. The rate of bubble formation depends sensitively on the details of the theory - the decay of the false vacuum can be very fast or very slow or something in between!

The energy stored in the Higgs fields would produce high energy particles which would collide and create other particles. The product would be a hot soup of particles at high temperature, exactly as we assumed earlier for the starting point of the big bang cosmology. The excess of matter over antimatter could be established immediately after the decay of the false vacuum by standard GUT processes.

The phenomenon of exponential expansion was called *inflation*.

So the supercooled phase transition dramatically affect the expansion rate of the universe. Was this good or bad? Would this inflation period of cosmological evolution lead to some catastrophe implying that the universe could never undergone such supercooling?

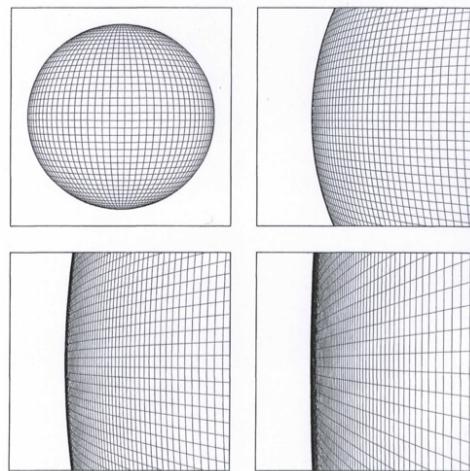
To the contrary, inflation not only dealt with the monopole problem but it also solve the *flatness problem* that had arisen. The flatness problem concerns the quantity Ω which is the ratio of the actual mass density of the universe to the critical density. The problem is caused by the instability of the situation in which $\Omega = 1$, which is like a pencil balanced on its point. If Ω is exactly equal to 1, it will remain exactly 1 forever (according to GR). But if Ω differed from 1 by a small amount in the early universe, then the deviation would grow with time and at the present time Ω would be very far from 1. Today, experiment says that Ω lies between 0.1 and 2 implying that one second after the big bang Ω must have been between 0.99999999999999 and 1.000000000000001. The big

bang cosmology, however, gives no explanation of why Ω began close to 1.

Inflation deals with the flatness problem. The effect of gravity is reversed during the period of inflation so all the equations describing the evolution of the universe are changed. In stead of Ω being driven away from 1 as it is during the rest of the history of the universe, during the period of inflation Ω is driven towards 1. In fact, it is driven towards 1 with incredible swiftness. In 100 doubling times, the difference between Ω and 1 decrease by a factor of 10^{60} . With inflation it is no longer necessary to postulate that the universe began with a value of Ω incredibly close to 1. Before inflation, it could have any value and due to the exponential nature of the inflation it would be incredibly close to 1 at the end of the inflation period. How does inflation do it?

According to GR, the mass density of the universe not only slows the cosmic expansion, but it causes the universe to curve. If we neglect the cosmological constant, then any mass density higher than the critical density causes space to curve back on itself forming a spatially closed universe. In such a universe the sum of the angles of a triangle would be more than 180° . Any mass density higher than the critical density causes space is curved in the opposite way forming a spatially open universe. In such a universe the sum of the angles of a triangle would be less than 180° . On the borderline between these two cases, when the mass density equals the critical density, the space is not curved at all meaning that ordinary Euclidean geometry is valid and the sum of the angles of a triangle is exactly 180° . If we accept this relation between Ω and geometry implied by GR, then we only need to understand why inflation drives the universe toward a state of geometric flatness.

The answer is as obvious as blowing up a balloon. The more we inflate the balloon the flatter the surface becomes as shown below.



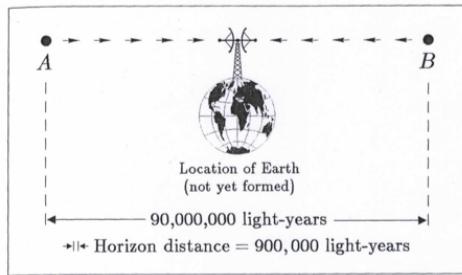
Inflation makes the universe look flat for the same reason that the surface of the earth looks flat even though we know the earth is a sphere. We only see a small region of the surface and the curvature is imperceptible.

The standard cosmological evolution would resume at the end of inflation so any deviation from flatness would begin to grow. The universe, however, would be so close to flat at the end of inflation that it would remain essentially flat even at the present time. Inflation thus leads to an experimental prediction - the present value of Ω should be very precisely equal to 1.

It turns out that inflation also solves the *horizon problem*. We have already mentioned the existence of horizons. Remember the horizon is related to the maximum distance light can travel in a given time interval. The most persuasive statement of the horizon problem, which we already mention in our discussion of Weinberg, focuses on the cosmic background radiation. Remember that until about 300000 years after the big bang, the photons of the cosmic background radiation were constant being scattered by collisions with the electrons in the hot plasma that filled the universe. At about 300000 years, however, the universe cooled enough so that the electrons combined with nuclei to form atoms. This gas is highly transparent to photons so most of the photons of the cosmic background radiation have been traveling in a straight line since that time. These photons therefore provide us with a picture of the universe at an age of 300000 years. The cosmic background radiation shows us, among other things, that the universe at 300000 years was incredibly uniform since the temperature of the radiation is found to be the same in all directions to an accuracy of 1 part in 100000. Can we understand how such extreme uniformity was established?

The general tendency of objects to come to uniform temperature (thermal equilibrium) is well understood - *zeroth law of thermodynamics*. The speed with which heat can transfer from place to place is limited by the speed of light so the transfer of heat in the early universe is limited by the horizon distance, which at 300000 years was about 900000 light-years, where the extra factor of 3 comes from the expansion of the universe - the photons make extra progress during the early period when the universe was small. If we consider two photons arriving today from opposite directions in the sky, then the mathematics of the big bang can be used to trace back the trajectories to 300000 years after the big bang. This calculation takes into account the expansion. It shows that the photons were emitted from two point about 90 million light-years apart as shown below.

In the figure, A and B label the points at which the two photons were emitted. The uniformity of the cosmic background radiation temperature implies that the temperature was the same at points A and B (to an accuracy of 1 part in 100000) yet they were separated from each other by about 100 times the horizon distance. Since nothing travels faster than light, in the context of the standard big bang theory there is no physical process that can bring these two points to the same temperature by 300000 years after the big bang.

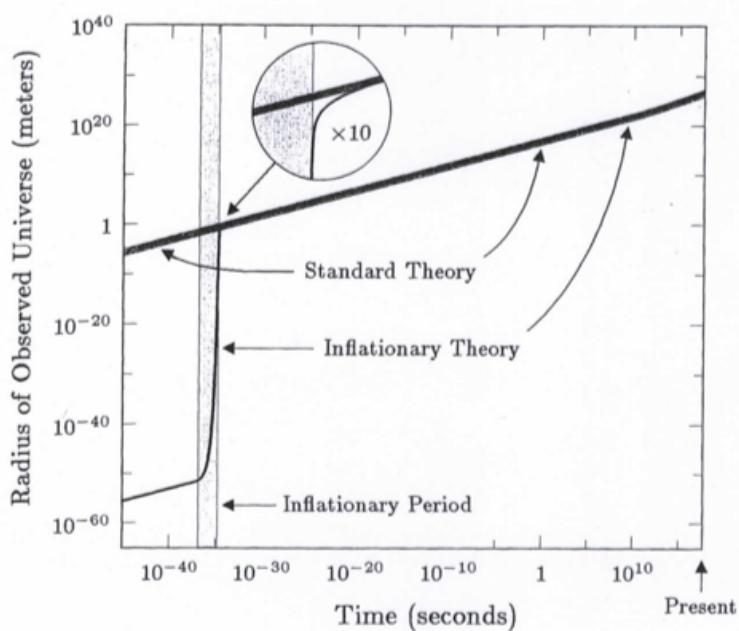


The horizon problem is not a failure of the standard big bang theory in the strict sense, since it is neither an internal contradiction nor an inconsistency between observation and theory. The uniformity of the observed universe is built into the theory by postulating that the universe began in a state of uniformity. As long as uniformity is present at the start, the evolution of the universe will preserve it. The problem, instead, is one of predictive power. One of the most important features of the observed universe - its large scale uniformity - cannot be explained by the big bang theory - it must be assumed as an initial condition!

To understand how inflation eliminates the horizon problem, the first step is to recognize that the size of the present universe is what it is - independent of any theory of how the universe evolved. Whether we believe in the standard big bang theory or inflation, the most distant objects we can detect are about 20-30 billion light-years away (any more distant objects, if they exist, are not part of the observed universe). To trace the history of the presently observed universe, however, we need to adopt a theory of how the universe evolved. The figure below shows the history of the universe in both the big bang theory and in the inflationary theory.

The two theories describe identical evolutions for all times after the end of the inflationary period, so the two curves describing the radius of the observed universe coincide for all times later than 10^{-35} seconds. During the brief period of inflation, however, the inflationary theory describes an enormous burst of expansion that is not predicted by the other theory. Thus, if we consider times earlier than the inflation period, the size of this region in inflationary theory is much smaller than in the standard theory.

In the inflationary theory the universe started out incredibly small. Before inflation, the radius of the observed universe shown is only 10^{-52} meters. The horizon problem therefore evaporates since the speed of light imposes no barrier for such a small region. There was plenty of time for such a small region to come to a uniform temperature in the same way that hot coffee in a cup cools to room temperature. Then once the uniformity was established in this small region, inflation stretched it to become large enough to encompass the entire observed universe. Thus, the uniformity in temperature throughout the observed universe is a natural consequence of inflation.



Implicit in the figure is a remarkable prediction of the inflationary theory. Due to the enormous expansion during the inflationary period, the size of the *observed* universe before inflation was absurdly small. There is no reason, however, to suppose that the size of the *entire* universe was this small. While the inflationary theory allows a wide variety of assumptions concerning the state of the universe before inflation, it seems very plausible that the size of the universe was about equal to the speed of light times its age. If the universe were much smaller than this, then it almost certainly would have already collapsed into a crunch. Applying this to the sample numbers in the figure one finds that the entire universe is expected to be at least 10^{23} times larger than the observed universe. This number arises as follows. The binning of inflation was about 10^{-37} seconds after the big bang. The speed of light times this age is 3×10^{-29} meters. The observed universe was 10^{-52} meters at this time which is 3×10^{23} times smaller. This ratio would persist so today the entire universe would be 3×10^{23} times larger than the observed universe. Thus, if the inflationary theory is correct, then the observed universe is only a minute speck in a universe that is many orders of magnitude larger.

5.5 Implications and Remaining Problems of the Inflationary Theory

Although inflation seem to be a solution for all of the problems of cosmology, there was still an important unresolved issue - how exactly does inflation end?

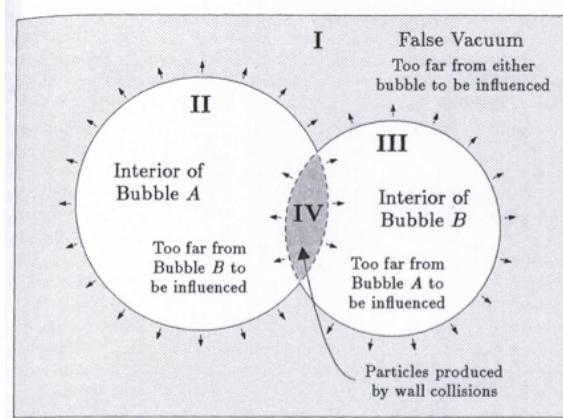
The distribution of matter in the early history of the real universe is known to have been extremely uniform as we can tell from the remarkable uniformity of the cosmic background radiation. One of the most attractive features of inflation was its ability to explain this uniformity. As inflation proceeds, the matter that was present at the beginning would be diluted to irrelevance, while space becomes filled with the exquisitely uniform mass density of the false vacuum. The complication, however, is that inflation must end. The energy of the false vacuum must be released to produce the ordinary matter that exists in the present universe. Would the uniformity produced by inflation survive the ending of inflation?

The false vacuum is an unstable state, which decays in a manner that is very similar to the way water boils. Small bubbles of normal matter (not false vacuum) would form in the midst of false vacuum (same as bubbles of steam in the midst of water heated to it boiling point) Once a bubble of normal matter forms, it immediately starts to grow. The bubble wall moves outward at a speed rapidly approaching the speed of light. If these bubble could smoothly coalesce, the uniformity of the false vacuum could be preserved. As the bubbles form and grow, the large energy density of the false vacuum (energy of the Higgs fields) is released. The released energy, however, is not distributed uniformly through space. It is concentrated, instead, in the bubble walls, which acquire more and more energy as the bubbles expand. It is only when the bubble walls collide that the energy can spread uniformly through space. The collision convert the energy to spurts of particles ejected in all directions which in turn collide with each other. Through their random motion the particles can perhaps spread to fill space uniformly. Therefore we need to understand the bubble collisions.

If the space were static or expanding moderately, then bubble collisions would be frequent. Without the exponential expansion of inflation, all the bubble walls would soon collide with other walls or with the particles produced by wall collisions. The energy from the bubble walls would be converted to normal matter and the decay of the false vacuum would be rapidly completed. There would be plenty of time for the particles to spread evenly through space, producing the uniform hot soup of matter that had been assumed as the starting point for the big bang theory.

When bubbles form during the exponential expansion of inflation, however, there is an important complication - as bubbles form randomly in space, the space between the bubbles continues to rapidly expand. Since the walls of a bubble

move outward at essentially the speed of light, and nothing can move faster, the effects of a bubble cannot extend beyond its own wall. Even when bubbles start to collide, the part of space not yet reached by any bubbles remains in the false vacuum state. The space outside the bubbles continues to exponentially expand, as if no bubbles had ever formed. An example of a collision is illustrated below.



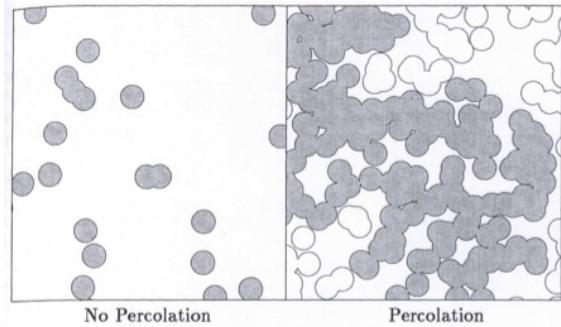
The collision is described by dividing space into four zones. Zone I (lightly shaded) is the outer region, too far from the center of either bubble to feel their influence. It is still exponentially expanding in the false vacuum state., oblivious to the formation of bubbles. Zones II and III (white) are each close enough to one of the bubbles to be influenced by it, but beyond the range of influence of the other bubble. These zones contain an almost perfect vacuum - not a false vacuum, but an ordinary vacuum - since virtually all the energy released by the decay of the false vacuum is deposited in the bubble wall. The complicated spray of particles produced by the the collision of bubble walls is restricted to Zone IV (darkly shaded), which is close enough to the centers of both bubbles to be influenced by them. The boundaries of this region are shown as dashed lines, tracing the motion the bubble walls would have followed if they had not been broken up by the collision.

If two bubbles form very near each other, the bubble walls would soon collide, and the energy in these colliding walls can be spread in all directions. But if two bubbles are not near each other when they form, then the space between them would expand so fast that the bubble walls, even moving at the speed of light, would never meet each other. The exponential expansion, therefore, drastically suppresses the collision of bubbles. It was not clear, therefore, whether these collisions would be frequent enough to spread the matter uniformly through the universe.

While false vacuum(region outside the bubbles) is decaying exponentially, those

parts that have not decayed are continuing to expand exponentially. The exponential expansions is always much faster than the exponential decay. If we followed a region of false vacuum for a time period during which half should decay(called the half-life), we would have 1/2 of the false vacuum remaining but the volume of this half would be larger than the volume of the whole region at the start - even while the false vacuum is decaying, its volume is increasing! The false vacuum never seems to completely disappear - it turns out however, it does not need to completely disappear to solve the monopole, flatness and horizon problems.

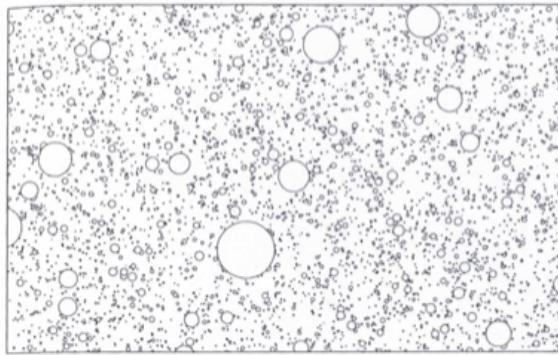
The next thing that was looked at was the way in which the growing bubbles merge into clusters. It turns out that if one placed equal-sized spheres randomly in space, allowing them to overlap, then a very curious phenomenon occurs as the density of spheres is increased. At very low densities, just a few spheres overlap with other spheres, while the majority are isolated. As the density is increased, clusters of 2 or 3 overlapping spheres become common - see left part of figure below.



If the density is further increased, the clusters continue to grow, at a rate that becomes precipitous as the fraction of space covered by the spheres approaches 29%. When the 29% threshold is crossed, the average size of a cluster literally becomes infinite. Some fraction of the spheres link together to form an infinite cluster, extending throughout all of space(see right side of above figure). This phenomenon is called *percolation*.

The bubbles that form during inflation materialize randomly at different times and then start to grow. At any instant, however, the distribution of bubbles is a collection of randomly placed spheres, so we could ask whether or not the configuration has percolated, i.e., whether or not the bubbles have fused into an infinite cluster. If the answer were yes, we still would have to investigate whether the bubble collision would suffice to spread the energy uniformly through space. However, if the answer were no - if the bubbles remained exclusively in finite-size clusters - then it would seem clear that the bubbles could never merge to form the huge region of uniformity needed for a theory of our universe.

It is hard to guess whether the bubbles would percolate, because there are arguments in both directions. On the positive side, we know that percolation occurs for equal-sized bubbles when the fraction of the volume covered by spheres is 29%. Since the bubbles in the early universe very quickly cover 99.9999999% or more, it seems very likely that they would percolate. On the other hand, the bubbles of the early universe are not all the same size. The early bubbles would grow rapidly and thus be spectacularly larger than later bubbles. This might change the percolation argument. Theorists quickly figured out that the bubbles form in the early universe would not percolate - they would remain in finite clusters as shown below.



Even when the fraction filled with bubbles exceeds 99.999999% the clusters never merge into an infinite block. The problem seems to be that the newly materializing bubbles are much smaller than those that have already been growing. As these tiny bubbles appear randomly in the gaps between previous bubbles, they cover an ever-increasing fraction of the volume without ever closing the gaps.

Each cluster, on average, is dominated by a single, largest bubble - the bubble that formed first and had the most time to grow. As the large bubble grows it collides with other bubbles, but these are typically much smaller, creating a thin fuzz around the boundaries of the large bubble. Since the bubble wall gains energy as it grows, the wall of the larger, older bubble is far more energetic than the walls of the small bubbles with which it typically collides. The wall of the large bubble rips almost undistorted through the smaller bubbles. The bulk of the energy in the universe, therefore, remains locked in the walls of the large bubbles. There would be no energy available to produce the hot soup of particles needed to start the big bang - this was named the *graceful exit problem*.

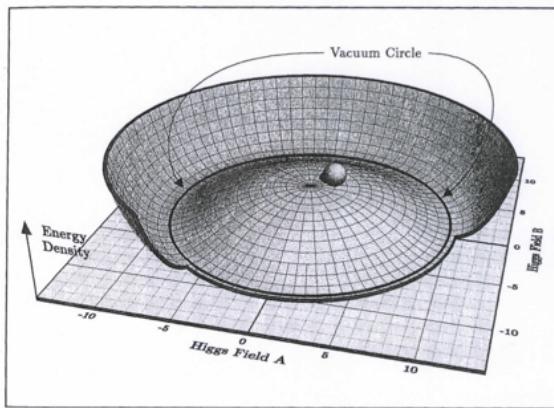
Inflation seems to be *almost* the perfect theory for describing how our universe began. It accounted for the origin of virtually all the matter in the universe and also explained why the early universe had a density so close to critical, why the cosmic background radiation is so uniform and why the universe is not inundated with magnetic monopoles. But we are left asking - could inflation end without destroying the exquisite uniformity that it creates?

5.6 A New Inflationary Theory

We need a workable ending for inflation. One effort centered on whether the universe might be contained within a single bubble. The question of merging then becomes irrelevant. First calculations, which depended critically on the detailed properties of the Higgs fields that cause the phase transition, of the inside of an isolated bubble seemed to indicate that that such a universe would be essentially empty. These calculations used what were thought to be typical properties of Higgs fields (actual properties are not well known). For example, the cosmic background radiation would have temperature of $10-29^{\circ}K$ (instead of $2.7^{\circ}K$) and Ω the ratio of the actual mass density to the critical mass density would be $10-86!$ Not very good!

Then another try(still assuming the universe is inside a bubble) used Higgs field properties that are not so *typical*. Instead of the sombrero or dented sombrero, a was a significantly different shape as shown below.

The Higgs fields that drive the inflation are a theoretical invention - the nature of these fields cannot be deduced from known physics. While the qualitative properties of the fields are modeled after the Higgs field of the EW theory(which works), the detailed properties have to be guessed (postulated). The new in-



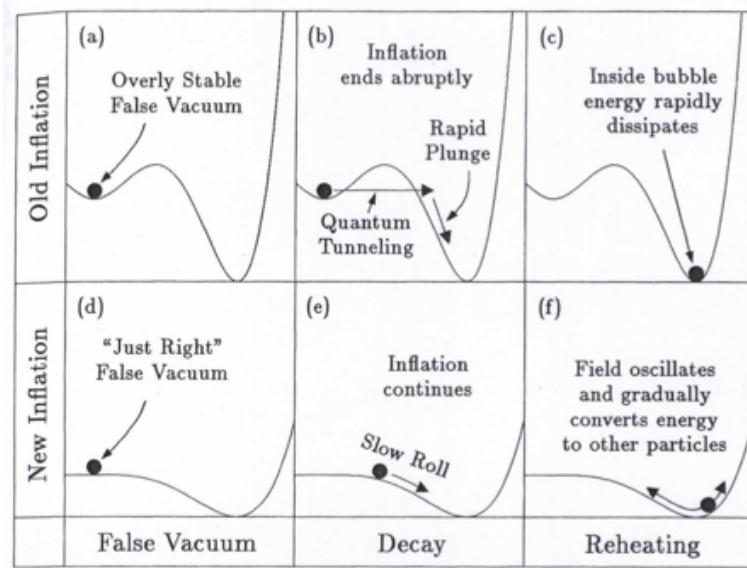
flation model chose to use the new shape for the Higgs fields - like a flattened sombrero - the middle is a flat gentle plateau rather than a rounded mountain.

The evolution of the Higgs field can again be visualized as the motion of a ball rolling on the hill of the energy diagram. For the steep hill of the old inflation model, the ball tends to rapidly move to the bottom, starting to oscillate about the vacuum circle. For the flattened mountain, however, the rolling is very slow. If the plateau is almost level near the center, then the ball will hover sluggishly before it finally begins to wander towards the vacuum circle.

As long as the ball is near the top of the hill, the energy density of the Higgs field remains high. Although this state is not as stable as the false vacuum state proposed in the old inflationary theory, it has essentially the same properties. The old false vacuum state was made more stable than needed - the ball dawdling near the center of the plateau is stable enough as it turns out. This state can also be called false vacuum and it can also drive inflation.

The figure below compares the old and new inflation theories.

In the old model, the phase transition follows the paradigm of boiling water. The Higgs field in a small, spherical region undergoes a process called quantum tunneling, which takes it from the false vacuum value to a point on the other side of the energy barrier (b). After the tunneling, the Higgs field lies on a steep part of the energy diagram hill. In the central region of the bubble, the Higgs field plummets to the bottom of the hill, quickly terminating the period of inflation. The field oscillates back and forth about the trough of the vacuum circle, but the oscillations are soon damped out (c), by a process similar to friction. The energy of the Higgs field is converted into a gas of many kinds of particles, but their density is rapidly diluted by the growth of the bubble. More energy is released as the bubble grows, since the Higgs field near the edge is plunging from the high energy value it has outside the bubble to the low energy value it has inside. The energy produced, however, moves outward with the bubble



wall, so the interior of the bubble remains essentially barren.

If the energy density of the Higgs fields is described by the flattened mountain, however, then the evolution is completely different. It is a much gentler, gradual phase transition, more like the congealing of Jell-O than the boiling of water. Inflation continues as the Higgs field begins to slowly drift away from the center of the plateau, so the energy density remains high while the bubble enlarges by many orders of magnitude (e). When the Higgs field finally slips off the plateau, the central region of the bubble has become large enough to easily contain the observed universe. As in (f) the Higgs field throughout this huge region oscillates and converts its energy to a hot soup of particles, exactly as required for the standard hot big bang model.

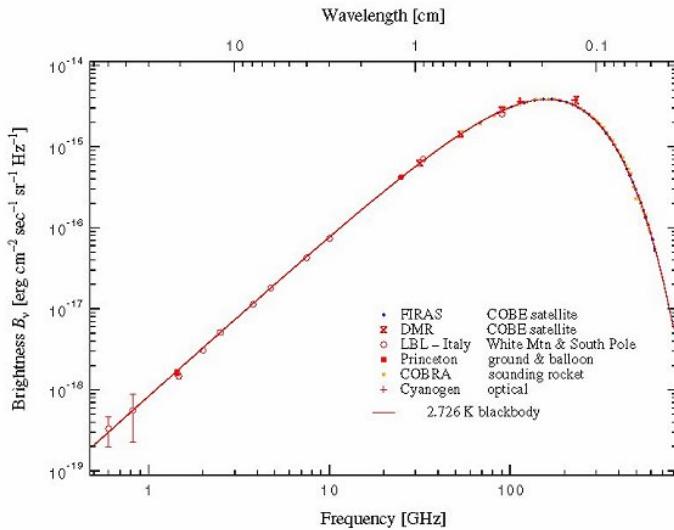
All the other successes of the old inflation are preserved by the new inflation and the graceful exit problem disappears! The universe as we observe it is inside one bubble!

Chapter 6

Latest Developments

6.1 Cosmic Background Radiation

It was discovered in 1965 that there is a feeble microwave radiation emanating uniformly from all directions in the sky. It contributes about one percent to the static on a television screen that is not tuned to a local channel. This is the cosmic microwave background radiation (CMBR). The CMBR spectrum is identical to a blackbody radiation of 2.726°K as shown in the figure below.

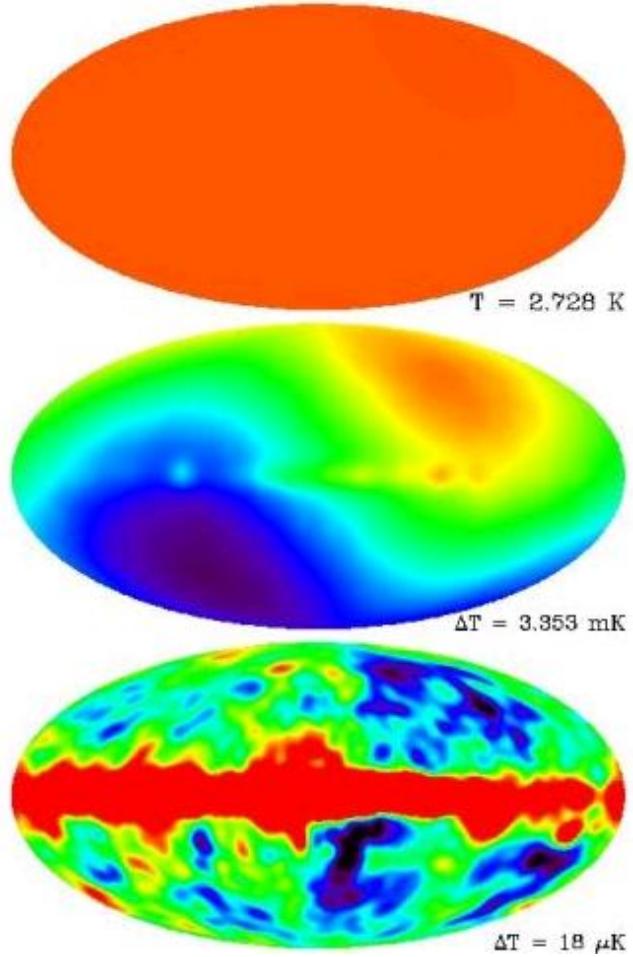


The solid line represents the blackbody radiation spectrum that has been computed from theory. The data being represented by various symbols are collected from various measurements. The agreement between observation and theory is

remarkable.

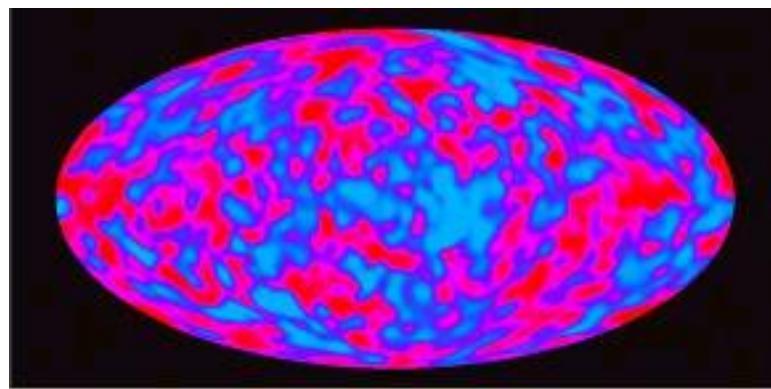
CMBR Fluctuations

The all sky maps in the figure below



shows the three views of CMBR (in false colors) with increasing sensitivity (in temperature variation). To a first approximation the sky is uniform (top). At a sensitivity level of 1 part in 1000, it reveals a shift in wavelength to blue and red (middle) . The pattern is caused by the motion of the earth relative to the frame of the CMBR. When this shift is subtracted off, fluctuations are visible at a sensitivity level of 1 part in 100000 (bottom). The red band in the middle is the emission from the Milky Way.

The figure below shows the final view after all the corrections have been applied.

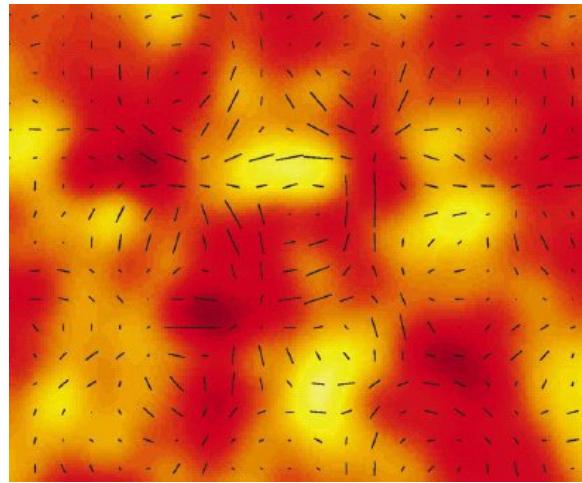


The slight variation in temperature then takes on a blotchy appearance with each patch a little above or below the average temperature of $2.726^{\circ}K$.

All the above-mentioned observations can be interpreted in a consistent way by the Big Bang Theory. According to this theory the CMBR was emitted about 380000 years after the Big Bang when neutral atoms (such as the hydrogen atoms) started to form. As the neutral atoms interact much less to the radiation, they became free and escaped the fireball at a blackbody temperature of about $4000^{\circ}K$. It takes about 14 billion years to reach us and has since been cooled down to $2.726^{\circ}K$ by the cosmic expansion.

CMBR Polarization

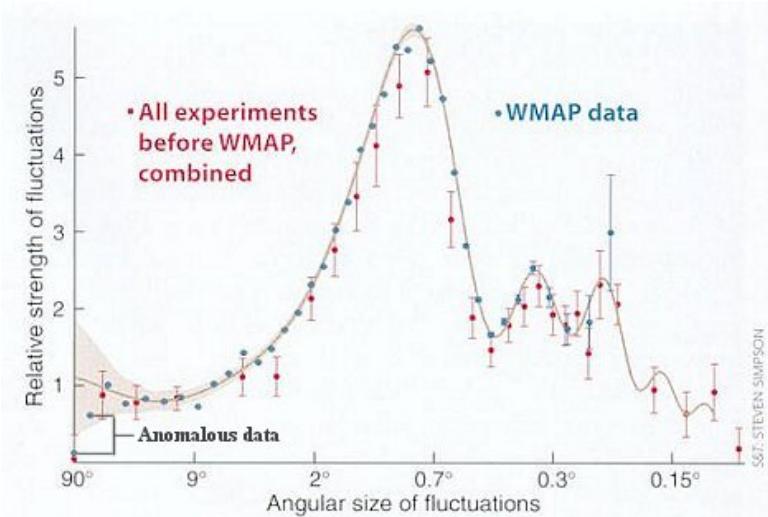
Recently in 2002, the Degree Angular Scale Interferometer (DASI) has detected partial polarization of the CMBR at the sensitivity level of one part in a million. In the figure below



the temperature fluctuations are represented by yellow for hotter, red for colder regions. Superimposed is the polarization measured by DASI. The polarization at each point is represented by a black line, whose orientation and length correspond to the direction and amount of polarization, respectively. The magnitude of the polarization on small angular scales depends on the anisotropy being in place at recombination but on large angular scales, the polarization patterns were formed at the beginning of the re-ionization era, when the first starlight began ionizing the cold hydrogen that filled the universe after the Big Bang cooled. Measurement by WMAP indicates that the first stars were born about 100 to 400 million years after the Big Bang. New polarization data (white bars in figure below) from WMAP in 2006 provide further evidence that the first stars formed some 400 million years after the Big Bang, which was followed by a period of inflation.

CMBR Power Spectrum

Theoretical physicists use the power spectrum plot to determine the cosmological parameters by the observational data. Essentially, the power spectrum is a plot of the amount of fluctuation against the angular (or linear) size. The fluctuation is the difference in the two measurements at the corresponding points. It can be the fluctuation of temperature or density or any other kind of measurable quantity. The figure below

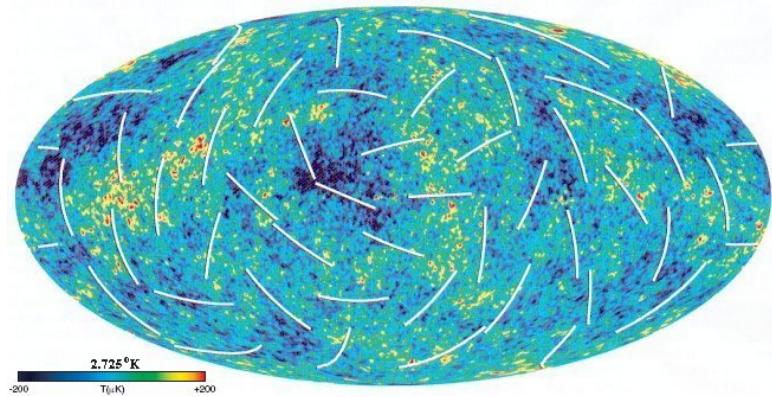


The most revealing display of the ripple data is a “power spectrum” showing how strong the ripples are at each size (dots with error bars). The observed power spectrum can be compared to the predictions made by different theoretical models of the universe; the solid line is the model that best fits the WMAP data (the tan areas reflect uncertainty in the model’s predictions). By tweaking basic cosmic parameters, the model can be fine-tuned to match the data.

shows just one example with the WMAP observational data superimposed on a theoretical curve. The theoretical curve varies with several parameters such as the total cosmic density, the baryon density (luminous matter) and the Hubble’s constant. The best fit model is the lambda cold dark matter model with an initial inflation, a period of galaxies formation induced by cold dark matter, and then the speedup of the cosmic expansion. However, none of the theoretical models based on inflation can account for the anomalous data in the figure at large angular size. Double checking the instruments and analyzing procedures also fails to explain the anomalies. More observations are needed to resolve the puzzle.

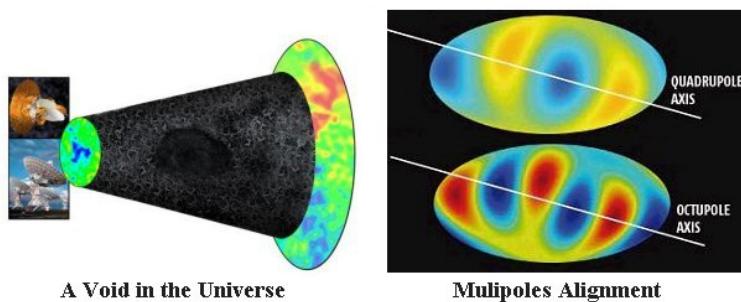
Wilkinson Microwave Anisotropy Probe (WMAP)

The Wilkinson Microwave Anisotropy Probe (WMAP) team has released the first detailed full-sky map of the oldest light in the universe on February 11, 2003. The figure below



shows the measurements with red indicates "warmer" and blue indicates "cooler" spots. The patterns in the map are tiny temperature differences within an extraordinarily evenly dispersed microwave radiation bathing the Universe, which now averages a frigid 2.73 degrees above absolute zero temperature. WMAP resolves the slight temperature fluctuations, which vary by only millionths of a degree. Analyses of this microwave radiation emitted only 380,000 years after the Big Bang appear to define our universe more precisely than ever before. Measurements from WMAP resolve several long-standing disagreements in cosmology rooted in less precise data. Specifically, present analyses of the WMAP all-sky image indicate that the universe is 13.7 billion years old (accurate to 1 percent), composed of 73 percent dark energy, 23 percent cold dark matter, and only 4 percent atoms, is currently expanding at the rate of 71 km/sec/Mpc (accurate to 5 percent), underwent episodes of rapid expansion called inflation, the geometry of the Universe is flat¹, and will expand forever. The Wilkinson Microwave Anisotropy Probe was launched on June 30, 2001. It is designed to operate for four years.

Further analysis of the WMAP data in 2007 reveals two oddities - see figure below.

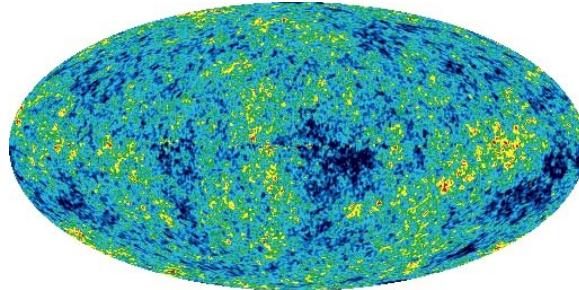


1. It has been deduced from the absence of radio sources that there is a big

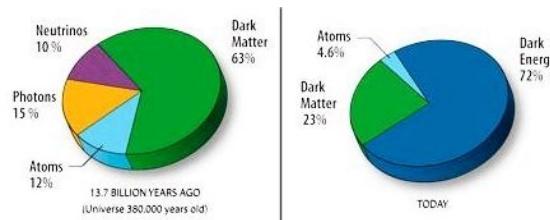
hole in the sky devoid of both normal and dark matter in the direction of the constellation Eridanus. Its size is nearly a billion light years across at a distance 6 - 10 billion light years away (40 times larger in volume than the previous record holder). The void coincides with an extra large cold spot in the WMAP map covering a few degrees of the sky (many times more than the full moon). The temperature of the void is between 20 and 45% lower than the average. It is suggested that the discovery of the void ties in neatly with the WMAP cold spot and the existence of dark energy as the photons would lose energy passing through an empty space.

2. WMAP's temperature variations can be decomposed into set of patterns called multipoles. The lowest multipoles are the largest-area, continent-and ocean-size undulations on the temperature map. Higher multipoles are like successively smaller-area plateaus, mountains and hills (and trenches and valleys) inserted on top of the larger features. As shown in the figure both the quadrupole and the octupole are aligned along an *axis* which standard cosmology cannot explain. This could happen by chance only about 0.1% of the time. Critics have considered a variety of possibilities. One explanation involves some kind of imperfection in WMAP's detector that introduces the patterns, but there is no evidence for this.

The NASA/WMAP Science Team presents the cosmic microwave temperature fluctuations from the 5-year WMAP data as show in the figure below.



The composition of the early universe has been measured from the data as shown in the figure below.



It is obvious by comparing with the composition in the current epoch that it varies as the universe expands. It appears that the dark energy density does

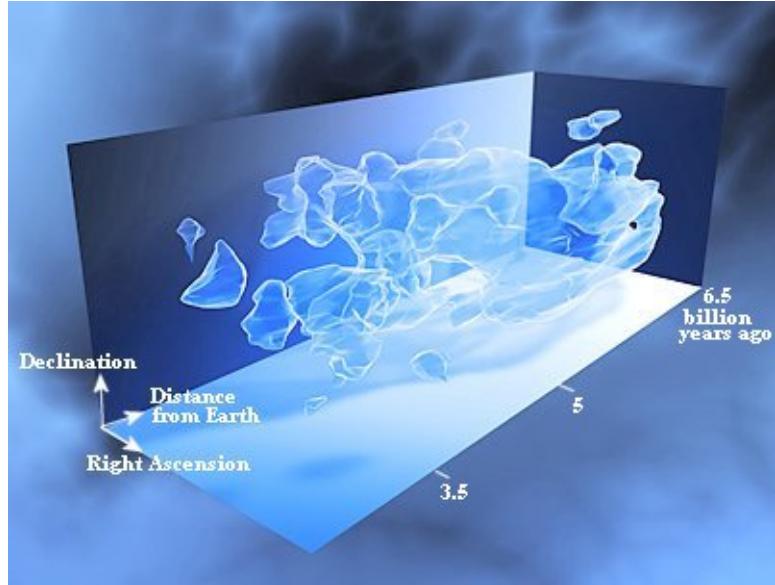
not decrease at all, so it now dominates the universe even though it was a tiny fraction 13.7 billion years ago. Other major findings include:

1. New evidence that a sea of cosmic neutrinos permeates the universe. Cosmic neutrinos existed in such huge numbers they affected the universes early development. That, in turn, influenced the microwaves that WMAP observes.
2. The first stars took more than a half-billion years to create a cosmic fog. The data provide crucial new insights into the end of the *dark ages*, when the first generation of stars began to shine. The glow from these stars created a thin fog of electrons in the surrounding gas that scatters microwaves.
3. The new WMAP data places tight constraints on the theory of inflation. Some versions of the inflation theory now are eliminated. Others have picked up new support.

6.2 Dark Matter

It seems that the Big Bang Theory has been validated conclusively with all these supporting evidences. However, recent observations in the last few years reveal that there is something amiss. It is noticed that even though there is not enough mass to hold the stars, galaxies and galaxy clusters in place, they are still moving around and would not disperse. It looks as if there is some kind of invisible force (gravity from the dark matter) to hold them together. The situation is similar to a puppet show, where the audience can safely assume that someone behind is manipulating the movements. It is suggested that the mass of dark matter within the lunar orbit can be computed by subtracting the total mass (Earth + Dark Matter) within the lunar orbit from the mass of the Earth measured by a gravity-sensing satellite. It turns out to be no more than $1.5 \times 10^{15} kg$ or about one billion times lower than the mass of the Earth. It means that the difference is too small to be measured by the 2008 technology. All that can be found is the upper bound, which is just another way of saying that there is no difference up to the current level of accuracy.

The figure below



shows the large-scale distribution of dark matter mapped by the Hubble's Cosmic Evolution Survey in early 2007. Since light will follow the deformed path created by massive object, the quantity and location of the dark matter can be estimated by the amount of the bending. However, it should be cautioned that such image represents only a small facet of the whole picture. Just like representing the distortion of space-time as a piece of stretched rubber sheet, or the probability density of an electron (in the atom) by some foggy orbital, the dark matter distribution map does not include the other interesting properties such as its lack of interaction with other matter except via gravity. Note the increasing clumpiness from distant past to more recent epoch in the picture.

The map of dark matter forms a filamentous 'skeleton' upon which visible matter congregates, eventually producing stars and galaxies. Baryonic structures are expected to form only inside the dark-matter scaffold. But as shown in the figure below



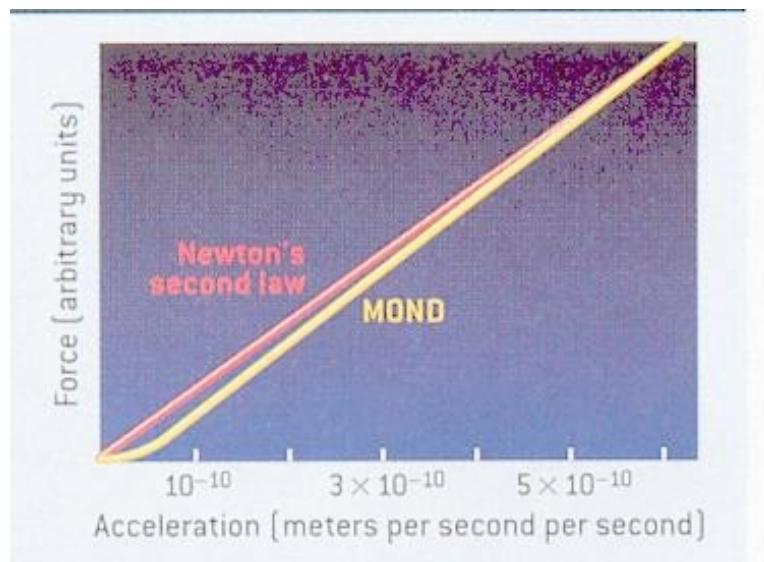
the concentrations of dark matter (mapped in contours) usually - but not always - match up with normal matter (colored). The discrepancies could be a simple error resulting from the way the observations were made. Alternatively it is suggested that dark matter, if the clump is small enough, could have any accumulating visible matter blown out of it by a high-energy phenomenon such as a quasar or a supernova, for example. The collision of two galaxies could also blow an amount of visible matter out as a faint satellite galaxy that has no associated dark matter.

The nature of dark matter has been a mystery defying all attempts to explain as summarized below:

- Neutrino - It is a reasonable candidate for dark matter because of its unreactive nature. Theoretical calculations indicate that there should be as many as 100 million neutrinos for every atom in the universe. However, the recent estimates of neutrino mass is so slight that it could account for only about 0.1 - 7 per cent of the mass of the universe.
- WIMP - Many new particles with heavy mass appear in the supersymmetry formulation. These are referred to as weakly interacting massive particles, or WIMPs. For example, the photino (the fermionic partner of photon) has a mass about 10 to 100 times that of the proton. Most of these electrically neutral particles would, like neutrino, go straight through Earth. On rare occasion, however, one might interact with an atom in the material they pass through. So far, the only claimed detection of a dark matter particle (by an Italian team in 2000) has been strongly disputed.
- Nonluminous matter - Ordinary hidden matter consists of atoms that emit little or no light. It includes a host of celestial objects such as planets, dark gas clouds, brown dwarfs, neutron stars, and black holes. The Massive Compact Halo Objects Project (MACHO) has been looking for them in the halo of the Milky Way. A search for microlensing has turned up four

candidates toward the Large Magellanic Cloud and 45 toward the Galactic Bulge.

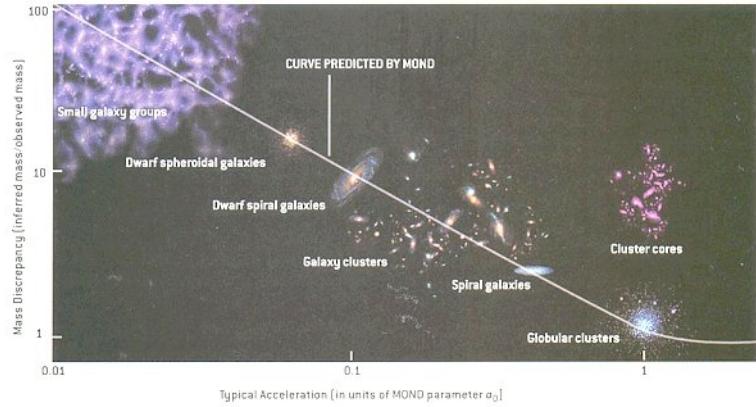
- MOND - It is proposed that instead of looking for the dark matter, slight altering of Newton's second law will account for the discrepancy. The formula $F = ma$ is amended in such a way that smaller gravitational force or mass can impart a given acceleration in a certain range as shown in the figure below.



According to MOND (Modified Newtonian Dynamics), a test particle at a distance r from a large mass M is subject to the acceleration a by the following formula:

$$\mu \left(\frac{a}{a_0} \right) a = \frac{GM}{r^2}$$

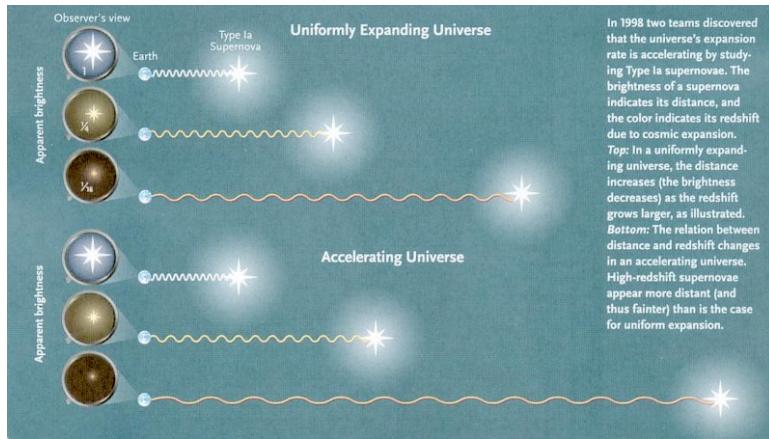
where G is the gravitational constant, $a_0 \approx cH_0 \approx 10^{-8} \text{ cm/sec}^2$ is the MOND parameter, H_0 is the Hubble constant, and $\mu(a/a_0)$ is a function of a/a_0 such that $\mu(a/a_0) \approx a/a_0$ for $a/a_0 \ll 1$ and $\mu(a/a_0) \approx 1$ for $a/a_0 \gg 1$. The figure below



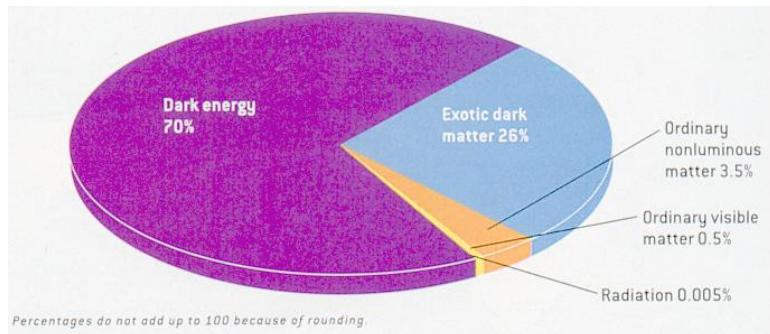
shows the MOND prediction on mass discrepancy for many astronomical objects. Its main failure occurs in the cores of large galaxy clusters. Since its proposal in 1983, MOND has become a controversial subject among astronomers. It is considered as a rather ad hoc invention to fit this special problem of dark matter. Even if it is correct, the new formula should be derived from a more fundamental theory. There is now a candidate theory - Moffat's non-linear gravity.

Dark Energy

The other problem with modern cosmology is related to the use of the Type Ia supernovae as *standard candles* to measure the distance of remote objects. The measurements imply that the cosmic expansion is accelerating as shown in the figure below



which shows that the supernova appears to be dimmer than expected from an uniformly expanding universe. It is proposed that there is some kind of repulsive "dark energy" to induce the acceleration. The figure below



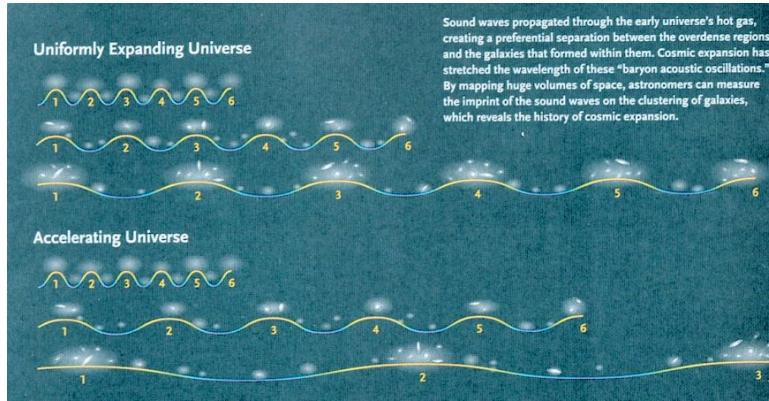
shows the proportion of the various matter-energy components in the Universe. Most of the matter-energy content is in the form of *dark energy*. The composition of the Universe is listed in the table below.

Material	Representative Particles	Particle Mass or Energy (ev)	No. of Particles in Observed Universe	Probable Contribution to Mass of Universe	Sample Evidence
Ordinary matter	Protons, electrons	10^6 to 10^9	10^{78}	5%	Direct observation, inference from element abundances
Radiation	photons	10^{-4}	10^{87}	0.005%	Microwave telescope observations
Hot dark matter	Neutrinos	< 1	10^{87}	0.3%	Neutrino measurements, cosmic structure
Cold dark matter	Supersymmetric particles?	10^{11}	10^{77}	25%	Inference from galaxy dynamics
Dark energy	Scalar particles?	10^{-33}	10^{118}	70%	Supernova observations of accelerated cosmic expansion

New data in 2006 further refine the universe's contents to: 4% ordinary matter, 22% dark matter, and 74% dark energy.

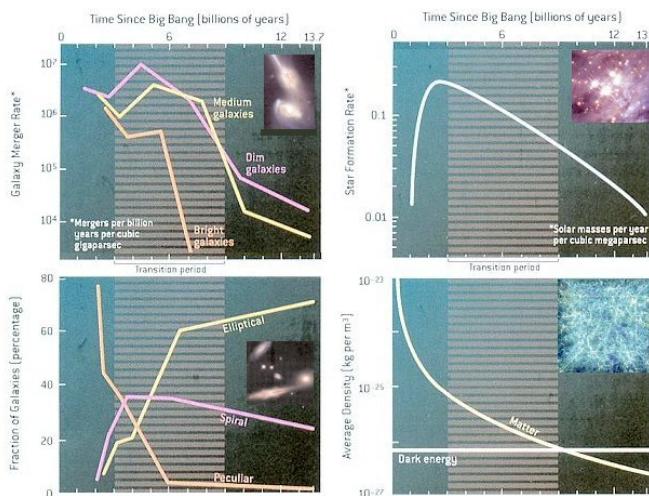
Acceleration of the cosmic expansion is placed on a firmer footing when it is observed in 2003 that the CMBR becomes slightly hotter after going through a galaxy, which forms a gravitational (potential) well. Dark energy, being gravitationally repulsive, makes a gravitational well shallower as a photon passes through, so the photon exits with slightly more energy than it had when it entered.

Another method to verify the cosmic acceleration is by detecting the bunching intervals of the clusters of galaxies. Whereas Type Ia supernovae behave like standard candles, the spacing between clusters of galaxies acts like a standard ruler. The bunching was generated by the cosmic sound wave, which compressed matter to higher density at its peaks. According to different scenarios of cosmic expansion, the amount of stretching is different as shown in the figure below.



In the primordial gas, the incoherent acoustic oscillations created peaks at intervals of 436000 light years, today the spacing should be about 500 million light years depending on the kind of cosmic model.

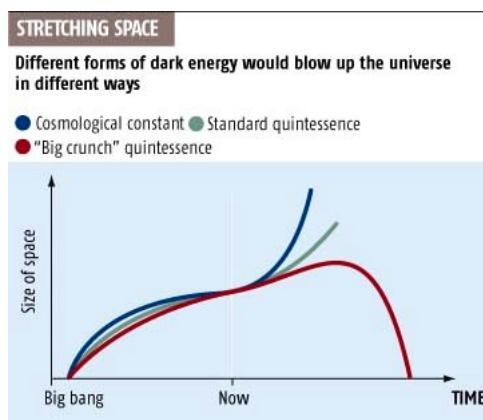
Gradually, it dawns on some astronomers that dark energy could be responsible for turning off galaxy and star formation in the latter half of the cosmic history at redshift $z \sim 0.75$ (~ 6.5 billion years since the big bang). The central piece of evidence is the rough coincidence in timing between the end of most galaxy and cluster formation and the onset of the domination of dark energy. Both happened when the universe was about half its present age. The influence of dark energy include stopping the merger of galaxies, sorting out the types of galaxies, lowering the rate of star formation, and preventing the growth of galaxy clusters as shown in the figure below.



Such idea has been confirmed in 2008 by NASA's Chandra X-ray Observatory. The X-ray results on the hot gas in dozens of galaxy clusters some of which are relatively close and others are more than halfway across the universe reveal that accelerated expansion stifles the growth of galaxy clusters. It also tentatively identifies the cosmic constant as the dark energy.

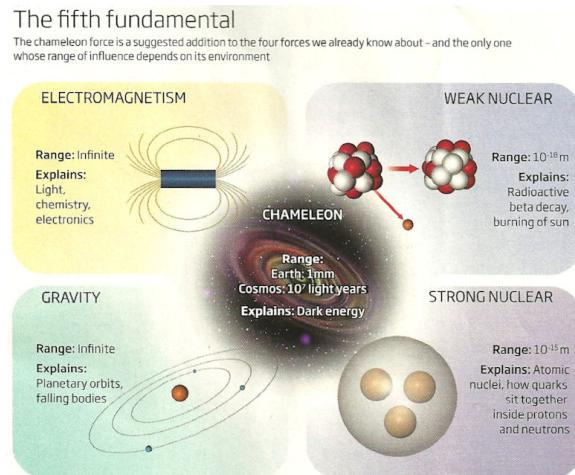
The nature of *dark energy* is still the subject of intense research observationally and theoretically. Some of the suggestions are listed below:

- Cosmological constant - Einstein had introduced a term with the cosmological constant in the gravitational field equation to keep a static universe from collapsing. This additional repulsive force is no longer necessary when the cosmic expansion became apparent. It has become fashionable again with the new discovery of cosmic acceleration. It is very tempting to identify the cosmological constant with the vacuum energy of the various quantum fields. However, the simplest versions of quantum theory predict far too much energy - 10^{120} higher than the observed value by one estimate. One explanation involves the cancellation between the boson positive contributions and the fermion negative contributions to almost zero, leaving only a residual trace corresponding to the observed dark energy.
- Quintessence - This hypothetical form of dark energy permeates all space. Like inflation, quintessence is thought to have somehow originated when the universe was just 10-35 sec old. It is driven by a scalar field whose energy varies gradually. The difference is the energy and time scale: inflation occurred quickly at very high energies, whereas the scalar field responsible for quintessence operates at much lower energies over a much longer time frame. One of its key differences from the cosmological constant is that it can vary depending on time and place. The figure below



portrays the fates of the universe according to different scenarios of dark energy. Research in 2006 found that the dark energy's influence on the universe in the current epoch is the same as 9 billion years ago. The discovery rules out quintessence models that change too rapidly.

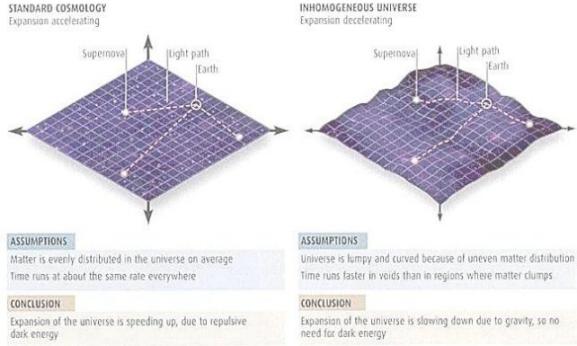
- Chameleon Theory - The theory is a variation of the Quintessence, which varies in time. This fifth force however depends on space in such a way that the range of the force is shorter in an environment of higher density. For example, the ratio of density from the vicinity of the Earth to the void of the Cosmos is about $1/10^{-28}$ leading to a ratio of the force ranges of about $1/10^{26}$ as shown below



The virtue of this theory is that it can be tested by a number of observations including the modification of fine structure constant, change of the ratio between the mass of electron and proton, additional polarization of star light, variation on the age of the universe estimation, ... because of chameleon-photon oscillation (switching). Such flexibility now becomes the problem for validating the theory. Since this theory is conceived to fit observations and has yet to be derived from anything more fundamental, it is very easy to adjust its parameters to fit the available data.

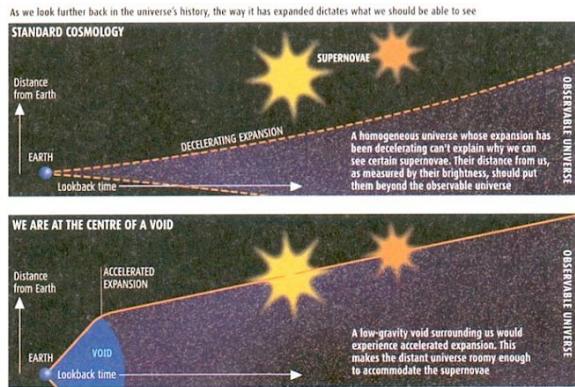
- Illusion - This explanation asserts that the apparent acceleration is just an illusion. Sky survey reveals that matter is distributed unevenly on large scales with gigantic super-clusters and huge voids in between. Because we live in a gravitationally bound system (the Milky way), our clocks run more slowly than they would in a void. In addition, space is negatively curved in the void, so the volume for a given radius is larger than in the relatively flat space. In effect, our estimate of volume is too small and the estimate of time is too slow giving the wrong impression of acceleration as in the figure below.

Observations of light from receding supernovae have led cosmologists to conclude that the expansion of the universe is accelerating. They explain this by invoking dark energy. But this assumes standard cosmology. The uneven distribution of matter in the cosmos might mean our time and space calculations are wrong and the expansion is decelerating after all. There would be no need for dark energy.



However, slight increase of the WMAP temperature associated with intervening regions of superclusters shows that the dark energy is real. Since the CMBR photons gained a small amount of additional energy as they re-emerge from a gravitational well while the cosmic expansion is accelerating.

- **Void** - If the Earth is located in a vast cosmic void (with less matter than the average) of the size between 300 million to 3 billion light years, then object outside the void would be further away (than envisioned from a homogeneous universe) because the void would expand faster with less gravitational retardation as in the figure below.



The problem with this explanation is the requirement that the Earth has to be in the middle of the void (in violation of the Copernican Principle); otherwise it would be inconsistent with the WMAP data, which is isotropic. Measurement of the rate of cosmic expansion over time can be used to check against this hypothesis. The observational sensitivity required to record the tiny changes is currently beyond astronomers' capabilities. But it should become feasible with a new generation of ultra-

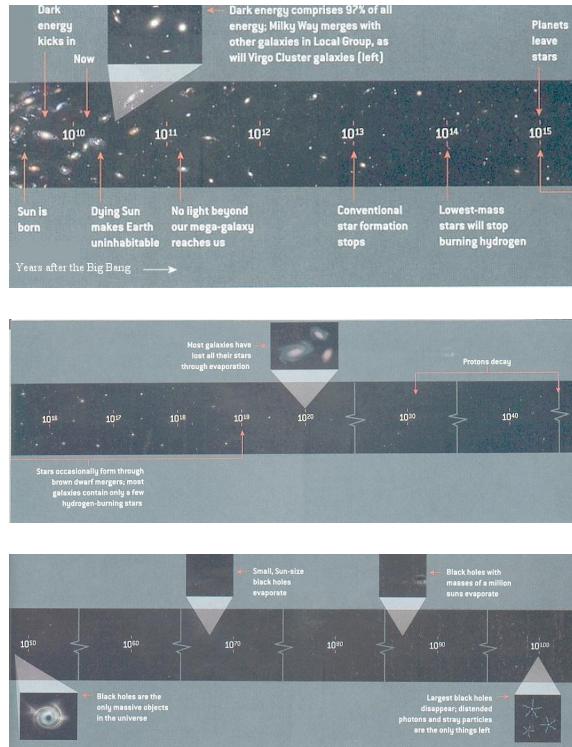
sensitive telescopes.

Citing some seemingly inexplicable anomalies in astronomical observations, it is suggested that the *Void Theory* would remain viable even the Earth is slightly off center (up to 50 million light years). It has been shown that the CMBR is a bit lopsided - hotter in one direction than in the other. This asymmetry is usually attributed to the motion of the Solar system through space but could also be a sign of a lumpy universe. Furthermore, small fluctuations in the CMBR appear to align in the specific direction (the *axis of evil*). This alignment picks out a preferred direction in the sky, which, though hard to imagine in a Copernican universe, might be explained in terms of displacement of the Earth from the center of a void. A preferred direction would also have other effects, such as large-scale coherent motions of galaxies and galaxy clusters. Several observations have claimed detection of such *dark flow*, but it remains controversial. And then there is the argument that the *Void Theory* would not violate the Copernican principle (that the Earth is not special) if the region under consideration is very large and containing many more voids such as the inhomogeneous universe shown in the figure below



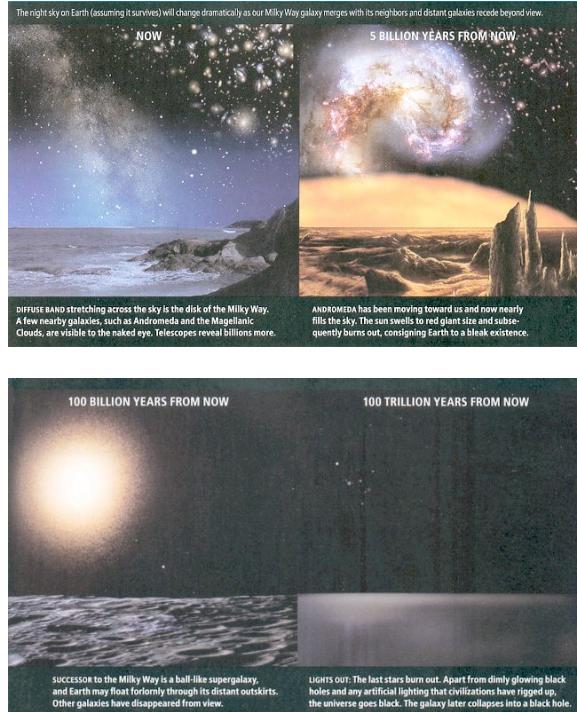
which also depicts the different explanations with dark energy and void.

The effect of dark energy became dominant only at an epoch about 8×10^9 years after the Big Bang. If the acceleration persists in the future, it will impose a horizon surrounding a galaxy like the Milky way - a distance beyond which light cannot reach us. The figures below depict the sequence of events for the future of the universe with cosmic acceleration according to a computer simulation.



The model assumes that the dark energy permeating the vacuum has a positive, constant value - similar to the cosmological constant, as Einstein once posited.

A study on the consequence of cosmic acceleration concludes that while most of the galaxies move away beyond the cosmic horizon, the local group of galaxies will collapse into a supergalaxy by gravitational attraction. Eventually, the universe goes black when the last stars burn out. Future civilizations (if there is any left) will have a very different perspective of the universe. Our descendants will observe an island of stars (the supergalaxy) embedded in a vast emptiness. It will resemble the de Sitter universe originally envisioned by Einstein. The figure below shows the sequence of events according to an artist's rendition.



Since we don't feel the effect of dark energy and dark matter around us except through the gravitational influence on large scale, a model has been constructed with no other interactions between each other or with ordinary matter. It fits the observational data such as the high-redshift supernovae, the microwave background radiation, the distribution of large-scale structure, and the dynamic of celestial objects very well. But if 96% of the Universe is in the form of unseen substances, does this not mean that there is the possibility of hidden structure? Might the dark sector be a fascinating place, with its own intricate interactions perhaps even a kind of intelligent life? Is there a *dark light* that we do not see, radiating and absorbing in the dark Universe? Such possibility suggests that human beings are extremely unimportant in the grand scheme of the universe as portrayed in a 1985 movie called "Insignificance" in which Einstein and Monroe explores relativity and our place in the universe.



Are physicists just making up dark energy?

I finish with an article by physicist David Goldberg.

In this week's "Ask a Physicist," I get defensive about dark energy. It dominates the universe; it's completely absurd; and it's apparently absolutely necessary.

I've been putting off talking about dark energy for a while. I was afraid that you'd judge me, and all cosmologists, as being charlatans. But I can't put it off any longer, so for this week's "Ask a Physicist" our (metaphorical) shame will be revealed by Donovan, who asks:

I've never seen someone try to explain dark energy as anything other than a feature of our universe. Instead of explaining dark energy as a mysterious force inside of our universe that is pushing out and inflating us, has anyone ever tried to explain it by postulating some kind of "vacuum" outside of the universe that is "pulling" us out?

You see? This is why I'm so defensive. The idea of dark energy is so ridiculous that almost every question is based on trying to make it go away. And believe me, I share your concerns. I don't want to believe in dark energy, but I have no choice.

When I was a student in the mid 1990's, cosmologists thought we had it all figured out. Dark matter had been known (or at least suspected) since the 1920's, and most respectable physicists simply assumed that the universe was made up of a combination of ordinary and dark matter. Since gravity is attractive, this should presumably act to pull the universe together.

At the same time, several groups were observing distant supernova explosions. Supernovas (Type Ia's, if you must be precise) are really useful probes of the universe because a) they are very bright, which means that you can see them

from very far away, and b) they are "standard candles" which means that if you understand them well enough, you can figure out exactly how bright they really are, and from that, you can determine their distance. As I discussed previously, we can measure the redshifts of the supernovas to tell how much the universe has expanded since they blew up. From the combination of expanding universe and distance voila! We can determine how quickly the universe is slowing down.

Only it isn't.

In 1998, the High-z Supernova Search Team followed quickly by the Supernova Cosmology Project announced that based on their observations, the universe is, in fact, accelerating. Subsequent observations have confirmed this, and the culprit has been dubbed "Dark Energy." Like "Dark Matter," the name is meant to obscure the fact that we have no frakking clue what it really is.

Dark energy, as you almost certainly know if you're an avid pop-sci reader, is a "mysterious substance" (it's always called mysterious) which causes the universe to accelerate. This is not as ridiculous as it would seem at first blush, since when it comes to gravity, you've probably been lied to. You probably already know that mass creates a gravitational field, but general relativity shows that any form of energy (including a big box of photons) will do the trick. Stranger still is that gravity gets an extra power-up if there is pressure involved. Under normal circumstances, we don't notice this, since even in the center of the sun, the pressure is tiny compared to the energy density.

Dark energy is a weird case. The idea is that the pressure is negative kind of like elastic which means that the net gravity is repulsive. This being 10^{100} , I'd be remiss if I didn't point out that dark energy is the closest thing that we have to anti-gravity. It's not anti-gravity, mind you, but if you have your heart set on writing it into your story, it's the best you're going to do.

And there's a lot of it; our current best estimate is that dark energy accounts for about 70% of the total energy in the universe. Take a moment for that to sink in. For all of you who would give me grief about dark matter simply because we've never captured a particle of the stuff, consider the fact that at least dark matter has the good grace to behave something like everyday particles. With dark energy, we have three times as much, and even our best models don't involve particles that we might see in a detector. But that doesn't mean that we've never seen anything like dark energy.

Quantum Electrodynamics is, besides a potentially awesome name for a band, one of the most successful theories ever. It basically unifies electromagnetism and quantum mechanics, and has predicted everything from the detailed structure of the atom to the magnetic strength of the electron to fantastic precision. It also has a well-earned bad reputation for producing lots of infinities in calculations. This is bad, by and large, but we can normally get around it by

subtracting one infinity from another. Yes, it's a cheat. Yes, it makes me feel dirty. But it also works, and I guarantee that you're not going to make me feel any worse about myself than my quantum field theory students did when I distracted them with puppies and quickly just erased the infinities.

One of the infinities that pop out of this theory is related to the particles and antiparticles that are constantly being created out of "the vacuum." There are two cool things and one really crappy thing about the "vacuum energy" of these temporary particles. The first cool thing is that it has exactly the negative pressure needed to make dark energy. The second cool thing is that it isn't just made up. We can observe the "Casimir Effect" in a lab. Two metal plates in a vacuum will be pulled together because there are more vacuum fluctuations outside the plates than between them.

The really crappy thing, however, is that any realistic calculation of the vacuum energy gives an energy density about 10¹⁰⁰ times larger than the density actually measured by the accelerations of supernovas (even if you round down the "infinity" a bit). This is not a small problem. As far as I am concerned, it is the worst problem in physics, and one of the reasons that Donovan and other readers have posed questions trying to get around the whole thing.

One possibility is to suggest that the universe isn't accelerating at all. Perhaps the most distant supernovas are just different from the ones near by, and that effect oddly makes the universe look like it's accelerating? I had this thought myself, at first, but the supernova evidence isn't the only reason to be a believer.

One of the best pieces of evidence that we have is the measurement of the Cosmic Microwave Background, the radiation from when the universe was only about 380,000 years old. We can observe the bumps and wiggles of hot and cold which behaved like water waves do today. Since the light from these peaks had to travel for almost the entire history of the universe before reaching our telescopes, we can use this information to figure out the shape of the universe to incredible accuracy. It's "flat," by the way. What this means for us is that since we know that dark and ordinary matter combined only add up to about 30% of the energy needed to make our universe flat, the rest must be made up by something else: dark energy. But that's not all. From the distribution of galaxies, to gravitational lensing, to the number of distant galaxy clusters, every piece of evidence points to a model with this dark energy feature in it.

Basically, if you want to get rid of dark energy, you have to get rid of relativity. You're welcome to try, but for my money, it seems like a better bet to try to figure out what dark energy is.

The problem is that we honestly don't know. One possibility is that Einstein's "cosmological constant" was right all along, and for some reason, it's simply

hard-wired into our theory of gravity. Another is that the dark energy is a fluid, with a pressure that might change from place to place, and time to time. So far, there's no evidence that this is the case, but the error bars are still pretty big.

But if the vacuum energy is so large, why is dark energy so small? We just don't know, which is why I got all defensive early on. One possibility (and this is a possibility likely to piss off many physicists) is that it's simply the anthropic principle at work. Maybe all universes have more or less dark energy, but the ones with more than us accelerated so quickly that no stars or complex structures could form. This would mean, of course, no stars, which means no human beings, or ALF for that matter, to be having this conversation.

So back to Donovan's original question: Couldn't the acceleration of the universe be real, but basically be a pull from the outside? The problem is that the universe doesn't have an outside. Wherever you go, there you are. But I'll play along. Let's imagine that our universe does have a crispy crust, and some cosmic pizza-maker is pulling it from all sides. What happens then? The edges may be accelerating away from you, but the nearby mushrooms would more or less sit still. That's because gravity is a local curvature in space-time, and it takes something let's call the dark energy yeast to push all the mushrooms away. Hopefully the whole pizza digression will distract you from the fact that while physicists are pretty damn sure that something is playing the role of dark energy, we have no idea what it is.

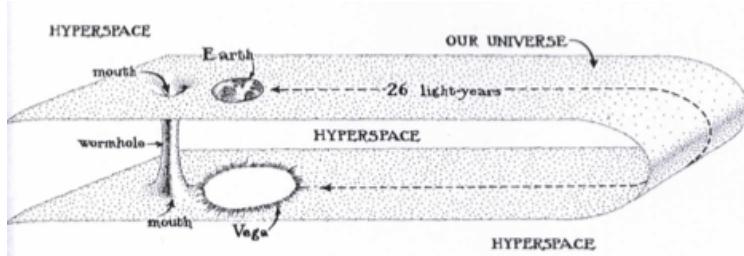
And now you know our shameful, shameful secret.

Chapter 7

WormHoles and Time Machines

We ask this question: can highly advanced civilizations build wormholes through hyperspace for rapid interstellar travel and machines for traveling backward in time?

A wormhole is a hypothetical shortcut for travel between distant points in the universe. The wormhole has two entrances called *mouths*, one (for example) near Earth, and the other (for example) in orbit around Vega, 26 light-years away. The mouths are connected to each other by a tunnel through hyperspace (the wormhole) that might be only a kilometer long. If we enter the near-Earth mouth, we find ourselves in the tunnel. By traveling just one kilometer down the tunnel we reach the other mouth and emerge near Vega, 26 light-years away as measured in the external Universe. The figure below



depicts such a wormhole in something called an embedding diagram. Let us digress from our story a bit to figure out how these embedding diagrams work. **Embedding Diagrams** A helpful way to visualize warped space is to use an *embedding diagram*. The full Schwarzschild solution for a point mass is represented by the interval relation (worked out earlier)

$$ds^2 = - \left(1 - \frac{r^*}{r}\right) c^2 dt^2 + \frac{1}{\left(1 - \frac{r^*}{r}\right)} dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2$$

where

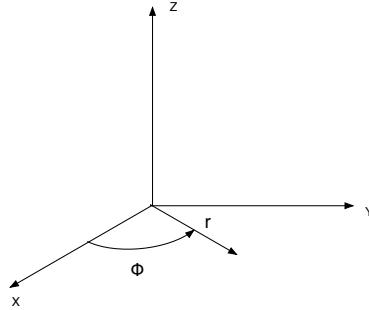
$$r^* = \frac{2GM}{c^2}$$

Since it is difficult to work with the full 3-dimensional curved space, we concentrate on a 2-dimensional slice where we fix $\theta = \pi/2$ and consider only one instant of time. In the absence of gravity we then have

$$ds^2 = -c^2 dt^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \rightarrow dr^2 + r^2 d\phi^2$$

The embedding procedure goes as follows:

1. We imagine this 2-dimensional surface as a part of a 3-dimensional Euclidean space with a cylindrical coordinate system (r, ϕ, z) as shown below



with $x = r \cos \phi$ and $y = r \sin \phi$.

2. The relationship $z(r)$ between the new coordinate z introduced to create the fake Euclidean space and the normal coordinate r is determined by assuming a flat space interval expression

$$ds^2 = dz^2 + dr^2 + r^2 d\phi^2$$

Now we have from calculus

$$dz = \frac{dz}{dr} dr$$

so that we obtain

$$ds^2 = \left(\left(\frac{dz}{dr} \right)^2 + 1 \right) dr^2 + r^2 d\phi^2$$

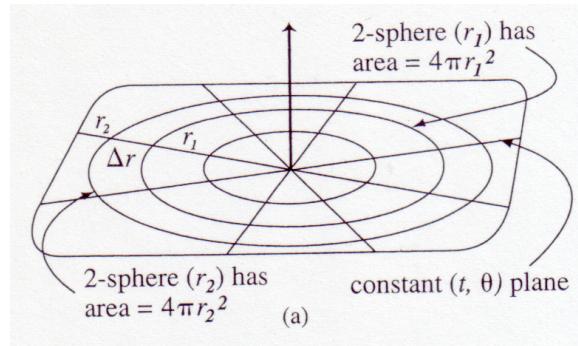
3. Comparison with the true interval expression in our special case says

$$ds^2 = \left(\left(\frac{dz}{dr} \right)^2 + 1 \right) dr^2 + r^2 d\phi^2 = dr^2 + r^2 d\phi^2$$

or that

$$\frac{dz}{dr} = 0 \rightarrow z = \text{constant}$$

This is just the flat plane as shown below.



If instead we carry out the same procedure with the Schwarzschild solution it goes as follows.

$$ds^2 = \left(\left(\frac{dz}{dr} \right)^2 + 1 \right) dr^2 + r^2 d\phi^2 = \frac{1}{(1 - \frac{r^*}{r})} dr^2 + r^2 d\phi^2$$

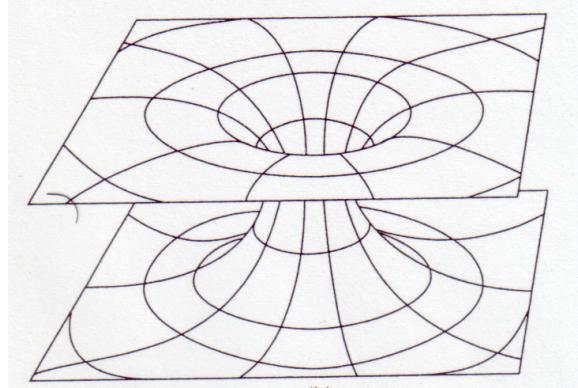
or

$$\frac{dz}{dr} = \pm \frac{r^*}{(r - r^*)^{1/2}}$$

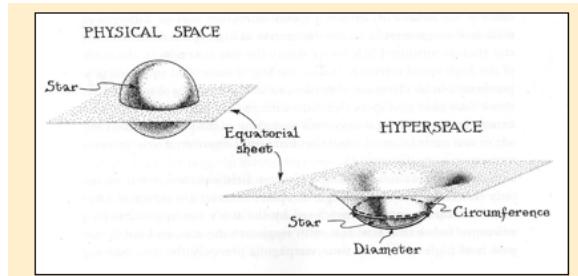
Using calculus we then have

$$z = \pm 2r^*(r - r^*)^{1/2} \rightarrow z^2 = 4r^*(r - r^*)$$

This is a sideways parabola in the (r, z) plane for a given value of ϕ . It is the same for any ϕ value so the 3-dimensional representation is as shown below.

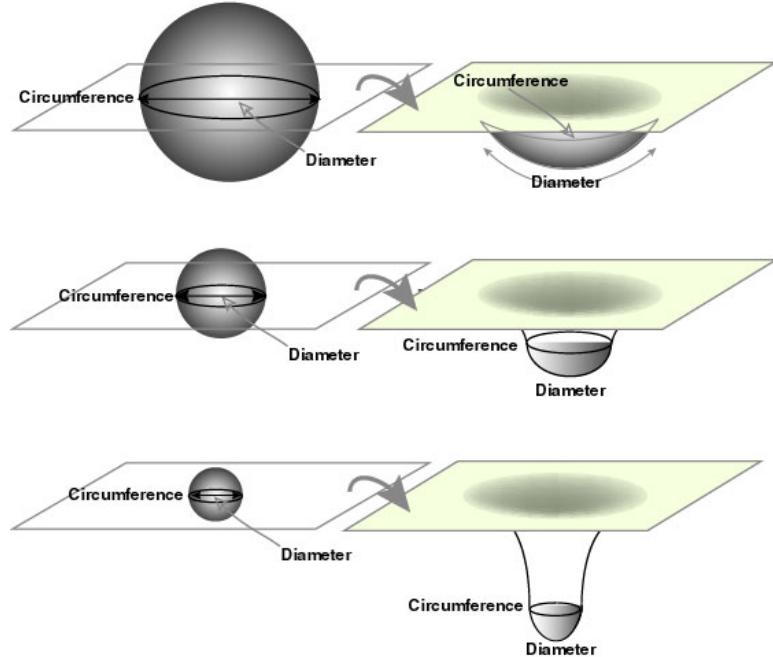


This is called an *Einstein-Rosen bridge*. A comparison of physical space and embedded space is shown below.



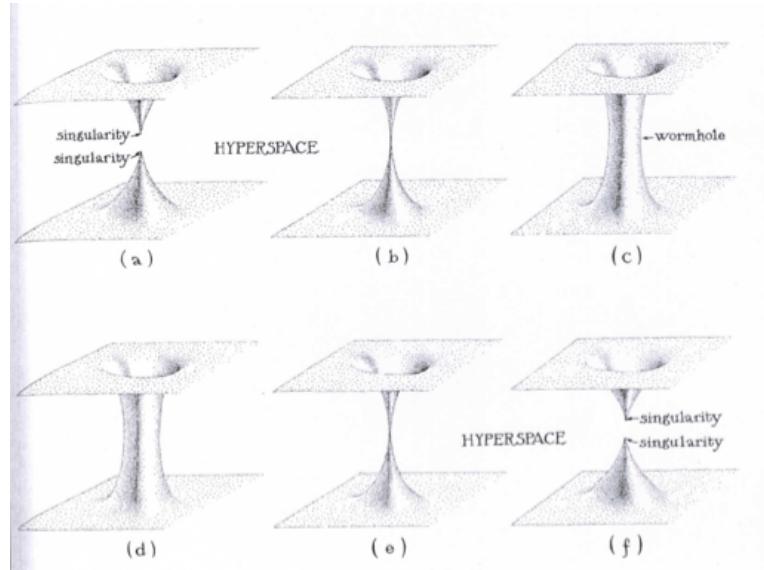
and as we vary the radius for fixed mass.

STARS WITH THE SAME MASS, BUT DIFFERENT SIZES: HOW CURVED?



In the diagram the space of our Universe is depicted as a two-dimensional sheet. Just as an ant crawling over a sheet of paper is oblivious to whether the paper is lying flat or is gently folded, so we in our Universe are oblivious to whether our Universe is lying flat in hyperspace or is gently folded, as in the diagram. However, the gentle fold is important; it permits the Earth and Vega to be near each other in hyperspace so they can be connected by the short wormhole. With the wormhole in place, we, like an ant or worm crawling over the embedding diagrams surface, have two possible routes from Earth to Vega: the long, 26-light-year route through the external Universe, and the short, 1-kilometer route through the wormhole. What would the wormholes mouth look like, if it were on Earth, in front of us? In the diagrams two-dimensional universe the wormholes mouth is drawn as a circle; therefore, in our three-dimensional universe it would be the three-dimensional analogue of a circle; it would be a sphere. In fact, the mouth would look something like the spherical horizon of a nonrotating black hole, with one key exception: The horizon is a *one-way* surface; anything can go in, but nothing can come out. By contrast, the wormhole mouth is a *two-way* surface; we can cross it in both directions, inward into the wormhole, and back outward to the external Universe. Looking into the spherical mouth, we can see light from Vega; the light has entered the other mouth near Vega and has traveled through the wormhole, as though the wormhole were a light pipe.

or optical fiber, to the near-Earth mouth, where it now emerges and strikes us in the eyes. Wormholes are not mere figments of a science fiction writers imagination. They were discovered by Ludwig Flamm mathematically, as a solution to Einsteins field equation, in 1916, just a few months after Einstein formulated his equation; Einstein and Nathan Rosen explored them in the 1950s; and John Wheeler and his research group studied them extensively, by a variety of mathematical calculations, in the 1950s. However, none of the wormholes that had been found as solutions of Einsteins equation was suitable for because none of them could be traversed safely. Each and every one of them was predicted to evolve with time in a very peculiar way: The wormhole is created at some moment of time, opens up briefly, and then pinches off and disappears-and its total life span from creation to pinch-off is so short that nothing whatsoever (no person, no radiation, no signal of any sort) can travel through it, from one mouth to the other. Anything that tries will get caught and destroyed in the pinch-off. The figure below shows a simple example.



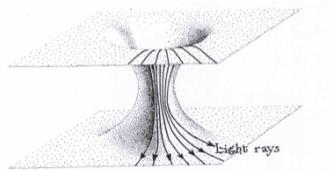
The process is as follows

- Initially there is no wormhole; instead there are two singularities, one near Earth and one near Vega.
- Then at some moment of time, the two singularities reach out through hyperspace, find each other, annihilate each other, and in the annihilation they create the wormhole.
- The wormhole grows in circumference.
- Then begins to contract.

- (e) The pinches off.
- (f) Creating two singularities similar to those in which the wormhole was born - but with one exception. Each initial singularity (a) is like that of the big bang; time flows out of it, so it can give birth to something - the universe in the case of the big bang, and the wormhole in this case. Each final singularity (f), by contrast, is like that of the big crunch; time flows into it, so things get destroyed in it - the universe in the case of the big crunch, and the wormhole in this case. Anything that tries to cross the wormhole during its brief life gets caught in the pinch-off and, along with the wormhole itself, gets destroyed in the final singularities (f).

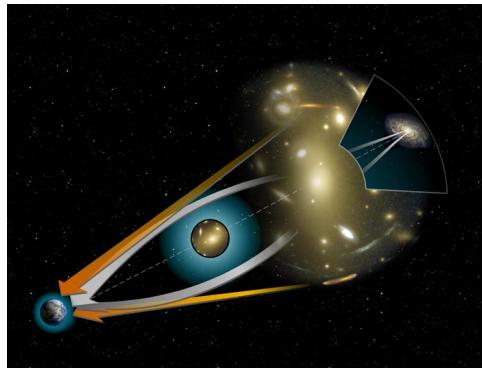
Most physicist colleagues have been skeptical of wormholes. Not only does Einsteins field equation predict that wormholes live short lives if left to their own devices; their lives are made even shorter by random infalling bits of radiation: The radiation gets accelerated to ultra-high energy by the wormholes gravity, and as the energized radiation bombards the wormholes throat, it triggers the throat to recontract and pinch off far faster than it would otherwise - so fast, in fact, that the wormhole has hardly any life at all. There is another reason for skepticism. Whereas black holes are an inevitable consequence of stellar evolution (massive, slowly spinning stars, of just the sort that astronomers see in profusion in our galaxy, will implode to form black holes when they die), there is no analogous, natural way for a wormhole to be created. In fact, there is no reason at all to think that our Universe contains today any singularities of the sort that give birth to wormholes (as in the above figure), and even if such singularities did exist, it is hard to understand how two of them could find each other in the vast reaches of hyperspace, so as to create a wormhole in the manner shown above. Wormholes, despite the skepticism about them seem to valid. Perhaps there is some way that infinitely advanced civilization could hold a wormhole open, that is, prevent it from pinching off, so that one could travel through it from Earth to Vega and back. If one calculates, making the calculations easy by idealizing the wormhole as precisely spherical (so in the earlier figure, where one of our Universes three dimensions is suppressed, it is precisely circular in cross section). Then, we find three things:

1. The only way to hold the wormhole open is to thread the wormhole with some sort of material that pushes the wormhole's walls apart, gravitationally. Such material is called exotic because, as we shall see, it is quite different from any material that any human has ever yet met.
2. Just as the required exotic material must push the wormholes walls outward, so also, whenever a beam of light passes through the material, the material will gravitationally push outward on the beams light rays, prying them apart from each other. In other words, the exotic material will behave like a *defocusing lens*; it will gravitationally defocus the light beam. So any spherical wormhole through which a beam of light can travel must gravitationally defocus the light beam. To see that this is so, imagine (as



drawn below)

that the beam is sent through a converging lens before it enters the wormhole, thereby making all its rays converge radially toward the wormholes center. Then the rays will always continue to travel radially (how else could they possibly move?), which means that when they emerge from the other mouth, they are diverging radially outward, away from the wormholes center, as shown. The beam has been defocused. The wormholes spacetime curvature, which causes the defocusing, is produced by the *exotic* material that threads through the wormhole and holds the wormhole open. Since spacetime curvature is equivalent to gravity, it in fact is the exotic materials gravity that defocuses the light beam. In other words, the exotic material gravitationally repels the beams light rays, pushing them away from itself and hence away from each other, and thereby defocuses them. This is precisely the opposite to what happens in a gravitational lens



There light from a distant star is focused by the gravitational pull of an intervening star or galaxy or black hole; here the light is defocused.

3. One learns from the Einstein field equation that, in order to gravitationally defocus light beams and gravitationally push the wormholes walls apart, the exotic material threading the wormhole must have a negative average energy density as seen by a light beam traveling through it. This requires a bit of explanation. Recall that gravity (spacetime curvature) is produced by mass and that mass and energy are equivalent ($E = Mc^2$). This means that gravity can be thought of as produced by energy. Now, take the energy density of the material inside the wormhole (its energy per cubic centimeter), as measured by a light beam, that is, as measured by someone who travels through the wormhole at (nearly) the speed of light and average that energy density along the light beams trajectory. The resulting averaged energy density must be negative in order for the material to be able to defocus the light beam and hold the wormhole open, that is, in order for the wormholes material to be *exotic*.

This does not necessarily mean that the exotic material has a negative energy as measured by someone at rest inside the wormhole. Energy density is a relative concept, not absolute; in one reference frame it may be negative, in another positive. The exotic material can have a negative energy density as measured in the reference frame of a light beam that travels through it, but a positive energy density as measured in the wormholes reference frame. Nevertheless, because almost all forms of matter that we humans have ever encountered have positive average energy densities in everyones reference frame, physicists have long suspected that exotic material cannot exist. Presumably the laws of physics forbid exotic material, we physicists have conjectured, but just how the laws of physics might do so was not at all clear.

Perhaps our prejudice against the existence of exotic material is wrong. Perhaps exotic material can exist. A key to the answer had been provided in the 1970s by Stephen Hawking. In 1970, when proving that the surface areas of black holes always increase, Hawking had to assume that there is no exotic material near any black holes horizon. If exotic material were in the horizons vicinity, then Hawkings proof would fail, his theorem would fail, and the horizons surface area could shrink. Hawking didnt worry much about this possibility, however; it seemed in 1970 a rather safe bet that exotic material cannot exist. Then, in 1974, came a great surprise: Hawking inferred as a by-product of his discovery of black-hole evaporation that vacuum fluctuations near a hole's horizon are exotic: They have negative average energy density as seen by outgoing light beams near the holes horizon. In fact, it is this exotic property of the vacuum fluctuations that permits the holes horizon to shrink as the hole evaporates, in violation of Hawkings area-increase theorem. Because exotic material is so important for physics, I shall explain this in greater detail.

The origin and nature of vacuum fluctuations goes like this. When one tries to

remove all electric and magnetic fields from some region of space, that is, when one tries to create a perfect vacuum, there always remain a plethora of random, unpredictable electromagnetic oscillations - oscillations caused by a tug-of-war between the fields in adjacent regions of space. The fields *here* borrow energy from the fields *there*, leaving the fields there with a deficit of energy, that is, leaving them momentarily with negative energy. The fields there then quickly grab the energy back and with it a little excess, driving their energy momentarily positive, and so it goes, onward and onward.

Under normal circumstances on Earth, the average energy of these vacuum fluctuations is zero. They spend equal amounts of time with energy deficits and energy excesses, and the average of deficit and excess vanishes. Not so near the horizon of an evaporating black hole, Hawking's 1974 calculations suggested. Near a horizon the average energy must be negative, at least as measured by light beams, which means that the vacuum fluctuations are exotic. The horizon distorts the vacuum fluctuations away from the shapes they would have on Earth, and by this distortion it makes their average energy density negative, that is, it makes the fluctuations exotic.

Under what other circumstances will vacuum fluctuations be exotic? Can they ever be exotic inside a wormhole, and thereby hold the wormhole open? First, it was proved that in flat spacetime, that is, far from all gravitating objects, vacuum fluctuations can never be exotic - they can never have a negative average energy density as measured by light beams. On the other hand, it was then proved that in curved spacetime, under a very wide variety of circumstances, the curvature distorts the vacuum fluctuations and thereby makes them exotic.

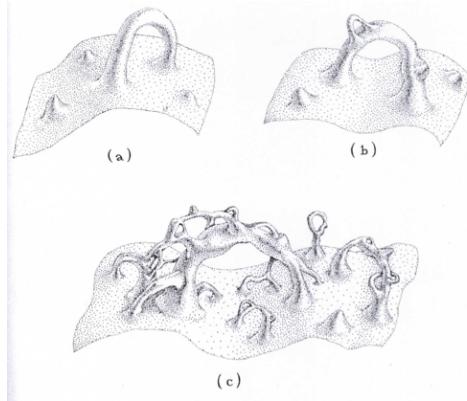
Is a wormhole that is trying to pinch off such a circumstance? Can the curvature of the wormhole, by distorting the vacuum fluctuations, make them exotic and enable them to hold the wormhole open? We still do not know. We now ask. What things do the laws of physics permit an infinitely advanced civilization to do, and what things do the laws forbid? (By an *infinitely advanced civilization*, we mean one whose activities are limited only by the laws of physics, and not at all by ineptness, lack of know-how, or anything else).

We physicists, I believe, have tended to avoid such questions because they are so close to science fiction. While many of us may enjoy reading science fiction or may even write some, we fear ridicule from our colleagues for working on research close to the science fiction fringe. We therefore have tended to focus on two other, less radical, types of questions: What kinds of things occur naturally in the Universe? (for example, do black holes occur naturally? and do wormholes occur naturally?). And what kinds of things can we as humans, with our present or near-future technology, do? (for example, can we produce new elements such as plutonium and use them to make atomic bombs? and can we produce high-temperature superconductors and use them to lower the power bills for levitated trains and Superconducting Super-collider magnets?)

By 1990s it seemed clear that we physicists had been much too conservative in our questions. Already, one question was beginning to bring a payoff. By asking, Can an infinitely advanced civilization maintain wormholes for rapid interstellar travel? physicists had identified exotic material as the key to wormhole maintenance, and triggered a somewhat fruitful effort to understand the circumstances under which the laws of physics do and do not permit exotic material to exist. Suppose that our Universe was created (in the big bang) with no wormholes at all. Then eons later, when intelligent life has evolved and has produced a (hypothetical) infinitely advanced civilization, can that infinitely advanced civilization construct wormholes for rapid interstellar travel? Do the laws of physics permit wormholes to be constructed where previously there were none? Do the laws permit this type of change in the topology of our Universes space? We physicists want to know whether and how the universes topology can be changed now, within the confines of physical law.

We can imagine two strategies for constructing a wormhole where before there was none: a quantum strategy and a classical strategy.

The quantum strategy relies on gravitational vacuum fluctuations, that is, the gravitational analogue of the electromagnetic vacuum fluctuations discussed earlier - random, probabilistic fluctuations in the curvature of space caused by a tug-of-war in which adjacent regions of space are continually stealing energy from each other and then giving it back. Gravitational vacuum fluctuations are thought to be everywhere, but under ordinary circumstances they are so tiny that no experimenter has ever detected them. Just as an electrons random motions become more vigorous when one confines the electron to a smaller and smaller region so also gravitational vacuum fluctuations are more vigorous in small regions than in large, that is, for small wavelengths rather than for large. By combining the laws of quantum mechanics and the laws of general relativity in a tentative and crude way, it was deduced that in a region the size of the Planck-Wheeler length 1.62×10^{-33} centimeter or smaller, the vacuum fluctuations are so huge that space as we know it *boils* and becomes a froth of quantum foam - the same sort of quantum foam as makes up the core of a spacetime singularity as we have already seen with the bubbles of the inflation theory. See figure below.



The figure show embedding diagrams illustrating quantum foam. The geometry and topology of space are not definite; instead, they are probabilistic. They might have, for example, 0.1% probability for the form shown in (a), a 0.4% probability for (b), a 0.2% probability for (c), and so on.

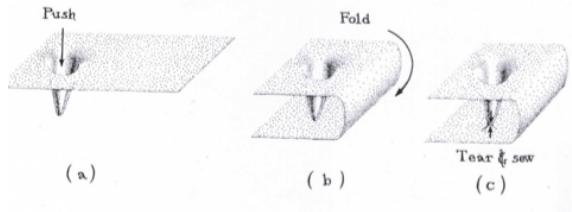
Quantum foam, therefore, is everywhere: inside black holes, in interstellar space, in the room where you sit, in your brain. But to see the quantum foam, one would have to zoom in with a (hypothetical) supermicroscope, looking at space and its contents on smaller and smaller scales. One would have to zoom in from the scale of you and me (hundreds of centimeters) to the scale of an atom (10^{-13} centimeter), to the scale of an atomic nucleus (10^{-15} centimeter), and then on downward by twenty factors of 10 more, to 10^{-33} centimeter. At all the early, *large* scales, space would look completely smooth, with a very definite (but tiny) amount of curvature. As the microscopic zoom nears, then passes 10^{-32} centimeter, however, one would see space begin to writhe, ever so slightly at first, and then more and more strongly until, when a region just 10^{-33} centimeter in size fills the supermicroscope's entire eyepiece, space has become a froth of probabilistic quantum foam.

Since the quantum foam is everywhere, it is tempting to imagine an infinitely advanced civilization reaching down into the quantum foam, finding in it a wormhole (say, a *big* one with its 0.4 percent probability as we will see), and trying to grab that wormhole and enlarge it to classical size. In 0.4 percent of such attempts, if the civilization were truly infinitely advanced, they might succeed. Or would they?

We do not yet understand the laws of quantum gravity well enough to know. One reason for our ignorance is that we do not understand the quantum foam itself very well. We aren't even 100% sure it exists. However, the challenge of this type of thought experiment - an advanced civilization pulling wormholes out of the quantum foam - might be of some conceptual help in the coming

years, in efforts to firm up our understanding of quantum foam and quantum gravity. So much for the quantum strategy of wormhole creation. What is the classical strategy?

In the classical strategy, our infinitely advanced civilization would try to warp and twist space on macroscopic scales (normal, human scales) so as to make a wormhole where previously none existed. It seems fairly obvious that, in order for such a strategy to succeed, one must tear two holes in space and sew them together - the figure below shows an example.



The strategy for making a wormhole illustrated here is

- (a) A *sock* is created in the curvature of space.
- (b) Space outside the sock is gently folded in hyperspace.
- (c) A small hole is torn in the toe of the sock, a hole is torn in space just below the hole, and the edges of the holes are *sewn* together.

This strategy looks classical (macroscopic) at first sight. However, the tearing produces, at least momentarily, a spacetime singularity, that is, a sharp boundary at which spacetime ends and which is governed by the laws of quantum gravity, so this strategy is really a quantum one.

We will not know Whether it is permitted until we understand the laws of quantum gravity. Is there no way out? Is there no Way to make a wormhole without getting entangled with the ill-understood laws of quantum gravity - no perfectly classical way?

Somewhat surprisingly, there is - but only if one pays a severe price. In 1966, Geroch showed that one can construct a wormhole by a smooth, singularity-free warping and twisting of spacetime, but one can do so only if, during the construction, time also becomes twisted up as seen in all reference frames. More specifically, While the construction is going on, it must be possible to travel backward in time, as well as forward; the *machinery* that does the construction, whatever it might be, must function briefly as a time machine that carries things from late moments of the construction back to early moments (but not back to moments before the construction began). The universal reaction to Gerochs theorem was - *surely the laws of physics forbid time machines, and thereby they*

will prevent a wormhole from ever being constructed classically, that is, without tearing holes in space.

Therefore, we must now ask - Do the laws of physics really forbid time machines, and if so, how? How might the laws enforce such a prohibition? Before proceeding with these questions, let us pause and take stock. Our best understanding of wormholes is this: If no wormholes were made in the big bang, then an infinitely advanced civilization might try to construct one by two methods, quantum (pulling it out of the quantum foam) or classical (twisting space-time without tearing it). We do not understand the laws of quantum gravity well enough to deduce whether the quantum construction of wormholes is possible. We do understand the laws of classical gravity (general relativity) well enough to know that the classical construction of wormholes is permitted only if the construction machinery, whatever it might be, twists time up so strongly, as seen in all reference frames, that it produces, at least briefly, a time machine. We also know that, if an infinitely advanced civilization somehow acquires a wormhole, then the only way to hold the wormhole open (so it can be used for interstellar travel) is by threading it with exotic material. We know that vacuum fluctuations of the electromagnetic field are a promising form of exotic material: They can be exotic (have a negative average energy density as measured by a light beam) in curved spacetime under a wide variety of circumstances. However, we do not yet know whether they can be exotic inside a wormhole and thereby hold the wormhole open.

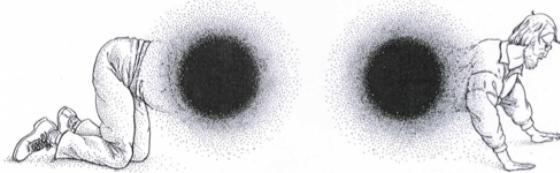
From now on we shall assume that an infinitely advanced civilization has somehow acquired a wormhole and is holding it open by means of some sort of exotic material; and I shall ask what other uses, besides interstellar travel, the civilization might find for its wormhole.

Time Machines

If a wormhole can really be held open, then it will permit one to travel over interstellar distances far faster than light. Doesn't this mean that one can also use a wormhole to travel backward in time? It was figured out how to construct a time machine using two wormholes that move at high speeds relative to each other. We shall not describe that time machine here, because it is a bit complicated and there is a simpler, more easily described time machine to which I shall come shortly. A bothersome problem at this point is reflected in the question: How does time decide how to hook itself up through a wormhole? To make this question more concrete, think about this example.

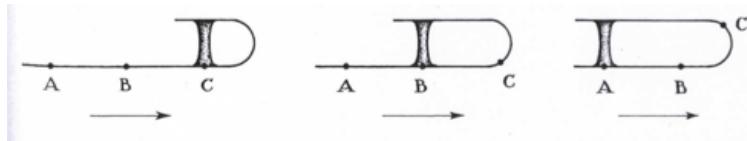
Suppose that we have a very short wormhole, one whose tunnel through hyperspace is only 50 centimeters long, and suppose that both mouths of the wormhole - two spheres, each 2 meters in diameter - are sitting in my Swarthmore office. And suppose that I climb through the wormhole, head first. From my viewpoint, I must emerge from the second mouth immediately after I enter

the first, with no delay at all; in fact, my head is coming out of the second mouth while my feet are still entering the first. Does this mean that one of my students, Ariana, sitting there, will also see my head emerging from the second mouth while my feet are still climbing into the first, as shown below?



If so, then time *hooks up through the wormhole* in the same manner as it hooks up outside the wormhole. On the other hand, I asked myself, isn't it possible that, although the trip through the wormhole takes almost no time as seen by me, Ariana must wait an hour before she sees me emerge from the second mouth; and isn't it also possible that she sees me emerge an hour before I entered? If so, then time would be hooked up through the wormhole in a different manner than it hooks up outside the wormhole. What could possibly make time behave so weirdly?

On the other hand, why shouldnt it behave in this way? Only the laws of physics know the answer. Somehow, we ought to be able to deduce from the laws of physics just how time will behave. As an aid to understanding how the laws of physics control times hookup, think about a more complicated situation. Suppose that one mouth of the wormhole is at rest in my office and the other is in interstellar space, traveling away from Earth at nearly the speed of light. And suppose that, despite this relative motion of its two mouths, the wormholes length (the length of its tunnel through hyperspace) remains always fixed at 50 centimeters. This is explained as follows. See figure below.

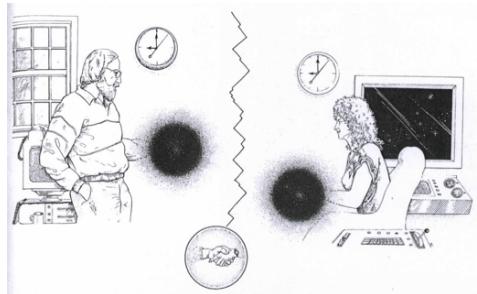


Each of the diagrams is an embedding diagram seen in profile. The diagrams are a sequence of snapshots that depict motion of the universe and the wormhole *relative to hyperspace*. Relative to hyperspace, the bottom part of our universe is sliding rightward in the diagrams, while the wormhole and the top part of our universe remain at rest. Correspondingly, as seen in our universe, the mouths of the wormhole are moving relative to each other (they are getting farther apart), but as seen through the wormhole they are at rest with respect to each other; the wormhole's length does not change. Thus, the figure shows how it is possible for the length of the wormhole to remain fixed while its mouths, as seen in the external Universe, move relative to each other.

As seen in the external Universe, the two mouths are in different reference frames, frames that move at a high speed relative to each other; and the mouths therefore must experience different flows of time. On the other hand, as seen through the wormholes interior, the mouths are at rest with respect to each other, so they share a common reference frame, which means that the mouths must experience the same flow of time. From the external viewpoint they experience different time flows, and from the internal viewpoint, the same time flow; how confusing!

Gradually, the confusion subsides and all becomes clear. The laws of general relativity predict, unequivocally, the flow of time at the two mouths, and they predict, unequivocally, that the two time flows will be the same when compared through the wormhole, but will be different when compared outside the wormhole. Time, in this sense, hooks up to itself differently through the wormhole than through the external universe, when the two mouths are moving relative to each other. And this difference of hookup implies that from a single wormhole, an infinitely advanced civilization can make a time machine. There is no need for two wormholes. How? Easy, if you are infinitely advanced.

To explain how, we consider a thought experiment in which we humans are infinitely advanced beings. My friend and I find a very short wormhole, and put one of its mouths in my office and the other outside on the lawn. Now, as this thought experiment will show, the manner in which time is hooked up through any wormhole actually depends on the wormholes past history. For simplicity, I shall assume that when my friend and I first acquire the wormhole, it has the simplest possible hookup of time: the same hookup through the wormholes interior as through the exterior universe. In other words, if I climb through the wormhole, My friend, I, and everyone on Earth will agree that I emerge from the mouth on the lawn at essentially the same moment as I entered the mouth in my office. Having checked that time is, indeed, hooked up through the wormhole in this way, my friend and I then make a plan: I will stay in my office with the one mouth, while my friend takes the other mouth on a very high speed trip out into the universe and back. Throughout the trip, we will hold hands through the wormhole as shown below.



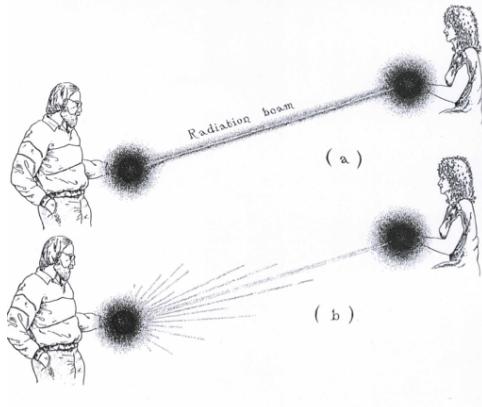
My friend departs at 9:00 A.M. on 1 January 2001, as measured by herself, by me, and by everybody else on Earth. My friend zooms away from Earth at nearly the speed of light for 6 hours as measured by her own time; then she reverses course and zooms back, arriving on the front lawn 12 hours after her departure as measured by her own time. I hold hands with her and watch her through the wormhole throughout the trip, so obviously I agree, while looking through the wormhole, that she has returned after just 12 hours, at 9:00 P.M. on 1 January 2001. Looking through the wormhole at 9:00 P.M., I can see not only my friend; I can also see, behind her, the lawn. Then, at 9:01 P.M., I turn and look out the window-and there I see an empty lawn. The vehicle is not there; my friend and the other wormhole mouth are not there. Instead, if I had a good enough telescope pointed out the window, I would see my friends spaceship flying away from Earth on its outbound journey, a journey that as measured on Earth, *looking through the external universe*, will require 10 years (This is the standard *twins paradox*, the high-speed *twin* who goes out and comes back (my friend) measures a time lapse of only 12 hours, while the *twin* who stays behind on Earth (me) must wait 10 years for the trip to be completed.

I then go about my daily routine of life. For day after day, month after month, year after year, I carry on with life, waiting until finally, on 1 January 2011, my friend returns from her journey and lands on the lawn. I go out to meet her, and find, as expected, that she has aged just 12 hours, not 10 years. She is sitting there in the vehicle, her hand thrust into the wormhole mouth, holding hands with somebody. I stand behind her, look into the mouth, and see that the person whose hand she holds is myself, 10 years younger, sitting in our living room on 1 January 2001. The wormhole has become a time machine. If I now (on 1 January 2011) climb into the wormhole mouth in the vehicle, I will emerge through the other mouth in my office on 1 January 2001, and there I will meet my younger self. Similarly, if my younger self climbs into the mouth in the living room on 1 January 2001, he will emerge from the mouth in the vehicle on 1 January 2011. Travel through the wormhole in one direction takes me backward 10 years in time; travel in the other direction takes me 10 years forward.

Neither I nor anyone else, however, can use the wormhole to travel back in time beyond 9:00 P.M., 1 January 2001. It is impossible to travel to a time earlier than when the wormhole first became a time machine. The laws of general relativity are unequivocal. If wormholes can be held open by exotic material, then these are general relativity predictions.

Soon a problem arises. It seems that a wormhole would be automatically destroyed whenever an advanced civilization tries to convert it into a time machine? Let me explain. Imagine that my friend is zooming back to Earth with one wormhole mouth in her vehicle and I am sitting in my office on Earth with the other. Shortly after she turns around and begins zooming home, it suddenly becomes possible for radiation (electromagnetic waves) to use the wormhole for

time travel: Any random bit of radiation that leaves my office in Swarthmore traveling at the speed of light toward her vehicle can arrive at the spacecraft after 5 years time (as seen on Earth), enter the wormhole mouth there, travel back in time by 5 years (as seen on Earth), and emerge from the mouth on Earth at precisely the same moment as it started its trip. The radiation piles right on top of its previous self, not just in space but in spacetime, doubling its strength. What's more, during the trip each quantum of radiation (each photon) got boosted in energy due to the relative motion of the wormhole mouths (a *Doppler-shift* boost). After the radiation's next trip out to the vehicle then back through the wormhole, it again returns at the same time as it left and again piles up on itself, again with a Doppler-boosted energy. Again and again this happens, making the beam of radiation infinitely strong as shown below (part (a)).



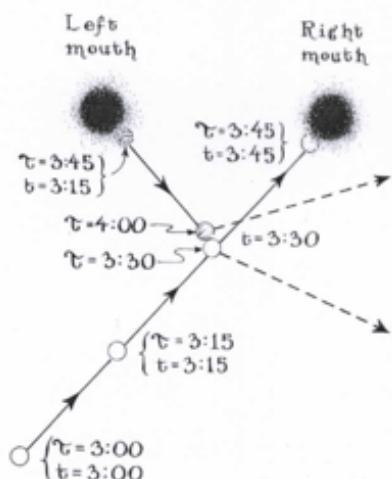
In this way, beginning with an arbitrarily tiny amount of radiation, a beam of infinite energy is created, coursing through space between the two wormhole mouths. As the beam passes through the wormhole it will produce infinite spacetime curvature and probably destroy the wormhole, thereby preventing the wormhole from becoming a time machine.

Thinking about it, the wormhole would not be destroyed. We have overlooked a crucial fact - every time the beam of radiation passes through the wormhole, the wormhole defocuses it in the manner as discussed earlier. After the defocusing, the beam emerges from the mouth on Earth and spreads out over a wide swath of space, so that only a tiny fraction of it can get caught by the mouth on the vehicle and transported through the wormhole back to Earth to *pile up* on itself (part (b)). By adding up all the radiation from all the trips through the wormhole (a tinier and tinier amount after each defocusing trip), it turns out that the final beam would be weak; far too weak to destroy the wormhole. This brush with wormhole destruction should warn us that unexpected dangers await any maker of time machines.

The Matricide Paradox Among possible controversies, the most vigorous was over what is called the *matricide paradox*. If we have a time machine (wormhole-based or otherwise), I should be able to use it to go back in time and kill my mother before I was conceived, thereby preventing myself from being born and killing my mother. Central to the matricide paradox is the issue of free will - do I, or do I not, as a human-being, have the power to determine my own fate? Can I really kill my mother, after going backward in time, or (as in so many science fiction stories) will something inevitably stay my hand as I try to stab her in her sleep?

Now, even in a universe without time machines, free will is a terribly difficult thing for physicists to deal with. We usually try to avoid it. It just confuses issues that otherwise might be lucid. With time machines, all the more so. It was a conjecture at this point to suggest that there would never be unresolvable paradoxes for any inanimate object that passes through the wormhole and the controversy would only occur when humans go through. No such luck!

Consider this elegant and simple variant of the matricide paradox - a variant that is not entangled with free will and that we therefore should work OK if the conjecture were valid. Take a wormhole that has been made into a time machine, and place its two mouths at rest near each other, out in interplanetary space as shown below.

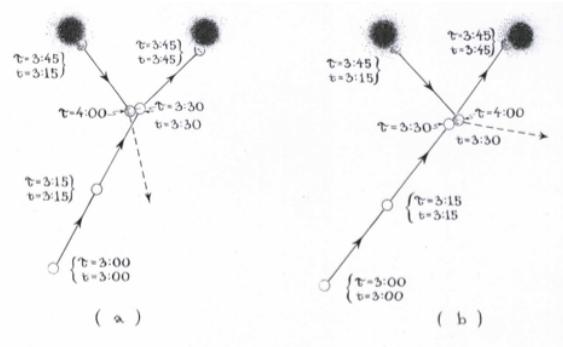


The details of the experiment are: The wormhole is very short and has been made into a time machine, so that anything that enters the right mouth emerges as measured on the outside, 30 minutes before it went in. The flow of time outside the mouth is denoted by the symbol t ; the flow of time as experienced by the billiard ball itself is denoted by τ . The billiard ball is launched at $t = 3 : 00 P.M.$ from the indicated location and with just the right velocity

to enter the right mouth at $t = 3 : 45 P.M.$. The ball emerges from the left mouth 30 minutes earlier at $t = 3 : 15 P.M.$, and then hits its younger self at $t = 3 : 30 P.M.$, knocking itself off track so it cannot enter the right mouth and hit itself.

Thus, if a billiard ball is launched toward the right mouth from an appropriate initial location and with an appropriate initial velocity, the ball will enter the right mouth, travel backward in time, and fly out of the left mouth before it entered the right (as seen by you and me outside the wormhole), and it will then hit its younger self, thereby preventing itself from ever entering the right mouth and hitting itself. This situation, like the matricide paradox, entails going back in time and changing history. In the matricide paradox, I go back in time and, by killing my mother, prevent myself from being born. In this paradox, the billiard ball goes back in time and, by hitting itself, prevents itself from ever going back in time. Both situations are nonsensical. Just as the laws of physics must be logically consistent with each other, so also the evolution of the universe, as governed by the laws of physics, must be fully consistent with itself - or at least it must be so when the universe is behaving classically (non-quantum mechanically); the quantum mechanical realm is a little more subtle. Since both I and a billiard ball are highly classical objects (that is, we can exhibit quantum mechanical behavior only when one makes exceedingly accurate measurements on us) there is no way that either I or the billiard ball can go back in time and change our own histories.

So what happens to the billiard ball? To find out we focus our attention on the balls initial conditions, that is, its initial location and velocity. We ask - For the same initial conditions as led to the above paradox, is there any other billiard ball trajectory that, unlike the one in the above figure, is a logically self-consistent solution to the physical laws that govern classical billiard balls? Yes. There indeed is a fully self-consistent billiard ball trajectory that begins with initial data in the paradox and satisfies all the laws of physics that govern classical billiard balls. In fact, there are two such trajectories. They are shown in in the figure below.



Let us describe each of these trajectories in turn, from the viewpoint of the ball itself. On trajectory (a) (left half of figure), the ball, young, clean, and pristine, starts out at time $t = 3 : 00 P.M.$, moving along precisely the same route as in the original paradox statement, a route taking it toward the wormholes right mouth. A half hour later, at $t = 3 : 30 P.M.$, the young, pristine ball gets hit on its left, rear side, by an older-looking, cracked ball (which will turn out to be its older self). The collision is gentle enough to deflect the young ball only slightly from its original course, but hard enough to crack it. The young ball, now cracked, continues onward along its slightly altered trajectory and enters the wormhole mouth at $t = 3 : 45 P.M.$, travels backward in time by 30 minutes, and exits from the other mouth at $t = 3 : 15 P.M.$. Because its trajectory has been altered slightly by comparison with the original paradoxical trajectory, the ball, now old and cracked, hits its younger self a gentle, glancing blow on the left, rear side at $t = 3 : 30 P.M.$, instead of the vigorous, highly deflecting blow of the original arrangement. The evolution thereby is made fully self-consistent.

Trajectory (b), the right half of the figure, is the same as (a), except that the geometry of the collision is slightly different, and correspondingly the trajectory between collisions is slightly different. In particular, the old, cracked ball emerges from the left mouth on a different trajectory than in (a), a trajectory that takes it in front of the young, pristine ball (instead of behind it), and produces a glancing blow on the young balls front, right side (instead of left, rear side).

One can show that both trajectories, (a) and (b), satisfy all the physical laws that govern classical billiard balls, so both are possible candidates to occur in the real universe(if the real universe can have wormhole-based time machines).

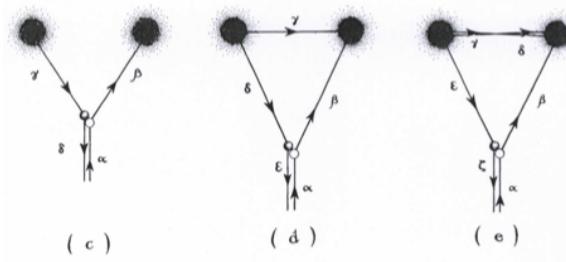
This is most disquieting. Such a situation can never occur in a universe without time machines. Without time machines, each set of initial conditions for a billiard ball gives rise to one and only one trajectory that satisfies all the classical laws of physics. There is a unique prediction for the balls motion. The time machine has ruined this. There now are two, equally good predictions for the balls motion. Actually, the situation is even worse than it looks at first sight: The time machine makes possible an infinite number of equally good predictions for the balls motion, not just two. We now show a simple example.

The Billiard Ball Crisis: An Infinity of Trajectories

If a billiard ball is fired between the two mouths of a wormhole-based time machine, there are two trajectories on which it can travel. See figure below.



On one (a), it hurtles between the mouths unscathed. On the other (b), as it is passing between the two mouths, it gets hit and knocked rightward, toward the right mouth; it then goes down the wormhole, emerges from the left mouth before it went down, hits itself, and flies away. Later, a third trajectory was found that satisfies all the laws of physics, the trajectory (c) below:



The collision, instead of occurring between the mouths, occurs before the ball reaches the mouths vicinity.

In fact, the collision could be made to occur earlier and earlier, as in (d) and (e) above, if the ball travels through the wormhole several times between its two visits to the collision event. For example, in (e), the ball travels up route α gets hit by its older self and knocked along β and into the right mouth; it then travels through the wormhole (and backward in wormhole again (and still farther back in time), emerging from the left mouth on γ , which takes it through the wormhole yet again (and even farther back in time), emerging along ϵ , which takes it to the collision event, from which it is deflected down ζ .

Evidently, there are an infinite number of trajectories (each with a different number of wormhole traversals) that all satisfy the classical (non-quantum) laws of physics, and all begin with identically the same initial conditions (the same initial billiard ball location and velocity). One is left wondering whether physics has gone crazy, or whether, instead, the laws of physics can somehow tell us which trajectory the ball ought to take.

Do time machines make physics go crazy? Do they make it impossible to predict how things evolve? If not, then how do the laws of physics choose which trajectory, out of the infinite allowed set, a billiard ball will follow?

In search of an answer we must turn from the classical laws of physics to the quantum laws. Why the quantum laws? Because they are the Ultimate Rulers of our Universe. For example, the laws of quantum gravity have ultimate control over gravitation and the structure of space and time. Einsteins classical, general relativistic laws of gravity are mere approximations to the quantum gravity laws - approximations with excellent accuracy when one is far from all singularities and looks at spacetime on scales far larger than 10^{-33} centimeter, but approximations nevertheless.

Similarly, the classical laws of billiard ball physics, which we have used in studying the paradox, are mere approximations to the quantum mechanical laws. Since the classical laws seem to predict *nonsense* (an infinity of possible billiard ball trajectories), one hoped that the quantum mechanical laws would give a deeper understanding. The *rules of the game* are very different in quantum physics than in classical physics. When one provides the classical laws with initial conditions, they predict what will happen afterward (for example, what trajectory a ball will follow); and, if there are no time machines, their predictions are unique. The quantum laws, by contrast, predict only probabilities for what will happen, not certainties (for example, the probability that a ball will travel through this, that, or another region of space). In light of these rules of the quantum mechanical game, the answer was obtained from the quantum mechanical laws is not surprising. It said that if the ball starts out moving along the original paradoxical trajectory ($time = 5 : 00 P.M.$), then there will be a certain quantum mechanical probability - say, 48 percent - for it subsequently to follow trajectory (a) above, and a certain probability - say, also 48 percent - for trajectory (b), and a certain (far smaller) probability for each of the infinity of other classically allowed trajectories. In any one *experiment*, the ball will follow just one of the trajectories that the classical laws allow; but if we perform a huge number of identical billiard ball experiments, in 48 percent of them the ball will follow trajectory (a), in 48 percent trajectory (b), and so forth.

This conclusion is somewhat satisfying. It suggests that the laws of physics might accommodate themselves to time machines fairly nicely. There are surprises, but there seem not to be any outrageous predictions, and there is no sign of any unresolvable paradox.