Quest for Unification†

H.C. Lee
Theoretical Physics Branch
Atomic Energy of Canada Limited
Chalk River Nuclear Laboratories
Chalk River, Ontario, Canada K0J 1J0
and
Department of Applied Mathematics
University of Western Ontario
London, Ontario, Canada N6A 5B9

Abstract

The motivation and meaning for the long march towards a unification of all fundamental laws of physics are discussed. Progress made in the modern era, including the rise to prominence of string and superstring theories, is described. Abstract concepts used by theorists, such as those involving extra dimensional spaces and their topological properties, to gain a deeper understanding of the physical laws are explained. The presentation is nonmathematical and intended for nonspecialists.

FUNDAMENTAL LAWS OF PHYSICS

The total body of physical phenomena is vast and infinitely complex, and no one could hope to understand all of it. But we believe the set of fundamental laws governing these phenomena is finite, perhaps even quite small. Newton's simple law of gravity

$$F = G\,\frac{M_1 M_2}{R^2}$$

allows one to understand ballistics (ICBM nowadays), tidal waves, planetary motions, quite a bit of astro-physics and numerous other things. Principles underlying dynamos, radio, TV, radar, microwaves and, with a little bit of quantum mechanics thrown in, atomic physics, lasers, etc. are all based on the set of four Maxwell's equations for electromagnetism. Fundamental laws govern simple as well as complex systems. We use the simplest system to study the laws, and apply the laws on complex systems. Often the degree of complexity of a system is so high that the connection between its behaviour and a fundamental law is extremely remote — thus biology will probably always remain a largely empirical science.

The laymen's and specialists' views of the four force laws that form the body of our knowledge of the fundamental laws of phsyics are shown in Tables 1 and 2, respectively.

As we progress to a deeper and deeper understanding of the fundamental laws of physics, we are forced to study simpler and simpler systems. Galileo watched stars; Newton, apples; and now we study isolated electrons, photons, quarks and gluons, all invisible to the naked eye. Paradoxically the instruments required for such studies have become even bigger, and more powerful and expensive to build, while the laws involve ideas that, at least to the uninitiated, have become more abstract and complex.

TABLE 1. LAYMAN'S VIEW OF THE FOUR KNOWN FORCES

FORCE	PHENOMENA
ELECTROMAGNETIC	ALL THINGS ELECTRIC RADIO, TV . . . LIGHT, OPTICS, LASER BIOLOGY CHEMISTRY ATOMIC PHYSICS CONDENSED MATTER PHYSICS
WEAK	β-DECAY (LONG LIVED RADIOACTIVITY) CONTROL OF NUCLEAR REACTIONS BURNING OF STARS
STRONG	BURNING OF STARS NUCLEAR POWER STRUCTURE OF NUCLEUS OF MATTER NEUTRON STARS
GRAVITATIONAL	WEIGHT AND FITNESS LARGE TERRESTRIAL STRUCTURE PLANETARY MOTION LARGE SCALE STRUCTURE OF UNIVERSE BLACK HOLES

TABLE 2. PARTICLE PHYSICISTS' VIEW OF THE FOUR KNOWN FORCES

FORCE	CARRIER (BOSONS)	AFFECTED ELEMENTARY PARTICLES*
ELECTRO-MAGNETIC	PHOTON	ALL CHARGED LEPTONS (ELECTRON, POSITRON, $\cdots$) ALL QUARKS, $W^{\mp}$ BOSONS
WEAK	$W^{\mp}$, Z° BOSONS	ALL LEPTONS (ELECTRONS, NEUTRONS $\cdots$) ALL QUARKS, $W^{\mp}$ AND Z^0
STRONG	GLUONS	ALL QUARKS, GLUONS
GRAVITA-TIONAL	GRAVITON (?)	ALL PARTICLES WITH MASS

BOSONS	FERMIONS (LEPTONS & QUARKS)
↓	↓
OBEY BOSE-EINSTEIN STATISTICS	OBEY FERMI-DIRAC STATISTICS
↓	↓
GENERATE FORCE FIELDS	CONSTITUENTS OF MATTER

† Work supported in part by a grant from the Natural Sciences & Engineering Research Council of Canada.

* Particles interact with each other by the emission and absorption of carriers.

SPACETIME SYMMETRY AND INTERNAL STYMMETRY

The theories[a] that we use to express our understanding of the laws of physics possess many symmetries, derived from the fact that they are invariant (i.e., do not change) under certain transformations. Gravity, or general relativity, is built on the principle of invariance under general coordinate transformations in spacetime. Often there is a direct relation between a symmetry of a theory and a conserved quantity. In spacetime the symmetry of rotational invariance leads to the conservation of angular momentum and translational invariance to the conservation of linear momentum. There are however symmetries not apparently related to spacetime transformations. These are called internal symmetries, the implication being that they may reflect the symmetries of transformations in some internal space. Invariance of the electron wave function χ under the phase transformation

$$\chi(x) \rightarrow e^{i\Lambda}\chi(x)$$

leads to conservation of the number of electrons. That under the local gauge transformation

$$\chi(x) \rightarrow e^{i\lambda(x)}\chi(x)$$

leads to the conservation of electric charge. Note that Λ is a constant while $\lambda(x)$ is an arbitrary function of spacetime coordinate x (which is why we call the gauge transformation local). Were there not such an arbitrariness, there would not be the guarantee that a person measuring the electric force between two electrons in Iqaluit will find exactly the same result as another person making the measurement on Queen Charlotte Island.

Gauge invariance (by which from now on we shall mean local gauge invariance) also has a causal relation with the vanishing of the photon mass, responsible for the unscreened coulomb force having an infinite range. The modern view of electromagnetism is that it is a theory based on the principle of invariance under gauge transformations characterized by a single (but arbitrary) function, just as gravity is the theory based on the principle of invariance under general coordinate transformations. The symmetry group for the latter is GL(3,1), which acts on the four-dimensional spacetime. The symmetry group for the former is the one parameter U(1), which acts on a space equivalent to a circle which may not have any physical meaning but which we like to think of as an internal space. The three fundamental forces other than gravity all have an underlying principle of gauge invariance. In the case of quantum chromodynamics, the theory for the strong interaction, the gauge transformation on the quark wavefunction $\chi(x)$ is

$$\chi(x) \rightarrow \exp\left(\sum_a i\lambda_a(x)t_a\right)\chi(x)$$

where t_a, $a = 1, \cdots 8$, are the eight generators of SU(3) group, λ_a are eight arbitrary functions, and $\chi(x)$ transforms as the fundamental representation of SU(3). Quarks are to quantum chromodynamics what electrons are to electromagnetism: in both cases these particles, collectively known as fermions, interact with each other by exchanging bosons, or the force carriers; gluons in the case of chromodynamics and photons in the case of electromagnetism. There is very strong indirect evidence that protons and neutrons, which constitute the core (or nucleus) of all matter on earth, and most likely that in the whole universe, are each made of three quarks. In chromodynamics the conserved quantity associated with the SU(3) gauge invariance is the analogy of the electric charge, the "color" of strong interaction.

[a] We call a theory the minimum set of propositions constructed to precisely describe, following established mathematical rules and logical conventions, the cause and effect of a set of physical phenomena. Thus for the electromagnetic force, the theory of electromagnetism, and so on.

UNIFICATION OF GAUGE THEORIES

The theory for the weak force[1] is a little bit more complicated. Actually, the weak force was without a real theory until the early seventies, even though an effective theory existed and was sufficiently reliable to enable man to utilize it to build reactors, treat cancer and, alas, also build bombs. There is still not (and never will be) a theory for the weak force by itself. Rather it is part of a theory that unites it with the electromagnetic force under one principle of gauge invariance, with the gauge group being SU(2) × U(1). Here nature only reveals the symmetry in a badly broken form; the $W^{\mp}$ and Z^{o} bosons — carriers of the weak force, instead of being massless as the photon is, are about 100 times as heavy as the proton. Broken symmetry is a familiar phenomenon. In superconducting matter, translation invariance is broken when temperature drops below a critical value, and paired electrons develop an energy gap that induces superconductivity. Above the critical temperature translation symmetry is restored, the energy gap disappears and superconductivity is lost.

In the study of fundamental law of physics temperature is commonly measured in terms of energy — the SU(2) × U(1) symmetry uniting the electromagnetic and weak forces is broken at about 100 GeV (about 10^{15} °C). Below that energy the $W^{\mp}$ and Z^{o} boson become very massive and the strength of the weak force is drastically reduced in comparison to that of the electromagnetic force carried by the massless photon. Above that energy the $W^{\mp}$ and Z^{o} also become massless and the two forces are united.

The broken SU(2) × U(1) symmetry makes our universe a superconductor for neutrinos. In a normal superconductor an electron is prevented from interacting with its surroundings unless it is sufficiently energetic to overcome the energy gap; the result is that most electrons travel unimpeded. In the same way, when the symmetry of the unified electroweak force is broken, the neutrino is effectively prevented from interacting with anything unless its energy is close to the masses of the $W^{\mp}$ and Z^{0} bosons; low energy neutrinos therefore travel through the cosmos unimpeded[b] (see Figure 1). A lucky break, since it enables us to decode the reaction taking place at the core of the sun by studying solar neutrinos that traverse the interior of the sun, the space between the sun and the earth and the atmosphere, and reach our laboratories practically untouched. No other particle would have survived such a journey.

Like many successful theories before it, the SU(2) × U(1) theory made testable predictions: it predicted the existence of a new type of (neutral) weak force mediated by Z^{o}, verified experimentally[2] in 1973, and predicted the values of the masses of the Z^{o} and $W^{\mp}$ bosons, verified[3] in 1983. The success of the SU(2) × U(1) electroweak theory suggests that all the three forces based on the principle of gauge invariance might be united in a theory with an underlying gauge group containing

$$\underset{\text{strong}}{SU(3)} \times \underset{\text{electroweak}}{SU(2) \times U(1)}$$

The most important prediction of this grand unification is that proton is not stable, but must decay (or disintegrate) in processes such as

$$\text{proton} \rightarrow \text{electron} + \text{pi-meson}$$

with a lifetime calculated to be of the order of 10^{31} years. If our universe has lived for only 20 billion (2×10^{10}) or so years, how would it be possible to measure a lifetime that is 10^{31} years? One will not think the effort futile if one understands lifetime is an expression of probability — the

[b] Not completely. In potential theory, the neutrino can still interact with others by a tunnelling effect. In field theory, it interacts by exchanging virtual $W^{\mp}$ and Z^{o} boson with other particles.

10^{31} year lifetime also means that one in 10^{31} protons should disintegrate in any given year. A water tank five times the size of an olympic-size swimming pool filled with water (10^4 cubic meters) has about 10^{31} protons, so in order to test the prediction of proton decay, it is sufficient to be able to detect the decay signal generated in such a water tank, or in some other detector built on the same principle. Intensive and elaborate searches for proton decay carried out in the last several years has so far been unsuccessful, yielding the conclusion that if the strong and electroweak forces are united, it must be in such a way that the proton lifetime is longer than about 10^{32} years. The water-tank proton-decay detectors can also serve as neutrino detectors (for some purposes it would be better to replace the water with heavy water). Since the universe is virtually a superconductor to neutrinos, large neutrino detectors could be the astronomical observatories of the future, allowing us to probe further and deeper into the cosmos than ever before. Indeed, tremendous excitement[4] was recently generated by the detection, at several proton-decay detectors, of bursts of signals believed to be triggered by neutrinos from the brightest supernova seen this century and, at 170,000 light years away so close it is almost within our own galaxy.

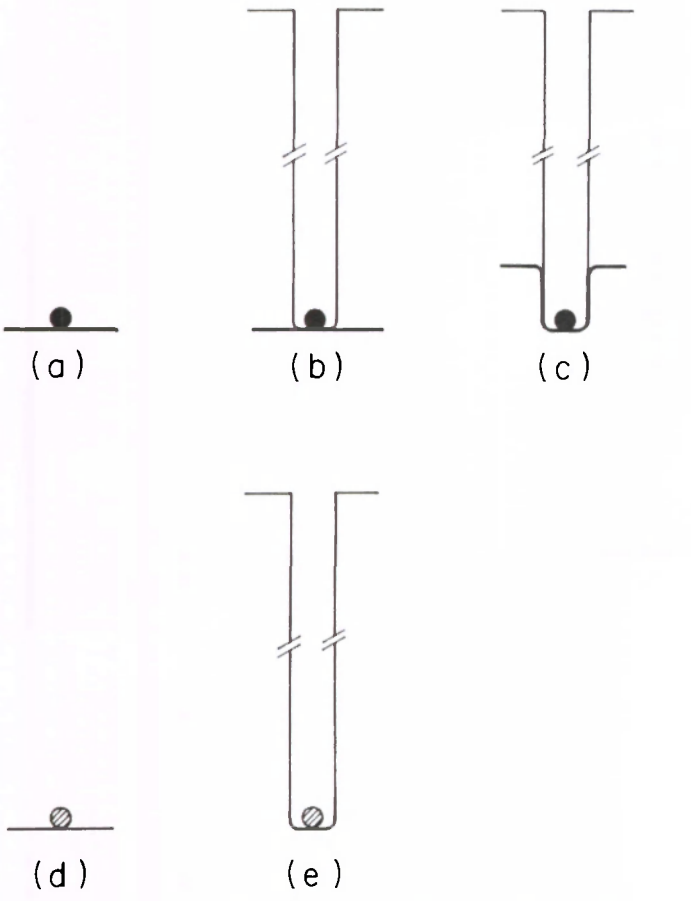

Fig. 1. The universe is a superconductor to neutrinos. In an environment with a temperature of about 10^{15}°C, as was our universe an instant after the "big bang", the SU(2) x U(1) symmetry of the electroweak force is unbroken. In this case both the electron (a) and the neutrino (d) are free to interact with other particles, thereby losing energy. When temperature cools to below 10^{15}°C, the SU(2) x U(1) symmetry is broken, and an "energy gap" of about 10^{11} electron volts is formed that prevents weak interaction. This keeps the neutrino (e), which is oblivious to the electromagnetic force, from any interaction, so that the universe becomes a superconductor to neutrinos. This is not the case for the electron (b), whose electromagnetic interaction is still unimpeded. When the temperature is further cooled down to below the critical temperature for the material in which the electron resides (about −250°C in conventional superconductors and perhaps up to 14°C in the newly discovered superconductors) translation symmetry is broken and an energy gap of 3×10^{-3} to 3×10^{-2} electron volts is formed (c) that prevents the electron from interaction, making the material a superconductor to electrons.

INTERNAL SYMMETRY AND EXTRA DIMENSIONS

The internal symmetry of gauge transformations and the spacetime symmetry of general coordinate transformations are dynamical symmetries, meaning that they are symmetries out of which theories for forces emerge. There are also symmetries which are not dynamical, and some of these, just as gauge symmetries, are internal. A well known and extremely useful example is the isospin in nuclear physics. In isospin space the proton and the neutron are just the two "magnetic" substates, spin-up and spin-down, of a single isospin-1/2 nucleon state. If we believe in quarks, then isospin is just the manifestation of a larger internal symmetry among quarks, called flavor. What are the spaces on which internal symmetries act? It would be more satisfying to us if they acted on some internal but nevertheless real spaces, rather than just on abstract ones. This would imply spacetime is not just four-dimensional, but rather has extra dimensional compact[c] spaces — the internal symmetry spaces. There would be no conflict with reality provided these internal spaces are curled up to such a small size that their effect, other than the internal symmetries, have so far escaped detection (see Figure 2).

The idea of higher (i.e. more than 4) dimensional spacetime is an old one. Many years ago Kaluza[5] pointed out that if spacetime were 5-dimensional, with the extra compact dimensional space being a circle of radius R (see Figure 3), and if the law of physics in the 5-dimensional spacetime were just Einstein's gravity in five dimensions, then in the limit of small R the low energy approximation of the theory would be just Einstein's gravity and Maxwell's electromagnetism in 4-dimensional Minkowski spacetime. Expanding on this idea, Klein[6] showed that the charge of the induced electromagnetism should be quantized by virtue of the finiteness

Fig. 2. A point (a) is zero-dimensional. The circle (b) is a one-dimensional compact space known as a one-sphere, or S^1. The surface of the sphere (c) is a two-dimensional compact space known as a two-sphere, or S^2. Impossible-to-draw generalizations are the n-spheres, S^n, n = 3,4,···. When the size of the compact space (in the case of S^n, its radius) is much smaller than any instrument can measure, the space becomes indistinguishable from a point (d), except that objects built on such "points" may have internal symmetries arising from the topological (roughly, global geometric) properties of the compact space. In the other extreme, when the size of the compact space is large compared to the measuring scale, then the neighborhood of any point on the space can be viewed as being flat.

[c] Very roughly, a compact space is a bounded space that includes its boundary, if there is one. Many interesting compact spaces have no boundaries. An example is the surface of a sphere, see Figure 2(c).

of R, in which case R must be of the order of 10^{-31} cm, which is certainly far, far smaller than anything that could be measured directly. There is no good reason that the Kaluza radius would not vibrate, thereby generating excitation modes with frequencies of the order of 1/R, or 10^{17} GeV. Since the masses of known elementary particles are all either much smaller or of the order of 100 GeV, it is concluded that they must belong to the zero-energy modes of the particle spectrum of the Kaluza-Klein universe.

From 1921 up to the late 1970's, the status of the Kaluza-Klein theory was essentially that of a mathematical curiosity. Among other things, the idea of an extra dimension was too farfetched; the smallness of its radius rendered any measurement unimaginable. There has been a very significant revival of Kaluza-Klein theories in the last few years. Now-a-days the implications of the extra dimensions are taken seriously. After all the radius of Kaluza's circle is not much smaller than the length scale of grand unification, which is the object of many intensive experimental tests (proton decay, monopoles, neutrino mass, etc.). Going to higher dimensions seems to be the best way we know to begin to unite gravity, not only with electromagnetism, but with all other forces of nature, and the extra-dimensional compact spaces are still our best bet to understand the origins of internal symmetries[7] (see Figure 4).

TROUBLES WITH QUANTUM GRAVITY

The Kaluza-Klein approach to unification was ultimately unsuccessful for several reasons. One is related to how fermions, or matter particles, are incorporated into the theory. The fermions (electrons, quarks, etc.) we know have definite handedness, and the left-handed ones are distinct from the right-handed ones. A theory with distinct left and right-handed fermions is said to have chiral[d] symmetry and can only be constructed in spaces with certain properties (for example, the total number of extra dimensions must be even and the Euler number of the extra dimensional space must not vanish) which are difficult to satisfy in a Kaluza-Klein context. The handedness of fermions can be empirically verified by showing that the mirror image (in the normal sense of the word) of a fermion is distinct from itself, and chiral symmetry is what we believe to be the symmetry that protects fermions from acquiring masses of the order of the Kaluza-Klein scale (about 10^{17} GeV). Another reason is whereas forces based on the gauge principle can be quantized, gravity cannot, and remains unquantizable even after a Kaluza-Klein unification. The quantum effects of gauge theories are firmly established experimental facts, so it would be highly unsatisfactory if we were forced to concede that gravity alone must remain a classical theory. This would amount to saying that the uncertainty principle applies to gauge theories, but not to gravity.

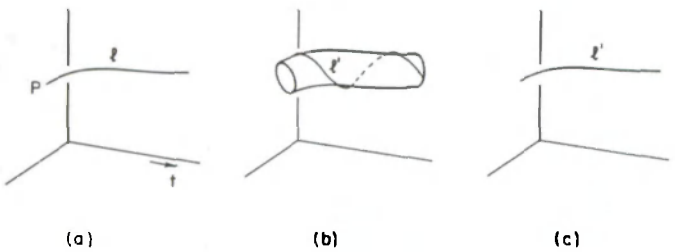

Fig. 3. A point P in motion in Minkowski spacetime M traces a world-line ℓ (a). In Kaluza's five-dimensional spacetime, which is the product space M x S^1, each point in M is replaced by a circle so that, locally, world-lines such as ℓ' lie on the surface of a tube (b). When the radius of the Kaluza circle is far smaller than any instrument can measure, the world-line again appears to move in normal spacetime (c), except that its points have internal structure.

The connection between the uncertainty principle and difficulty in quantizing gravity is as follows. If a photon has energy exceeding twice the electron mass (about 1.1. MeV) then it has a certain probability of converting into an electron-positron pair. Conversely such a pair can always mutually annihilate and change into a photon. The uncertainty principle allows the photon-pair conversion even when the photon is less energetic than 1.1 MeV, in which case the pair will exist only for a very short time period before changing back to the original photon. This process

PHOTON → (virtual) e^+e^- PAIR → PHOTON

is called vacuum polarization; the energy of the pair can be arbitrarily high, corresponding to an arbitrarily short wave length and, according to the uncertainty principle, the time period for existence of the pair must be correspondingly short. By their very nature all quantum theories permit vacuum polarization, and give precise rules for computing its probability.

Invariably on first trial the calculated amplitude for vacuum polarization is infinitely large, giving a singularity to the theory. In gauge theories such infinities can be absorbed by redefinitions of the wave functions, interaction strengths and masses of the particle, so that (at least in principle) a finite prediction is made of the value of any physically measureable quantity. This scheme, known as renormalization, has been empirically tested many, many times for gauge theories, and has yet to be shown to be flawed.

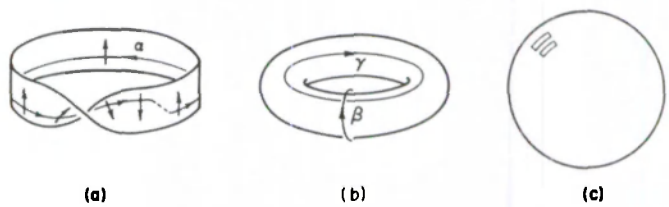

Fig. 4. Spaces classified by their topological properties are called manifolds. The Moebius strip M^2 (a) is formed by cutting a strip and gluing it back together after giving it a single twist. It facilitates an intrinsic integer-counting system based on the number of loops α wound around the strip. The strip also has a group with two elements: e, the identity and g, $g^2 = 1$. This group structure is revealed when an upward-pointing arrow changes to downward pointing after it is transported once around the strip. The action of the group element g on the arrow is to flip it from up to down, or vice versa, while e leaves the arrow unchanged. The torus T^2 (b) is the product S^1 x S^1. It facilitates two independent integer-counting systems: one counting the winding number around the small circle β and the other the winding number around the big circle γ. The two-sphere S^2 (c) cannot count loops, since any loop drawn on it can shrink to a point. However, it counts layers of wrapping covering the whole sphere. M^2, T^2 and S^2, as do most compact spaces, also have other topological group structures which are more difficult to visualize. Because these topological groups are independent of the size of the spaces and are unchanged by any distortion affected on the spaces, they are attractive candidates as the bases of internal symmetries of elementary particles.

[d] From the Greek word chiro, hand.

RENORMALIZATION

The concept of *renormalization* can be understood by considering the following functional (function of a function),

$$L[f,m,g] = (d/dx)^2f^2(x) - m^2f^2(x) + gf^4(x)$$

which one may think of as the Lagrangian describing the dynamics of the self-interacting particle with "wavefunction" f, with mass m and coupling constant g. Vacuum polarization may induce new elements into the dynamics of the system which, let us suppose, may be incorporated into a new Lagrangian

$$L_{v.p.} = Z(d/dx)^2f^2(x) - m_1^2f^2(x) + g_1f^4(x)$$

The dynamics of the system including vacuum polarization is now given by the sum of L and $L_{v.p.}$. However this change in L can also be expressed alternatively, by absorbing it in redefinitions of f, m and g, as follows:

$$f \rightarrow f' = (1 + Z)^{1/2}f$$

$$m \rightarrow m' = (1 + Z)^{-1}(m + m_1)$$

$$g \rightarrow g' = (1 + Z)^{-2}(g + g_1)$$

It is easy to check that the original L with the new wavefunction f', new mass m' and new coupling constant g' satisfies

$$L[f',m',g'] = L[f,m,g] + L_{v.p.}$$

Thus the effect of vacuum polarization is included by a renormalization of L. The rules of the game allow the quantities Z, m_1 and g_1 to be infinite. Since only f', m' and g' are the quantities that can be directly or indirectly measured in the laboratory, it is sufficient that these be finite. In particular the original quantities f, m and g need not be finite. All that is needed to eliminate the infinities induced by vacuum polarization is for the original quantities to have corresponding infinities but with the opposite signs, so that the infinities are exactly cancelled in the new quantities.

It is easy to see that the renormalization programme outlined above can always be carried out if vacuum polarization does not introduce terms not contained in the original Lagrangian. Actually a much stronger statement can be made: renormalization is possible provided vacuum polarization introduces only a finite number of new terms not included the original Lagrangian. To see that this is indeed true one only has to recognize that all the new terms can be thought of as also being present in the original Lagrangian, but with zero coefficients.

Renormalizability is thus reduced to the issue of ascertaining that only a finite number of terms can appear in the Lagrangian. Usually this assurance is provided by a symmetry that is sufficiently restrictive. In gauge theories this job is precisely done by the gauge symmetry. However, the presence of a symmetry alone is not sufficient to guarantee renormalization, and gravity, built on the principle of invariance under general co-ordinate transformation, is unfortunately just such a case.

A quantum version of gravity also permits vacuum polarization of the type

GRAVITON → PARTICLE — ANTIPARTICLE PAIR → GRAVITON

but here, because of the short-distance property of the theory the infinities are so innumerable that renormalization breaks down. The innumerability of infinities in gravity is a direct consequence of the fact that Newton's gravitational constant G has dimension -2 (ie., inversely proportional to momentum squared), unlike the coupling constants in gauge theories which are just dimensionless constant numbers. The dimensionality of G assures that the Langrangian density for Einstein's action, $G^{-1}R$ where R is the curvature and has dimension 2, has the correct overall dimension, namely 4. The lowest level of vacuum polarization in quantum gravity induces new infinite terms to the Lagrangian density having the form

$$G^{-1}(G\partial_\mu\partial_\nu)T^{\mu\nu}$$

where ∂_μ is a derivative with respect to x^μ and $T^{\mu\nu}$ can be any dimension 2, rank-2 tensor such that $\partial_\mu\partial_\nu T^{\mu\nu}$ is general coordinate transformation invariant. Note that the presence of the two derivatives exactly cancels the dimension of G so that the new term has overall the required dimension 4. The next level of vacuum polarization will induce yet another set of terms having the form

$$G^{-1}(G^2\partial_\mu\partial_\nu\partial_\lambda\partial_\kappa)T^{\mu\nu\lambda\kappa}$$

where $T^{\mu\nu\lambda\kappa}$ is rank-4 tensor but otherwise satisfies the same criteria as $T^{\mu\nu}$. The nth-level vacuum polarization will induce new terms that can be schematically written as

$$G^{-1}(G\partial\partial)^n\, T^{(2n)}$$

where $T^{(2n)}$ is a rank-2n tensor. Since new tensors of higher ranks can always be constructed, the process of new terms appearing at each successive higher level is unending and therefore quantum gravity is unrenormalizable. Note how the dimensionality of G has been instrumental in allowing new tensors to come into play. Had G been dimensionless, like in gauge theories, then $(G\partial\partial)^n$ would have been replaced by $(G)^n$ in the last expression, and $T^{(2n)}$ by $T^{(0)}$, that is scalars of dimension 2, of which there are a finite number, the theory would have been renormalizable.

What is the relation between the unrenormalizability of gravity and its short distance behavior? In units of length[e], G is equal to $(10^{-33}\ \text{cm})^2$. This implies that variations of $T^{\mu\nu}$ at the length scale of 10^{-33} cm (the Planck length) are important and strongly coupled to the original Lagrangian density, since for such small scale variations

$$(G\partial_\mu\partial_\nu)T^{\mu\nu} \sim R.$$

In other words, because $T^{\mu\nu}$ (as well as the other $T^{(2n)}$'s) describe spacetime, the operator $(G\partial\partial)$ is sensitive to the structure of spacetime at the Planck scale.

SUPERSYMMETRY AND SUPERGRAVITY

The dynamical symmetries — gauge invariance and general coordinate transformation invariance — that have given us a deeper understanding of the fundamental laws impose very strict constraints on the force carriers. The bosons of gauge theories have one-to-one correspondence with the generators of gauge transformations and the graviton (of Einstein's gravity) is just the metric of Minkowski spacetime. On the other hand, these symmetries impose rather loose constraints on the fermions; it is sufficient that the fermions transform as representations of the gauge groups. This is not at all restrictive since every group has an infinite number of representations. The fermions have too much freedom and a guiding principle is needed to reduce it. Supersymmetry is such a principle: it assigns to each boson a supersymmetric fermionic partner. Supersymmetry has several other attractive features: (a) A supersymmetric theory with spacetime dependent supersymmetric transformations automatically contains gravity, such a theory is called supergravity. (b) If gauge theories are united with gravity, then as a result of finite corrections from vacuum polarization, massless or almost massless particles (a category to which all known elementary particles belong) will acquire masses of the Planck mass scale (10^{19} GeV) unless they are "protected" by a symmetry principle; supersymmetry can serve as such a principle. (c) Supersymmetric boson and fermion partners always make opposite-sign contributions to infinities induced by vacuum polarization. Indeed, hope was raised considerably when a

[e] Physicists find it convenient to convert a dimensionful quantity into powers of length units by attaching to it factors of $\hbar$ and c. Thus $G \rightarrow G\hbar/c^3 = 2.7 \times 10^{-66}\ \text{cm}^2$.

supersymmetric theory was found in which all vacuum polarization-infinities cancel exactly[8]. Does there exist a supergravity in which such a cancellation also occurs? This would give us a theory uniting gauge theories with a quantizable gravity. So far all searches for such a supergravity have failed.

STRINGS

If the uncontrollable infiniteness associated with vacuum polarization in quantum gravity is caused by the short distance property of the theory, one may try to avoid the difficulty by altering the short distance structure of the theory. A radical and so far quite successful approach is to assume that particles are not points moving in spacetime, but are strings. String theories were accidentally discovered in the late 1960's when attempts were made to explain a phenomenon, known as duality, in collisions among strongly interacting particles (such as nucleons and mesons). Roughly speaking, duality in this context describes the equivalence between the amplitudes, apart from trivial kinematic factors, for the two reactions

$$A + B \rightarrow C + D$$

$$A + \text{anti-}C \rightarrow \text{anti-}B + D.$$

where A, B, C and D stand for particles. The equivalence extends to reactions in which the positions of particles are permuted in other ways, provided a particle is changed to its anti-particle whenever it is brought across the arrow, and vice versa. Models for the strong interaction having this property, called dual models, were intensely studied in the mid-sixties. Around 1970 it was discovered that the quantum theory of a vibrating string[9] (in the normal sense of the word; mathematically it is a one-dimensional extended object) shares the same underlying algebraic structure with dual models, and therefore also has the property of duality. Viewed as a point field theory, string theory is a theory of spinless particles moving in a two-dimensional (one space, one time) spacetime. Since it is not possible to argue consistently that spacetime is two-dimensional, an alternative interpretation is needed.

The interpretation commonly adopted is the following: the (wavefunctions of the) spinless particles are viewed as spacetime coordinates which are however not points, but rather strings. Recall that when a point particle moves, it traces out a world-line. Classical mechanics follows from the action principle requiring that the world-line describes the geodesic, or the path of shortest distance between two given points. In complete analogy, the motion of a string sweeps out a world-sheet, and the dual model follows from the action principle requiring the world-sheet to be the two-dimensional surface having the smallest area (see Figure 5).

If a consistent theory is to be built from string coordinates, then the dynamical tensor variables associated with linear and rotational transformations of the coordinates must satisfy among themselves the Poincaré algebra set down by Dirac long ago. The consequence is surprising: the algebra can be satisfied only if spacetime is twenty-six dimensional! The early string theory also suffered from two serious maladies: it predicted a incorrect spectrum for the strongly interacting particles and admitted the existence of particles with negative masses, known as tachyons[f]. This plus the discovery in 1973 of quantum chromodynamics, which was very quickly demonstrated to be a far superior theory for the strong interaction than the string theory, expedited the almost complete abandonment of the latter by physicists soon afterwards, although by then it had already been pointed out that the spectrum of string theory contains particles that may be identified as gravitons, so that string theory could be viewed as a candidate for a unified theory, instead of a theory for the strong interaction. The observation was eventually instrumental in motivating the recent revival of the string theory, this time as a unified theory, after all other attempts at unification had failed.

The fact that string coordinates are extended objects means they can have excitations associated with vibrations and other possible contortions of strings. Although the mathematical articulation of the string theory is often very complex, the computation of the spectrum of these excitations is essentially that of a harmonic oscillator. Because the wavelength of these oscillations cannot be longer than the length of the string, which will most likely be of the order of the Planck length, even the lowest excitations will have energies that are far too high for any of the known particles. Thus the latter must be in the ground state, i.e. they must belong to the zero energy modes of the string. The physics of these zero energy modes, which behave like point particles, just like the normal modes of any oscillating system behave as point particles, is called the low energy limit of string theory. String theory is viewed as a unified theory including gravity because it already contains all the known particles — the spin-2 gravitons, the spin-1 vector bosons and the spin-1/2 fermions — in its low energy limit. This relegates all the point theories we have to the status of mere effective theories which can be exempt from rigorous requirements such as renormalizability, and renders irrelevant the fact that Einstein's gravity cannot be quantized and is unrenormalizable; it is sufficient that string theory can be quantized and is renormalizable.

SUPERSTRINGS

What happened to the tachyons that were present in the old string theories? These are removed from the unified string theories by the introduction of supersymmetry: corresponding to the spinless string coordinates obeying the usual bosonic commutation relations are now introduced spinor (spin-1/2) coordinates obeying fermionic anticommutation relations. Strings with both kinds of coordinates are called superstrings[10]. Supersymmetry ensures that the tachyons generated respectively by the bosonic and spinor strings exactly cancel. The condition that they dynamic variables associated with the superstring coordinates satisfy the generalized Poincaré algebra — super-Poincaré algebra — can only be satisfied in a ten dimensional spacetime, with one temporal and nine spatial dimensions[g].

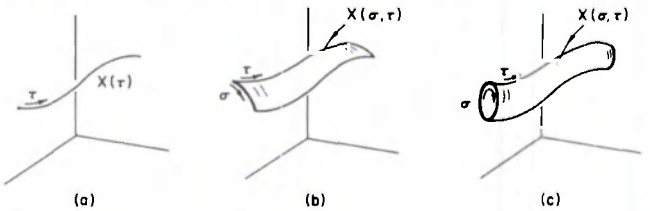

Fig. 5. (a) A world-line x(τ) formed by a point moving in spacetime can be parametrized by a single time-like variable τ. (b) A moving open string parametrized by σ sweeps out a two-parameter world-sheet X(σ,τ), which is treated as a spacetime coordinate in string theory. A consistent theory requires that X(σ,τ) be reparametrization invariant. (c) For closed strings the world-sheet coordinates are tubes. Note the different interpretations given to the Kaluza-tube in Figure 3b and the world-tube depicted here.

[f] From the Greek word tachy, meaning swift. It can be argued that if a particle has negative mass, then it must travel faster than the speed of light, hence the name.

[g] Spacetimes with more than one temporal dimensions do not obey causality. For in such spacetimes it will be possible to have a world-line whose projection on, say, two of the temporal dimensions form a closed curve, thus implying that an event taking place at one spatial point may occur before itself at another spatial point. However, the ancient Taoist who asserted that time ran in a circle might have known something that we don't.

What about renormalizability? The sometimes made pronouncement that superstring theories are finite and renormalizable is premature, even though there are many reasons to be optimistic: a) Unlike the point theory of gravity, which as explained earlier is manifestly unrenormalizable, there is not a dimensionful coupling constant in superstring and thus no a priori reason to believe that it should not be renormalizable. (b) On the contrary, if vacuum polarization preserves super-Poincaré invariance, then superstrings would be renormalizable because the set of allowed terms in the string action is restricted by the invariance. In this case vacuum polarization would not introduce new terms in the action, but would at most induce such changes in existing terms that could be absorbed by redefinition of the coordinates and other quantities. This is analogous to gauge theories for which preservation of gauge invariance guarantees renormalizability. However, it has not been proved that super-Poincaré invariance is in fact preserved by string vacuum polarization (see Figure 6). (c) So far only a limited number of calculations for the lowest order vacuum polarization effects in the various types of superstrings have been carried out, and results are either finite or infinite but renormalizable. (d) The requirement that vacuum polarization (in the lowest order) preserves super-Poincaré invariance imposes very restrictive symmetries on the zero energy modes of the superstring, symmetries which can nevertheless be satisfied.

In short, the renormalizability of superstrings is reduced to the issue of the preservation of super-Poincaré invariance, and so far there is no sign to suggest all is not well, although practicable techniques for calculating high order vacuum polarizations are still under development, and a proof of invariance to all orders is lacking.

Assuming for the moment that superstrings are renormalizable, what is the progress made in terms of uniqueness? In this regard superstring theory is an improvement over Kaluza-Klein theory. In Kaluza-Klein theory one begins with a set of pre-chosen (gauge) symmetries and then proceeds to find a space(time) with extra dimensions that possesses such symmetries. In the superstring approach super-Poincaré invariance uniquely determines the number of extra dimensions to be six, and the space thus determined must possess a symmetry large enough to accommodate all known particles. There is unfortunately still an infinite number of six-dimensional compact spaces. Even the number of such spaces with the right symmetries is likely to be very large, and we still do not have a principle to help us pick out either the phenomenologically required symmetries or the right space.

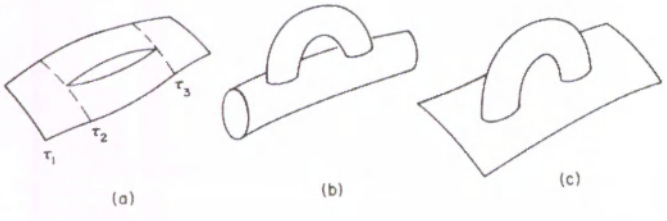

Fig. 6. (a) **The open string propagates from τ_1 to τ_2, at which "time" it splits into two strings, which then fuse to form a single string once again at τ_3. Because string is an extended object, it admits vacuum polarization even in the absence of a string field theory. (b) A similar process for a closed-string is represented by a tube with a handle.** (c) **A world-sheet for an open-string can also grow a handle. String theories admitting only closed-strings are simpler; they involve only tubes with handles. Theories with open-strings must also have closed-strings; they involve surfaces with slits, handles and other more complicated growths and formations.**

There is also the question: if spacetime is indeed ten dimensional, then what causes six of the dimensions to be compact and the other four noncompact? Indeed, why not five and five, eight and two, all compact or all noncompact? We would like to believe that questions such as these have rational answers, that the six-and-four combination is the solution of some master equation derived from the same set of principles that give rise to string theories, or is the consequence of the dynamics of string theories. At the same time the possibility cannot be ruled out that the six-and-four combination is not a unique solution of these principles or dynamics, that different solutions exist, each associated with its own probability, thus giving rise to different possible universes with different laws of physics, and that we are by chance living in just one of the possible universes.

STRING FIELD THEORY

Even though string theory contains gravity and gauge theories in its low energy limit, a consensus for a string theory principle, in the sense of the equivalence principle for gravity and the principle of local gauge invariance for gauge theories, has not yet been established. The string theory we have talked about so far is just the quantum mechanics of strings, not the quantum field theory of strings. The difference between the two is analogous to the difference between Bohr's patchwork quantum theory of the 1920's and the elegant quantum electrodynamics that eventually replaced it in the late 1940's.

String field theory is inherently more complex than point field theory. Central to the latter are fields $\phi(x)$, which (mathematically) create and/or annihilate particles, that are functions of spacetime coordinates x. In comparison string fields $\Phi[X]$, which create and annihilate strings, are functions of spacetime coordinates $X(\sigma,\tau)$, which are themselves functions of the two parameters σ and τ (this makes the string fields functionals, that is, functions of functions). String field theories including interaction among strings have been found only in special cases that do not reveal the full symmetry and geometric structure of the theory and permit only perturbation expansions. They are therefore not very useful.

A very different approach, geometrically based but also giving only a perturbation expansion, is characterized by a sum over all world-sheet surfaces that are topologically distinct.[11] For closed strings this boils down to summing all surfaces with different numbers of handles (see Figure 6), equivalent to summing all levels of vacuum polarization in point field theories.

The search for an underlying string principle is currently the most intensely pursued topic in theoretical particle physics. Many proponents think success is imminent. Whether this will come to pass only time will tell. What has already occurred with the rise of string theory are a vastly expanded horizon for theoretical physics, a new awareness of the links between physics and several branches of mathematics, hitherto thought to be purely abstract, and a much deeper understanding of the laws of physics.

TESTING UNIFIED THEORY

When a new theory is put forward, especially a theory as radical as string theory where the existence of extra dimensions is proclaimed, it is essential that it be experimentally verified. Einstein's general relativity was tested against the precession of the perihelion of Mercury and the bending of light; quantum electrodynamics against the magnetic moment of the electron; the unified electroweak theory against the detection of neutral weak forces and the observation of the W and Z bosons; quantum chromodynamics against the discovery of charm and beauty resonances and the observation of gluon jets. What are the testable predictions of string

theories? The answer is embarrassing but true: none has yet been identified. Critics are not slow to point out that string theories appear to have tremendous "postdictive" power, but nothing much else.

The lack of testable predictions from string theories is essentially an unavoidable consequence of a theory uniting gravity with the other forces, and originates from the fact the dimensionful gravitational (Newton's) constant G is equal to the square of the Planck length. This immediately implies that, unless there is some very subtle and as yet undiscovered mechanisms at work, spacetime (in the unified theory) is nontrivially structured at the scale of (at most a few orders of magnitude greater than) the Planck length. Equivalently, it implies that spacetime has nonvanishing curvature at the Planck scale. This of course could not happen to the Minkowski spacetime that we know, for otherwise the resultant gravitational force would be unimaginably stronger than what it is. Therefore it is the extra-dimensional spaces that are nontrivial at the Planck scale. In any case, any irrefutable verification of string theory would invariably involve some measurement at the Planck scale, which of course is far smaller than anything we can ever hope to measure directly.

One idea crucial to string theories could be tested in the near future, however. If superstring is to be at all realistic, then there should exist for each known elementary particle an as yet undiscovered superpartner. The new generation of accelerators at Fermilab near Chicago, SLAC in Stanford and CERN in Geneva should give us a verdict on the existence of these particles within the next few years. As well, the case for unification, though not necessarily for string theory, will receive a tremendous boost if evidence of proton decay is eventually established at one of the many underground detectors around the world.

If neither supersymmetry nor proton decay can be confirmed long after the new accelerators have come into operation, and this should bring us well into the next decade, then our hopes for a unified theory would unavoidably be severely dampened. Even then, it would be difficult to imagine theorists to be complacent with the status quo, to accept that unification of the electromagnetic and weak forces is just an accident, that the internal quantum numbers of elementary particles do not have a geometric basis, that gravity, alone among all the fundamental forces, is exempt from the uncertainty principle.

If, by chance, both supersymmetry and proton decay are to be confirmed, then work must still continue to assure that other crucial predictions of the unified theory are not contradicted. Since none of the tests would likely be a direct test at the Planck scale, this verification process could go on for a long time.

REFERENCES

1. S. Weinberg, Phys. Rev. Lett. **19** (1967) 1264; A. Salam, Proc. of Eighth Nobel Symposium, ed. N. Svartholm (Almqvist & Wilsells, 1968) 367.
2. F.J. Hassert et al., Phys. Lett. **46B** (1973) 121.
3. C. Rubbia, Rev.Mod. Physl. **57** (1985) 699.
4. Nature, **326** (1987) 11.
5. Th. Kaluza, Sitzungober. Preuss, Akad. Wiss. Berlin, Math. Phys. **K1** (1921) 1966.
6. O. Klein, Z. Phys. **37** (1926) 895.
7. E. Witten, in "Unified String Theories", eds. M. Green and D. Gross (World Scientific, 1986) p. 400.
8. L. Brink, O. Lindgren and B.E. Nilsson, Nucl. Phys. B212(1983)401; S. Mandelstam, Nucl. Phys. **B213** (1983) 149.
9. P. Goddard, J. Goldstone, C. Rebbi and C.B. Thorn, Nucl. Phys. B56 (1973) 109; J. Scherk, Rev. Mod. Phys. **47** (1975) 123.
10. J.H. Schwarz, Phys. Rep. 89 (1982) 223; M.B. Green and J.H. Schwarz, Phys. Lett. **149B** (1984) 117; **151B** (1985) 21; M.B. Green, in "Unified String Theories", loc. cit. p. 294.
11. A.M. Polyakov, Phys. Lett. **103B** (1981) 207, 211.