Physics 1978

PETER LEONIDOVITCH KAPITZA

for his basic inventions and discoveries in the area of low-temperature physics

ARNO A PENZIAS and ROBERT W WILSON

for their discovery of cosmic microwave background radiation

THE NOBEL PRIZE FOR PHYSICS

Speech by Professor LAMEK HULTHÉN of the Royal Academy of Sciences. Translation from the Swedish text

Your Majesties, Your Royal Highnesses, Ladies and Gentlemen,

This year's prize is shared between Peter Leonidovitj Kapitza, Moscow, "for his basic inventions and discoveries in the area of low-temperature physics" and Arno A. Penzias and Robert W. Wilson, Holmdel, New Jersey, USA, "for their discovery of cosmic microwave background radiation".

By low temperatures we mean temperatures just above the absolute zero, −273°C, where all heat motion ceases and no gases can exist. It is handy to count degrees from this zero point: "degrees Kelvin" (after the British physicist Lord Kelvin) E.g. 3 K (K = Kelvin) means the same as -270°C.

Seventy years ago the Dutch physicist Kamerlingh-Onnes succeeded in liquefying helium, starting a development that revealed many new and unexpected phenomena. In 19 11 he discovered *superconductivity* in mercury: the electric resistance disappeared completely at about 4 K. 1913 Kamerlingh-Onnes received the Nobel prize in physics for his discoveries, and his laboratory in Leiden ranked for many years as the Mekka of low temperature physics, to which also many Swedish scholars went on pilgrimage.

In the late twenties the Leiden workers got a worthy competitor in the young Russian Kapitza, then working with Rutherford in Cambridge, England. His achievements made such an impression that a special institute was created for him: the Royal Society Mond Laboratory (named after the donor Mond), where he stayed until 1934. Foremost among his works from this period stands an ingenious device for liquefying helium in large quantities-a pre-requisite for the great progress made in low temperature physics during the last quarter-century.

Back in his native country Kapitza had to build up a new institute from scratch. Nevertheless, in 1938 he surprised the physics community by the discovery of the *superfluidity* of helium, implying that the internal friction (viscosity) of the fluid disappears below 2.2 K (the so-called lambda-point of helium). The same discovery was made independently by Allen and Misener at the Mond Laboratory. Later Kapitza has pursued these investigations in a brilliant way, at the same time guiding and inspiring younger collaborators, among whom we remember the late Lev Landau, recipient of the physics prize 1962 "for his pioneering theories for condensed matter, especially liquid helium". Among Kapitza's accomplishments we should also mention the method he developed for producing very strong magnetic fields.

Kapitza stands out as one of the greatest experimenters of our time, in his domain the uncontested pioneer, leader and master.

We now move from the Institute of Physical Problems, Moscow, to Bell Telephone Laboratories, Holmdel, New Jersey, USA. Here Karl Jansky, in the beginning of the thirties, built a large movable aerial to investigate sources of radio noise and discovered that some of the noise was due to radio waves coming from the Milky Way. This was the beginning of radio astronomy that has taken such an astounding development after the second World War-as an illustration let me recall the discovery of the pulsars, honoured with the physics prize 1974.

In the early 1960ies a station was set up in Holmdel to communicate with the satellites Echo and Telstar. The equipment, including a steerable horn antenna, made it a very sensitive receiver for microwaves, i.e. radio waves of a few cm wavelength. Later radio astronomers Arno Penzias and Robert Wilson got the chance to adapt the instrument for observing radio noise e.g. from the Milky Way. They chose a wave length c. 7 cm where the cosmic contribution was supposed to be insignificant. The task of eliminating various sources of errors and noise turned out to be very difficult and time-consuming, but by and by it became clear that they had found a background radiation, equally strong in all directions, independent of time of the day and the year, so it could not come from the sun or our Galaxy. The strength of the radiation corresponded to what technicians call an antenna temperature of 3 K.

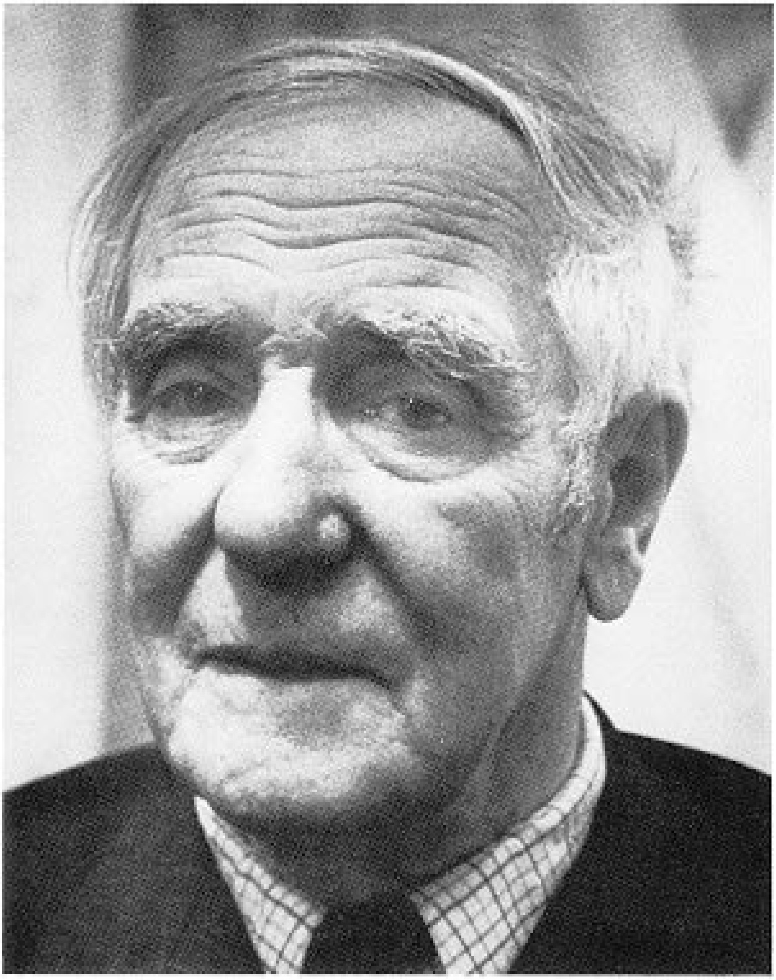
Continued investigations have confirmed that this background radiation varies with wave length in the way prescribed by wellknown laws for a space, kept at the temperature 3 K. Our Italian colleagues call it “la luce fredda”-the cold light.

But where does the cold light come from? A possible explanation was given by Princeton physicists Dicke, Peebles, Roll and Wilkinson and published together with the report of Penzias and Wilson. It leans on a cosmological theory, developed about 30 years ago by the Russian born physicist George Gamow and his collaborators Alpher and Herman. Starting from the fact that the universe is now expanding uniformly, they concluded that it must have been very compact about 15 billion years ago and ventured to assume that the universe was born in a huge explosion-the “Big Bang”. The temperature must then have been fabulous: 10 billion degrees, perhaps more. At such temperatures lighter chemical elements can be formed from existing elementary particles, and a tremendous amount of radiation of all wave lengths is released. In the ensuing expansion of the universe, the temperature of the radiation rapidly goes down. Alpher and Herman estimated that this radiation would still be left with a temperature around 5 K. At that time, however, it was considered out of the question, that such a radiation would ever be possible to observe. For this and other reasons the predictions were forgotten.

Have Penzias and Wilson discovered “the cold light from the birth of the universe” ? It is possible-this much is certain that their exceptional perse-

verance and skill in the experiments led them to a discovery, after which cosmology is a science, open to verification by experiment and observation.

Piotr Kapitsa, Arno Penzias, Robert Wilson, In accordance with our tradition I have given a brief account in Swedish of the achievements, for which you share this year's Nobel prize in Physics. It is my privilege and pleasure to congratulate you on behalf of the Royal Swedish Academy of Sciences and ask you to receive your prizes from the hands of His Majesty the King!



PJOTR LEONIDOVICH KAPITZA

Pjotr Leonidovich Kapitza was born in Kronstadt, near Leningrad, on the 9th July 1894, son of Leonid Petrovich Kapitza, military engineer, and Olga Ieronimovna née Stebnitskaia, working in high education and folklore research.

Kapitza began his scientific career in A.F. Ioffe's section of the Electromechanics Department of the Petrograd Polytechnical Institute, completing his studies in 1918. Here, jointly with N.N. Semenov, he proposed a method for determining the magnetic moment of an atom interacting with an inhomogeneous magnetic field. This method was later used in the celebrated Stern-Gerlach experiments.

At the suggestion of A.F. Ioffe in 1921 Kapitza came to the Cavendish Laboratory to work with Rutherford. In 1923 he made the first experiment in which a cloud chamber was placed in a strong magnetic field, and observed the bending of alfa-particle paths. In 1924 he developed methods for obtaining very strong magnetic fields and produced fields up to 320 kilogauss in a volume of 2 cm^3. In 1928 he discovered the linear dependence of resistivity on magnetic field for various metals placed in very strong magnetic fields. In his last years in Cambridge Kapitza turned to low temperature research. He began with a critical analysis of the methods that existed at the time for obtaining low temperatures and developed a new and original apparatus for the liquefaction of helium based on the adiabatic principle (1934).

Kapitza was a Clerk Maxwell Student of Cambridge University (1923-1926), Assistant Director of Magnetic Research at Cavendish Laboratory (1924-1932), Messel Research Professor of the Royal Society (1930-1934), Director of the Royal Society Mond Laboratory (1930-1934). With R.H. Fowler he was the founder editor of the International Series of Monographs on Physics (Oxford, Clarendon Press).

In 1934 he returned to Moscow where he organized the Institute for Physical Problems at which he continued his research on strong magnetic fields, low temperature physics and cryogenics.

In 1939 he developed a new method for liquefaction of air with a low-pressure cycle using a special high-efficiency expansion turbine. In low temperature physics, Kapitza began a series of experiments to study the properties of liquid helium that led to discovery of the superfluidity of helium in 1937 and in a series of papers investigated this new state of matter.

During the World War II Kapitza was engaged in applied research on

the production and use of oxygen that was produced using his low pressure expansion turbines, and organized and headed the Department of Oxygen Industry attached to the USSR Council of Ministers.

Late in the 1940's Kapitza turned his attention to a totally new range of physical problems. He invented high power microwave generators - planotron and nigotron (1950-1955) and discovered a new kind of continuous high pressure plasma discharge with electron temperatures over a million K.

Kapitza is director of the Institute for Physical Problems. Since 1957 he is a member of the Presidium of the USSR Academy of Sciences. He was one of the founders of the Moscow Physico-Technical Institute (MFTI), and is now head of the department of low temperature physics and cryogenics of MFTI and chairman of the Coordination Council of this teaching Institute. He is the editor-in-chief of the Journal of Experimental and Theoretical Physics and member of the Soviet National Committee of the Pugwash movement of scientists for peace and disarmament.

He was married in 1927 to Anna Alekseevna Krylova, daughter of Academician A.N. Krylov. They have two sons, Sergei and Andrei.

Honorary degrees

D.Phys.-Math.Sc., USSR Academy of Sciences, 1928; DSc., Algiers University, 1944; Sorbonne, 1945; D.Ph., Oslo University, 1946; D.Sc., Jagellonian University, 1964; Technische Universitat Dresden, 1964; Charles University, 1965; Columbia University, 1969; Wroclaw Technical University, 1972; Delhi University, 1972; Université de Lausanne, 1973; D.Ph., Turku University, 1977.

Honorary memberships

Member of the USSR Academy of Sciences, 1939 (corresponding member - 1929); Fellow of the Royal Society, London, 1929; French Physical Society, 1931; Institute of Physics, England, 1934; International Academy of Astronautics, 1964; Honorary Member of the Moscow Society of Naturalists, 1935; the Institute of Metals, England, 1943; the Franklin Institute, 1944; Trinity College Cambridge, 1925; New York Academy of Sciences, 1946; Indian Academy of Sciences, 1947; the Royal Irish Academy, 1948; National Institute of Sciences of India, 1957; German Academy of Naturalists "Leopoldina", 1958; International Academy of the History of Science, 197 1; Tata Institute of Fundamental Research, Bombay, India, 1977. Foreign Member of Royal Danish Academy of Sciences and Letters, 1946; National Academy of Sciences, USA, 1946; Indian National Sciences Academy, 1956; Polish Academy of Sciences, 1962; Royal Swedish Academy of Sciences, 1966; American Academy of Arts and Sciences, 1968; Royal Netherlands Academy of Sciences, 1969; Serbian Academy of Sciences and Arts, 197 1; Finnish Academy of Arts and Sciences, 1974. Honorary Fellow of Churchill College Cambridge, 1974.

Awards

Medal of the Liege University, 1934; Faraday Medal of the Institute of Electrical Engineers, 1942; Franklin Medal of the Franklin Institute, 1944; Sir Devaprasad Sarbadhikary Gold Medal of the Calcutta University, 1955; Kothenius Gold Medal of the German Academy of Naturalists "Leopoldina", 1959; Frederic Joliot-Curie Silver Medal of the World Peace Committee, 1959; Lomonosov Gold Medal of the USSR Academy of Sciences, 1959; Great Gold Medal of the USSR Exhibition of Economic Achievements, 1962; Medal for Merits in Science and to Mankind of the Czechoslovak Academy of Sciences, 1964; International Niels Bohr Medal of Dansk Ingeniwvørening, 1964; Rutherford Medal of the Institute of Physics and

Physical Society, England, 1966; Golden Kamerlingh Onnes Medal of the Netherlands Society of Refrigeration, 1968; Copernir Memorial Medal of the Polish Academy of Sciences, 1974.

USSR State Prize - 1941, 1943; Simon Memorial Award of the Institute of Physics and Physical Society, England, 1973; Rutherford Memorial Lecture, Royal Society of London; Bernal Memorial Lecture, Royal Society of London, 1976.

Order of Lenin - 1943, 1944, 1945, 1964, 1971, 1974; Hero of Socialist Labour, 1945, 1974; Order of the Red Banner of Labour, 1954;
Order of the Jugoslav Banner with Ribbon, 1967.

Publications

Collected Papers of P.L. Kapitza, 3 vol., Pergamon Press, Oxford, 1964- 1967:
High Power Microwave Electronics, Pergamon Press, 1964.
Experiment. Theory. Practice. “Nauka”, Moscow, 1977.
Le livre du probleme de physique, CEDIC. Paris, 1977.

Professor Kapitza died in 1984.

PLASMA AND THE CONTROLLED THERMONUCLEAR REACTION

Nobel Lecture, 8 December, 1978

bY
P. L. KAPITZA
Institute for Physical Problems of the Academy of Sciences, Moscow, USSR

The choice of the theme for my Nobel lecture presents some difficulty for me. Usually the lecture is connected with work recognized by the prize. In my case the prize was awarded for work in low temperature physics, at temperatures of liquid helium, a few degrees above absolute zero. It so happened that I left this field some 30 years ago, although at the Institute under my directorship low temperature research is still being done. Personally I am now studying plasma phenomena at those very high temperatures that are necessary for the thermonuclear reaction to take place. This research has led to interesting results and has opened new possibilities, and I think that as a subject for the lecture this is of more interest than my past low temperature work. For it is said, "les extremes se touchent".

It is also well recognized that at present the controlled thermonuclear reaction is the process for producing energy that can effectively resolve the approaching global energy crisis, resulting from the depletion of fossil fuels used now as our principal energy source.

It is also well known that intensive research on fusion is done in many countries and is connected with fundamental studies of high temperature plasmas. The very possibility of fusion is well beyond doubt, for it takes place in the explosion of hydrogen bombs. We also have a detailed theoretical understanding of nuclear fusion reactions that is in agreement with experiments. But in spite of the great effort and large sums spent up to now, it is impossible to conduct the process of fusion as to make it a useful source of energy. This certainly is a cause for some bewilderment.

One could expect that during the decades of experimental and theoretical plasma work in studying the conditions for fusion we would have reached a sufficient understanding of the various facts that hinder us from setting up a controlled thermonuclear reaction. It could be expected that we should have discovered and revealed the main difficulties that bar our progress. In this lecture I hope to clarify, what are these difficulties, and what are the chances that these difficulties will be resolved. I will also try to explain the divergence of opinions of different scientists on the practical possibilities for obtaining useful thermonuclear energy.

Before embarking into this subject I would like to speak on the practical importance of obtaining energy from nuclear sources.

The reality of the approaching global energy crisis is connected with the unavoidable lack of raw materials: gas, oil, coal. This is now generally appreciated. It is also known that the GNP (gross national product) that

determines the wellbeing of people is proportional to the expenditure of energy. Energy resources depletion will inevitably lead to general impoverishment.

Two possible ways out of the approaching energy crisis are discussed. The first, maybe the more attractive, is to extensively use the inexhaustable sources of energy: hydroelectric power, the power of wind, solar energy, geothermal energy. The second way is to use nuclear energy discovered by man less than a hundred years ago. At present heavy element fission power is already cheaper than energy from some nonexhaustable sources.

It is well known that the main fuel in these reactors is uranium. It has been shown that as used at present there is enough uranium for only a hundred years. If in the future uranium will be more fully used in breeder reactors, it will last 50 times longer, for a few thousand years. Many consider that uranium dissolved in sea water may also be efficiently used for cheap energy production. Thus it may seem that the processes now used in modern nuclear reactions may resolve the approaching energy crisis. But there are important reasons against using uranium as a source of energy. These arguments are mainly connected with security.

In the first place, the use of uranium leads to accumulation of longlived radioactive wastes and the problem of safely storing a growing amount of these waste materials. This is a problem that at present has not been definitely solved.

In the second place, in a large energy-producing nuclear power plant a vast amount of radioactive material is accumulated, so that in a hypothetical accident, the dispersion of this material might lead to a catastrophe comparable in scale to that of Hiroshima.

I think that eventually modern technology will resolve these two dangers. But there is still a third hazard, even more grave. This is the danger, that the construction of great numbers of nuclear power stations will inevitably lead to such a huge amount of radioactive material disseminated around the world, that an efficient control on its proper uses will be practically impossible. In the long run not only a small country, but even a wealthy man or a large industrial organization will be able to build its own atomic bomb. There is at present no secret of the bomb. The necessary amount of plutonium, especially if breeders are to be widely built, will be readily available. Thus recently in India a small bomb was built and exploded. With the present system of international organizations there is nobody with sufficient authority that could execute the necessary control of the peaceful use of uranium as a source of energy. Moreover, it is not clear now how such an organization could be set up. This is the main reason why it is most important to obtain energy by the third way, through the process of thermonuclear fusion.

It is common knowledge, that this process will not lead to generation of large amounts of radioactive wastes and thus to a dangerous accumulation of radioactive material, and mainly it does not open any chances for a feasible nuclear explosion. This is the main reason why the solution of the

scientific and technical problems involved in controlled thermo-nuclear fusion is considered of prime importance by many physicists.

The conditions for the thermonuclear reaction for energy production are well known and firmly established. There are two reactions of importance: the D + D and D + T process. The first one is the reaction between two nuclei of deuterium. The second occurs in the interaction of deuterium with tritium. In both cases fast neutrons are emitted, whose energy may be used. As a small amount of deuterium is present in water and is easy to extract, an abundant source of fuel is available. Free tritium practically does not exist in nature and tritium has to be produced, as it is usually done, through the interaction of neutrons with lithium.

The thermonuclear reaction is to take place in a high temperature plasma. So as to practically use the energy of neutrons the production of energy has to be greater than the power used to sustain the high plasma temperature. Thus the energy, obtained from the neutrons, has to be much greater than the bremsstrahlung radiation of the electron gas in the plasma. Calculations show that for useful energy production for the D + D reaction the necessary ion temperature is 10 times greater than in the case of the D + T reaction. Although the D + T reaction works at a lower temperature, it is hampered by the necessity to burn lithium, whose amount in nature is limited. Moreover, it seems that the use of lithium greatly complicates the design of the reactor. Calculations show that for obtaining useful energy the temperature of ions in a plasma for the D + D reaction should be about 10^9K and for the D + T reaction about 10^8K.

From research in plasma and nuclear physics it is thus known that for practical energy generation purposes, the technical problem of realising a controlled thermonuclear reaction is reduced to obtaining a plasma ion temperature at least 10^8K with a density 10^{13} - 10^{14} c m^{-3}. It is obvious that the containment of a plasma in this state by any ordinary vessel cannot be done, as there is no material that can withstand the necessary high temperature.

A number of methods for the containment of plasma and its thermal isolation have been suggested.

The most original and promising method was the “Tokamak” proposed in the Soviet Union and under development for more than a decade /(1) page 15/. The principle of its operation can be seen from the design shown in fig. 1. The plasma is confined by a magnetic field, generated in a toroidal solenoid. The plasma has the form of a ring of a radius R and a cross section of the radius a, placed in the coil. The plasma has a pressure of a few atmospheres. As it expands in the magnetic field, currents are excited that retard this expansion. The plasma is surrounded by a vacuum insulation, This is necessary to sustain the sufficiently high temperatures at which thermonuclear reactions take place. It is obvious that this method of confinement is limited in time. Calculations show that due to the low thermal capacity of the plasma, the energy for initial plasma heating, even in cases when the plasma exists for a few seconds, will be small as com-

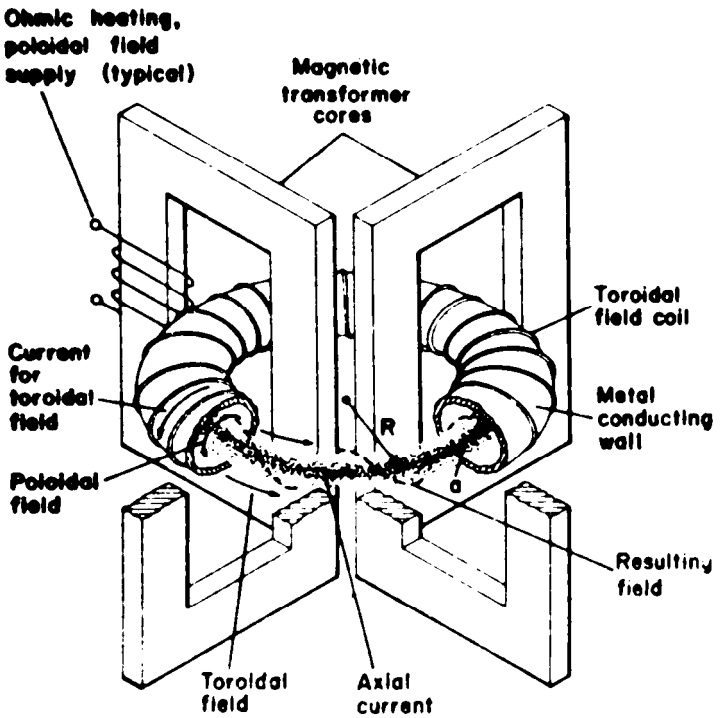


Fig. 1. Main features of a Tokamak

pared with the thermonuclear energy. Thus a reactor of this type may effectively work only in a pulsed mode. The Tokamak is started as a betatron: by discharging condensers through the coils of the transformer yoke. In practice plasma confinement by this method is not simple. In the first place there are difficulties in stabilising the plasma ring in the magnetic field. With the growth of the cross section radius a and moreover of the torus radius R, the ring loses its proper form and becomes unstable. This difficulty may be circumvented by choosing the appropriate ratio of R to a, and by properly designing the magnetic field, although at present the time for plasma confinement is only a small fraction of a second. It is assumed that with scaling the Tokamak up this time will be proportional to the square of the size of the machine.

But the main difficulty is due to reasons, not fully appreciated in the beginning. For the thermonuclear reaction one has to heat the D and T ions. The main difficulty in passing heat to them is due to the fact that the plasma is heated by an electric field. In this case all the energy is transferred to the electrons and is only slowly transferred to the ions because of their large mass as compared to the mass of the electrons. At higher temperatures this heat transfer gets even less efficient. In the Tokamak the plasma is heated by the betatron current induced through the condensor discharge. Thus, all the energy for heating the plasma is confined to

the electrons and is transferred to the ions by collisions. To heat the ions to the desired temperatures the necessary time A t is much longer than the time during which we may maintain heating of the plasma by an electric current. The calculations that are usually done are complicated, as attempts were made to do them as exactly as possible, and so they lose in clarity. It is easy to estimate the lower time limit in which the ion heating may be made by the following simple formula /(2) page 24 expression 14/

$$\Delta t > -2.5 \cdot 10^{2} \frac{f}{\Lambda} \frac{T_e^{3/2}}{n} ln\,(1-\frac{T_i}{T_e})$$

We assume that during heating the plasma density n,

$$n = \frac{7.3 \cdot 10^{21} P}{T_e},$$

the pressure P (atm) and the electron temperature T_e are constant.

The coefficient f is equal to the ratio of the ion mass to that of the proton, A is the well known logarithmic factor /2 in (4)/, T_i-the ion temperature. For modern Tokamaks, operating with the D + T reaction and at plasma temperatures $T_i = 5.10^8$ and $n = 3.10^{13} cm^{-3}$ (with an initial electron temperature $T_e = 10^9 K$) the time necessary to heat the ions to nuclear process temperatures is more than 22 seconds, at least two orders of magnitude more than confinement times in the modern Tokamaks. The plasma confinement time may be made greater only by building a larger machine, as it seems that the time A t is proportional to the square of the size. From this formula it also follows that the time A t for the D + D reaction is greater by another two orders of magnitude and then A t-2. 10^3 sec. The difficulties with the time for heating the ions is now fully recognized, although one cannot see how to shorten this time and how a Tokamak may work if, before the plasma ions have been heated, all the betatron energy from the condensers will be fully radiated by the electrons. That is why in the current Tokamak projects extraneous energy sources are envisaged that are greater than the energy of the betatron process, used only for initially firing the plasma.

Extra energy must be transferred to the ions by a more efficient way than Coulomb scattering of electrons on ions. There are two possible processes for this. The first /(1) page 20/ already used, consists in injecting into the plasma ring atoms of deuterium or tritium, already accelerated to temperatures necessary for the thermonuclear reaction. The second process of heating is through exciting radial Alfvèn magnetoacoustic waves in the external magnetic field by the circulating high frequency current. It is known /(3)/ that the energy dissipated by magnetoacoustic waves is directly passed into the ions and the transmitted power is sufficient to heat the ions and sustain their temperature for a sufficiently long time. Thus the problem of heating the ions may be solved, although the mode of operation of Tokamak will be more complicated than at first suggested. The design of the Tokamak becomes more complicated and its efficiency diminishes.

In all nuclear reactors the power generated is proportional to the volume of the active zone and the losses are proportional to its surface. Therefore the efficiency of nuclear reactors is greater for larger sizes and there exists a critical size for a nuclear reactor after which it may generate useful power. The practically necessary dimension is determined not by scientists but by the engineers who design the machine in general with proper choice of all the auxiliaries and the technology necessary for energy production. The following development is to a great measure determined by the talent and inventive ability of the design engineers. That is why the critical size of the Tokamak will be mainly determined by the proposed designs. Personally I think that the existing published design solutions lead us to a critical size for Tokamaks that make them unfeasible. But certainly life does show that the ingenuity of man has no limits and therefore one cannot be sure that a practically useful critcial size of Tokamaks may not be reached in the future.

One must note that although the main difficulty for obtaining a thermonuclear reaction in Tokamaks is the heating of deuterium and tritium ions, there is a difficulty of still another kind that does not have a well defined solution. In a Tokamak, for example, the plasma attracts and absorbs impurities extracted from the walls of the container. These impurities greatly lower the reaction rate. The plasma emits neutral atoms that hit and erode the wall. Moreover, the extraction of energy from neutrons also complicates the design of the Tokamak and leads to a larger critical size. Will we be able to bring the critical dimension of the Tokamak to a practically possible size? Even if it will eventually happen, of course we have no means to say when it will happen. Now we may only state that there are no theoretical reasons why in a Tokamak controlled thermonuclear reactions are not feasible, but the possiblity to release useful energy is as yet beyond the scale of our current practice.

Among other approaches to controlled thermonuclear fusion serious considerations should be given to pulsed methods without magnetic confinement /(1) page 33/. The idea is to heat a D + T pellet about 1 mm in diameter in a short time so as it will not have time to fly apart. For this very high pressures are necessary, that ensure intensive heat transfer between ions and electrons. It is assumed that in this way the thermonuclear reaction in a D + T pellet may fully take place. For this it is necessary to have a very powerful source of focussed laser light that should heat the pellet from all sides simultaneously in about a nanosecond. This heating is a complicated process, but using modern computers one may calculate all necessary conditions. If we illuminate a pellet by a well focussed laser beam, this may lead to a surplus of thermonuclear energy. But when one considers this process in detail, it is not clear how one can possibly resolve the technical and engineering difficulties. How, for instance, can one ensure uniform and simultaneous illumination and how can one usefully exploit the neutron energy?

In this case one may also say that the basic theoretical idea is sound, but the consequent engineering development with current technology is beyond our reach. Once again one cannot completely exclude a solution to this problem, although the design for laser implosion seems to me even less probable than the pulsed magnetic methods like the Tokamak.

The third approach to a thermonuclear reactor is based on continuously heating the plasma. Up to now this method has been developed only at our Institute. Our work was described 9 years ago /(4). Since then this type of reactor has been studied in detail, and now we see the main difficulties which we have to encounter. I will describe here in general terms what are the problems demanding a scientific solution.

As distinct from Tokamaks and the laser implosion method for producing conditions for the thermonuclear process, our method was not specially invented, but while developing a high power CW microwave generator accidentally we discovered a hot plasma phenomenon. We constructed an efficient microwave generator operating at 20 cm wave length with a power of a few hundred kW. This generator was called the "Nigotron" and its principles are described in (5) where full details of its construction with operating characteristics are given. In the process of its development beginning in 1950, during tests of our early model, high power microwave radiation was passed through a quartz sphere, filled with helium at 10 cm Hg pressure. We observed a luminiscent discharge with well defined boundaries. The phenomenon was observed only for a few seconds, as the quartz sphere in one place melted through.

These observations led us to the suggestion that the ball lightening may be due to high frequency waves, produced by a thunderstorm cloud after the conventional lightening discharge. Thus the necessary energy is produced for sustaining the extensive luminosity, observed in a ball lightening. This hypothesis was published in 1955 (‘7). After some years we were in a position to resume our experiments. In March 1958 in a spherical resonator filled with helium at atmospheric pressure under resonance conditions with intense H_{ol} oscillations we obtained a free gas discharge, oval in form. This discharge was formed in the region of the maximum of the electric field and slowly moved following the circular lines of force.

We started to study this type of discharges where the plasma was not in direct contact with the walls of the resonator. We assume that this plasma may be at a high temperature. During a number of years we studied this interesting phenomenon in various gases and at different pressures, up to some tens of atmospheres at different power levels, reaching tens of kW. We also studied the effect of a magnetic field reaching 2,5 T in our experiments. This work is described in detail (4). A sketch of our setup is shown in fig. 2.

The plasma discharge has a cord-like form 10 cm long, equal to half the wavelength. Intense microwave oscillations E_{ol} are excited in a cylindrical resonator (1). The cord of the discharge is situated at the maximum of the electric field and its stability along the longitudinal axis was due to the high

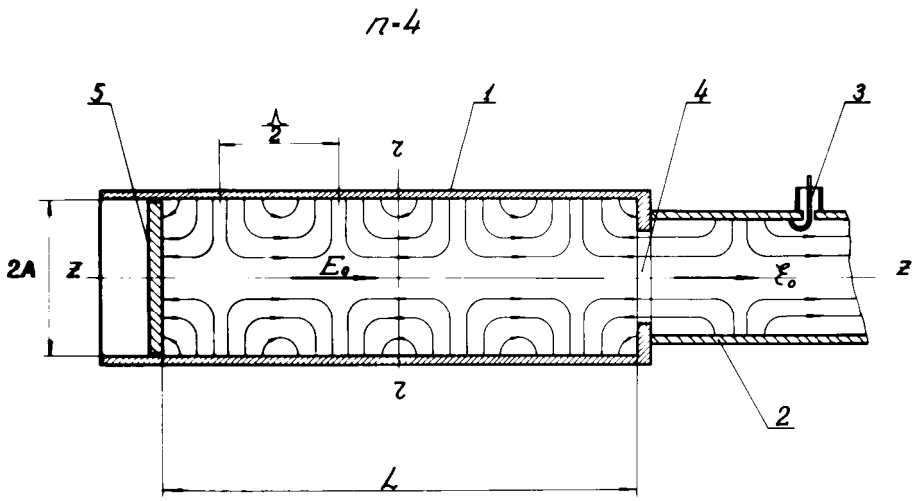


Fig. 2. Structure of the HF field in a resonator for E_{01} oscillations

frequency electric field. In a radial direction the stability was provided by rotating the gas. The discharge in hydrogen or deuterium was of great interest. At low powers the discharge did not have a well defined boundary and its luminosity was diffuse. At higher power the luminosity was greater and the diameter of the discharge increased. Inside the discharge a well defined filamentary cord-like nucleus was observed. In our initial experiments the power dissipated in the discharge was up to 15 kW and the pressure reached 25 atm. The higher the pressure, the more stable was the discharge with a well defined shape. A photograph of the discharge is shown in fig. 3. By measuring the conductivity of the plasma and by using passive and active spectral diagnostics we could firmly establish that the central part of the discharge had a very high temperature - more than a million K. So at the boundary of the plasma cord in the space of a few millimeters we had a discontinuity of temperature more than a million K. This meant that at its surface there was a layer of very high heat isolation. At first some doubt was expressed about the existence of such a layer. Various methods of plasma diagnostics were used, but they all and always confirmed the high temperature - more than a million K. Later we found out how it is possible to explain the physical nature of this temperature jump. It is easy to show that at these high temperatures electrons scattered at the boundary and freely diffusing into the surrounding gas will carry away a power of hundreds of kW. The lack of such a thermal flux may be explained by assuming the existence of electrons reflected without losses at the boundary of a double layer. The occurrence of a similar phenomenon is well known as such a layer exists in hot plasmas surrounded by dielectric walls, say, of glass or ceramics.

It is well known that in these conditions even at high pressures the electrons may have a temperature of many ten thousands of K and not

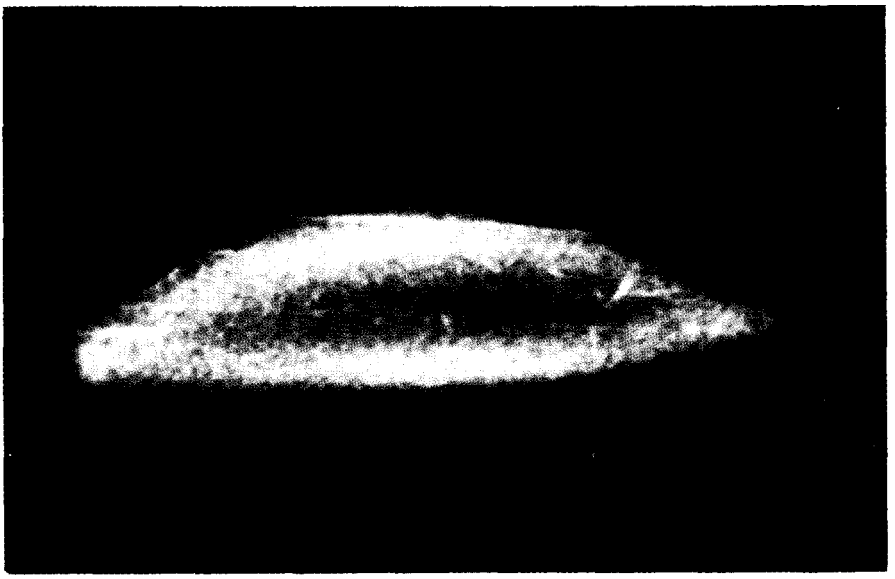


Fig. 3. Photograph of a cord discharge in deuterium with an admixture of 5 % argon at high power P = 14, 7 kW and high pressure p = 3,32 atm. Length of the discharge ~ 10 cm. The left edge of the discharge is blocked by the window. Oscillations of E_{01} type (1969)

markedly heat the walls. This phenomenon is well explained by the existence of a double layer on the dielectric surface. The mechanism leading to its formation is simple. When the electron hits the surface, due to its greater mobility it penetrates the dielectric to a greater depth than the ions and leads to the formation of an electric double layer, the electric field of which is so directed that it elastically reflects the hot electrons. The low electron heat conductance at the surface of plasmas is widely used in gas discharge lamps and the method of plasma heat insulation was first suggested by Langmuir. We assume that at a sufficiently high pressure a similar mechanism of heat insulation may take place in our hot plasma. The existance of a double layer in the plasma on the boundary of the cord discharge as a discontinuity in density was experimentally observed by us. This mechanism for a temperature discontinuity may obviously exist only if the ion temperature is much lower than the electron temperature and not much above the temperature at which the plasma is noticibly ionised. But this is only necessary at the boundary of the discharge. In the central part of the discharge the ion temperature may reach high values. As we will see further, the difference in temperatures inside the core and at the surface is determined by the value of the thermal flux and the heat conductivity of the ion gas. Usually the heat conductivity is high, but in a strong magnetic field the transverse heat conductivity may become very small. Thus we may expect that in a strong magnetic field the ion temperature in the core will not differ from the electron temperature and may be sufficiently high to obtain in a deuterium or tritium plasma a thermonucle-

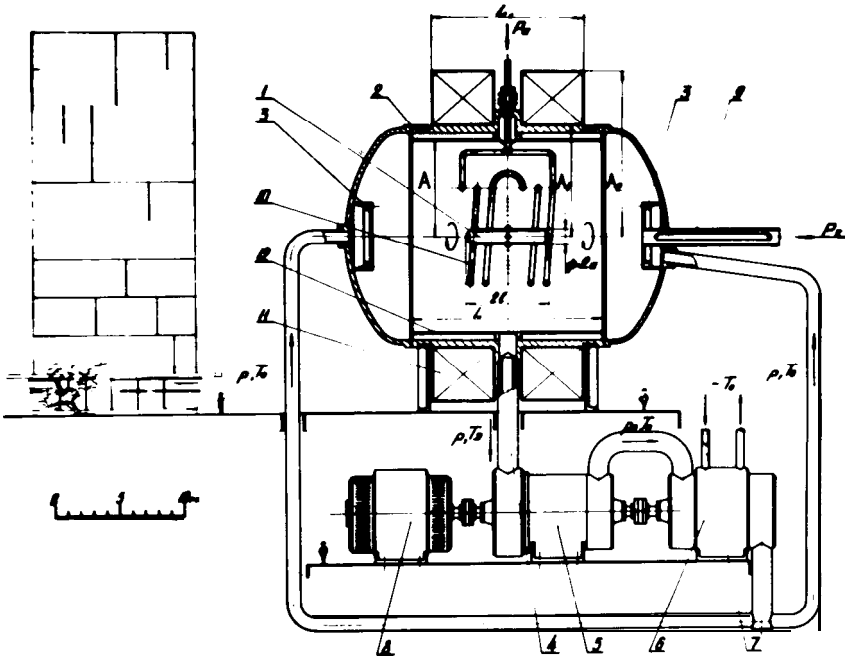


Fig. 4. Drawing of the construction of a thermonuclear reactor operating on a closed cycle. I - cord discharge, 2 - cylindrical container of the reactor, 3 - inclined nozzles, 4 - pipe connecting the container of the reactor with the gas turbine, 5 - gas turbine, 6 - isothermal compressor, 7 - cooling water, 8 - generator, 9 - coaxial waveguide, 10 - coil for the alternating magnetic field, 11 - solenoid, 12 - copper wall of the resonator, L - length of the resonator, L_1 - length of the solenoid, Pa - power of magnetoacoustic oscillations, P_r - high-frequency power, A - radius of the resonator, A_1 - internal radius of the winding, A_2 - external radius of the winding, 2 1 - length of the cord discharge, 2a - diameter of the cord discharge, h - distance between the wall of the container and the resonator.

ar reaction. This is the basis for designing a thermonuclear reactor to produce useful energy, and this has been worked out (8). The general outlay and the description of the reactor are shown in fig. 4.

The cord discharge (1) takes place in a confining vessel and resonator (2). The deuterium pressure is 30 atm, the magnetic field 1 T, produced by an ordinary solenoid. The design shows how the neutron energy is used. The gas heated by the neutrons passes through a gas turbine (5) where it adiabatically expands. Next it passes through a turbocompressor (6) and is isothermally compressed. The excess power is consumed in the generator (8). The cord discharge is heated by a high frequency field as it is done in cylindrical resonators (see fig. 2). The difference is in the coil surrounding the discharge and used to excite magnetoacoustic waves so as to raise the plasma ion temperature/(4) page 1003/. This design and pertinent calculations were published in 1970/(8) page 200/so as to demonstrate the expected parameters of our thermonuclear reactor, working with our plasma cord.

During the past time we have considerably increased our understanding of the processes in the plasma. We have mainly improved the microwave

diagnostics and it is now possible to measure with 5% accuracy the radial density distribution, its dependence on the magnetic field, pressure and supplied microwave power. The necessary stability conditions have been established. All this has allowed us to raise the microwave power by many times and in this way increase the electron temperature up to 50 million K. If we could establish temperature equilibrium between the electrons and ions in this case even without the extra heating of the plasma by magneto-acoustic oscillations, we could have reached the D + T reaction. The design of the reactor is simpler and its size is smaller. In this case the thermonuclear reactor would be not only easier to build but the neutron energy is easier to convert to mechanical power. Thus we escape the main difficulties on the way to building pulsed thermonuclear reactors.

But still we have also some unresolved difficulties which merit most serious consideration, because they might make the whole problem unsolvable. The main difficulty is the following. Now we can obtain in our installation a high frequency discharge at a pressure of 25 atmoshperes and continuously maintain the electrons at a temperature of 50 million K, and going to a greater size of our discharge even more. At present the size is limited only by the power conveyed to it. Thus we have permanently an electron gas with a record high temperature, even higher than the electron temperature inside the Sun. The main problem is to heat the ions to the same temperature, for although the electron gas interacts with the ions in the entire volume of the discharge, it is not easy to raise this temperature in such a way.

The temperature equalisation proceeds in two steps. In the first step the energy is passed from the electrons to the ions. This is simply due to the collisions of electrons with ions, and in this case it is obvious that the heat transfer will be proportional to the volume. The next stage is the transfer of energy from the ion gas to the surrounding media. This flux will be proportional to the surface of the plasma cord. At a given thermal conductivity of the ion gas the temperature will increase for larger sizes of the cross section o the plasma cord. Thus at a certain heat conductivity there will be a critical size for the diameter of the plasma cord, when the ion temperature will reach a value close to that of the electrons and the required D + D or D + T reaction can take place. If we know the heat conductivity of the plasma, then it is easy to calculate the critical dimension. If, for example, we make this calculation for ordinary ion plasma in the absence of a magnetic field, when the heat conductivity is determined by the mean free path, we will find that the plasma must have an unrealizably large size of many km. One can lower this cross section only by decreasing the heat conductivity of the ion gas by placing it in a magnetic field as it is done in the reactor shown in fig. 4. The heat conductivity of an ion gas in a magnetic field is markedly decreased and it is determined not by the mean free path but by the radius of Larmor orbits the size of which is inversely proportional to the magnetic field. The thermal conductivity of ion gas in a magnetic field is easy to calculate.

It is thus seen that the critical diameter of the cord is inversely proportional to the magnetic field and at a field of a few tesla the diameter of the cord to get thermal neutrons will be 5-10 cm, that can readily be provided for. For this we need a plasma installation considerably greater than the one in which we at present study the nature of the electron gas in the plasma. In the conditions of our laboratory this installation is quite feasible and is now under construction.

It may be shown that the thermonuclear reactor we have described makes it possible to obtain conditions not only for the D + T reaction but also for D + D, if it were not for yet another factor that could eventually make the whole process unfeasible.

We determined the heat conductivity of the ion gas by considering the mean free path of the ion, assuming it to be equal to the Larmor orbit radius, having not taken into account the effect of convection fluxes of heat in a gas. It is well known that even in ordinary gases the convection heat transfer is much larger then the heat conduction due to molecular collisions. It is also known that unfortunately it is virtually impossible to calculate theoretically the heat transfer by convective currents even for the simple case of random turbulent motion in an ordinary gas. In this case we usually can, by dimensional considerations, estimate the thermal conductivity in a similar case and then generalize it for a special case, determining the necessary coefficients empirically. In the case of plasma the process depends on many more parameters and the problem of determining the convectional thermal conductivity is even more complicated than in an ordinary gas. But theoretically we may estimate, which factors have most influence on the rate of convection. To sustain convection one must supply energy. In a gas this energy is drawn from the kinetic energy of flow and leads to loss of heat.

In a quiescent plasma there is no such source of energy. But in an ionized plasma there may be another source of energy that will excite convection. This source is connected with temperature gradients and some of the thermal energy flux could produce convection. Quantitatively this process is described by internal stresses and was first studied by Maxwell (9). Maxwell had shown that internal stresses are proportional to the square of viscosity and derivative of the temperature gradient. In an ordinary gas they are so small that up to now they have not yet been experimentally observed. This is because the viscosity, which is proportional to the mean free path, at normal pressures equals to $\sim 10^{-5}$ cm and so at low temperature gradients, the stresses are small.

In the plasma the mean free path of electrons and ions is of the order of cm and the temperature gradients are high. In this case the internal stresses following Maxwell's formula are 10 orders of magnitude greater than in a gas and we may expect both convection currents and turbulence. The presence of a magnetic field certainly can have effect on this phenomenon, and with additional effect of an electric field on convection it makes even a rough theoretical approach to estimating the magnitude of convec-

tion very unreliable. In this case there is only one alternative: to study these processes experimentally and this is what we are now doing.

In any case convectional thermal conductivity will lower the heating of ions and will lead to a greater critical cross section for the thermonuclear plasma cord. Correspondingly the size of the reactors for useful energy production will be greater.

If this size will be out of our practical reach, then we should consider methods to decrease convectional heat transfer. This may be done by creating on the boundary of the plasma a layer without turbulence, as it happens in fluids where we have the Prandtl boundary layer. This possibility has been theoretically considered /(4) page 10021.

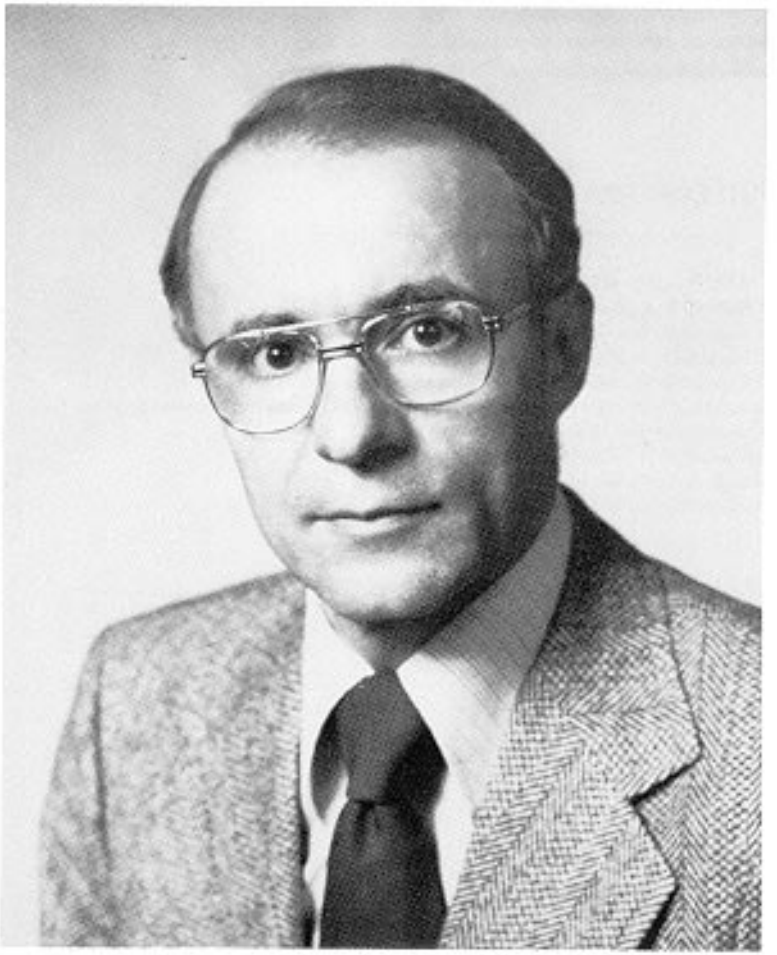
In conclusion we may say that the pulsed method used in Tokamaks can now be fully worked out theoretically, but the construction of a thermonuclear reactor, based on this method, leads to a large and complicated machine. On the other hand, our thermonuclear reactor is simple in construction, but its practical means of realisation and size depend on convection heat transfer processes, that cannot be treated purely theoretically.

The main attraction in scientific work is that it leads to problems, the solution of which it is impossible to foresee, and that is why for scientists research on controlled thermonuclear reactions is so fascinating.

LITERATURE

1. Ribe, F. I.., Rev. of Modern Physics, 47, 7, 1975.
2. Kapitza, P. L., JETP Lett., 22 (1), 9, 1975.
3. Kapitza, P. I., Piraevskii L. P., Sov. Phys. -JETP, 40 (4), 701, 1975.
4. Kapitza, P. L., Soviet Phys. -JETP, 30, (6), 973, 1970.
5. Kapitza, P. L., High-Power Microwave Electronics, Pergamon Press, Oxford, 1964.
6. Капица, П. Л., Филимонов, С. И., Капица, С. П., Сборник «Электроника больших мощностой, № 6, «Наука», стр. 7, 1969.
7. Kapitza, P. L., Collected papers, vol. 2, 776, Pergamon Press, Oxford, 1965.
8. Kapitza, P. L., Sov. Phys. -JETP, 31, (2), 199, 1970.
9. Maxwell, J. C., Phil. Trans. R. S., 170, 231, 1879.

The English translation from the Russian original text is authorized by the laureate.



Arno A Penzias

ARNO A. PENZIAS

I was born in Munich, Germany, in 1933. I spent the first six years of my life comfortably, as an adored child in a closely-knit middle-class family. Even when my family was rounded up for deportation to Poland it didn't occur to me that anything could happen to us. All I remember is a long train trip and scrambling up and down three tiers of narrow beds attached to the walls of a very large room. After some days of back and forth we were returned to Munich. All the grown-ups were happy and relieved, but I began to realize that there were bad things that my parents couldn't completely control, something to do with being Jewish. I learned that everything would be fine if we could only get to "America".

One night, shortly after my sixth birthday, my parents put their two boys on a train for England; we each had a suitcase with our initials painted on it and a bag of candy. They told me to be sure and take care of my younger brother. I remember telling him, "jetzt sind wir allein" as the train pulled out.

My mother received her exit permit a few weeks before the war broke out andjoined us in England. My father had arrived in England almost as soon as the two of us, but we didn't see him because he was interned in a camp for alien men. The only other noteworthy event in the six or so months we spent in England awaiting passage to America occurred when I found that I could read my school books.

We sailed for America toward the end of December 1939 on the Cunard liner Georgic using tickets that my father had foresightedly bought in Germany a year and half earlier. The ship provided party hats and balloons for the Christmas and New Year's parties, as well as lots of lifeboat drills. The grey three-inch gun on the aft deck was a great attraction for us boys.

We arrived in New York in January of 1940. My brother and I started school and my parents looked for work. Soon we became "supers" (superintendents of an apartment building). Our basement apartment was rent free and it meant that our family would have a much-needed second income without my mother having to leave us alone at home. As we got older and things got better, we left our "super" job and my mother got a sewing job in a coat factory; my father's increasing wood-working skills helped him land ajob in the carpentry shop of the Metropolitan Museum of Art. As the pressures on him eased, he later found time to hold office in a fraternal insurance company as well as to serve as the president of the local organization of his labor union.

It was taken for granted that I would go to college, studying science, presumably chemistry, the only science we knew much about. "College" meant City College of New York, a municipally supported institution then beginning its second century of moving the children of New York's immigrant poor into the American middle class. I discovered physics in my freshman year and switched my "major" from chemical engineering. Graduation, marriage and two years in the U.S. Army Signal Corps, saw me applying to Columbia University in the Fall of 1956. My army experience helped me get a research assistantship in the Columbia Radiation Laboratory, then heavily involved in microwave physics, under I. I. Rabi, P. Kusch and C. H. Townes. After a painful, but largely successful struggle with courses and qualifying exams, I began my thesis work under Professor Townes. I was given the task of building a maser amplifier in a radio-astronomy experiment of my choosing; the equipment-building went better than the observations.

In 1961, with my thesis complete, I went in search of a temporary job at Bell Laboratories, Holmdel, New Jersey. Their unique facilities made it an ideal place to finish the observations I had begun during my thesis work. "Why not take a permanent job? You can always quit," was the advice of Rudi Kompfner, then Director of the Radio Research Laboratory. I took his advice, and have remained here ever since.

Since the large horn antenna I had planned to use for radio-astronomy was still engaged in the ECHO satellite project for which it was originally constructed, I looked for something interesting to do with a smaller fixed antenna. The project I hit upon was a search for line emission from the then still undetected interstellar OH molecule. While the first detection of this molecule was made by another group, I learned quite a bit from the experience. In order to make some reasonable estimate of the excitation of the molecule, I adopted the formalism outlined by George Field in his study of atomic hydrogen. To make sure that I had it right, I took my calculation to him for checking. One of the factors in the calculation was the radiation temperature of space at the line wavelength, 18-cm. I used 2 K, a somewhat larger value than he had used earlier, because I knew that at least two measurements at Bell Laboratories had indications of a sky noise temperature in excess of this amount, and because I had noticed in Hertzberg's Diatomic Molecule book that interstellar CN was known to be excited to this temperature. The results of the calculation were used and forgotten. It was not until Dr. Field reminded me of them in December of 1966, that I had any recollection of the earlier connection. So much for the straight-line view of the progress of science!

The successful detection of OH at MIT made me look for a larger antenna. At the invitation of A. E. Lilley, I took key parts of my equipment to the Harvard College Observatory and spent several months participating in various OH observations. In the meantime, the horn antenna was pressed into service for another satellite project. A new Bell System satellite, TELSTAR, was due to be launched in 'mid-1962. While the primary

earth station at Andover, Maine, was more-or-less on schedule, it was feared that the European partners in the project would not be ready at launch time, leaving Andover with no one to talk to. As it turned out, fitting the Holmdel horn with a 7-cm receiver for TELSTAR proved unnecessary; the Europeans were ready at launch time. This left the Holmdel horn and its beautiful new ultra low-noise 7-cm traveling wave maser available for radio astronomy. This stroke of good fortune came at just the right moment. A second radio astronomer, Robert Wilson, came from Caltech on a job interview, was hired, and set to work early in 1963.

In putting our radio astronomy receiving system together we were anxious to make sure that the quality of the components we added were worthy of the superb properties of the horn antenna and maser that we had been given. We began a series of radio astronomical observations. They were selected to make the best use of the careful calibration and extreme sensitivity of our system. Among these projects was a measurement of the radiation intensity from our galaxy at high latitudes which resulted in the discovery of the cosmic microwave background radiation, described in Wilson's lecture.

When our 7-cm program was accomplished, we converted the antenna to 2 1 -cm observations including another microwave background measurement as well as galactic and intergalactic atomic hydrogen studies. As time went on, the amount of front line work that we could do became increasingly restricted. Much larger radio telescopes existed and they were being fitted with low-noise parametric amplifiers whose sensitivity began to approach that of our maser system. As a result we began looking for other things to do. An investigation of the cosmic abundance of deuterium was clearly an important problem. However nature had put the deuterium atomic line in an all but inaccessible portion of the long wavelength radio spectrum. I remember saying, to Bob Dicke, something to the effect that I didn't relish giving three years of my professional life to the measurement of the atomic deuterium line. He immediately replied, "Finding deuterium is worth three years". Fortunately, a better approach to the measurement of deuterium in space soon became available to me.

Up through the late 1960's the portion of the radio spectrum shortward of l-cm wavelength was not yet available for line radio astronomy owing to equipment limitations. At Bell Laboratories, however, many of the key components required for such work had been developed for communications research purposes. With Keith Jefferts, a Bell Labs atomic physicist, Wilson and I assembled a millimeter-wave receiver which we carried to a precision radio telescope built by the National Radio Astronomy Observatory at Kitt Peak, Arizona, early in 1970. This new technique enabled us to discover and study a number of interstellar molecular species. Millimeter-wave spectral studies have proved to be a particularly fruitful area for radio astronomy, and are the subject of active and growing interest, involving a large number of scientists around the world. The most personally satisfying portion of this work for me was the discovery in 1973 of a

deuterated molecular species, DCN. Subsequent investigations enabled us to trace the distribution of deuterium in the galaxy. This work provided us with evidence for the cosmological origin of this important substance, which earned the nickname "Arno's white whale" during this period.

From the first, I made it my business to engage in the communications work at Bell Labs in addition to my astronomical research. It seemed only reasonable to contribute to the pool of technology from which I was drawing. Similarly, Bell Labs has always been a contributor to, as well as a user of, the store of basic knowledge, as evidenced by their hiring of a radio astronomer in the first place.

As time went on, the applied portion of my efforts included administrative responsibilities. In 1972 I became the Head of the Radio Physics Research Department upon the retirement of A. B. Crawford, the brilliant engineer who built the horn antenna Wilson and I used in our discovery. In 1976, 1 became the Director of the Radio Research Laboratory, an organization of some sixty people engaged in a wide variety of research activities principally related to the understanding of radio and its communication applications.

Early in 1979, my managerial responsibilities increased once again when I was asked to assume responsibility for Bell Labs' Communications Sciences Research Division. While I continued the personal research which traced the effects of nuclear processing in the Galaxy through the study of interstellar isotopes, pressure from other interests curtailed my entry into a new area - the nature and distribution of molecular clouds in interstellar space. Instead, I barely managed to introduce this subject to two of my graduate students who explored it in their PhD theses.

Then, toward the end of 1981, an unexpected event imposed an abrupt end to my career as a research scientist, when AT&T and the US Department of Justice decided to settle their anti-trust suit by breaking up the Bell System. In the process, I received yet another promotion - this time to Vice-President of Research - at a moment when two-thirds of the traditional research funding base moved off with the newly-divested local telephone companies.

As a result, I found myself facing several issues at once: What sort of research organization did the new AT&T require? How to create this new organization without destroying the world's premier industrial research laboratory in the process? Would the people in this large and tradition bound organization accept and support the changes needed to adapt to new economic and technological imperatives? Needless to say, such matters kept me quite busy.

In retrospect, the research organization which emerged from the decade following the Bell System's breakup deploys a far richer set of capabilities than its predecessor. In particular, our work features a growing software component, even as we strive to improve our hardware capabilities in areas such as lightwave and electronics. The marketplace upheaval brought forth

by increased competition has helped speed the pace of technological revolution, and forced change upon the institutions all industrialized national, Bell Labs included. While change is rarely comfortable, I am happy to say that we not only survived but also grew in the process.

Except for two or three papers on interstellar isotopes, my tenure as Bell Labs' Vice-President of Research brought my personal research in astrophysics to an end. In its place, I have developed an interest in the principles which underlie the creation and effective use of technology in our society, and eventually found time to write a book on the subject *Ideas and Information,* published by W. W. Norton in 1989. In essence, the book depicts computers as a wonderful tool for human beings but a dreadful role model. In other words, if you don't want to be replaced by a machine, don't act like one. The warm reception this book received in the US, and the ten other countries which published it in various translations has given me much satisfaction.

I have also been a visiting member of the Astrophysical Sciences Department at Princeton University from 1972 to 1982. My occasional lecturing and research supervision were more than amply repaid by stimulating professional and personal relationships with faculty members and students.

Finally, most important of all is the love and support of my family, my wife Anne, our children and grandchildren.

BSc., City College of New York, 1954
M.A., Columbia University, 1958
Ph. D., Columbia University, 1962

Docteur Honoris Causa, Paris Observatory, 1976
Henry Draper Medal, National Academy of Sciences, 1977
Herschel Medal, Royal Astronomical Society, 1977

Member, National Academy of Sciences
Fellow, American Academy of Arts and Sciences
Fellow, American Physical Society
Member, International Astronomical Union
Member, International Union of Radio Science.

THE ORIGIN OF ELEMENTS

Nobel Lecture, 8 December, 1978

by
ARNO A. PENZIAS
Bell Laboratories, Holmdel, N. J. USA

Throughout most of recorded history, matter was thought to be composed of various combinations of four basic elements; earth, air, fire and water. Modern science has replaced this list with a considerably longer one; the known chemical elements now number well over one hundred. Most of these, the oxygen we breathe, the iron in our blood, the uranium in our reactors, were formed during the fiery lifetimes and explosive deaths of stars in the heavens around us. A few of the elements were formed before the stars even existed, during the birth of the universe itself.

The story of how the modern understanding of the origin of the chemical elements was acquired is the subject of this review. A good place to begin is with Lavoisier who, in 1789, published the first scientific list of the elements. Five of the twenty or so elements in Lavoisier's list were due to the work of Carl Wilhelm Scheele of Köping. (He was rewarded with a pension by the same Academy to whom the present talk is adressed, more than a century before Alfred Nobel entrusted another task of scientific recognition to it.) Toward the end of the last century the systematic compilation of the elements into Mendeleev's periodic table carried with it the seeds of hope for a systematic understanding of the nature of the elements and how they came to be.

The full scientific understanding of the origin of the elements requires a description of their build-up from their common component parts (e.g., protons and neutrons) under conditions known to exist, or to have existed, in some accessible place. Thus, the quest for this understanding began with nuclear physics. Once plausible build-up processes were identified and the conditions they required were determined, the search for appropriate sites for the nuclear reactions followed. Although this search was begun in earnest in the nineteen thirties, it was only toward the end of the nineteen sixties that the full outlines of a satisfactory theoretical framework emerged. In the broad outlines of the relevant scientific thought during this period one can discern an ebb and flow between two views. In the first, the elements were thought to have been made in the stars of our galaxy and thrust back out into space to provide the raw material for, among other things, new suns, planets and the rock beneath our feet. In the second view, a hot soup of nuclear particles was supposed to have been cooked into the existing' elements before the stars were formed. This pre-stellar state was generally associated with an early hot condensed stage of the expanding universe.

Historically, the first quantitative formulations of element build-up were

attempted in the nineteen thirties; they were found to require conditions then thought to be unavailable in stars. As a consequence, attention turned in the 1940's to consideration of a pre-stellar state as the site of element formation. This effort was not successful in achieving its stated goal, and in the 1950's interest again turned to element formation in stars. By then the existence of a wide range of stellar conditions which had been excluded in earlier views had become accepted. Finally, the 1960's saw a reawakened interest in the idea of a pre-stellar state at the same time that decisive observational support was given to the "Big Bang" universe by the discovery of cosmic microwave background radiation and its identification as the relict radiation of the initial fireball.

Given the benefit of hindsight, it is clear that the process of understanding was severely impeded by limitations imposed by the narrow range of temperature and pressure then thought to be available for the process of nuclear build-up in stars. The theory of stellar interiors based upon classical thermodynamics (Eddington, 1926) seemed able to explain the state of the then known stars in terms of conditions not vastly different from those in our sun. The much higher temperatures and pressures suggested by the nuclear physics of element formation were thought to be possible only under conditions of irreversible collapse (i.e. the theory lacked mechanisms for withstanding the tremendous gravitational forces involved); hence no material produced under those conditions could have found its way back into the interstellar medium and ordinary stars. The arguments and mechanisms required to depict the formation of heavy elements and their ejection into space are subtle ones. In describing them, S. Chandrasekhar wrote, ". . one must have faith in drawing the consequences of the existence of the white dwarf limit. But that faith was lacking in the thirties and forties for reasons set out in my (to be published) article 'Why are the Stars as they are?." Thus, our story of a forty-year-long journey begins with the absence of sufficient faith.

The nuclear physics picture of element formation in an astrophysical setting was the subject of von Weizsäcker's "Über Elementumwandlungen im Innern der Sterne" (1937, 1938). (Interested readers can find a guide to earlier literature in Alpher and Herman's 1950 review.) The central feature of von Weizsäcker's work is a "build-up hypothesis" of neutrons and intervening β-decays; the direct build-up from protons would be blocked by the Coulomb repulsion of the positively charged nuclei of the heavier elements. Quantitative predictions that follow from this hypothesis can be obtained from the general features of empirical abundance-stability data through use of thermodynamic equilibrium relations like those used in the study of chemical reactions.

Consider the reversible exothermic reaction of two elements A and B combining to form a stable compound AB with an energy of formation ΔE, i. e.,

$$A+B \rightarrow AB+\Delta E. \qquad (1)$$

Using square brackets to indicate concentration, we can compute relative abundances at thermal equilibrium from the relation

$$\frac{[A]x[B]}{[AB]} \propto exp\ (-\Delta E/kT). \tag{2}$$

where k is Boltzmann's constant.

The stable isotopes of the lighter elements have approximately equal numbers of neutrons and protons (fig. 1). The sequential addition of neutrons to a nucleus, ^{16}O say, results in heavier isotopes of the same element, ^{17}O and then ^{18}O in this case, until the imbalance of neutrons and protons is large enough to make the nucleus unstable. (^{19}O β - decays to ^{19}F in ~ 29 seconds.) A measure of the stability of an isotope is the increment in binding energy due to the last particle added. In the case of ^{17}O, for example, we have for this increment,

$$\Delta E(17) = [M(16)+M(n)-M(17)]c^2, \tag{3}$$

where M(16), M(n) and M(17) are the masses of ^{16}O, a neutron and ^{17}O, respectively, and c^2 is the square of the speed of light. In our example, the mass of ^{17}O is 17.004533 A.M.U., that of the neutron is 1.008986 and that of ^{16}O is 16.00000. Substituting in eqn (3) we find the binding energy increment to be .004453 A.M.U. or 6.7×10^{-6} ergs. We can get some idea of the temperatures involved in the addition of a neutron to ^{16}O from the use of relation (2). Because of the exponential nature of this relation, we can

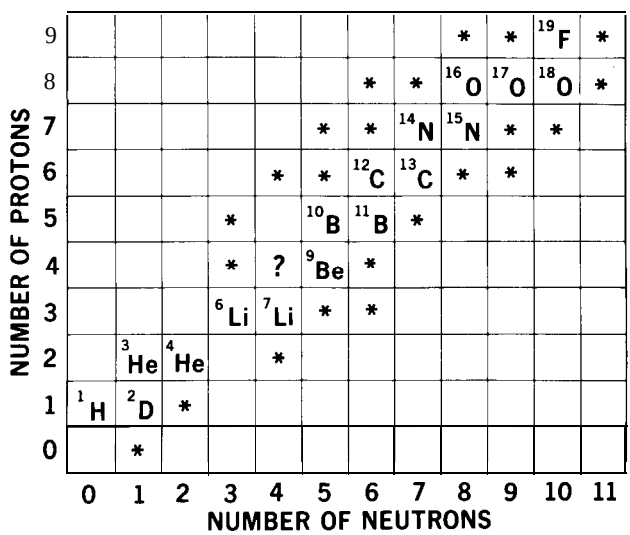


Fig. 1 ***The Elements, Hydrogen Through Flourine.*** The stable nuclei are plotted as a function of the number of protons and neutrons they contain. Radioactive combinations are indicated by an asterisk, an empty box indicates that the corresponding combination of protons and neutrons doesn't exist. (Note that both ***mass-5*** boxes are empty.) The question mark indicates ^{8}Be; it can exist under special conditions as a metastable combination of two ^{4}He nuclei, thus providing the key stepping-stone in the transformation of three ^{4}He's into ^{12}C.

expect $\triangle E$ and kT to be of comparable magnitude for a wide range of relative isotopic abundances. Thus, from the approximation,

$$\triangle E \approx kT$$

we find that 6.7x10ergs corresponds to a temperature of 5×10^{10}K.

Following earlier workers, von Weizsäcker applied the above relations to the relative abundance of the isotopes of a given element having three stable isotopes, (^{16}O, ^{17}O and ^{18}O for example) in a state of equilibrium established by thermal contact with a bath of neutrons at temperature T. If [^{16}O], [^{17}O], [^{18}O] and [n] are the concentrations of the two oxygen nuclei and the neutrons respectively, we may use the relations (2) and (3) to write

$$\frac{[^{16}O][n]}{[^{17}O]} \propto exp\ (\Delta E(17)/kT)$$

as well as

$$\frac{[^{17}O][n]}{[^{18}O]} \propto exp\ (\Delta E(18)/kT)$$

Thus the relative abundances of the three isotopes yield a pair of expressions involving the neutron density and temperature which permit the separate determination of these two quantities from the oxygen abundance data alone. (The abundances of several hundred stable nuclei -fig. 2 - had been determined from terrestrial samples supplemented by stellar spectra and meteorites.)

Using this three-isotope method, Chandrasekhar and Henrich (1942) obtained thermal equilibrium neutron densities and temperatures for five elements. Not surprisingly, in view of previous work, each element required a different temperature and neutron density. While the range of the temperature values was relatively small, between $2.9 \times 10^{9}$for neon and $12.9 \times 10^{9}$for silicon, the neutron densities ranged from ~ 10^{31}c $m^{-3}$for silicon to - 10^{19}c $m^{-3}$for sulphur, some twelve orders of magnitude! The high values of the temperatures and pressures derived as well as their lack of element-to-element consistency shows the shortcomings of this thermal equilibrium picture of stellar element formation.

Another problem with this neutron build-up picture was the simultaneous requirement of very rapid neutron capture in the formation of elements such as uranium and thorium, and very slow neutron capture for the formation of others. The “slow” elements require the capture sequence of neutrons to be slow enough to permit intervening p-decays, while others require *rapid* sequential neutron capture in order to permit their formation from a series of short-lived nuclei. The elements formed by these slow and rapid processes correspond, respectively, to the s and r peaks of fig. 2 [A concise early discussion of this problem is presented in the final chapter of Chandrasekhar’s 1939 text.]

Another approach to the element formation problem provided an enormous contribution to understanding the nuclear physics of stars. In a

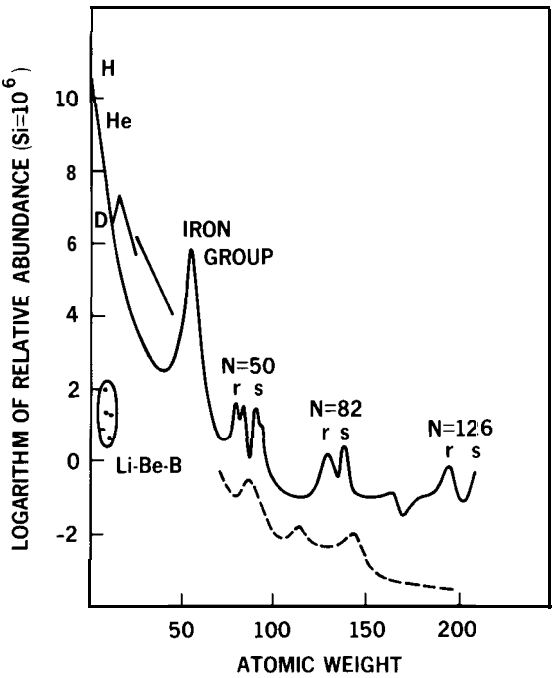


Fig. *2* ***Relative Abundances of the Elements:*** Smoothed curves representing the abundances of various groups of elements, after Burbidge et. al, 1957, who presented a total of eight processes to fit this data (See Clayton 1968 for a more modern treatment.) Lithium, Beryllium and Boron (circled) are not formed in the build-up process which goes from helium to carbon. The small amounts of these elements found in nature are fragments from the break-up of heavier elements.

beautiful paper entitled, "Energy Production in Stars", Bethe (1939) considered the individual nuclear reactions of the light nuclei, from hydrogen through oxygen. This paper established the role of the fusing of hydrogen into helium by two processes and demonstrated their quantitative agreement with observations. In the first process, protons combine to form a deuteron which is then transformed into ^{4}He by the further capture of protons. In the second, carbon and nitrogen are used as catalysts, viz

$$^{12}C+H = {}^{13}N+\gamma,\ {}^{13}N = {}^{13}C+e^+$$
$$^{13}C+H = {}^{14}N+\gamma$$
$$^{14}N+H = {}^{15}O+\gamma,\ {}^{15}O = {}^{15}N+e^+$$
$$^{15}N+H = {}^{12}C+{}^4He$$

(The notation and format are taken from the cited reference.)

As to the build-up of the heavier elements, however, no stable build-up process beyond the mass-4 nucleus had been found; a mass-4 nucleus cannot be combined with any other nucleus to form a heavier nucleus. In particular, no stable mass-5 nucleus exists, so the addition of a neutron or proton to ^{4}He doesn't work. Bethe wrote, "The progress of nuclear physics

in the past few years makes it possible to decide rather definitely which processes can and which cannot occur in the interior of stars . . under present conditions, no elements heavier than helium can be built up to any appreciable extent". In an attempt to bypass the mass-4 barrier, Bethe considered, and correctly rejected, the direct formation of ^{12}C from the simultaneous collision of three helium nuclei. He also noted that the formation of ^{8}Be from two helium nuclei was prevented by the fact that this nucleus was known to be unstable, having a negative binding energy of "between 40 and 100 keV". This energy difference corresponds to a temperature of some $10^{9}K$, again to be compared with the $\sim 2x10^{7}K$ which was then thought to be the allowed stellar temperature. It was not realized at that time that it is possible to form ^{8}Be from ^{4}He at a sufficiently high ^{4}He density and temperature and so bypass the mass-4 barrier. So it was that recognition of the crucial role of ^{8}Be in the build-up of the elements had to await the acceptance, in the early 1950's, of a new understanding of the physics of stellar interiors.

In the intervening decade, therefore, attention was diverted toward processes which could have occurred before the formation of the stars, namely a hot dense state associated with the birth of the universe. The formalism associated with the birth of the universe had been laid out by Friedman (1922), Lemaitre (1927) and Einstein and deSitter (1932). The applicability of this formalism to the real world was established by the beautiful simplicity of Hubble's (1929) powerful result that the observed velocities of the "extragalactic nebulae" [i.e. the galaxies which make up the universe] were proportional to their distances from the observer. In its simplest form, the most distant galaxy is moving away at the fastest rate and the nearest at the slowest. This is exactly what one would expect if all the galaxies had begun their flight from a common origin and, at a common starting time, had been given their start in a trememdous explosion.

Not widely popular among respectable scientists of the time, this idea of an expanding universe was taken up in the 1940's in part because the theories of the stellar origin of the elements had failed in the 1930's. (The expanding universe picture was generally ignored again in the 1950's when the wide variety of stellar phenomena became understood. It was only in the 1960's that a more balanced view emerged, but that comes later in our story.) The title of Chandrasekhar and Henrich's 1942 paper "An Attempt to Interpret the Relative Abundances of the Elements and Their Isotopes" reflects the tentative and unsatisfactory nature of the state of understanding at that time. The paper begins, "It is now generally agreed that the chemical elements cannot be synthesized under conditions *now believed* (emphasis added) to exist in stellar interiors." As an alternative, the authors suggested that the expansion and cooling of the early universe might be a possible site for the processes. In this view, each of the elements had its abundance "frozen out" at an appropriate stage of the expansion of the hot ($\geqslant 10^{9}K$), dense ($\geqslant 10^{6}gr/cm^{3}$) universe.

As was shown by George Gamow (1946), however, the formation of elements in the early universe could not have occurred through these equilibrium processes. He accomplished this demonstration by a straightforward calculation of the time scales involved. (The interested general reader can find more on this and related points in the mathematical appendices of S. Weinberg's (1977) delightful book "The First Three Minutes".)

Consider a point mass m located on the surface of an expanding sphere with mass density ϱ. The energy E of the mass with respect to the center of the sphere is a fixed quantity, the sum of its kinetic and potential energies (the latter is a negative quantity), viz

$$E = const = \frac{mv^2}{2} - \frac{Gm(4\pi\rho R^3/3)}{R} \tag{5}$$

where G is the constant of gravitation, ϱ, the density, R, the radius of the sphere, and v, the outward velocity of the point mass, are all functions of time. Since $4\pi\varrho R^3/3$, the mass within the sphere, is not an increasing function of R, the far right-hand term must become arbitrarily large for sufficiently small values of R(t), i.e., at early times in the expansion. Under this "early time approximation" both right-hand terms must become very large because the difference between them is fixed. Thus we can regard the two terms as essentially equal at early times and, upon simple rearrangement, obtain

$$\frac{R(t)^2}{v(t)^2} \approx \frac{3}{8\pi\rho(t)G} \tag{6}$$

Now R/v is a characteristic time scale for the expansion: it is the reciprocal of Hubble's constant and is referred to as the Hubble age in cosmology. (Hubble's "constant" is constant in the spatial sense; it varies in time.) Putting numerical values in (6), we have

$$Age \approx \sqrt{\frac{10^6}{\rho}}\ sec. \tag{7}$$

where ϱ is expressed in gr/cm^3. Thus, as Gamow pointed out, a neutron density of $10^{30} cm^{-3}$ (about $10^6 gr/cm3$) would exist for less than one second in the early universe. Since the β-decays necessary to establish the appropriate equalities between protons and neutrons are typically measured in minutes, it is clear that the time period needed to establish equilibrium with neutrons at the high densities required simply was not available in the early expanding universe.

This demonstration set the stage for the consideration of nonequilibrium processes. Fortunately, two timely developments for the undertaking of such a study had just occurred. The first was the publication of the values of neutron capture cross-sections in the open literature after the end of World War II. The second was a bright graduate student in need of a thesis topic. Lifshitz (1946) solved the problem that Gamow's student,

R. A. Alpher, had originally selected for a thesis topic, one having to do with turbulence and galaxy formation in the early universe. As a result, Alpher soon set to work on a new topic, the nonequilibrium formation of the elements by neutron capture. Since not all cross-sections were available, Alpher fitted a smooth curve through the published points, and used this curve for his calculations. The results of Alpher's calculation were introduced to the scientific world in a brief letter whose list of authors makes it part of the folklore of physics (Alpher, Bethe and Gamow 1948).

At this point the trail divides. Two different paths of investigation must be followed before they merge again into final results. We proceed to follow one of them with the understanding that we must return here later to follow the other.

In presenting his thesis results Alpher initiated a series of interactions between scientists which led to a succession of results very different from what he might have expected. First, Enrico Fermi, present at a seminar given by Alpher, soon raised an important objection: The straight line interpolation of capture cross-sections leads to a serious error in the case of the light nuclei. The neutron capture cross-section of a mass-4 particle is known to be essentially zero, whereas Alpher's curve was fitted to the average cross-sections of the nearby nuclei, which are quite large. Fermi had his student Turkevitch redo Alpher's calculations using explicit measured values for the cross-sections. Fermi and Turkevitch's results, never published separately but merely sent directly to Alpher, showed what Gamow and his co-workers knew and admitted privately, that their mechanism could produce nothing heavier than mass-4 from neutrons alone.

Second, Fermi pressed his friend Martin Schwartzschild for observational evidence of the formation of the heavy elements in stars. Together with his wife Barbara, Schwartzschild amply fulfilled this request. In one of the classic papers of observational astronomy (Schwartzschild and Schwartzschild 1950) they measured the faint spectra of two groups of stars of the same stellar type, F dwarfs, stars with long uneventful lifetimes. A separation into two groups, Population I and Population II, was done on the basis of velocity. This distinction, due to Baade, makes use of the fact that interstellar gas is almost totally confined to the galactic plane because vertical (i.e., perpendicular to the plane) gas motions are quickly damped out by cloud-to-cloud collisions. Thus, new stars born from this gas are to be found in the plane, without appreciable vertical motion. (These stars, which are easier to find, were found first and hence are called Population I.) Old stars, formed before the formation of the galactic disc retain the high velocities of the gas from which they were formed because dissipative encounters between stars are negligibly rare. Consequently, older (Pop II) stars can be distinguished by their higher velocities. The Schwartzschilds' comparison of the spectra of the two populations provided a clear answer: the younger Population I stars had the greater abundance of iron and other metals, thus revealing the enrichment of the interstellar medium between the times that the older and younger stars were formed.

This unmistakable evidence of metal production by stars during the lifetime of the galaxy removed the need for a pre-stellar mechanism for element formation. Only the path around the mass-4 barrier for element build-up in stars still had to be found. This was the third and final step.

Martin Schwartzschild presented this challenge to a young nuclear physicist, Ed Salpeter. Salpeter set to work, having a much wider range of accepted stellar conditions to work with than did Bethe in his earlier investigation. He soon found (Salpeter 1952) that ^{8}Be, unstable though it is, can be present in the hot dense cores of red giant stars in sufficient quantities to provide a convenient stepping stone for the formation of ^{12}C through the addition of a helium-4 nucleus.

With both observational support and the theoretical path around the mass-4 barrier, the triumph of stellar element formation now seemed complete. Fred Hoyle dismissed all pre-stellar theories of element build-up as "requiring a state of the universe for which we have no evidence" (Burbidge, et. al. 1957). So much for Alpher and Gamow's theory! "If the curve is simple the explanation must be simple" Gamow (1950) had said. But the curve of elemental abundances is not a simple one (Fig. 2). Burbidge et. al. presented no less than seven separate processes to account for the data, and left room for more under an eighth heading to fill in the few remaining gaps of their picture.

Ironically, it was Fred Hoyle himself who found a gap that could not be filled in the stellar picture, a gap in the best-understood process of them all, the formation of helium from hydrogen. Although the burning of hydrogen into helium provides the sun and the other stars with their energy and with building blocks for the formation of the heavier elements, Hoyle concluded that about ninety percent of the helium found in stars must have been made before the birth of the galaxy. The basis for this conclusion was an energy argument: the total amount of energy released by the formation of all the observed helium is some ten times greater than the energy radiated by the galaxies since their formation. Thus, "it is difficult to suppose that all the helium has been produced in ordinary stars" (Hoyle and Taylor 1964). Instead, attention was turned to helium formation in the early stages of an expanding universe, reviving work begun by George Gamow some sixteen years earlier. As indicated above, our description of Gamow's work was deferred in order to first follow the progress of the stellar picture of element build-up. We can now follow the second path.

Despite the problems inherent in Alphers treatment, (see, e.g., Alpher and Herman 1950), it provided the basis for a statement of profound simplicity and great power (Gamow 1948). Although wrong in almost every detail, Gamow's new insight pointed the way for others to follow. He noted that nuclear build-up cannot take place in the hottest, most condensed, state of the early universe because thermal photons at high temperatures $\geqslant 10^{10}$K are energetic enough to break up bound particle groups. Only when the temperature has cooled to $\sim 10^{9}$K, can nuclear

reactions begin. Any build-up, however, must be completed during the few hundred seconds before all the free neutrons decay into protons. Gamow considered a cylinder (Fig. 3) swept out by a neutron with a 10^9K thermal velocity during its lifetime. The cross-section of the cylinder was the capture cross-section for deuteron formation. If there was to have been appreciable element build-up in the early universe, Gamow reasoned, some fraction, say one half, of the initial neutrons had to have collided with protons to form deuterons before they had time to decay. Thus, half of Gamow's sample cylinders should contain a proton. This statement determines the number of protons per unit volume. From this result, the mass of the proton, and his estimate of the fraction of matter that was in the form of protons (roughly one half), Gamow obtained the mass density of matter in the universe at 10^9K, about 10^{-6}gm/cm^3.

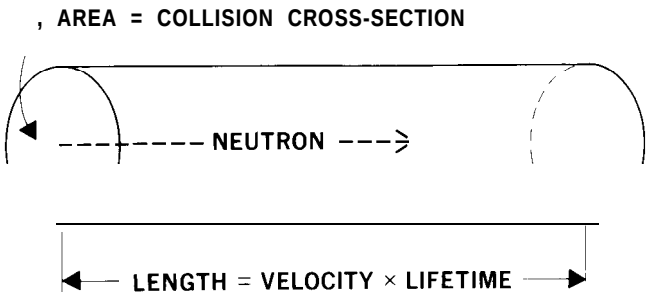


Fig. 3 Gamows ***Sample Cylinder;*** The volume swept out by a neutron in the early universe. The length of the cylinder is the product of the neutron's thermal velocity (at 10^9K) and its decay time. The cross-sectional area is the neutron-proton collision cross-section for deuteron formation. The fraction of neutrons forming deuterons is equal to the probability that the cylinder contains a proton.

Gamow then noted that the mass density of radiation at 10^9K (i.e., its energy density divided by c^2) was about 10gr/cm^3, as compared with only 10^{-6}gr/cm^3 for matter. This makes radiation the dominant component in the entropy of the early universe, permitting it to cool during the expansion as if the matter were not present. In that case, the temperature varies inversely with the radius of the expanding volume element (Tolman 1934, Peebles 1971) i.e.,

$$T \propto R^{-1}. \tag{8}$$

Now since ϱ, the density of matter, varies inversely as the cube of the radius, we can replace (8) with

$$T \propto \sqrt[3]{\rho}. \tag{9}$$

or

$$\frac{T_1}{T_2} = \sqrt[3]{\frac{\rho_1}{\rho_2}}.$$

This neat relation between temperature and matter density holds as long as radiation remains the dominant component. When the temperature drops below ~ 3×10^3K, the matter is too cool to remain ionized, and once it becomes neutral it is essentially transparent to the radiation. The radiation is then no longer coupled to the matter, it is free to expand forever in untroubled isolation, and eqn (9) continues to apply.

Gamow was only interested in tracing the radiation to the epoch when the matter becomes neutral and decouples from the radiation. From that point on, the matter has only its own thermal energy to support itself against gravitational collapse, so it fragments and condenses to form galaxies. Gamow used eqn (9) to find the density of matter at $3X\ 10^3$K and the Jeans criterion to determine the size of the collapsing fragments. Thus he was able to obtain a relation for the mass of galaxies containing only fundamental constants and the single assumption that half the initial neutrons collided to form deuterons. This was quite a trick, even for him!

Gamow's paper inspired his former student, Alpher and his collaborator Robert Herman to do the calculations more rigorously (Alpher and Herman 1949). Most importantly they replaced the "early-time" approximation Gamow used with a more exact formulation and traced the temperature of the relict primordial radiation to the present epoch. Taking the present matter density of the universe to be 10^{-30}gm/cm^3, they concluded that the present energy density of the relict radiation should correspond to a temperature of a few degrees Kelvin. Although mention of this prediction persisted in Gamow's popular writing, it was only repeated explicitly in a few of their subsequent scientific works. As for detection, they appear to have considered the radiation to manifest itself primarily as an increased energy density (Alpher and Herman 1949, pg. 1093). This contribution to the total energy flux incident upon the earth would be masked by cosmic rays and integrated starlight, both of which have comparable energy densities. The view that the effects of three components of approximately equal additive energies could not be separated may be found in a letter by Gamow written in 1948 to Alpher (unpublished, and kindly provided to me by R. A. Alpher from his files). "The space temperature of about 5° K is explained by the present radiation of stars (C-cycles). The only thing we can tell is that the residual temperature from the original heat of the Universe is *not* higher than 5° K." They do not seem to have recognized that the unique spectral characteristics of the relict radiation would set it apart from the other effects.

The first published recognition of the relict radiation as a detectable microwave phenomen appeared in a brief paper entitled "Mean Density of Radiation in the Metagalaxy and Certain Problems in Relativistic Cosmology", by A. G. Doroshkevich and 1. D. Novikov (1964a) in the spring of 1964. Although the English translation (1964b) appeared later the same year in the widely circulated "Soviet Physics-Doklady", it appears to have escaped the notice of the other workers in this field. This remarkable paper not only points out the spectrum of the relict radiation as a black-

body microwave phenomenon, but also explicity focuses upon the Bell Laboratories twenty-foot horn reflector at Crawford Hill as the best available instrument for its detection! Having found the appropriate reference (Ohm 1961), they misread its results and concluded that the radiation predicted by the "Gamow Theory" was contradicted by the reported measurement.

Ohm's paper is an engineering report on a low-noise microwave receiving system. The reported noise of this system contained a residul excess of almost exactly three degrees! Ohm had measured a total system noise temperature of some 22K including the contribution of the receiver, the antenna, the atmosphere and the sky beyond. Separate measurements of each of the components of this noise temperature, except the sky beyond the atmosphere, totalled - 19K. (From an analysis of his measurement errors, Ohm concluded that both sets of measurements, the total and the sum of individual contributions, could be consistent with an intermediate value). The atmospheric contribution was measured by moving the antenna in elevation and fitting the change in system temperature to a cosecant relation, a standard procedure which is described by Wilson (1978). To avoid confusion with other quantities, the atmospheric contribution thus derived was denoted T_{sky}, the "sky temperature". Ohm's value of 2.3K for this quantity was in good agreement with atmospheric attenuation theory. The background contribution due to the relict radiation has no elevation dependence and cannot be detected by this technique. Perhaps due to the unfortunate name, Doroshkevitch and Novikov regarded T_{sky}, as containing the background radiation and therefore leading to a null result. The disappointment is reflected in Section IV of Zeldovitch's concurrent (1965) review.

The year 1964 also marked a reawakened interest in the "Gamow Theory" by Hoyle and Taylor (1964) as well as the first unambiguous detection of the relict radiation. The rough outlines of Gamow's initial treatment had long since been refined by the work of others. For example, it was pointed out by Hayashi (1950) that the assumption of an initial neutron material was incorrect. The radiation field at $T > 10^9$K generates electron- positron pairs which serve to maintain quasi-thermal equilibrium between neutrons and protons (see also Chandrasekhar and Henrich, 1942, who made the same point). Alpher, Follin and Herman (1953) incorporated this process into their rigorous treatment of the problem. Their work benefited from the availability of what was, by the standard of those days, a powerful electronic computer which permitted them to include the dynamic effects of expansion and cooling upon collisional and photo-disintegrated processes. Their results, which have not been substantially altered by subsequent work, are chiefly marked by (1) conversion of some 15 %' of the matter into helium, with the exact amount dependent only slightly upon the density at $T \approx 10^9$K and (2) production of deuterium whose surviving abundance is sensitively dependent upon the initial temperature/density relation. The same ground was covered in Hoyle and

Taylor's 1964 paper, which cited Alpher, Follin and Herman's paper and noted the agreement with the earlier results. Neither paper made any mention of surviving relict radiation.

Shortly thereafter, P. J. E. Peebles treated the same subject for a different reason. R.H. Dicke had, with P.G. Roll and D.T. Wilkenson, set out to measure the background brightness of the sky at microwave wavelengths. At his suggestion, Peebles began an investigation of the cosmological constraints that might be imposed by the results of such a measurement. Peebles' paper, which was submitted to the Physical Review and circulated in preprint form in March of 1965. This paper paralleled the above light element production picture and included Hoyle and Taylor (1964) among its references. In addition, it explicitly delineated the surviving relict radiation as a detectable microwave phenomenon. At about the same time, microwave background radiation was detected at Bell Laboratories and its extragalactic origin established. No combination of the then known sources of radio emission could account for it. Receipt of a copy of Peebles' preprint solved the problem raised be this unexplained phenomenon. Eddington tells us: "Never fully trust an observational result until you have at least one theory to explain it. "The theory and observation were then brought together in a pair of papers (Dicke et al, 1965, Penzias and Wilson 1965) which led to decisive support for evolutionary cosmology and further renewal of interest in its observational consequences.

The existence of the relict radiation established the validity of the expanding universe picture with its cosmological production of the light elements, deuterium, helium-3 and helium-4 during the hot early stages of the expansion. The build-up of the heavier elements occurs at a much later stage, after the stars have formed. In stars, the cosmologically produced helium-4, together with additional amounts of helium produced by the stars themselves, is converted (via beryllium-8) into carbon-12 from which the heavier elements are then built. The stellar process described by Burbidge et al (1957) have been supplemented and, in some cases, replaced by processes whose existence was established trough later work, of which explosive nucleosynthesis is the most significant one. (See Clayton 1968 for a review.) Much of the build-up of the heavier elements goes on in a few violent minutes during the life of massive stars in which their outer shells are thrown outward in supernova explosions. This mechanism accounts both for the formation of the heavy elements as well as for their introduction into interstellar space. Thus, the total picture seems close to complete but puzzling gaps remain, such as the absence of solar neutrinos (Bahcall and Davis, 1976). One thing is clear however, observational cosmology is now a respectable and flourishing science.

Acknowledgment

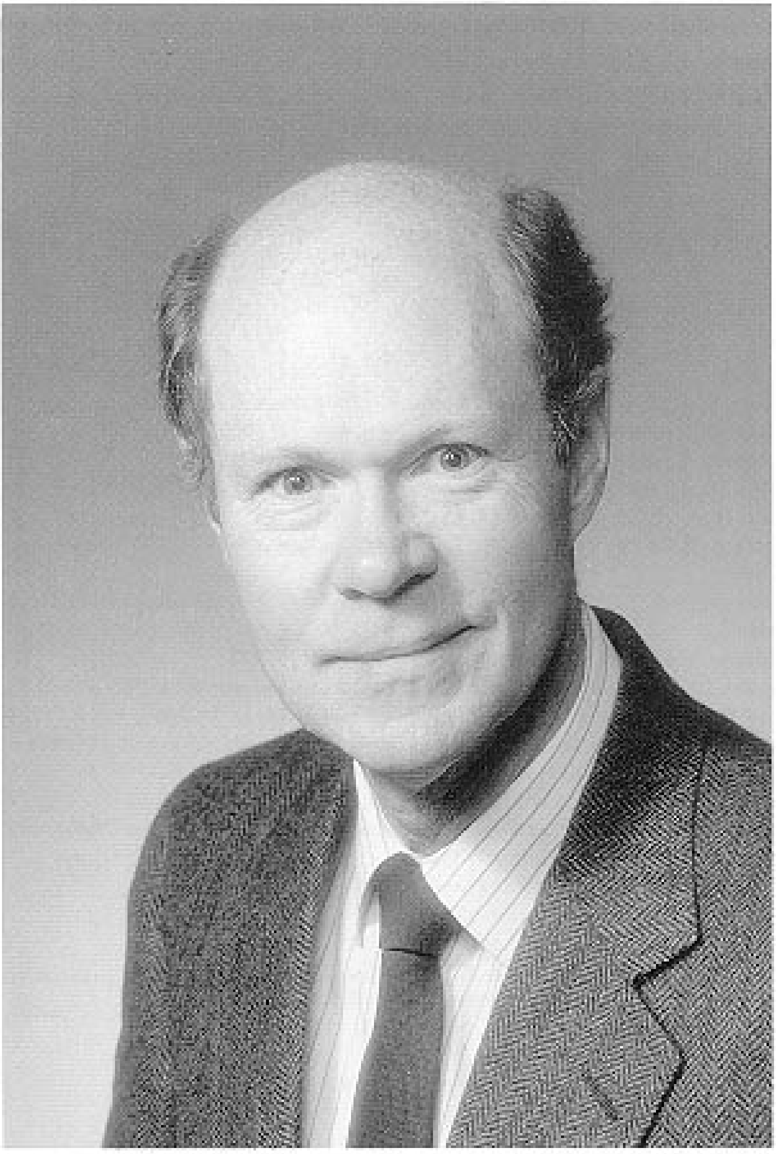
My first thanks must go to the members of the Academy for the great honor they have bestowed upon me. The work which resulted in the occasion for this talk is described in an accompanying paper by my friend

and colleague Robert W. Wilson. I am profoundly grateful for his unfailing help throughout our fifteen years of partnership.

The preparation of the talk upon which this manuscript is based owes much to many people. Conversations with R. A. Alpher, John Bahcall, S. Chandrasekhar and Martin Schwartzschild were particularly helpful. I am also grateful to A. B. Crawford, R. H. Dicke, G. B. Field, R. Kompfner, P. J. E. Peebles, D. Sciama, P. Thaddeus and S. Weinberg for earlier help given personally and through their published work.

REFERENCES

1. Alpher, R. A., Bethe, H. A., and Gamow, G., 1949, Phys. Rev. 73. 803.
2. Alpher, R. A. and Herman, R. C., 1949, Phys. Rev. 75, 1089
3. Alpher, R. A. and Herman, R. C., 1950, Rev. Mod. Phys.22. 153.
4. Alpher. R. A., Follin, J. W. and Herman, R. C., 1953, Phys. Rev. 92, 1347.
5. Bethe, H. A., 1939, Phys. Rev. 55, 434.
6. Burbidge, E. M., Burbidge, G. R., Fowler, W. A., and Hoyle, F., 1957, Rev. Mod. Phys. 29, 547.
7. Chandrasekhar, S., 1939, ***An Introduction to the Study*** of ***Stellar Structure,*** (University of Chicago).
8. Chandrasekhar, S. and Henrich, L. R., 1942, Ap. J. 95, 228.
9. Clayton, D. D., 1968 ***Principles of Stellar Evolution and Nucleosynthesis*** (McGraw-Hill).
10. Dicke, R. H., Peebles, P. J. E., Roll, P. G. and Wilkinson, D. T., 1965, Ap. J. 142, 414.
11. Doroshkevich. A. G. and Novikov, I. D.. 1964a Dokl. Akad. Navk. SSR 154, 809.
12. Doroshkevich, A. G. and Novikov, I. D., l964b Sov. Phys-Dokl. 9. 111.
13. Einstein, A. and deSitter, W., 1932 Proc. Nat. Acad. Sci. ***18,*** 312.
14. Eddington, A. ***S.,*** 1926, ***The Internal Constitution of the Stars,*** (Cambridge University Press).
15. Fowler, R. H., 1926, M.N.R.A. 87, 114.
16. Friedmann, A., 1922. Zeits. Fur Physik 10, 377.
17. Gamow, G., 1946, Phys. Rev. 70, 572.
18. Gamow, G., 1948, Nature ***162, 680.***
19. Gamow, G., 1950, Physics Today 3, No. 8, 16
20. Hayashi, C., 1950, Prog. Theo. Phys. (Japan) 5, 224.
21. Hoyle, F. and Taylor, R. J., 1964, Nature 203, 1108.
22. Hubble, E. P., 1929, Proc. N.A.S. 15, 168.
23. Lemaitre, G., 1927, Ann. Soc. Sci. Brux. ***A47, 49.***
24. Lifschitz, E., 1946, J. Phys. USSR ***10, 116.***
25. Ohm. E. A., 1961, Bell Syst. Tech. J. 40, 1065.
26. Peebles, P. J. E., 1971, Physical Cosmology (Princeton University Press).
27. Penrias, A. A. and Wilson, R. W., 1965, Ap. J. 142, 419.
28. Salpeter, E. E., 1952. Ap. J. ***115, 326.***
29. Schwartzchild, B. and Schwartzschild, M. 1950, Ap. J. 212, 248.
30. Tolman, R. C., 1934, Relativity ***Thermodynamics and Cosmology,*** (Clarendon Press, Oxford).
31. von Weizsäcker, C. F., 1937, Physik. Zeits. 38, 176.
32. von Weizsäcker, C. F., 1938, Physik. Zeits. 39, 633.
33. Weinberg, S., The First Three Minutes, (Basic Books).
34. Wilson, R. W., 1978 Nobel Lecture
35. Zeldovich, 1965, Advances in Astr. and Ap.3. 241.



ROBERT W. WILSON

My grandparents moved to Texas from the South after the U.S. Civil War and settled on small farms in the Dallas-Ft. Worth area. Both families emphasized education as the way to improve their children's lives and both my parents managed to graduate from college. After receiving an M.A. in chemistry from Rice University, my father worked for an oil well service company in Houston. I was born on January, 10, 1936. Two sisters followed, three and seven years later.

I attended public school in Houston. I took piano lessons for several years, and in high school, I played trombone in the marching band. I remember especially enjoying two seasonal activities: ice skating with the Houston Figure Skating Club in the winter and visiting an aunt and uncle's farm in west Texas in the summer.

During my pre-college years I went on many trips with my father into the oil fields to visit their operations. On Saturday mornings I often went with him to visit the company shop. I puttered around the machine, electronics, and automobile shops while he carried on his business. Both of my parents are inveterate do-it-yourselfers, almost no task being beneath their dignity or beyond their ingenuity. Having picked up a keen interest in electronics from my father, I used to fix radios and later television sets for fun and spending money. I built my own hi-fi set and enjoyed helping friends with their amateur radio transmitters, but lost interest as soon as they worked.

My high school career was undistinguished except for math and science. However, having barely been admitted to Rice University, I found that I enjoyed the courses and the elation of success and graduated with honors in physics. I did a senior thesis with C. F. Squire building a regulator for a magnet for use in low-temperature physics. Following that I had a summer job with Exxon and obtained my first patent. It covered the high-voltage pulse generator for a pulsed neutron source in a down-hole well-logging tool.

Following Rice, I went to Caltech for a Ph.D in physics, without any strong idea of what I wanted to do for a thesis topic. For the first year I lived in the Athenaeum (faculty club) where I became acquainted with a small group of graduate students and visiting faculty members, with whom I often dined and went on weekend outings. When the end of my second quarter approached, I needed a trial research project. David Dewhirst, a Cambridge astronomer and one of the Athenaeum group, suggested that I see John Bolton and Gordon Stanley about radio astronomy. The situation seemed perfect for me. John had come to Caltech to build the Owens

Valley Radio Observatory, and the heavy construction was finished. Radio astronomy offered a nice mixture of electronics and physics.

My introduction to radio astronomy was, however, delayed for a summer. I returned to Houston to court and marry Elizabeth Rhoads Sawin, whose spirit and varied interests have added much to my happiness during our twenty-year marriage.

The following year I took my first astronomy courses and went to the observatory during school breaks. That summer John Bolton asked me to join him in observing some of the bright regions on a radio map of the Milky Way which had been made by Westerhaut. By the end of the summer, this project had expanded to making a complete map of that part of the Milky Way which was visible to us. When it was time to measure our chart records and start drawing contour maps from the data, John set up a drawing board in his office, and worked with me on the project. This was typical of John. Whatever the project, whether digging a hole, surveying, laying cables, observing, or reducing data, John would work along with the others. His interest in our map-making and the location of the drawing board kept me at the map-making task instead of designing the next piece of equipment, which would have been my natural inclination.

Our first son, Philip, was born during my fourth year at Caltech. He had many trips to the Owens Valley Radio Observatory, the first at the age of two weeks. He and Betsy were readily accepted at the observatory.

My thesis project was to have been hydrogen-line interferometry, but when the first plans for a local oscillator system didn't work out, I used the galactic survey as the basis for my thesis. John Bolton returned to Australia before I completed my Ph.D. Maarten Schmidt, who had previously done galactic research and was currently working on quasars, saw me through the last months of thesis work. I remained at Caltech for an additional year as a postdoctoral fellow to finish several projects in which I was involved.

The project of setting up and running the Owens Valley Radio Observatory was very much a community effort. At one time or another I worked with all of the staff and other students and learned from all of them. My collaborations with V. Radhakrishram and B. G. Clark were especially fruitful. I also had the opportunity to meet many of the world's astronomers who visited Caltech.

In 1961, H. E. D. Scovil at Bell Labs offered to help us make a pair of traveling-wave maser amplifiers for the interferometer. V. Radhakrishran got the job of going to Bell Labs to make our masers. I had wanted to go, but had not yet completed my degree work. I worked with Rad on that project, though, and developed a good feeling toward Bell Labs which was later a strong influence on my decision to take a job there.

I joined Bell Laboratories at Crawford Hill in 1963 as part of A. B. Crawford's Radio Research department in R. Kompfner's laboratory. I started working with the only other radio astronomer, Arno Penzias, who had been there about two years. Our early radio astronomy projects are described in my Nobel lecture.

With the creation of Comsat by U. S. Congress, Bell System satellite efforts and related space research were reduced. In 1965 Arno and I were told that the radio astronomy effort could only be supported at the level of one full-time staff member, even though Art Crawford and Rudi Kompfner strongly supported our astronomical research. Arno and I agreed that having two half-time radio astronomers was a better solution to our problem than having one full-time one, so we started taking on other projects. The first one was a joint project-a propagation experiment on a terrestrial path using a 10.6μ carbon dioxide laser as a source. Following that, I did two applied radio astronomy projects. For the first, I designed a device we called the Sun Tracker. It automatically pointed to the sun while it was up every day and measured the attenuation of the sun's cm-wave radiation in the earth's atmosphere. Since, as we expected, the attenuation was large for too much of the time for a practical satellite system, I next set up three fixed-pointed radiometers at spaced locations to check on the feasibility of working around heavy rains.

In I969 Arno suggested that we start doing millimeter wave astronomy. We could take the low noise millimeter-wave receivers which had been developed at Crawford Hill by C. A. Burrus and W. M. Sharpless for a waveguide communication system and make an astronomical receiver with them. We planned to use it at the National Radio Astronomy Observatory's new 36-foot radio telescope at Kitt Peak in Arizona. Our observations began in 1969 with a continuum receiver. The next year, K. B. Jefferts joined us, and with much help from C. A. Burrus at Crawford Hill and S. Weinreb at NRAO we made a spectral line receiver at 100-120 GHz. We were excited to discover unexpectedly large amounts of carbon monoxide in a molecular cloud behind the Orion Nebula. We quickly found that CO is widely distributed in our galaxy and so abundant that the rare isotopic species $^{13}C^{16}O$ and $^{12}C^{18}O$ were readily measurable. We soon observed a number of other simple molecules. Our major efforts were directed toward isotope ratios as a probe of nucleogenesis and understanding the structure of molecular clouds.

In 1972, S. J. Buchsbaum, who was our new executive director, revived an earlier proposal and suggested that we build a millimeter-wave facility at Crawford Hill. It was to be used partly for radio astronomy, and partly to monitor the beacons on the Comstar satellites which AT&T was planning to put up. I was project director for the design and construction of the antenna and was responsible for the equipment and programming necessary to make it a leading millimeter-radio telescope. The winter of 1977-78 was our first good observing season with the 7-meter antenna and I am looking forward to several more years of millimeter wave astronomy with it.

We still live in the house in Holmdel which we bought when I first came to Bell Laboratories. Our two younger children were born here, Suzanne in 1963, and Randal in 1967. We have come to enjoy the eastern woodlands and I now look forward to skiing and outdoor ice skating with my

family and associates in the winter. I spend many evenings reading or continuing the day's work, but I also enjoy playing the piano, jogging, and traveling with the family.

B.A. 1957 Rice University "with honors in Physics".
Ph.D. 1962 California Institute of Technology.
Married 1958 Elizabeth Rhoads Sawin.
Employment:
Caltech, Research Fellow 1962- 1963.
Bell Laboratories 1963 -
Member of Technical Staff 1963-1976.
Head Radio Physics Research Department 1976-
Adjunct Professor, State University of New York (SUNY) 1978
Member:
American Astronomical Society
International Astronomical Union
American Physical Society
International Union of Radio Sciences
American Academy of Arts and Sciences
Honors:
Phi Beta Kappa
Sigma Xi
Henry Draper Award 1977
Herschel Medal 1977

THE COSMIC MICROWAVE BACKGROUND RADIATION

Nobel Lecture, 8 December, 1978

by
ROBERT W. WILSON
Bell Laboratories
Holmdel, N.J. U.S.A.

1. INTRODUCTION

Radio Astronomy has added greatly to our understanding of the structure and dynamics of the universe. The cosmic microwave background radiation, considered a relic of the explosion at the beginning of the universe some 18 billion years ago, is one of the most powerful aids in determining these features of the universe. This paper is about the discovery of the cosmic microwave background radiation. It starts with a section on radio astronomical measuring techniques. This is followed by the history of the detection of the background radiation, its identification, and finally by a summary of our present knowledge of its properties.

II. RADIO ASTRONOMICAL METHODS

A radio telescope pointing at the sky receives radiation not only from space, but also from other sources including the ground, the earth's atmosphere, and the components of the radio telescope itself. The 20-foot horn-reflector antenna at Bell Laboratories (Fig. 1) which was used to discover the cosmic microwave background radiation was particularly suited to distinguish this weak, uniform radiation from other, much stronger sources. In order to understand this measurement it is necessary to discuss the design and operation of a radio telescope, especially its two major components, the antenna and the radiometer[1].

a. Antennas

An antenna collects radiation from a desired direction incident upon an area, called its collecting area, and focuses it on a receiver. An antenna is normally designed to maximize its response in the direction in which it is pointed and minimize its response in other directions.

The 20-foot horn-reflector shown in Fig. 1 was built by A. B. Crawford and his associate? in 1960 to be used with an ultra low-noise communications receiver for signals bounced from the Echo satellite. It consists of a large expanding waveguide, or horn, with an off-axis section parabolic reflector at the end. The focus of the paraboloid is located at the apex of the horn, so that a plane wave traveling along the axis of the paraboloid is focused into the receiver, or radiometer, at the apex of the horn. Its design emphasizes the rejection of radiation from the ground. It is easy to see

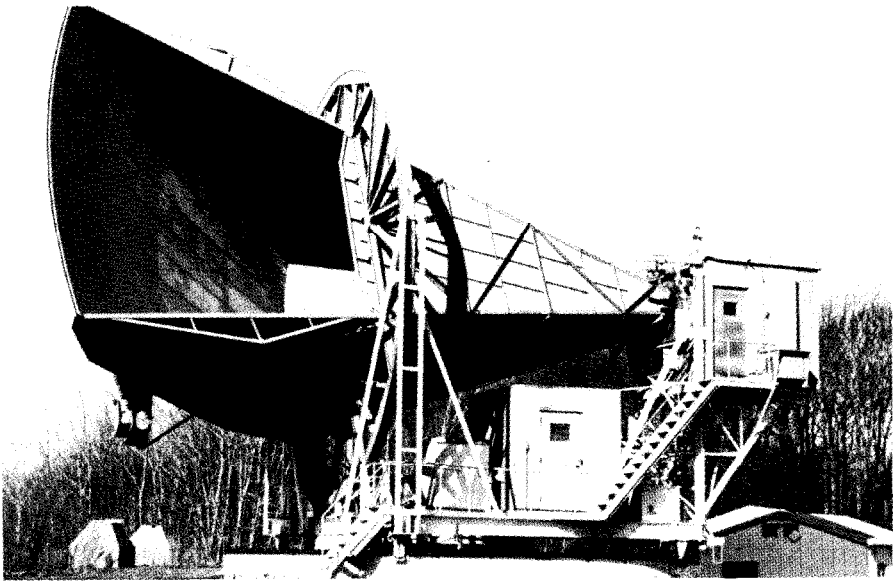


Fig. I The 20 foot horn-reflector which was used to discover the Cosmic Microwave Background Radiation.

from the figure that in this configuration the receiver is well shielded from the ground by the horn.

A measurement of the sensitivity of a small hornreflector antenna to radiation coming from different directions is shown in Fig. 2. The circle marked isotropic antenna is the sensitivity of a fictitious antenna which receives equally from all directions. If such an isotropic lossless antenna were put in an open field, half of the sensitivity would be to radiation from the earth and half from the sky. In the case of the hornreflector, sensitivity in the back or ground direction is less than 1/3000 of the isotropic antenna. The isotropic antenna on a perfectly radiating earth at 300 K and with a cold sky at 0° K would pick up 300 K from the earth over half of its response and nothing over the other half, resulting in an equivalent antenna temperature of 150 K. The horn-reflector, in contrast, would pick up less than .05 K from the ground.

This sensitivity pattern is sufficient to determine the performance of an ideal, lossless antenna since such an antenna would contribute no radiation of its own. Just as a curved mirror can focus hot rays from the sun and burn a piece of paper without becoming hot itself, a radio telescope can focus the cold sky onto a radio receiver without adding radiation of its own.

h. Radiometers

A radiometer is a device for measuring the intensity of radiation. A microwave radiometer consists of a filter to select a desired band of

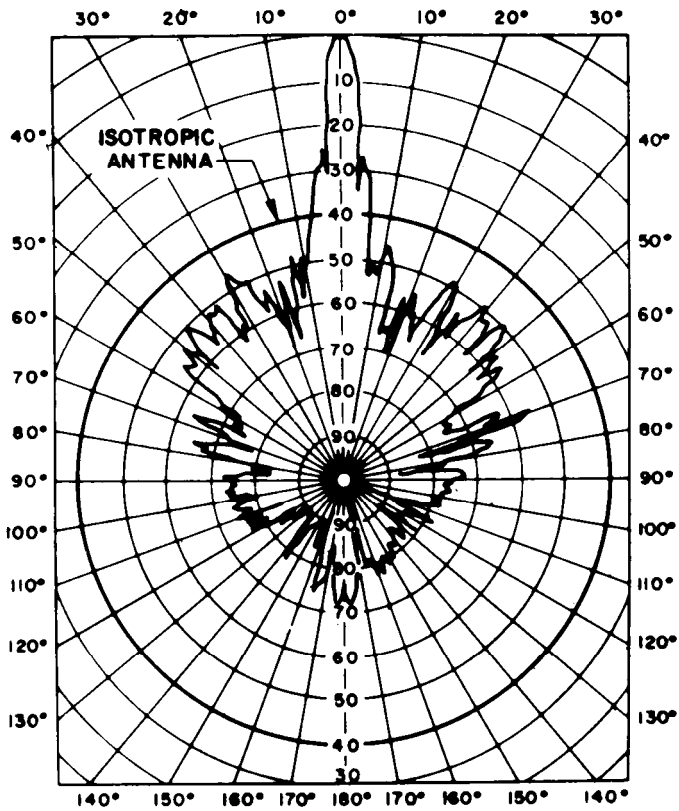


Fig. 2 Sensitivity pattern of a small horn-reflector antenna. This is a logarithmic plot of the collecting area of the antenna as a function of angle from the center of the main beam. Each circle below the level of the main beam represent a factor of ten reduction in sensitivity. In the back direction around 180 the sensitivity is consistently within the circle marked 70, corresponding to a factor of 10^{-7} below the sensitivity at 0.

frequencies followed by a detector which produces an output voltage proportional to its input power. Practical detectors are usually not sensitive enough for the low power levels received by radio telescopes, however, so amplification is normally used ahead of the detector to increase the signal level. The noise in the first stage of this amplifier combined with that from the transmission line which connects it to the antenna (input source) produce an output from the detector even with no input power from the antenna. A fundamental limit to the sensitivity of a radiometer is the fluctuation in the power level of this noise.

During the late 1950's, H. E. D. Scovil and his associates at Bell Laboratories, Murray Hill were building the world's lowest-noise microwave amplifiers, ruby travelling-wave masers[3]. These amplifiers were cooled to 4.2 K or less by liquid helium and contribute a correspondingly small amount of noise to the system. A radiometer incorporating these amplifiers can therefore be very sensitive.

Astronomical radio sources produce random, thermal noise very much

like that from a hot resistor, therefore the calibration of a radiometer is usually expressed in terms of a thermal system. Instead of giving the noise power which the radiometer receives from the antenna, we quote the temperature of a resistor which would deliver the same noise power to the radiometer. (Radiometers often contain calibration noise sources consisting of a resistor at a known temperature.) This "equivalent noise temperature" is proportional to received power for all except the shorter wavelength measurements, which will be discussed later.

c. Observations

To measure the intensity of an extraterrestrial radio source with a radio telescope, one must distinguish the source from local noise sources-noise from the radiometer, noise from the ground, noise from the earth's atmosphere, and noise from the structure of the antenna itself. This distinction is normally made by pointing the antenna alternately to the source of interest and then to a background region nearby. The difference in response of the radiometer to these two regions is measured, thus subtracting out the local noise. To determine the absolute intensity of an astronomical radio source, it is necessary to calibrate the antenna and radiometer or, as usually done, to observe a calibration source of known intensity.

III. PLANS FOR RADIO ASTRONOMY WITH THE 20-FOOT HORN-REFLECTOR

In 1963, when the 20-foot horn-reflector was no longer needed for satellite work, Arno Penzias and I started preparing it for use in radio astronomy. One might ask why we were interested in starting our radio astronomy careers at Bell Labs using an antenna with a collecting area of only 25 square meters when much larger radio telescopes were available elsewhere. Indeed, we were delighted to have the 20-foot horn-reflector because it had special features that we hoped to exploit. Its sensitivity, or collecting area, could be accurately calculated and in addition it could be measured using a transmitter located less than 1 km away. With this data, it could be used with a calibrated radiometer to make primary measurements of the intensities of several extraterrestrial radio sources. These sources could then be used as secondary standards by other observatories. In addition, we would be able to understand all sources of antenna noise, for example the amount of radiation received from the earth, so that background regions could be measured absolutely. Traveling-wave maser amplifiers were available for use with the 20-foot horn-reflector, which meant that for large diameter sources (those subtending angles larger than the antenna beamwidth), this would be the world's most sensitive radio telescope.

My interest in the background measuring ability of the 20-foot horn-reflector resulted from my doctoral thesis work with **J. G. Bolton at**

Caltech. We made a map of the 3 1cm radiation from the Milky Way and studied the discrete sources and the diffuse gas within it. In mapping the Milky Way we pointed the antenna to the west side of it and used the earth's rotation to scan the antenna across it. This kept constant all the local noise, including radiation that the antenna picked up from the earth. I used the regions on either side of the Milky Way (where the brightness was constant) as the zero reference. Since we are inside the Galaxy, it is impossible to point completely away from it. Our mapping plan was adequate for that project, but the unknown zero level was not very satisfying. Previous low frequency measurements had indicated that there is a large, radio-emitting halo around our galaxy which I could not measure by that technique. The 20-foot horn-reflector, however, was an ideal instrument for measuring this weak halo radiation at shorter wavelengths. One of my intentions when I came to Bell Labs was to make such a measurement.

In 1963, a maser at 7.35 cm wavelength[3] was installed on the 20-foot horn-reflector. Before we could begin doing astronomical measurements, however, we had to do two things: 1) build a good radiometer incorporating the 7.35 cm maser amplifier, and; 2) finish the accurate measurement of the collecting-area (sensitivity) of the 20-foot horn-reflector which D. C. Hogg had begun. Among our astronomical projects for 7 cm were absolute intensity measurements of several traditional astronomical calibration sources and a series of sweeps of the Milky Way to extend my thesis work. In the course of this work we planned to check out our capability of measuring the halo radiation of our Galaxy away from the Milky Way. Existing low frequency measurements indicated that the brightness temperature of the halo would be less than 0.1 K at 7 cm. Thus, a background measurement at 7 cm should produce a null result and would be a good check of our measuring ability.

After completing this program of measurements at 7 cm, we planned to build a similar radiometer at 21 cm. At that wavelength the galactic halo should be bright enough for detection, and we would also observe the 21 cm line of neutral hydrogen atoms. In addition, we planned a number of hydrogen-line projects including an extension of the measurements of Arno's thesis, a search for hydrogen in clusters of galaxies.

At the time we were building the 7-cm radiometer John Bolton visited us and we related our plans and asked for his comments. He immediately selected the most difficult one as the most important: the 21 cm background measurement. First, however, we had to complete the observations at 7 cm.

IV. RADIOMETER SYSTEM

We wanted to make accurate measurements of antenna temperatures. To do this we planned to use the radiometer to compare the antenna to a reference source, in this case, a radiator in liquid helium. I built a switch

which would connect the maser amplifier either to the antenna or to Arno's helium-cooled reference noise source[5] (cold load). This would allow an accurate comparison of the equivalent temperature of the antenna to that of the cold load, since the noise from the rest of the radiometer would be constant during switching. A diagram of this calibration system[6] is shown in Figure 3 and its operation is described below.

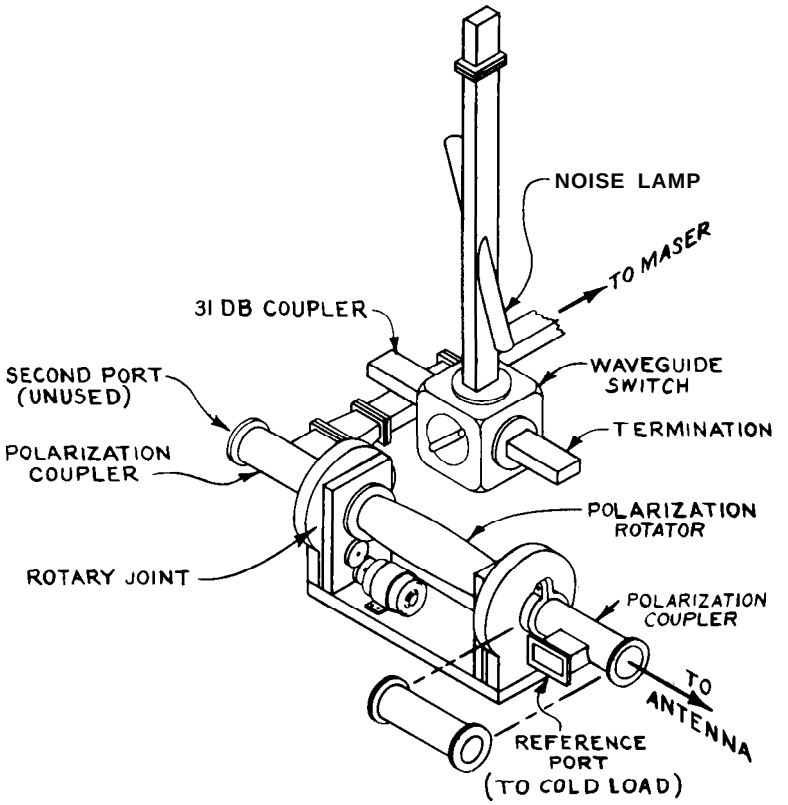


Fig. 3 The switching and calibration system of our 7.35 cm radiometer, The reference port was normally connected to the helium cooled reference source through a noise adding attenuator.

a. *Switch*

The switch for comparing the cold load to the antenna consists of the two polarization couplers and the polarization rotator shown in Fig. 3. This type of switch had been used by D. H. Ring in several radiometers at Holmdel. It had the advantage of stability, low loss, and small reflections. The circular waveguide coming from the antenna contains the two orthogonal modes of polarization received by the antenna. The first polarization coupler reflected one mode of linear polarization back to the antenna and substituted the signal from the cold load for it in the waveguide going to the rotator. The second polarization coupler took one of the two modes of linear polarization coming from the polarization rotator and coupled it to the rectangular (single-mode) waveguide going to the maser. The polarization rotator is the microwave equivalent of a half-wave plate in optics. It is a

piece of circular waveguide which has been squeezed in the middle so that the phase shifts for waves traveling through it in its two principal planes of linear polarization differ by 180 degrees. By mechanically rotating it, the polarization of the signals passing through it can be rotated. Thus either the antenna or cold load could be connected to the maser.

This type of switch is not inherently symmetric, but has very low loss and is stable so that its asymmetry of .05 K was accurately measured and corrected for.

b. *Reference Noise Source*

A drawing of the liquid-helium cooled reference noise source is shown in Figure 4. It consists of a 122 cm piece of 90 percent-copper brass waveguide connecting a carefully matched microwave absorber in liquid He to a room-temperature flange at the top. Small holes allow liquid helium to fill the bottom section of waveguide so that the absorber temperature could be known, while a mylar window at a 30” angle keeps the liquid out of the rest of the waveguide and makes a low-reflection microwave transition between the two sections of waveguide. Most of the remaining parts are for the cryogenics. The *gas* baffles make a counter-flow heat exchanger between the waveguide and the helium gas which has boiled off, greatly extending the time of operation on a charge of liquid helium. Twenty liters of liquid helium cooled the cold load and provided about twenty hours of operation.

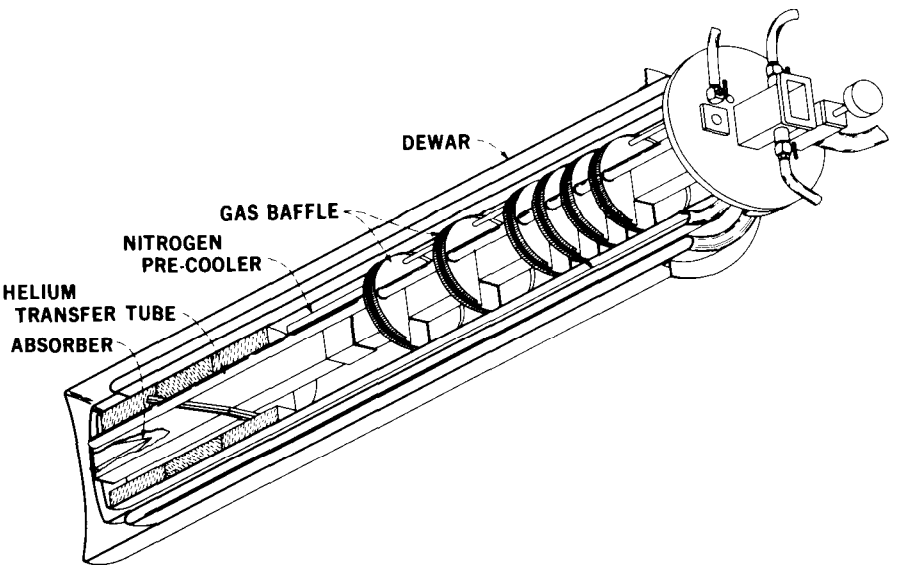


Fig. 4 The Helium Cooled Reference Noise Source.

Above the level of the liquid helium, the waveguide walls were warmer than 4.2 K. Any radiation due to the loss in this part of the waveguide would raise the effective temperature of the noise source above 4.2 K and

must be accounted for. To do so we monitored the temperature distribution along the waveguide with a series of diode thermometers and calculated the contribution of each section of the waveguide to the equivalent temperature of the reference source. When first cooled down, the calculated total temperature of the reference noise source was about 5 K, and after several hours when the liquid helium level was lower, it increased to 6 K. As a check of this calibration procedure, we compared the antenna temperature (assumed constant) to our reference noise source during this period, and found consistency to within 0.1 K.

c. Scale Calibration

A variable attenuator normally connected the cold load to the reference port of the radiometer. This device was at room temperature so noise could be added to the cold load port of the switch by increasing its attenuation. It was calibrated over a range of 0.11 dB which corresponds to 7.4 K of added noise.

Also shown in Fig. 3 is a noise lamp (and its directional coupler) which was used as a secondary standard for our temperature scale.

d. *Radiometer Backend*

Signals leaving the maser amplifier needed to be further amplified before detection so that their intensity could be measured accurately. The remainder of our radiometer consisted of a down converter to 70 MHz followed by I. F. amplifiers, a precision variable attenuator and a diode detector. The output of the diode detector was amplified and went to a chart recorder.

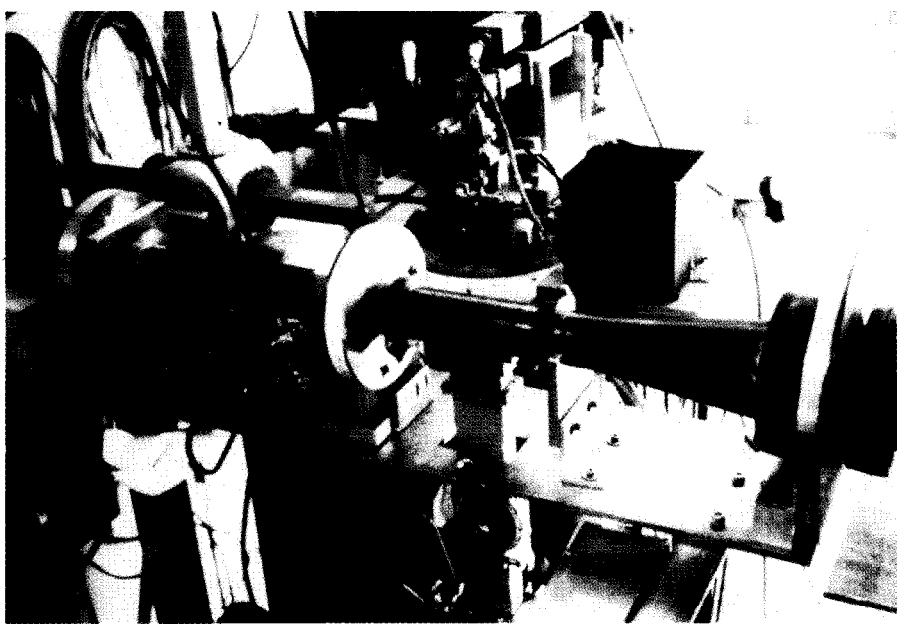


Fig. 5 Our 7.35 cm radiometer installed in the cab of the 20 foot horn-reflector.

e. Equipment Performance

Our radiometer equipment installed in the cab of the 20-foot horn-reflector is shown in Fig. 5. The flange at the far right is part of the antenna and rotates in elevation angle with it. It was part of a double-choke joint which allowed the rest of the equipment to be fixed in the cab while the antenna rotated. The noise contribution of the choke-joint could be measured by clamping it shut and was found to be negligible. We regularly measured the reflection coefficient of the major components of this system and kept it below 0.03 percent, except for the maser whose reflection could not be reduced below 1 percent. Since all ports of our waveguide system were terminated at a low temperature, these reflections resulted in negligible errors.

V. PRIOR OBSERVATIONS

The first horn-reflector-travelling-wave maser system had been put together by DeGrasse, Hogg, Ohm, and Scovil in 1959[7] to demonstrate the feasibility of a low-noise, satellite-earth station at 5.31 cm. Even though they achieved the lowest total system noise temperature to date, 18.5 K, they had expected to do better. Fig. 6 shows their system with the noise temperature they assigned to each component. As we have seen in Section IIa,

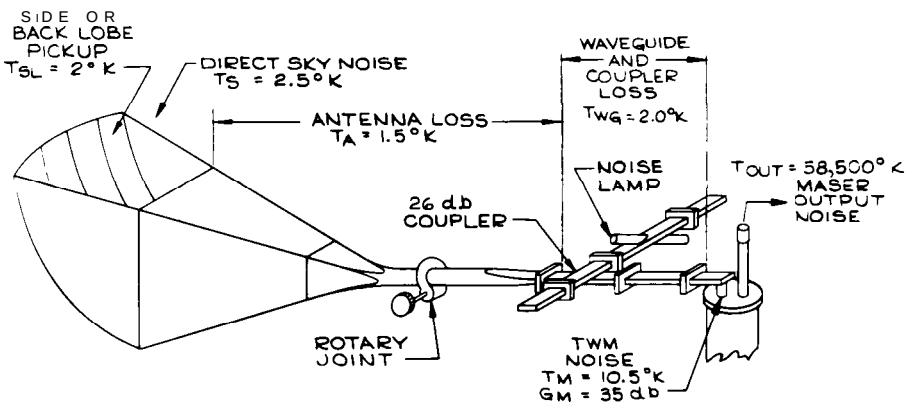


Fig. 6 A diagram of the low noise receiver used by deGrasse, Hogg, Ohm and Scovil to show that very low noise earth stations are possible. Each component is labeled with its contribution to the system noise.

the 2 K they assigned to antenna backlobe pickup is too high. In addition, direct measurements of the noise temperature of the maser gave a value about a degree colder than shown here. Thus their system was about 3 K hotter than one might expect. The component labeled T_s in Fig. 6 is the radiation of the earth's atmosphere when their antenna was aimed straight up. It was measured by a method first reported by R. H. Dicke[8]. (It is interesting that Dicke also reports an upper limit of 20 K for the cosmic microwave background radiation in this paper - the first such report.) If the antenna temperature is measured as a function of the angle above the

horizon at which it is pointing, the radiation of the atmosphere is at a minimum when the antenna is directed straight up. It increases as the antenna points toward the horizon, since the total line of sight through the atmosphere increases. Figure 7 is a chart recording Arno Penzias and I

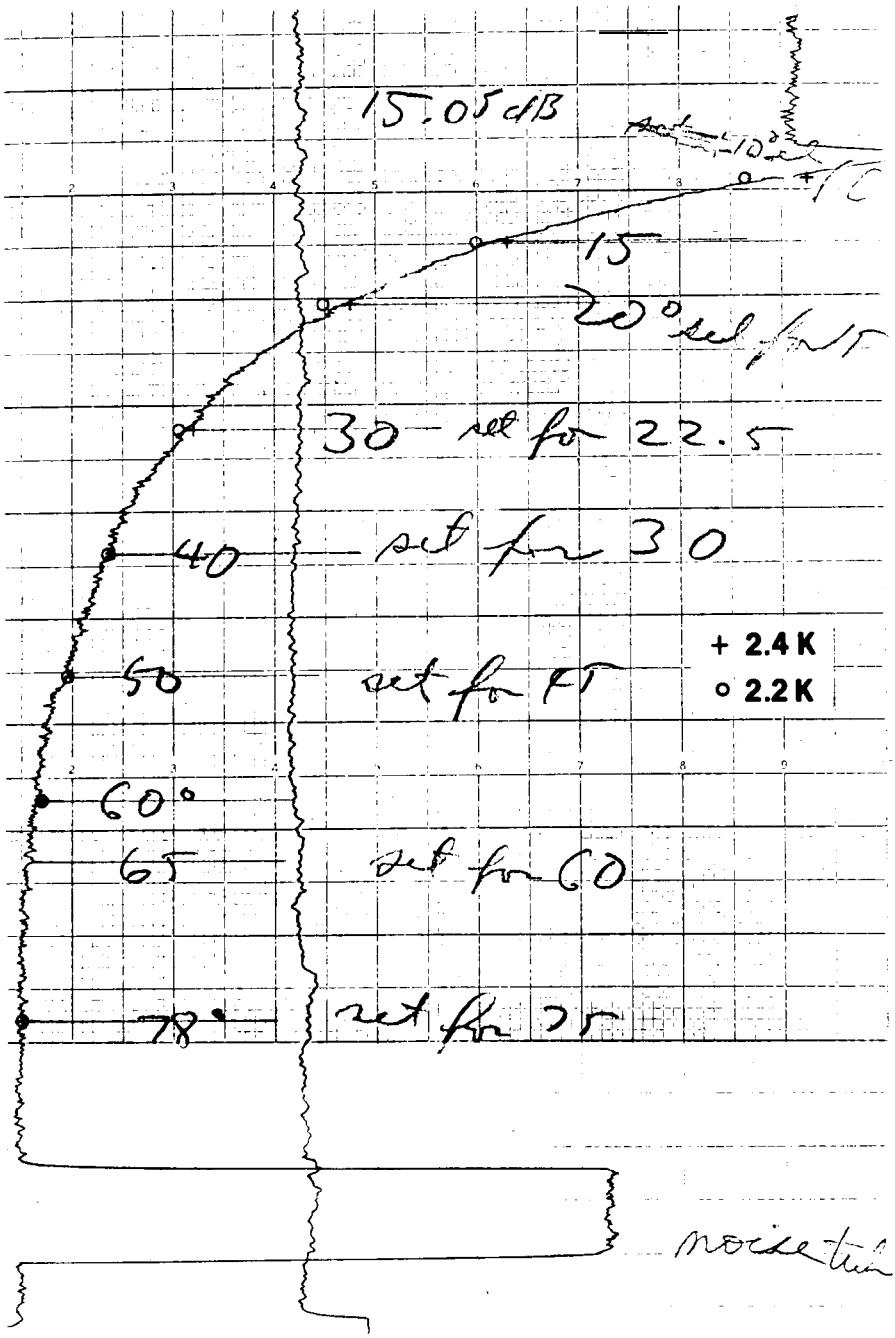


Fig. 7 A measurement of atmospheric noise at 7.35 cm wavelength with theoretical fits to the data for 2.2 and 2.4K Zenith atmospheric radiation.

made with the 20-foot horn-reflector scanning from almost the Zenith down to 10° above the horizon. The circles and crosses are the expected change based on a standard model of the earth's atmosphere for 2.2 and 2.4 K Zenith contribution. The fit between theory and data is obviously good leaving little chance that there might be an error in our value for atmospheric radiation.

Fig. 8 is taken from the paper in which E. A. Ohm[9] described the receiver on the 20-foot horn reflector which was used to receive signals bounced from the Echo satellite. He found that its system temperature was 3.3 K higher than expected from summing the contributions of the components. As in the previous 5.3 cm work, this excess temperature was smaller

TABLE II — SOURCES OF SYSTEM TEMPERATURE

Source	Temperature
Sky (at zenith)	2.30 ± 0.20°K
Horn antenna	2.00 ± 1.00°K
Waveguide (counter-clockwise channel)	7.00 ± 0.65°K
Maser assembly	7.00 ± 1.00°K
Converter	0.60 ± 0.15°K
Predicted total system temperature	18.90 ± 3.00°K

the temperature was found to vary a few degrees from day to day, but the lowest temperature was consistently 22.2 ±2.2°K. By realistically assuming that all sources were then contributing their fair share (as is also tacitly assumed in Table II) it is possible to improve the over-all accuracy. The actual system temperature must be in the overlap region of the measured results and the total results of Table II, namely between **20 and 21.9°K. The most likely minimum system temperature was therefore**

$$T_{system} = 21 \pm 1°K.*$$

The inference from this result is that the "+" temperature possibilities of Table II must predominate.

Fig. 8 An excerpt **from E. A. Ohm's article on the Echo receiver** showing **that his system temperature was 3.3K higher than predicted**

than the experimental errors, so not much attention was paid to it. In order to determine the unambiguous presence of an excess source of radiation of about 3 K, a more accurate measurement technique was required. This was achieved in the subsequent measurements by means of a switch and reference noise source combination which communications systems do not have.

VI. OUR OBSERVATIONS

Fig. 9 is a reproduction of the first record we have of the operation of our system. At the bottom is a list of diode thermometer voltages from which we could determine the cold load's equivalent temperature. The recorder trace has power (or temperature) increasing to the right. The middle part of this trace is with the maser switched to the cold load with various settings of the noise adding attenuator. A change of 0.1 dB corresponds to a temperature change of 6.6 K, so the peak-to-peak noise on the trace amounts to less than 0.2 K. At the top of the chart the maser is switched to

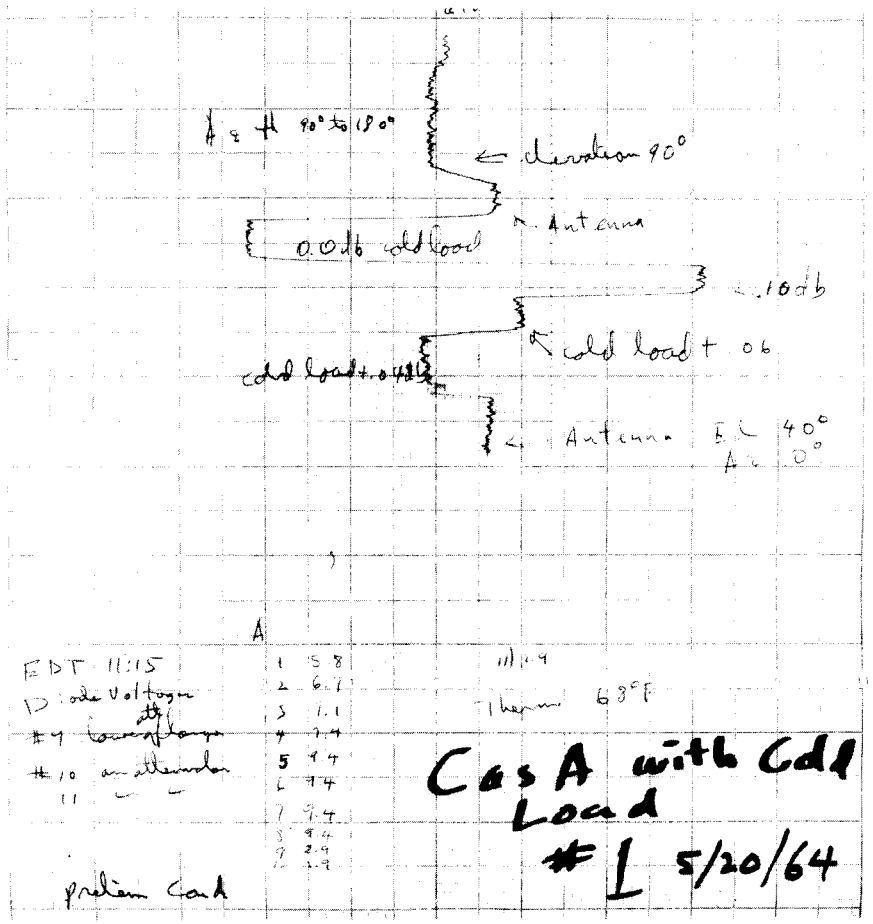


Fig. 9 The first measurement which clearly showed the presence of the microwave background. Noise temperature is plotted increasing to the right. At the top, the antenna pointed at 90° elevation is seen to have the samt noise temperature as the cold load with 0.04 db attenuation (about 7.5K). This is considerably above the expected value of 3.3K.

the antenna and has about the same temperature as the cold load plus .04 dB, corresponding to a total of about 7.5 K. This was a troublesome result. The antenna temperature should have been only the sum of the atmospheric contribution (2.3 K) and the radiation from the walls of the antenna and ground (1 K). The excess system temperature found in the previous experiments had, contrary to our expectations, all been in the antenna or beyond. We now had a direct comparison of the antenna with the cold load and had to assign our excess temperature to the antenna whereas in the previous cases only the total system temperature was measured. If we had missed some loss, the cold load might have been warmer than calculated, but it could not be colder than 4.2 K - the temperature of the liquid helium. The antenna was at least 2 K hotter than that. Unless we could understand our “antenna problem” our 21 cm galactic halo experi-

ment would not be possible. We considered a number of possible reasons for this excess and, where warranted, tested for them. These were:

a. At that time some radio astronomers thought that the microwave absorption of the earth's atmosphere was about twice the value we were using - in other words the "sky temperature" of Figs. 6 and 8 was about 5 K instead of 2.5 K. We knew from our measurement of sky temperature such as shown in Fig. **7** that this could not be the case.
b. We considered the possibility of man-made noise being picked up by our antenna. However, when we pointed our antenna to New York City, or to any other direction on the horizon, the antenna temperature never went significantly above the thermal temperature of the earth.
c. We considered radiation from our galaxy. Our measurements of the emission from the plane of the Milky Way were a reasonable fit to the intensities expected from extrapolations of low-frequency measurements. Similar extrapolations for the coldest part of the sky (away from the Milky Way) predicted about .02 K at our wavelength. Furthermore, any galactic contribution should also *vary* with position and we saw changes only near the Milky Way, consistent with the measurements at lower frequencies.
d. We ruled out discrete extraterrestrial radio sources as the source of our radiation as they have spectra similar to that of the Galaxy. The same extrapolation from low frequency measurements applies to them. The strongest discrete source in the sky had a maximum antenna temperature of 7 K.

Thus we seemed to be left with the antenna as the source of our extra noise. We calculated a contribution of 0.9 K from its resistive loss using standard waveguide theory. The most lossy part of the antenna was its small diameter throat, which was made of electroformed copper. We had measured similar waveguides in the lab and corrected the loss calculations for the imperfect surface conditions we had found in those waveguides. The remainder of the antenna was made of riveted aluminum sheets, and although we did not expect any trouble there, we had no way to evaluate the loss in the riveted joints. A pair of pigeons was roosting up in the small part of the horn where it enters the warm cab. They had covered the inside with a white material familiar to all city dwellers. We evicted the pigeons and cleaned up their mess, but obtained only a small reduction in antenna temperature.

For some time we lived with the antenna temperature problem and concentrated on measurements in which it was not critical. Dave Hogg and 1 had made a very accurate measurement of the antenna's gain[10], and Arno and 1 wanted to complete our absolute flux measurements before disturbing the antenna further.

In the spring of 1965 with our flux measurements finished[5], we thoroughly cleaned out the 20-foot horn-reflector and put aluminum tape over the riveted joints. This resulted in only a minor reduction in antenna temperature. We also took apart the throat section of the antenna, and checked it, but found it to be in order.

By this time almost a year had passed. Since the excess antenna temperature had not changed during this time, we could rule out two additional sources: 1) Any source in the solar system should have gone through a large change in angle and we should have seen a change in antenna temperature. 2) In 1962, a high-altitude nuclear explosion had filled up the Van Allen belts with ionized particles. Since they were at a large distance from the surface of the earth, any radiation from them would not show the same elevation-angle dependence as the atmosphere and we might not have identified it. But after a year, any radiation from this source should have reduced considerably.

VII. IDENTIFICATION

The sequence of events which led to the unravelling of our mystery began one day when Arno was talking to Bernard Burke of M.I.T. about other matters and mentioned our unexplained noise. Bernie recalled hearing about theoretical work of P. J. E. Peebles in R. H. Dicke's group in Princeton on radiation in the universe. Arno called Dicke who sent a copy of Peebles' preprint. The Princeton group was investigating the implications of an oscillating universe with an extremely hot condensed phase. This hot bounce was necessary to destroy the heavy elements from the previous cycle so each cycle could start fresh. Although this was not a new idea" Dicke had the important idea that if the radiation from this hot phase were large enough, it would be observable. In the preprint, Peebles, following Dicke's suggestion calculated that the universe should be filled with a relic blackbody radiation at a minimum temperature of 10 K. Peebles was aware of Hogg and Semplak's (1961)[12] measurement of atmospheric radiation at 6 cm using the system of DeGrasse et al., and concluded that the present radiation temperature of the universe must be less than their system temperature of 15 K. He also said that Dicke, Roll, and Wilkinson were setting up an experiment to measure it.

Shortly after sending the preprint, Dicke and his coworkers visited us in order to discuss our measurements and see our equipment. They were quickly convinced of the accuracy of our measurements. We agreed to a side-by-side publication of two letters in the *Astrophysical Journal-a* letter on the theory from Princeton[13] and one on our measurement of excess antenna temperature from Bell Laboratories[14]. Arno and I were careful to exclude any discussion of the cosmological theory of the origin of background radiation from our letter because we had not been involved in any of that work. We thought, furthermore, that our measurement was independent of the theory and might outlive it. We were pleased that the mysterious noise appearing in our antenna had an explanation of any kind, especially one with such significant cosmological implications. Our mood, however, remained one of cautious optimism for some time.

VIII. RESULTS

While preparing our letter for publication we made one final check on the antenna to make sure we were not picking up a uniform 3 K from earth. We measured its response to radiation from the earth by using a transmitter located in various places on the ground. The transmitter artificially increased the ground's brightness at the wavelength of our receiver to a level high enough for the backlobe response of the antenna to be measurable. Although not a perfect measure of the structure of the backlobes of an antenna, it was a good enough method of determining their average level. The backlobe level we found in this test was as low as we had expected and indicated a negligible contribution to the antenna temperature from the earth.

The right-hand column of Fig. 10 shows the final results of our measurement. The numbers on the left were obtained later in 1965 with a new throat on the 20-foot horn-reflector. From the total antenna temperature we subtracted the known sources with a result of 3.4 ± 1 K. Since the errors in this measurement are not statistical, we have summed the maximum error from each source. The maximum measurement error of 1 K was considerably smaller than the measured value, giving us confidence in the reality of the result. We stated in the original paper that "This excess temperature is, within the limits of our observations, isotropic, unpolarized, and free of seasonal variations". Although not stated explicitly, our limits on an isotropy and polarization were not affected by most of the errors listed in Fig. 10 and were about 10 percent or 0.3 K.

	New Throat		Old Throat
He Temp.	4.22	4.22	
Calculated Contribution from Cold Load Waveguide	.38	.70 ± 0.2	
Attenuator Setting for Balance	2.73	2.40 ± 0.1	
Total C.L.	7.33	7.32 ± 0.3	6.7 ± 0.3
Atmosphere	2.3 ± 0.3		2.3 ± 0.3
Waveguide and Antenna loss	1.8 ± 0.3		.9 ± 0.3
Back lobes	.1 ± 0.1		.1 ± 0.1
Total Ant.	4.2 ± 0.7		3.3 ± 0.7
Background	3.1 ± 1		3.4 ± 1

Fig. 10 Results of our 3965 measurements of the microwave background. "Old Throat" and "New Throat" refer to the original and a replacement throat section for the 20 foot horn-reflector.

At that time the limit we could place on the shape of the spectrum of the background radiation was obtained by comparing our value of 3.5 K with a 74 cm survey of the northern sky done at Cambridge by Pauliny-Toth and Shakeshaft, 1962[15]. The minimum temperature on their map was 16 K. Thus the spectrum was no steeper than $\lambda^{0.7}$ over a range of wavelengths that varied by a factor of 10. This clearly ruled out any type of radio source known at that time, as they all had spectra with variation in the range $\lambda^{2.0}$ to $\lambda^{3.0}$. The previous Bell Laboratories measurement at 6 cm ruled out a spectrum which rose rapidly toward shorter wavelengths.

IX. CONFIRMATION

After our meeting, the Princeton experimental group returned to complete their apparatus and make their measurement with the expectation that the background temperature would be about 3 K.

The first confirmation of the microwave cosmic background that we knew of, however, came from a totally different, indirect measurement. This measurement had, in fact, been made thirty years earlier by Adams and Dunham[16-21]. Adams and Dunham had discovered several faint optical interstellar absorption lines which were later identified with the molecules CH, CH^+, and CN. In the case of CN, in addition to the ground state, absorption was seen from the first rotationally excited state. McKellar[22] using Adams' data on the populations of these two states calculated that the excitation temperature of CN was 2.3 K. This rotational transition occurs at 2.64 mm wavelength, near the peak of a 3 K black body spectrum. Shortly after the discovery of the background radiation, G. B. Field[23], I. S. Shklovsky[24], and P. Thaddeus[25] (following a suggestion by N. J. Woolf), independently realized that the CN is in equilibrium with the background radiation. (There is no other significant source of excitation where these molecules are located). In addition to confirming that the background was not zero, this idea immediately confirmed that the spectrum of the background radiation was close to that of a blackbody source for wavelengths larger than the peak. It also gave a hint that at short wavelengths the intensity was departing from the $1/\lambda^2$ dependence expected in the long wavelength (Raleigh-Jeans) region of the spectrum and following the true blackbody (Plank) distribution. In 1966, Field and Hitchcock[23] reported new measurements using Herbig's plates of ζ Oph and ζ Per obtaining 3.22 ± 0.15 K and 3.0 ± 0.6 K for the excitation temperature. Thaddeus and Clauser[25] also obtained new plates and measured 3.75 ± 0.5 K in ς Oph. Both groups argued that the main source of excitation in CN is the background radiation. This type of observation, taken alone, is most convincing as an upper limit, since it is easier to imagine additional sources of excitation than refrigeration.

In December 1965 Roll and Wilkinson[26] completed their measurement of 3.0 ± 0.5 K at 3.2 cm, the first confirming microwave measurement. This was followed shortly by Howell and Shakeshaft's[27] value of 2.8 ± 0.6

K at 20.7 cm[22] and then by our measurement of 3.2 K ± 1 K at 21.1 cm[28]. (Half of the difference between these two results comes from a difference in the corrections used for the galactic halo and integrated discrete sources.) By mid 1966 the intensity of the microwave background radiation had been shown to be close to 3 K between 2 1 cm and 2.6 mm, almost two orders of magnitude in wavelength.

X. EARLIER THEORY

I have mentioned that the first experimental evidence for cosmic microwave background radiation was obtained (but unrecognized) long before 1965. We soon learned that the theoretical prediction of it had been made at least sixteen years before our detection. George Gamow had made calculations of the conditions in the early universe in an attempt to understand Galaxy formation[29]. Although these calculations were not strictly correct, he understood that the early stages of the universe had to be very hot in order to avoid combining all of the hydrogen into heavier elements. Furthermore, Gamow and his collaborators calculated that the density of radiation in the hot early universe was much higher than the density of matter. In this early work the present remnants of this radiation were not considered. However in 1949, Alpher and Herman[30] followed the evolution of the temperature of the hot radiation in the early universe up to the present epoch and predicted a value of 5 K. They noted that the present density of radiation was not well known experimentally. In 1953 Alpher, Follin, and Herman[31] reported what has been called the first thoroughly modern analysis of the early history of the universe, but failed to recalculate or mention the present radiation temperature of the universe.

In 1964, Doroshkevich and Novikov[32,33] had also calculated the relic radiation and realized that it would have a blackbody spectrum. They quoted E. A. Ohm's article on the Echo receiver, but misunderstood it and concluded that the present radiation temperature of the universe is near zero.

A more complete discussion of these early calculations is given in Arno's lecture.[34]

XI. ISOTROPY

In assigning a single temperature to the radiation in space, these theories assume that it will be the same in all directions. According to contemporary theory, the last scattering of the cosmic microwave background radiation occurred when the universe was a million years old, just before the electrons and nucleii combined to form neutral atoms ("recombination").The isotropy of the background radiation thus measures the isotropy of the universe at that time and the isotropy of its expansion since then. Prior to recombination, radiation dominated the 'universe and the Jeans mass, or mass of the smallest gravitationally stable clumps was larger than a cluster of Galaxies. It is only in the period following recombination that Galaxies could have formed.

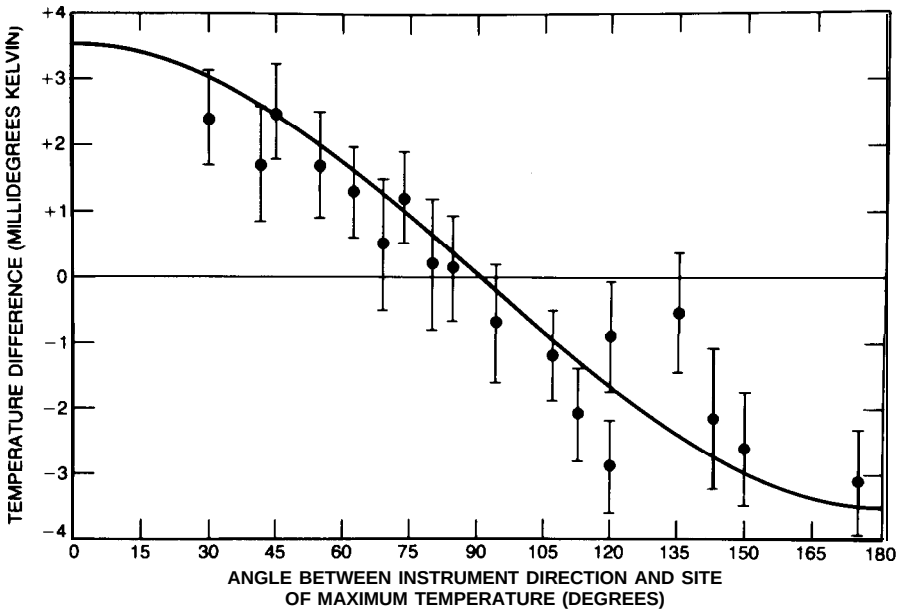


Fig. 11 Results of the large scale isotropy Experiment of Smoot, Gorenstein **and** Muller showing the clear cosine dependence of brightness expected from the relative velocity of the earth in the background radiation. The shaded area and arrows show the values allowed **by** the data of Woody **and Richards.** (This figure is reproduced with permission of Scientific American.)

In 1967 Rees and Sciama[35] suggested looking for large scale anisotropies in the background radiation which might have been left over from anisotropies of the universe prior to recombination.

In the same year Wilkinson and Partridge[36] completed an experiment which was specifically designed to look for anisotropy within the equatorial plane. The reported a limit of 0.1 percent for a 24 hour asymmetry and a possible 12 hour asymmetry of 0.2 percent. Meanwhile we had re-analyzed an old record covering most of the sky which was visible to us and put a limit of 0.1 K on any large scale fluctuations.[37]

Since then a series of measurements [38 39 40] have shown a 24-hour anisotropy due to the earth's velocity with respect to the background radiation. Data from the most sensitive measurement to date[41] are shown in Fig. 11. They show a striking cosine anisotropy with an amplitude of about .003 K, indicating that the background radiation has a maximum temperature in one direction and a minimum in the opposite direction. The generally accepted explanation of this effect is that the earth is moving toward the direction where the radiation is hottest and it is the blue shift of the radiation which increases its measured temperature in that direction. The motion of the sun with respect to the background radiation from the data of Smoot et al. is 390 ± 60 km/s in the direction 10.8^h R. A., 5° Dec. The magnitude of this velocity is not a surprise since 300 km/s is the orbital velocity of the sun around our galaxy. The direction, is different, however yielding a peculiar velocity of our galaxy of about 600 km/s. Since other nearby Galaxies; including the Virgo cluster,

have a small velocity with respect to our Galaxy, they have a similar velocity with respect to the matter which last scattered the background radiation. After subtracting the 24-hour anisotropy, one can search the data for more complicated anisotropies to put observational limits on such things as rotation of the universe[41]. Within the noise of .001 K, these anisotropies are all zero.

To date, no fine-scale anisotropy has been found. Several early investigations were carried out to discredit discrete source models of the background radiation. In the most sensitive experiment to date, Boynton and Partridge[42] report a relative intensity variation of less than 3.7×10^{-3} in an 80" Arc beam. A discrete source model would require orders of magnitude more sources than the known number of Galaxies to show this degree of smoothness.

It has also been suggested by Sunyaev and Zel'dovich[43] that there will be a reduction of the intensity of the background radiation from the direction of clusters of galaxies due to inverse Compton scattering by the electrons in the intergalactic gas. This effect which has been found by Birkinshaw and Gull[44], provides a measure of the intergalactic gas density in the clusters and may give an alternate measurement of Hubble's constant.

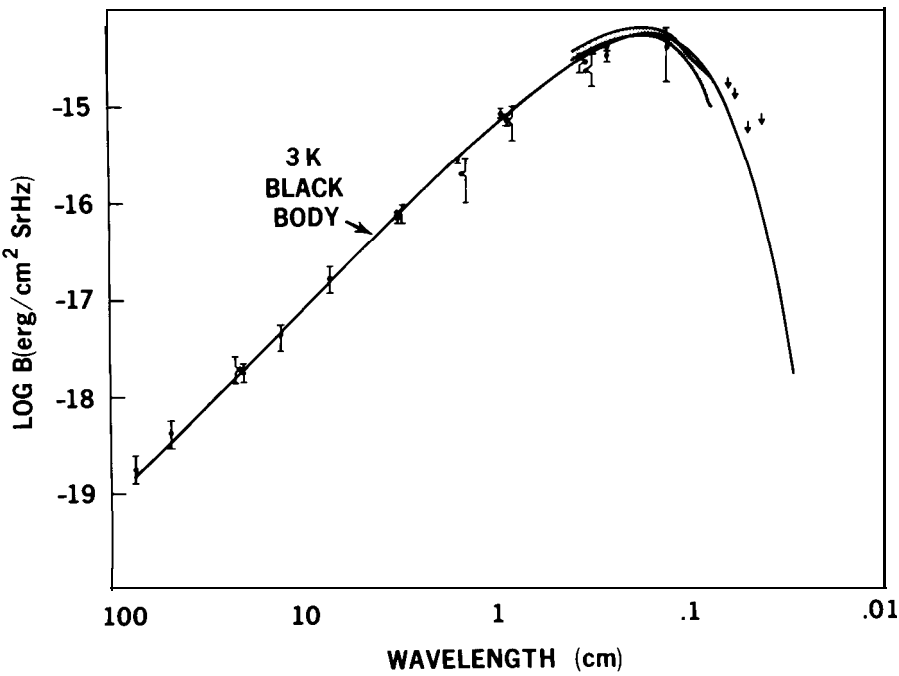


Fig. 12 Measurements of the spectrum of the cosmic microwave background radiation.

XII. SPECTRUM

Since 1966, a large number of measurements of the intensity of the background radiation have been made at wavelengths from 74 cm to 0.5 mm. Measurements have been made from the ground, mountain tops,

airplanes, balloons, and rockets. In addition, the optical measurements of the interstellar molecules have been repeated and we have observed their millimeter-line radiation directly to establish the equilibrium of the excitation of their levels with the background radiation[45]. Fig. 12 is a plot of most of these measurements[46]. An early set of measurements from Princeton covered the range 3.2 to .33 cm showing tight consistency with a 2.7 K black body [47-50]. A series of rocket and balloon measurements in the millimeter and submillimeter part of the spectrum have converged on about 3 K. The data of Robson, et al. [51] and Woody and Richards[52] extend to 0.8 mm, well beyond the spectral peak. The most recent experiment, that of D. Woody and P. Richards, gives a close fit to a 3.0 K spectrum out to 0.8 mm wavelength with upper limits at atmospheric windows out to 0.4 mm. This establishes that the background radiation has a blackbody spectrum which would be quite hard to reproduce with any other type of cosmic source. The source must have been optically thick and therefore must have existed earlier than any of the other sources, which can be observed.

The spectral data are now almost accurate enough for one to test for systematic deviations from a single-temperature blackbody spectrum which could be caused by minor deviations from the simplest cosmology. Danese and DeZotti[53] report that except for the data of Woody and Richards, the spectral data of Fig. 12 do not show any statistically significant deviation of this type.

XIII. CONCLUSION

Cosmology is a science which has only a few observable facts to work with. The discovery of the cosmic microwave background radiation added one -the present radiation temperature of the universe. This, however, was a significant increase in our knowledge since it requires a cosmology with a source for the radiation at an early epoch and is a new probe of that epoch. More sensitive measurements of the background radiation in the future will allow us to discover additional facts about the universe.

XIV. ACKNOWLEDGMENTS

The work which I have described was done with Arno A. Penzias. In our fifteen years of partnership he has been a constant source of help and encouragement. I wish to thank W. D. Langer and Elizabeth Wilson for carefully reading the manuscript and suggesting changes.

REFERENCES

1. A more complete discussion of radio telescope antennas and receivers may be found in several text books. Chapters 6 and 7 of J. D. Kraus “Radio Astronomy”, 1966, McGraw-Hill are good introductions to the subjects.
2. Crawford, A. B. Hogg, D. C. and Hunt, L. E. 1961, Bell System Tech. J., 40, 1095.

3. DeGrasse, R. W. Schultr-DuBois, E. 0. and Scovil, H. E. D. BSTJ 38, 30.5.
4. Tabor, W. J. and Sibilia, J. 1'. 1963, Bell System Tech. J., 42, 1863.
5. Penzias, A. A. 1965, Rev. Sci. Instr., 36. 68.
6. Penzias, A. A. and Wilson, R. W. 1965 Astrophysical Journal 142. 1149.
7. DeGrasse, R. W. Hogg, D. C. Ohm, E. A. and Scovil, H. E. D. 1950, Proceedings of the National Electronics Conference. 15, 370.
8. Dicke. R. Beringer R. Kyhl, R. L. and Vane, A. V. Phys. Rev. 70, 340 (1046).
9, Ohm, E. A. 1961 BSTJ, 40, 1065.
10. Hogg, D. C. and Wilson, R. W. Bell system Tech J., 44, 1019.
11. c.f. F. Hoyle and R. J. Taylor 1964. Nature 203, 1108. A less explicit discussion of the same notion occurs in [29].
12. Hogg, D. C. and Semplak, R. A. 1961, Bell System Technical Journal. 40, 1331 .
13. Dicke, R. H. Peebles, P. J. E. Roll, P. G. and Wilkinson, D. T. 1965, Ap. J., 142. 4 14.
14. Penzias, A. A. and Wilson R. W. 1965, Ap. J.. 142, 420.
15. Pauliny-Toth, I. I. K. and Shakeshaft, J. R. 1962, M. N. RAS.. 124, 61.
16. Adams, W. S. 1941, Ap. J. 93, 11.
17. Adams, W. S. 1943, Ap. J., 97, 105.
18. Dunham, T. Jr. 1937, PASP, 49, 26.
19. Dunham, T. Jr. 1939, Proc. Am. Phil. Soc., 81, 277.
20. Dunham, T. Jr. 1941, Publ. Am. Astron. Soc., 10, 123.
21. Dunham, T. Jr. and W. S. Adams 1937, Publ. Am. Astron. Soc., 9, 5.
22. McKellar, A. 1941, Publ. Dominion Astrophysical Observatory Victoria B. C. 7, 251.
23. G. B. Field, G. H. Herbig and J. L. Hitchcock, talk at the American Astronomical Society Meeting, 22-29 December 1965, Astronomical Journal 1966, ***71.*** 161; G. B. Field and J. I,. Hitchcock, 1966, Phys. Rev. Lett. ***16, 8*** 17.
24. Shklovsky, I. S. 1966, Astron. Circular No. 364, Acad. Sci. USSR.
25. Thaddeus, P. and Clauser, J. F. 1966, Phys. Rev. Lett. ***16, 8*** 19.
26. Roll, P. G. and Wilkinson, D. T. 1966, Physical Review Letters, ***16, 405.***
27. Howell, T. F. and Shakeshaft, J. R. 1966, Nature 210, 138.
28. Penzias, A. A. and Wilson R. W. 1967, Astron. J. 72, 315.
29. Gamow, G. 1948, Nature I62 680.
30. Alpher, R. .4. and Herman, R. C. 1949, Phys. Rev., 75, 1089.
31. Alpher, R. A. Follin, J. W. and Herman, R. C. 1953, Phys. Rev., 92, 1347.
32. Doroshkevich, A. G. and Novikov. I. D. 1964 Dokl. Akad. Navk. SSR 154, 809.
33. Doroshkevich, A. G. and Novikov, I. D. 1964 Sov. Phys. Dokl. 9, 111 .
34. Penzias, A. A. ***The Origin of the Elements,*** Nobel Prize lecture 1978.
3.5. Rees, M. J. and Sciama, D. W. 1967, Nature, 213, 374.
36. Partridge, R. B. and Wilkinson, D. T. 1967, Nature, ***215, 7*** 19.
37. Wilson, R. W. and Penzias, A. A. 1967, Science ***156, 1100.***
38. Conklin, E. K. 1969, Nature, 222, 97 I.
39. Henry, P. S. 1971, Nature, 231, 516.
40. Corey, B. E. and Wilkinson, D. T. 1976, Bull. Astron. Astrophys. Soc., 8, 351.
41. Smoot, G. F. Gorenstein, M. V. and Muller, R. A. 1977, Phys. Rev. Lett., 39, 898.
42. Boynton, P. E. and Partridge, R. B. 1973, Ap. J., 181, 243.
43. Sunyaev, R. A. and Zel'dovich, Ya B 1972, Comments Astrophys. Space Phys., 4, 173.
44. Birkinshaw, M. and Gull, S. F. 1978, Nature275 40.
45. Penzias, A. A. Jefferts, K. B. and Wilson, R. W'. 1972, Phys. Rev. Letters, 28, 772.
46. The data in Fig. 11 are all referenced by Danese and Dezotti[45] except for the 13 cm measurement of T. Otoshi 1975, IEEE Trans on Instrumentation and Meas. 24 174. I have used the millimeter measurements of Woody and Richards[48] and left off those of Robson et al[47] to avoid confusion.
47. Wilkinson, D. J. 1967, Phys. Rev. Letters. ***19,*** 1195.
48. Stokes, R. A. Partridge, R. B. and Wilkinson, D. J. 1967. Phys. Rev. Letters, 19, 1199.
49. Boynton, P. E. Stokes, R. A. **and** Wilkinson, D. J. 1968, Phys. Rev. Letters, 21, 462.
50. Boynton, P. E. and Stokes, R. A. 1974, Nature, 247, 528.
51. Robson, E. I. Vickers, D. G. Huizinga, J. S. Beckman.J. E. and Clegg, P. E. 1974, Nature ***251,*** 591.
52. Woody, D. P. and Richards, P. L. Private communication.
53. Danese, L. and DeZotti, G. 1978, Astron. **and** Astrophys.. 68, 157.

Physics 1979

SHELDON L GLASHOW, ABDUS SALAM and STEVEN WEINBERG

for their contributions to the theory of the unified weak and electromagnetic interaction between elementary particles, including inter alia the prediction of the weak neutral current

THE NOBEL PRIZE FOR PHYSICS

Speech by Professor BENGT NAGEL of the Royal Academy of Sciences. Translation from the Swedish text.

Your Majesties, Your Royal Highnesses, Ladies and Gentlemen,

This year's Nobel prize in Physics is shared equally between Sheldon Glashow, Abdus Salam and Steven Weinberg "for their contributions to the theory of the unified weak and electromagnetic interaction between elementary particles, including inter alia the prediction of the weak neutral current".

Important advances in physics often consist in relating apparently unconnected phenomena to a common cause. A classical example is Newton's introduction of the gravitational force to explain the fall of the apple and the motion of the moon around the earth. - In the 19th century it was found that electricity and magnetism are really two aspects of one and the same force, the electromagnetic interaction between charges. Electromagnetism, with the electron playing the leading part and the photon-the electromagnetic quantum of light-as the swift messenger, dominates technology and our everyday life: not only electrotechnics and electronics, but also atomic and molecular physics and hence chemical and biological processes are governed by this force.

When one began to study the atomic nucleus in the first decades of our century, two new forces were discovered: the strong and the weak nuclear forces. Unlike gravitation and electromagnetism these forces act only over distances of the order of nuclear diameters or less. The strong force keeps the nucleus together, whereas the weak force is responsible for the so called beta decays of the nucleus. Most radioactive substances used in medicine and technology are beta radioactive. The electron also participates in the weak interaction, but the principal part is played by the neutrino, a particle which is described as follows in a poem by the American writer John Updike:

"Cosmic Gall"

Neutrinos, they are very small.
They have no charge and have no mass
And do not interact at all.
The earth is just a silly ball
To them, through which they simply pass,
Like dustmaids down a drafty hall
Or photons through a sheet of glass.

At night, they enter at Nepal
And pierce the lover and his lass
From underneath the bed - you call
It wonderful; I call it crass.

The description is accurate, apart from the statement 'they do not interact at all'; they do interact through the weak force. The neutrinos of the poem, entering the earth at night at Nepal and exiting in the U.S. in a sort of reversed China syndrome, come to us from the centre of the sun. Solar energy, necessary for life on earth, is created when hydrogen is burnt to helium in the interior of the sun in a chain of nuclear reactions-even the advocates of "Solsverige" must ultimately rely on nuclear energy although it must be said that the fusion reactor Sun is well encapsuled and sufficiently relocated away from populated areas. The first ignating and moderating link in this chain, burning hydrogen to deuterium, is based on the weak force, which could then be called the Sunignator and Suntamer.

The theory which is awarded this year's prize, and which was developed in separate works by the prizewinners in the 60's, has extended and deepened our understanding of the weak force by displaying a close relationship to the electromagnetic force: these two forces emerge as different aspects of a unified electroweak interaction. This means e.g. that the electron and the neutrino belong to the same family of particles; the neutrino is the electron's little brother. Another consequence of the unified theory is that there should exist a new kind of weak interaction. It was formerly assumed that weak processes could occur only in connection with a change of identity of the electron to neutrino (or vice versa); such a process is said to proceed by a charged current, since the particle changes its charge. The theory implies that there should also be processes connected with a neutral current in which the neutrino-or else the electron-acts without changing identity. Experiments in the 70's have fully confirmed these predictions of the theory.

The importance of the new theory is first of all intrascientific. The theory has set a pattern for the description also of the strong nuclear force and for efforts to integrate further the interactions between elementary particles.

Let me end by giving an example of the intricate links which exist *between different branches of natural science.

Our body is to a large part constructed from "stardust": the elements besides hydrogen which build our cells have been formed in the interior of stars in nuclear reactions, which form a continuation of the processes taking place in our sun. According to the astrophysicists, certain heavy elements appearing in life-important enzymes and hormones -iodine and selenium are examples of such elements-can probably only be created in connection with violent explosions of giant stars, so called supernova explosions, which occur in our Galaxy once every one or two hundred

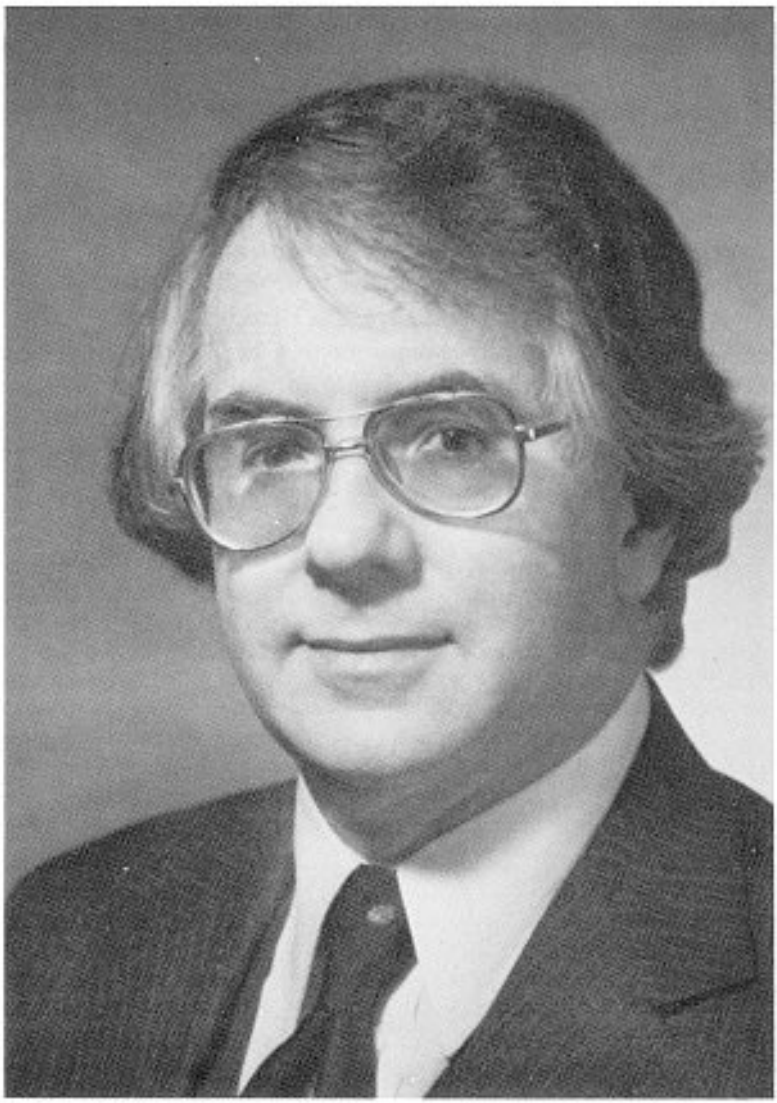
years. It is likely that neutrinos interacting via the neutral current play an important role in these explosions, in which a large part of the matter of the star is thrown out into space. Thus, for our functioning as biological beings we rely on elements formed milliards of years ago in supernova explosions, with the new kind of weak force predicted by the theory contributing in an important way; really a fascinating connection between biology, astrophysics and elementary particle physics.

Professors Sheldon Glashow, Abdus Salam, and Steven Weinberg,

In my talk I have tried to give a background to your great discoveries in the borderland between a strange but known country and the probably large unknown territory of the innermost structure of matter.

Our way of looking at this structure has changed radically in the last decade. The theory of electroweak interaction has been one of the most important forces to bring about this change of outlook.

It is a privilege and a pleasure for me to convey to you the warmest felicitations of the Royal Swedish Academy of Sciences and to invite you to receive your prizes from the hands of his Majesty the King.



Sheldon Lee Glashow

SHELDON LEE GLASHOW

My parents, Lewis Glashow and Bella née Rubin immigrated to New York City from Bobruisk in the early years of this century. Here they found the freedom and opportunity denied to Jews in Czarist Russia. After years of struggle, my father became a successful plumber, and his family could then enjoy the comforts of the middle class. While my parents never had the time or money to secure university education themselves, they were adamant that their children should. In comfort and in love, we were taught the joys of knowledge and of work well done. I only regret that neither my mother nor my father could live to see the day I would accept the Nobel Prize.

When I was born in Manhattan in 1932, my brothers Samuel and Jules were eighteen and fourteen years old. They chose careers of dentistry and medicine, to my parents' satisfaction. From an early age, I knew I would become a scientist. It may have been my brother Sam's doing. He interested me in the laws of falling bodies when I was ten, and helped my father equip a basement chemistry lab for me when I was fifteen. I became skilled in the synthesis of selenium halides. Never again would I do such dangerous research. Except for the occasional suggestion that I should become a physician and do science in my spare time, my parents always encouraged my scientific inclinations.

Among my chums at the Bronx High School of Science were Gary Feinberg and Steven Weinberg. We spurred one another to learn physics while commuting on the New York subway. Another classmate, Dan Greenberger, taught me calculus in the school lunchroom. High-school mathematics then terminated with solid geometry. At Cornell University, I again had the good fortune to join a talented class. It included the mathematician Daniel Kleitman who was to become my brother-in-law, my old classmate Steven Weinberg, and many others who were to become prominent scientists. Throughout my formal education, I would learn as much from my peers as from my teachers. So it is today among our graduate students.

I came to graduate school at Harvard University in 1954. My thesis supervisor, Julian Schwinger, had about a dozen doctoral students at a time. Getting his ear was as difficult as it was rewarding. I called my thesis "The Vector Meson in Elementary Particle Decays", and it showed an early commitment to an electroweak synthesis. When I completed my work in 1958, Schwinger and I were to write a paper summarizing our thoughts on

weak-electromagnetic unification. Alas, one of us lost the first draft of the manuscript, and that was that.

I won an NSF postdoctoral fellowship, and planned to work at the Lebedev Institute in Moscow with I. Tamm, who enthusiastically supported my proposal. I spent the tenure of my fellowship in Copenhagen at the Niels Bohr Institute (and, partly, at CERN), waiting for the Russian visa that was never to come. Perhaps all was for the best, because it was in these years (1958-60) that I discovered the SU(2) x U(1) structure of the electroweak theory. Interestingly, it was also in Copenhagen that my early work on charm with Bjorken was done. This was during a brief return to Denmark in 1964.

During my stay in Europe, I was "discovered" by Murray Gell-Mann. He presented my ideas on the algebraic structure of weak interactions to the 1960 "Rochester meeting" and brought me to Caltech. Then, he invented the eightfold way, which kept Sidney Coleman and me distracted for several years. How we found various electromagnetic formulae, yet missed the discovery of the Gell-Mann-Okubo formula and of the Cabibbo current is another story.

I became an assistant professor at Stanford University and then spent several years on the faculty of the University of California at Berkeley. During this time, I continued to exploit the phenomenological successes of flavor SU(3) and attempted to understand the departures from exact symmetry as a consequence of spontaneous symmetry breakdown. I returned to Harvard University in 1966 where I have remained except for leaves to CERN, MIT, and the University of Marseilles. Today, I am Eugene Higgins Professor of Physics at Harvard.

In 1969, John Iliopoulos and Luciano Maiani came to Harvard as research fellows. Together, we found the arguments that predicted the existence of charmed hadrons. Much of my later work was done in collaboration with Alvaro de Rújula or Howard Georgi. In early 1974, we predicted that charm would be discovered in neutrino physics or in e^+e^- annihilation. So it was. With the discovery of the J/Ψ particle, we realized that many diverse strands of research were converging on a single theory of physics. I remember once saying to Howard that if QCD is so good, it should explain the $\Sigma-\Lambda$ mass splitting. The next day he showed that it did. When we spoke, in 1974, of the unification of all elementary particle forces within a simple gauge group, and of the predicted instability of the proton, we were regarded as mad. How things change!

The wild ideas of yesterday quickly become today's dogma. This year I have been honored to participate in the inauguration of the Harvard Core Curriculum Program. My students are not, and will never be, scientists. Nonetheless, in my course "From Alchemy to Quarks" they seem to be as fascinated as I am by the strange story of the search for the ultimate constituents of matter.

I was married in 1972 to the former Joan Alexander. We live in a large old house with our four children, who attend the Brookline public schools.

Education

A.B. 1954 Cornell University
A.M. 1955 Harvard University
Ph. D. 1959 Harvard University
Married 1972 Joan Shirley Alexander

Employment

NSF Post-Doctoral Fellow 1958-60
Caltech Research Fellow 1960-61
Stanford University, Assistant Professor 1961-62
University of California, Berkeley, Associate Professor 1962-66
Harvard University, Professor 19661982
CERN, Visiting Scientist 1968
University of Marseilles, Visiting Professor 1970
MIT, Visiting Professor 1974
Brookhaven Laboratory, Consultant 1964
Texas A&M University, Visiting Professor 1982
University of Houston, Affiliated Senior Scientist, 1982-
Boston University, Distinguished Visiting Scientist, 1984-

Member

American Physical Society
Sigma Xi
American Association for the Advancement of Science
American Academy of Arts and Sciences
National Academy of Sciences

Honors

Westinghouse Science Talent Search Finalist 1950
Alfred P. Sloan Foundation Fellowship 1962-66
Oppenheimer Memorial Medal 1977
George Ledlie Award 1978

Honorary Degrees

Yeshiva University 1978
University of Aix-Marseille 1982
Adelphi University 1985
Bar-Ilan University 1988
Gustavus Augustus University 1989

TOWARDS A UNIFIED THEORY -THREADS IN A TAPESTRY

Nobel Lecture, 8 December, 1979

bY
SHELDON LEE GLASHOW
Lyman Laboratory of Physics Harvard University Cambridge, Mass., USA

INTRODUCTION

In 1956, when I began doing theoretical physics, the study of elementary particles was like a patchwork quilt. Electrodynamics, weak interactions, and strong interactions were clearly separate disciplines, separately taught and separately studied. There was no coherent theory that described them all. Developments such as the observation of parity violation, the successes of quantum electrodynamics, the discovery of hadron resonances and the appearance of strangeness were well-defined parts of the picture, but they could not be easily fitted together.

Things have changed. Today we have what has been called a "standard theory" of elementary particle physics in which strong, weak, and electromagnetic interactions all arise from a local symmetry principle. It is, in a sense, a complete and apparently correct theory, offering a qualitative description of all particle phenomena and precise quantitative predictions in many instances. There is no experimental data that contradicts the theory. In principle, if not yet in practice, all experimental data can be expressed in terms of a small number of "fundamental" masses and coupling constants. The theory we now have is an integral work of art: the patchwork quilt has become a tapestry.

Tapestries are made by many artisans working together. The contributions of separate workers cannot be discerned in the completed work, and the loose and false threads have been covered over. So it is in our picture of particle physics. Part of the picture is the unification of weak and electromagnetic interactions and the prediction of neutral currents, now being celebrated by the award of the Nobel Prize. Another part concerns the reasoned evolution of the quark hypothesis from mere whimsy to established dogma. Yet another is the development of quantum chromodynamics into a plausible, powerful, and predictive theory of strong interactions. All is woven together in the tapestry; one part makes little sense without the other. Even the development of the electroweak theory was not as simple and straightforward as it might have been. It did not arise full blown in the mind of one physic&t, nor even of three. It, too, is the result of the collective endeavor of many scientists, both experimenters and theorists.

Let me stress that I do not believe that the standard theory will long

survive as a correct and complete picture of physics. All interactions may be gauge interactions, but surely they must lie within a unifying group. This would imply the existence of a new and very weak interaction which mediates the decay of protons. All matter is thus inherently unstable, and can be observed to decay. Such a synthesis of weak, strong, and electromagnetic interactions has been called a "grand unified theory", but a theory is neither grand nor unified unless it includes a description of gravitational phenomena. We are still far from Einstein's truly grand design.

Physics of the past century has been characterized by frequent great but unanticipated experimental discoveries. If the standard theory is correct, this age has come to an end. Only a few important particles remain to be discovered, and many of their properties are alleged to be known in advance. Surely this is not the way things will be, for Nature must still have some surprises in store for us.

Nevertheless, the standard theory will prove useful for years to come. The confusion of the past is now replaced by a simple and elegant synthesis. The standard theory may survive as a part of the ultimate theory, or it may turn out to be fundamentally wrong. In either case, it will have been an important way-station, and the next theory will have to be better.

In this talk, I shall not attempt to describe the tapestry as a whole, nor even that portion which is the electroweak synthesis and its empirical triumph. Rather, I shall describe several old threads, mostly overwoven, which are closely related to my own researches. My purpose is not so much to explain who did what when, but to approach the more difficult question of why things went as they did. I shall also follow several new threads which may suggest the future development of the tapestry.

EARLY MODELS

In the 1920's, it was still believed that there were only two fundamental forces: gravity and electromagnetism. In attempting to unify them, Einstein might have hoped to formulate a universal theory of physics. However, the study of the atomic nucleus soon revealed the need for two additional forces: the strong force to hold the nucleus together and the weak force to enable it to decay. Yukawa asked whether there might be a deep analogy between these new forces and electromagnetism. All forces, he said, were to result from the exchange of mesons. His conjectured mesons were originally intended to mediate both the strong and the weak interactions: they were strongly coupled to nucleons and weakly coupled to leptons. This first attempt to unify strong and weak interactions was fully forty years premature. Not only this, but Yukawa could have predicted the existence of neutral currents. His neutral meson, essential to provide the charge independence of nuclear forces, was also weakly coupled to pairs of leptons.

Not only is electromagnetism mediated by photons, but it arises from the

requirement of local gauge invariance. This concept was generalized in 1954 to apply to non-Abelian local symmetry groups. [1] It soon became clear that a more far-reaching analogy might exist between electromagnetism and the other forces. They, too, might emerge from a gauge principle.

A bit of a problem arises at this point. All gauge mesons must be massless, yet the photon is the only massless meson. How do the other gauge bosons get their masses? There was no good answer to this question until the work of Weinberg and Salam [2] as proven by 't Hooft [3] (for spontaneously broken gauge theories) and of Gross, Wilczek, and Politzer [4] (for unbroken gauge theories). Until this work was done, gauge meson masses had simply to be put in ad hoc.

Sakurai suggested in 1960 that strong interactions should arise from a gauge principle. [5] Applying the Yang-Mills construct to the isospin-hypercharge symmetry group, he predicted the existence of the vector mesons ϱ and ω. This was the first phenomenological SU(2) X U(1) gauge theory. It was extended to local SU(3) by Gell-Mann and Ne'eman in 1961. [6] Yet, these early attempts to formulate a gauge theory of strong interactions were doomed to fail. In today's jargon, they used "flavor" as the relevant dynamical variable, rather than the hidden and then unknown variable "color". Nevertheless, this work prepared the way for the emergence of quantum chromodynamics a decade later.

Early work in nuclear beta decay seemed to show that the relevant interaction was a mixture of S, T, and P. Only after the discovery of parity violation, and the undoing of several wrong experiments, did it become clear that the weak interactions were in reality V-A. The synthesis of Feynman and Gell-Mann and of Marshak and Sudarshan was a necessary precursor to the notion of a gauge theory of weak interactions. [7] Bludman formulated the first SU(2) gauge theory of weak interactions in 1958. [8] No attempt was made to include electromagnetism. The model included the conventional charged-current interactions, and in addition, a set of neutral current couplings. These are of the same strength and form as those of today's theory in the limit in which the weak mixing angle vanishes. Of course, a gauge theory of weak interactions alone cannot be made renormalizable. For this, the weak and electromagnetic interactions must be unified.

Schwinger, as early as 1956, believed that the weak and electromagnetic interactions should be combined together into a gauge theory. [9] The charged massive vector intermediary and the massless photon were to be the gauge mesons. As his student, I accepted this faith. In my 1958 Harvard thesis, I wrote: "It is of little value to have a potentially renormalizable theory of beta processes without the possibility of a renormalizable electrodynamics. We should care to suggest that a fully acceptable theory of these interactions may only be achieved if they are treated together. . ." [10] We used the original SU(2) gauge interaction of Yang and Mills. Things had to be arranged so that the charged current, but not the neutral

(electromagnetic) current, would violate parity and strangeness. Such a theory is technically possible to construct, but it is both ugly and experimentally false. [11] We know now that neutral currents do exist and that the electroweak gauge group must be larger than SU(2).

Another electroweak synthesis without neutral currents was put forward by Salam and Ward in 1959. [12] Again, they failed to see how to incorporate the experimental fact of parity violation. Incidentally, in a continuation of their work in 196 1, they suggested a gauge theory of strong, weak, and electromagnetic interactions based on the local symmetry group SU(2) x SU(2). [13] This was a remarkable portent of the SU(3) x SU(2) x U(1) model which is accepted today.

We come to my own work [14] done in Copenhagen in 1960, and done independently by Salam and Ward. [15] We finally saw that a gauge group larger than SU(2) was necessary to describe the electroweak interactions. Salam and Ward were motivated by the compelling beauty of gauge theory. I thought I saw a way to a renormalizable scheme. I was led to the group SU(2) x U(1) by analogy with the approximate isospin-hypercharge group which characterizes strong interactions. In this model there were two electrically neutral intermediaries: the massless photon and a massive neutral vector meson which I called B but which is now known as Z. The weak mixing angle determined to what linear combination of SU(2) x U(1) generators B would correspond. The precise form of the predicted neutral-current interaction has been verified by recent experimental data. However, the strength of the neutral current was not prescribed, and the model was not in fact renormalizable. These glaring omissions were to be rectified by the work of Salam and Weinberg and the subsequent proof of renormalizability. Furthermore, the model was a model of leptons-it could not evidently be extended to deal with hadrons.

RENORMALIZABILITY

In the late 50's, quantum electrodynamics and pseudoscalar meson theory were known to be renormalizable, thanks in part to work of Salam. Neither of the customary models of weak interactions - charged intermediate vector bosons or direct four-fermion couplings - satisfied this essential criterion. My thesis at Harvard, under the direction of Julian Schwinger, was to pursue my teacher's belief in a unified electroweak gauge theory. I had found some reason to believe that such a theory was less singular than its alternatives. Feinberg, working with charged intermediate vector mesons discovered that a certain type of divergence would cancel for a special value of the meson anomalous magnetic moment.[16] It did not correspond to a "minimal electromagnetic coupling", but to the magnetic properties demanded by a gauge theory. Tzou Kuo-Hsien examined the zero-mass limit of charged vector meson electrodynamics.[17] Again, a sensible result is obtained only for a very special choice of the magnetic dipole moment and electric quadrupole moment, just the values assumed in a

gauge theory. Was it just coincidence that the electromagnetism of a charged vector meson was least pathological in a gauge theory?

Inspired by these special properties, I wrote a notorious paper.[18] I alleged that a softly-broken gauge theory, with symmetry breaking provided by explicit mass terms, was renormalizable. It was quickly shown that this is false.

Again, in 1970, Iliopoulos and I showed that a wide class of divergences that might be expected would cancel in such a gauge theory.[19] We showed that the naive divergences of order $(\alpha\Lambda^4)^n$ were reduced to "merely" $(\alpha\Lambda^2)^n$, where Λ is a cut-off momentum. This is probably the most difficult theorem that Iliopoulos or I had even proven. Yet, our labors were in vain. In the spring of 1971, Veltman informed us that his student Gerhart 't Hooft had established the renormalizability of spontaneously broken gauge theory.

In pursuit of renormalizability, I had worked diligently but I completely missed the boat. The gauge symmetry is an exact symmetry, but it is hidden. One must not put in mass terms by hand. The key to the problem is the idea of spontaneous symmetry breakdown: the work of Goldstone as extended to gauge theories by Higgs and Kibble in 1964.[20] These workers never thought to apply their work on formal field theory to a phenomenologically relevant model. I had had many conversations with Goldstone and Higgs in 1960. Did I neglect to tell them about my SU (2)xU (1) model, or did they simply forget?

Both Salam and Weinberg had had considerable experience in formal field theory, and they had both collaborated with Goldstone on spontaneous symmetry breaking. In retrospect, it is not so surprising that it was they who first used the key. Their SU (2)X U (1) gauge symmetry was spontaneously broken. The masses of the W and Z and the nature of neutral current effects depend on a single measurable parameter, not two as in my unrenormalizable model. The strength of the neutral currents was correctly predicted. The daring Weinberg-Salam conjecture of renormalizability was proven in 1971. Neutral currents were discovered in 1973[21], but not until 1978 was it clear that they had just the predicted properties.[22]

THE STRANGENESS-CHANGING NEUTRAL CURRENT

I had more or less abandoned the idea of an electroweak gauge theory during the period 1961- 1970. Of the several reasons for this, one was the failure of my naive foray into renormalizability. Another was the emergence of an empirically successful description of strong interactions - the SU(3) unitary symmetry scheme of Gell-Mann and Ne'eman. This theory was originally phrased as a gauge theory, with $\varrho, \omega,$ and K* as gauge mesons. It was completely impossible to imagine how both strong and weak interactions could be gauge theories: there simply wasn't room enough for commuting structures of weak and strong currents. Who could foresee the success of the quark model, and the displacement of SU(3)

from the arena of flavor to that of color? The predictions of unitary symmetry were being borne out - the predicted Ω^- was discovered in 1964. Current algebra was being successfully exploited. Strong interactions dominated the scene.

When I came upon the SU(2)xU(1) model in 1960, I had speculated on a possible extension to include hadrons. To construct a model of leptons alone seemed senseless: nuclear beta decay, after all, was the first and foremost problem. One thing seemed clear. The fact that the charged current violated strangeness would force the neutral current to violate strangeness as well. It was already well known that strangeness-changing neutral currents were either strongly suppressed or absent. I concluded that the Z^0 had to be made very much heavier than the W. This was an arbitrary but permissible act in those days: the symmetry breaking mechanism was unknown. I had "solved" the problem of strangeness-changing neutral currents by suppressing all neutral currents: the baby was lost with the bath water.

I returned briefly to the question of gauge theories of weak interactions in a collaboration with Gell-Mann in 1961.[23] From the recently developing ideas of current algebra we showed that a gauge theory of weak interactions would inevitably run into the problem of strangeness-changing neutral currents. We concluded that something essential was missing. Indeed it was. Only after quarks were invented could the idea of the fourth quark and the GIM mechanism arise.

From 196 1 to 1964, Sidney Coleman and I devoted ourselves to the exploitation of the unitary symmetry scheme. In the spring of 1964, I spent a short leave of absence in Copenhagen. There, Bjorken and I suggested that the Gell-Mann-Zweig-system of three quarks should be extended to four.[24] (Other workers had the same idea at the same time).[25] We called the fourth quark the charmed quark. Part of our motivation for introducing a fourth quark was based on our mistaken notions of hadron spectroscopy. But we also wished to enforce an analogy between the weak leptonic current and the weak hadronic current. Because there were two weak doublets of leptons, we believed there had to be two weak doublets of quarks as well. The basic idea was correct, but today there seem to be three doublets of quarks and three doublets of leptons.

The weak current Bjorken and I introduced in 1964 was precisely the GIM current. The associated neutral current, as we noted, conserved strangeness. Had we inserted these currents into the earlier electroweak theory, we would have solved the problem of strangeness-changing neutral currents. We did not. I had apparently quite forgotten my earlier ideas of electroweak synthesis. The problem which was explicitly posed in 1961 was solved, in principle, in 1964. No one, least of all me, knew it. Perhaps we were all befuddled by the chimera of relativistic SU(6), which arose at about this time to cloud the minds of theorists.

Five years later, John Iliopoulos, Luciano Maiani and I returned to the question of strangeness-changing neutral currents.[26] It seems incredible

that the problem was totally ignored for so long. We argued that unobserved effects (a large K_1K_2 mass difference; decays like $K \to \pi\nu\bar{\nu}$; etc.) would be expected to arise in any of the known weak interaction models: four fermion couplings; charged vector meson models; or the electroweak gauge theory. We worked in terms of cut-offs, since no renormalizable theory was known at the time. We showed how the unwanted effects would be eliminated with the conjectured existence of a fourth quark. After languishing for a decade, the problem of the selection rules of the neutral current was finally solved. Of course, not everyone believed in the predicted existence of charmed hadrons.

This work was done fully three years after the epochal work of Weinberg and Salam, and was presented in seminars at Harvard and at M. I. T. Neither I, nor my coworkers, nor Weinberg, sensed the connection between the two endeavors. We did not refer, nor were we asked to refer, to the Weinberg-Salam work in our paper.

The relevance became evident only a year later. Due to the work of 't Hooft, Veltman, Benjamin Lee, and Zinn-Justin, it became clear that the Weinberg-Salam *ansatz* was in fact a renormalizable theory. With GIM, it was trivially extended from a model of leptons to a theory of weak interactions. The ball was now squarely in the hands of the experimenters. Within a few years, charmed hadrons and neutral currents were discovered, and both had just the properties they were predicted to have.

FROM ACCELERATORS TO MINES

Pions and strange particles were discovered by passive experiments which made use of the natural flux of cosmic rays. However, in the last three decades, most discoveries in particle physics were made in the active mode, with the artificial aid of particle accelerators. Passive experimentation stagnates from a lack of funding and lack of interest. Recent developments in theoretical particle physics and in astrophysics may mark an imminent rebirth of passive experimentation. The concentration of virtually all high-energy physics endeavors at a small number of major accelerator laboratories may be a thing of the past.

This is not to say that the large accelerator is becoming extinct; it will remain an essential if not exclusive tool of high-energy physics. Do not forget that the existence of Z^0 at ~ 100 GeV is an essential but quite untested prediction of the electroweak theory. There will be additional dramatic discoveries at accelerators, and these will not always have been predicted in advance by theorists. The construction of new machines like LEP and ISABELLE is mandatory.

Consider the successes of the electroweak synthesis, and the fact that the only plausible theory of strong interactions is also a gauge theory. We must believe in the ultimate synthesis of strong, weak, and electromagnetic interactions. It has been shown how the strong and electroweak gauge groups may be put into a larger but simple gauge group.[27] Grand

unification - perhaps along the lines of the original SU (5) theory of Georgi and me - must be essentially correct. This implies that the proton, and indeed all nuclear matter, must be inherently unstable. Sensitive searches for proton decay are now being launched. If the proton lifetime is shorter than 10^{32} years, as theoretical estimates indicate, it will not be long before it is seen to decay.

Once the effect is discovered (and I am sure it will be), further experiments will have to be done to establish the precise modes of decay of nucleons. The selection rules, mixing angles, and space-time structure of a new class of effective four-fermion couplings must be established. The heroic days of the discovery of the nature of beta decay will be repeated.

The first generation of proton decay experiments is cheap, but subsequent generations will not be. Active and passive experiments will compete for the same dwindling resources.

Other new physics may show up in elaborate passive experiments. Today's theories suggest modes of proton decay which violate both baryon number and lepton number by unity. Perhaps this AB = ΔL = 1 law will be satisfied. Perhaps AB = − AL transitions will be seen. Perhaps, as Pati and Salam suggest, the proton will decay into three leptons. Perhaps two nucleons will annihilate in AB = 2 transitions. The effects of neutrino oscillations resulting from neutrino masses of a fraction of an election volt may be detectable. "Superheavy isotopes" which may be present in the Earth's crust in small concentrations could reveal themselves through their multi-GeV decays. Neutrino bursts arising from distant astronomical catastrophes may be seen. The list may be endless or empty. Large passive experiments of the sort now envisioned have never been done before. Who can say what results they may yield?

PREMATURE ORTHODOXY

The discovery of the J/Ψ in 1974 made it possible to believe in a system involving just four quarks and four leptons. Very quickly after this a third charged lepton (the tau) was discovered, and evidence appeared for a third Q= -1/3 quark (the b quark). Both discoveries were classic surprises. It became immediately fashionable to put the known fermions into families or generations:

$$\begin{pmatrix} u & \nu_e \\ d & e \end{pmatrix} \quad \begin{pmatrix} c & \nu_\mu \\ s & \mu \end{pmatrix} \quad \begin{pmatrix} t & \nu_\tau \\ b & \tau \end{pmatrix}.$$

The existence of a third Q = 2/3 quark (the t quark) is predicted. The Cabibbo-GIM scheme is extended to a system of six quarks. The three family system is the basis to a vast and daring theoretical endeavor. For example, a variety of papers have been written putting experimental constraints on the four parameters which replace the Cabibbo angle in a

six quark system. The detailed manner of decay of particles containing a single b quark has been worked out. All that is wanting is experimental confirmation. A new orthodoxy has emerged, one for which there is little evidence, and one in which I have little faith.

The predicted t quark has not been found. While the upsilon mass is less than 10 GeV, the analogous tt particle, if it exists at all, must be heavier than 30 GeV. Perhaps it doesn't exist.

Howard Georgi and I, and other before us, have been working on models with no t quark.[28] We believe this unorthodox view is as attractive as its alternative. And, it suggests a number of exciting experimental possibilities.

We assume that b and τ share a quantum number, like baryon number, that is essentially exactly conserved. (Of course, it may be violated to the same extent that baryon number is expected to be violated.) Thus, the b,τ system is assumed to be distinct from the lighter four quarks and four leptons. There is, in particular, no mixing between b and d or s. The original GIM structure is left intact. An additional mechanism must be invoked to mediate b decay, which is not present in the SU(3) x SU(2) x U(1) gauge theory.

One possibility is that there is an additional SU(2) gauge interaction whose effects we have not yet encountered. It could mediate such decays of b as these

$$b \rightarrow \tau^+ + (e^- \text{ or } \mu^-) + (d \text{ or } s).$$

All decays of b would result in the production of a pair of leptons, including a τ^+ or its neutral partner. There are other possibilities as well, which predict equally bizarre decay schemes for b-matter. How the b quark decays is not yet known, but it soon will be.

The new SU(2) gauge theory is called upon to explain CP violation as well as b decay. In order to fit experiment, three additional massive neutral vector bosons must exist, and they cannot be too heavy. One of them can be produced in e^+e^- annihilation, in addition to the expected Z^0. Our model is rife with experimental predictions, for example: a second Z^0, a heavier version of b and of τ, the production of τ b in e p collisions, and the existence of heavy neutral unstable leptons which may be produced and detected in e^+e^- or in up collisions.

This is not the place to describe our views in detail. They are very speculative and probably false. The point I wish to make is simply that it is too early to convince ourselves that we know the future of particle physics. There are too many points at which the conventional picture may be wrong or incomplete. The SU(3)xSU(2)xU(1) gauge theory with three families is certainly a good beginning, not to accept but to attack, extend, and exploit. We are far from the end.

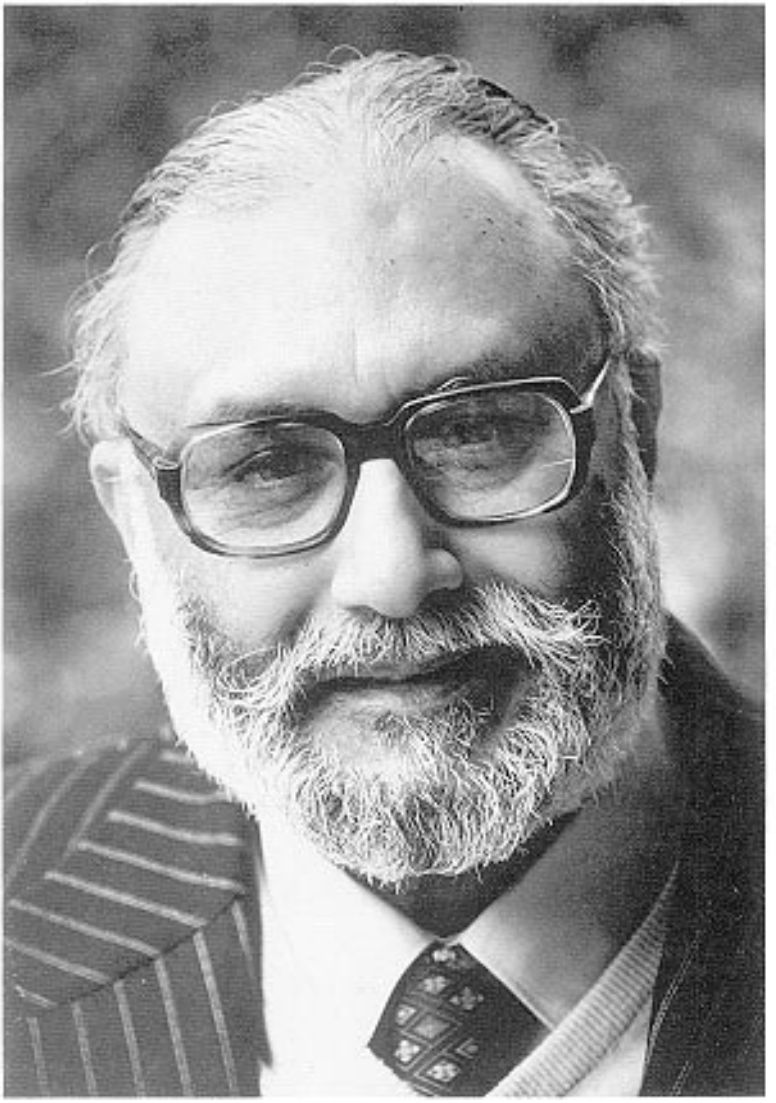
ACKNOWLEDGEMENTS

I wish to thank the Nobel Foundation for granting me the greatest honor to which a scientist may aspire. There are many without whom my work would never have been. Let me thank my scientific collaborators, especially James Bjorken, Sidney Coleman, Alvaro De Rújula, Howard Georgi, John Iliopoulos, and Luciano Maiani; the Niels Bohr Institute and Harvard University for their hospitality while my research on the electroweak interaction was done: Julian Schwinger for teaching me how to do scientific research in the first place; the Public School System of New York City, Cornell University, and Harvard University for my formal education; my high-school friends Gary Feinberg and Steven Weinberg for making me learn too much too soon of what I might otherwise have never learned at all; my parents and my two brothers for always encouraging a child's dream to be a scientist. Finally, I wish to thank my wife and my children for the warmth of their love.

REFERENCES

1. Yang, C. N. and Mills, R., Phys. Rev. 96, 191 (1954). Also, Shaw, R., unpublished.
2. Weinberg, S., Phys. Rev. Letters 19, 1264 (1967); Salam, A., in ***Elementary Particle Physics*** (ed. Svartholm, N.; Almqvist and Wiksell; Stockholm; 1968).
3. 't Hooft, G., Nuclear Physics B 33, 173 and B 35, 167 (1971); Lee, B. W., and Zinn-Justin, J., Phys. Rev. D 5, pp. 3121-3160 (1972); 't Hooft, G., and Veltman M., Nuclear Physics B 44, 189 (1972).
4. Gross, D. J. and Wilczek, F., Phys. Rev. Lett. 30, 1343 (1973); Politzer, H. D., Phys. Rev. Lett. 30, 1346 (1973).
5. Sakurai, J. J., Annals of Physics II, 1 (1960).
6. Cell-Mann, M., and Ne'eman, Y., ***The Eightfold Way*** (Benjamin, W. A., New York, 1964).
7. Feynman, R., and Cell-Mann, M., Phys. Rev. 109, 193 (1958); Marshak, R., and Sudarshan, E. C. G., Phys. Rev. 109, 1860 (1958).
8. Bludman, S., Nuovo Cimento Ser. 10 9, 433 (1958).
9. Schwinger, J., Annals of Physics 2, 407 (1958).
10. Glashow, S. L., Harvard University Thesis, p. 75 (1958).
11. Georgi, H., and Glashow, S. L., Phys. Rev. Letters 28, 1494 (1972).
12. Salam, A., and Ward, J., Nuovo Cimento ***11, 568*** (1959).
13. Salam, A., and Ward, J., Nuovo Cimento ***19,*** 165 (1961).
14. Glashow, S. L., Nuclear Physics 22, 579 (1961).
15. Salam, A., and Ward, J., Physics Letters 13, 168 (1964).
16. Feinberg, G., Phys. Rev. 110, 1482 (1958).
17. Tzou Kuo-Hsien, Comptes rendus 245, 289 (1957).
18. Glashow, S. L., Nuclear Physics ***10,*** 107 (1959).
19. Glashow, S. L., and Iliopoulos J., Phys. Rev. D 3, 1043 (197 1).
20. Many authors are involved with this work: Brout, R., Englert, F., Goldstone, J., Guralnik, G., Hagen, C., Higgs, P., Jona-Lasinio, G., Kibble, T., and Nambu, Y.
21. Hasert, F. J., et al., Physics Letters 46B, 138 (1973) and Nuclear Physics B 73, 1 (1974). Benvenuti, A., et al., Phys. Rev. Letters 32, 800 (1974).
22. Prescott, C. Y., et al., Phys. Lett. B 77, 347 (1978).
23. Cell-Mann, M., and Glashow, S. L., Annals of Physics 15, 437 (1961).
24. Bjorken, J., and Glashow, S. L., Physics Letters II, 84 (1964).

25. Amati, D., et al., Nuovo Cimento 34, 1732 (A 64); Hara, Y. Phys. Rev. ***134,*** B 701 (1964); Okun, L. B., Phys. Lett. ***12,*** 250 (1964); Maki, Z., and Ohnuki, Y., Progs. Theor. Phys. 32, 144 (1964); Nauenberg, M., (unpublished); Teplitz, V., and Tarjanne, P., Phys. Rev. Lett. ***11, 447 (1963).***
26. Glashow, S. L., Iliopoulos, J., and Maiani, L., Phys. Rev. D 2, 1285 (1970).
27. Georgi, H., and Glashow, S. L., Phys. Rev. Letters 33, 438 (1974).
28. Georgi, H., and Glashow, S. L., Harvard Preprint HUTP-79/A 053.



Abdus Salam عبدالسلام

ABDUS SALAM

Abdus Salam was born in Jhang, a small town in what is now Pakistan, in 1926. His father was an official in the Department of Education in a poor farming district. His family has a long tradition of piety and learning.

When he cycled home from Lahore, at the age of 14, after gaining the highest marks ever recorded for the Matriculation Examination at the University of the Panjab, the whole town turned out to welcome him. He won a scholarship to Government College, University of the Panjab, and took his MA in 1946. In the same year he was awarded a scholarship to St. John's College, Cambridge, where he took a BA (honours) with a double First in mathematics and physics in 1949 In 1950 he received the Smith's Prize from Cambridge University for the most outstanding pre-doctoral contribution to physics. He also obtained a PhD in theoretical physics at Cambridge; his thesis, published in 1951, contained fundamental work in quantum electrodynamics which had already gained him an international reputation.

Salam returned to Pakistan in 1951 to teach mathematics at Government College, Lahore, and in 1952 became head of the Mathematics Department of the Panjab University. He had come back with the intention of founding a school of research, but it soon became clear that this was impossible. To pursue a career of research in theoretical physics he had no alternative at that time but to leave his own country and work abroad. Many years later he succeeded in finding a way to solve the heartbreaking dilemma faced by many young and gifted theoretical physicists from developing countries. At the ICTP, Trieste, which he created, he instituted the famous "Associateships" which allowed deserving young physicists to spend their vacations there in an invigorating atmosphere, in close touch with their peers in research and with the leaders in their own field, losing their sense of isolation and returning to their own country for nine months of the academic year refreshed and recharged.

In 1954 Salam left his native country for a lectureship at Cambridge, and since then has visited Pakistan as adviser on science policy. His work for Pakistan has, however, been far-reaching and influential. He was a member of the Pakistan Atomic Energy Commission, a member of the Scientific Commission of Pakistan and was Chief Scientific Adviser to the President from 1961 to 1974.

Since 1957 he has been Professor of Theoretical Physics at Imperial College, London, and since 1964 has combined this position with that of Director of the ICTP, Trieste.

For more than forty years he has been a prolific researcher in theoretical elementary particle physics. He has either pioneered or been associated with all the important developments in this field, maintaining a constant and fertile flow of brilliant ideas. For the past thirty years he has used his academic reputation to add weight to his active and influential participation in international scientific affairs. He has served on a number of United Nations committees concerned with the advancement of science and technology in developing countries.

To accommodate the astonishing volume of activity that he undertakes, Professor Salam cuts out such inessentials as holidays, parties and entertainments. Faced with such an example, the staff of the Centre find it very difficult to complain that they are overworked.

He has a way of keeping his administrative staff at the ICTP fully alive to the real aim of the Centre - the fostering through training and research of the advancement of theoretical physics, with special regard to the needs of developing countries. Inspired by their personal regard for him and encouraged by the fact that he works harder than any of them, the staff cheerfully submit to working conditions that would be unthinkable here at the International Atomic Energy Agency in Vienna (IAEA). The money he received from the Atoms for Peace Medal and Award he spent on setting up a fund for young Pakistani physicists to visit the ICTP. He uses his share of the Nobel Prize entirely for the benefit of physicists from developing countries and does not spend a penny of it on himself or his family.

Abdus Salam is known to be a devout Muslim, whose religion does not occupy a separate compartment of his life; it is inseparable from his work and family life. He once wrote: "The Holy Quran enjoins us to reflect on the verities of Allah's created laws of nature; however, that our generation has been privileged to glimpse a part of His design is a bounty and a grace for with I render thanks with a humble heart."

The biography was written *by Miriam,Leti, now at IAEA, Vienna, who was at* one *time on the staff of ICTP (International Centre For Theoretical Physics, Trieste).*

Date of birth: 29 January, 1926
Place of birth: Jhang, Pakistan

Educational Career

Government College, Jhang and Lahore (1938- 1946)	M.A. (Panjab University)
Foundation Scholar, St. John's College, Cambridge (1946- 1949)	B.A. Honours Double first in Mathematics (Wrangler) and Physics
Cavendish Laboratory, Cambridge (1952)	Ph.D. in Theoretical Physics

Awarded Smith's Prize by the University of Cambridge for "the most outstanding pre-doctoral contribution to Physics" (1950)

DSc. Honoris Causa:

Panjab University, Lahore (1957)
University of Edinburgh (1971)
Panjab University, Lahore (Pakistan) (1957)
University of Edinburgh (UK) (1971)
University of Trieste (Italy) (1979)
University of Islamabad (Pakistan) (1979)
Universidad National de Ingenieria, Lima (Peru) (1980)
University of San Marcos, Lima (Peru) (1980)
National University of San Antonio Abad, Cuzco (Peru) (1980)
Universidad Simon Bolivar, Caracas (Venezuela) (1980)
University of Wroclow (Poland) (1980)
Yarmouk University (Jordan) (1980)
University of Istanbul (Turkey) (1980)
Guru Nanak Dev University, Amritsar (India) (1981)
Muslim University, Aligarh (India) (1981)
Hindu University, Banaras (India) (1981)
University of Chittagong (Bangladesh) (1981)
University of Bristol (UK) (1981)
University of Maiduguri (Nigeria) (1981)
University of the Philippines, Quezon City (Philippines) (1982)
University of Khartoum (Sudan) (1983)
Universidad Complutense de Madrid (Spain) (1983)
City College, City University of New York (USA) (1984)
University of Nairobi (Kenya) (1984)
Universidad National de Cuyo (Argentina) (1985)
Universidad National de la Plata (Argentina) (1985)
University of Cambridge (UK) (1985)
University of Goteborg (Sweden) (1985)
Kliment Ohridski University of Sofia (Bulgaria) (1986)
University of Glasgow (UK) (1986)
University of Science and Technology, Hefei (China) (1986)
The City University, London (UK) (1986)
Panjab University, Chandigarh (India) (1987)
Medicina Alternativa, Colombo (Sri Lanka) (1987)
National University of Benin, Cotonou (Benin) (1987)
University of Exeter (UK) (1987)
University of Gent (Belgium) (1988)
"Creation" International Association of Scientists and Intelligentsia (USSR) (1989)
Bendel State University, Ekpoma (Nigeria) (1990)
University of Ghana (Ghana) (1990)
University of Warwick (UK) (1991)
University of Dakar (Senegal) (1991)

University of Tucuman (Argentina) (1991)
University of Lagos (Nigeria) (1992)

Awards

Hopkins Prize (Cambridge University) for "the most outstanding contribution to Physics during 1957-1958"
Adams Prize (Cambridge University) (1958)
First recipient of Maxwell Medal and Award (Physical Society, London) (1961)
Hughes Medal (Royal Society, London) (1964)
Atoms for Peace Medal and Award (Atoms for Peace Foundation) (1968)
J. Robert Oppenheimer Memorial Medal and Prize (University of Miami) (1971)
Guthrie Medal and Prize (1976)
Matteuci Medal (Accademia Nazionale dei Lincei, Rome) (1978)
John Torrence Tate Medal (American Institute of Physics) (1978)
Royal Medal (Royal Society, London) (1978)
Einstein Medal (UNESCO, Paris) (1979)
Shri R.D. Birla Award (India Physics Association) (1979)
Josef Stefan Medal (Josef Stefan Institue, Ljublijana) (1980)
Gold Medal for Outstanding Contributions to Physics (Czechoslovak Academy of Sciences, Prague) (1981)
Lomonosov Gold Medal (USSR Academy of Sciences) (1983)
Copley Medal (Royal Society, London) (1990)

Appointments

Professor, Government College and Panjab University, Lahore (1951- 1954)
Elected Fellow St. John's College, Cambridge (1951-1956)
Member, Institute of Advanced Study, Princeton (1951)
Lecturer, Cambridge University (1954-1956)
Professor of Theoretical Physics, London University, Imperial College, London, since 1957
Director, International Centre for Theoretical Physics, Trieste, since 1964
Elected (First) Fellow of the Royal Society, London, from Pakistan (1959)
Elected, Foreign Member of the Royal Swedish Academy of Sciences (1970)
Elected, Foreign Member of the American Academy of Arts and Sciences (1971)
Elected, Foreign Member, USSR Academy of Sciences (1971)
Elected, Honorary Fellow St. John's College, Cambridge (1971)
Elected, Foreign Associate, USA National Academy of Sciences (Washington) (1979)
Elected, Foreign Member, Accademia Nazionale dei Lincei (Rome) (1979)
Elected, Foreign Member, Accademia Tiberina (Rome) (1979)
Elected, Foreign Member, Iraqi Academy (Baghdad) (1979)
Elected, Honorary Fellow, Tata Institute of Fundamental Research (Bombay) (1979)

Elected, Honorary Member, Korean Physics Society (Seoul) (1979)
Elected, Foreign Member, Academy of the Kingdom of Morocco (Rabat) (1980)
Elected, Foreign Member, Accademia Nazionale delle Scienze dei XL (Rome) (1980)
Elected, Member, European Academy of Science, Arts and Humanities (Paris) (1980)
Elected, Associate Member, Josef Stefan Institute (Ljublijana) (1980)
Elected, Foreign Fellow, Indian National Science Academy (New Delhi) (1980)
Elected, Fellow, Bangladesh Academy of Sciences (Dhaka) (1980)
Elected, Member, Pontifical Academy of Sciences (Vatican City) (1981)
Elected, Corresponding Member, Portuguese Academy of Sciences (Lisbon) (1981)
Founding Member, Third World Academy of Sciences (1983)
Elected, Corresponding Member, Yugoslav Academy of Sciences and Arts (Zagreb) (1983)
Elected, Honorary Fellow, Ghana Academy of Arts and Sciences (1984)
Elected, Honorary Member, Polish Academy of Sciences (1985)
Elected, Corresponding Member, Academia de Ciencias Medicas, Fisicas y Naturales de Guatemala (1986)
Elected, Fellow, Pakistan Academy of Medical Sciences (1987)
Elected, Honorary Fellow, Indian Academy of Sciences (Bangalore) (1988)
Elected, Distinguished International Fellow of Sigma Xi (1988)
Elected, Honorary Member, Brazilian Mathematical Society (1989)
Elected, Honorary Member, National Academy of Exact, Physical and Natural Sciences, Argentina (1989)
Elected, Honorary Member, Hungarian Academy of Sciences (1990)
Elected, Member, Academia Europaea (1990)

Orders and other Distinctions
Order of Andres Bello (Venezuela) (1980)
Order of Istiqlal (Jordan) (1980)
Cavaliere di Gran Croce dell'Ordine al Merito della Repubblica Italiana (1980)
Honorary Knight Commander of the Order of the British Empire (1989)

Awards for contributions towards peace and promotion of international scientific collaboration
Atoms for Peace Medal and Award (Atoms for Peace Foundation) (1968)
Peace Medal (Charles University, Prague) (1981)
Premio Umberto Biancomano (Italy) (1986)
Dayemi International Peace Award (Bangladesh) (1986)
First Edinburgh Medal and Prize (Scotland) (1988)
"Genoa" International Development of Peoples Prize (Italy) (1988)
Catalunya International Prize (Spain) (1990)

United Nations Assignments

Scientific Secretary, Geneva Conferences on Peaceful Uses of Atomic Energy (1955 and 1958)

Member, United Nations Advisory Committee on Science and Technology (1964-1975)

Member, United Nations Panel and Foundation Committee for the United Nations University (1970-1973)

Chairman, United Nations Advisory Committee on Science and Technology (1971-1972)

Member, Scientific Council, SIPRI, Stockholm International Peace Research Institute (1970)

Vice President, International Union of Pure and Applied Physics (1972-1978)

Pakistan Assignments

Member, Atomic Energy Commission, Pakistan (1958-1974)

Adviser, Education Commission, Pakistan (1959)

Member, Scientific Commission, Pakistan (1959)

Chief Scientific Adviser, President of Pakistan (1961-1974)

President, Pakistan Association for Advancement of Science (1961-1962)

Chairman, Pakistan Space and Upper Atmosphere Committee (1961-1964)

Governor from Pakistan to the International Atomic Energy Agency (1962-1963)

Member, National Science Council, Pakistan (1963-1975)

Member, Board of Pakistan Science Foundation (1973-1977)

Pakistani Awards

Sitara-i-Pakistan (S.Pk.)

Pride of Performance Medal and Award (1959)

GAUGE UNIFICATION OF FUNDAMENTAL FORCES

Nobel lecture, 8 December, 1979

bY
ABDUS SALAM
Imperial College of Science and Technology, London, England
and International Centre for Theoretical Physics, Trieste, Italy

Introduction: In June 1938, Sir George Thomson, then Professor of Physics at Imperial College, London, delivered his 1937 Nobel Lecture. Speaking of Alfred Nobel, he said: "The idealism which permeated his character led him to . . . (being) as much concerned with helping science as a whole, as individual scientists. . . . The Swedish people under the leadership of the Royal Family and through the medium of the Royal Academy of Sciences have made Nobel Prizes one of the chief causes of the growth of the prestige of science in the eyes of the world . . . As a recipient of Nobel's generosity, I owe sincerest thanks to them as well as to him."

I am sure I am echoing my colleagues' feelings as well as my own, in reinforcing what Sir George Thomson said-in respect of Nobel's generosity and its influence on the growth of the prestige of science. Nowhere is this more true than in the developing world. And it is in this context that I have been encouraged by the Permanent Secretary of the Academy - Professor Carl Gustaf Bernhard-to say a few words before I turn to the scientific part of my lecture.

Scientific thought and its creation is the common and shared heritage of mankind. In this respect, the history of science, like the history of all civilization, has gone through cycles. Perhaps I can illustrate this with an actual example.

Seven hundred and sixty years ago, *a* young Scotsman left his native glens to travel south to Toledo in Spain. His name was Michael, his goal to live and work at the Arab Universities of Toledo and Cordova, where the greatest of Jewish scholars, Moses bin Maimoun, had taught a generation before.

Michael reached Toledo in 1217 AD. Once in Toledo, Michael formed the ambitious project of introducing Aristotle to Latin Europe, translating not from the original Greek, which he did not know, but from the Arabic translation then taught in Spain. From Toledo, Michael travelled to Sicily, to the Court of Emperor Frederick II.

Visiting the medical school at Salerno, chartered by Frederick in 1231, Michael met the Danish physician, Henrik Harpestraeng - later to become Court Physician of King Erik Plovpenning. Henrik had come to Salerno to compose his treatise on blood-letting and surgery. Henrik's

sources were the medical canons of the great clinicians of Islam, Al-Razi and Avicenna, which only Michael the Scot could translate for him.

Toledo's and Salerno's schools, representing as they did the finest synthesis of Arabic, Greek, Latin and Hebrew scholarship, were some of the most memorable of international assays in scientific collaboration. To Toledo and Salerno came scholars not only from the rich countries of the East and the South, like Syria, Egypt, Iran and Afghanistan, but also from developing lands of the West and the North like Scotland and Scandinavia. Then, as now, there were obstacles to this international scientific concourse, with an economic and intellectual disparity between different parts of the world. Men like Michael the Scot or Henrik Harpestraeng were singularities. They did not represent any flourishing schools of research in their own countries. With all the best will in the world their teachers at Toledo and Salerno doubted the wisdom and value of training them for advanced scientific research. At least one of his masters counselled young Michael the Scot to go back to clipping sheep and to the weaving of woollen cloth.

In respect of this cycle of scientific disparity, perhaps I can be more quantitative. George Sarton, in his monumental five-volume History of Science chose to divide his story of achievement in sciences into ages, each age lasting half a century. With each half century he associated one central figure. Thus 450 BC-400 BC Sarton calls the Age of Plato; this is followed by half centuries of Aristotle, of Euclid, of Archimedes and so on. From 600 AD to 650 AD is the Chinese half century of Hsiian Tsang, from 650 to 700 AD that of I-Ching, and then from 750 AD to 1100 AD-350 years continuously-it is the unbroken succession of the Ages of Jabir, Khwarizmi, Razi, Masudi, Wafa, Biruni and Avicenna, and then Omar Khayam-Arabs, Turks, Afghans and Persians-men belonging to the culture of Islam. After 1100 appear the first Western names; Gerard of Cremona, Roger Bacon-but the honours are still shared with the names of Ibn-Rushd (Averroes), Moses Bin Maimoun, Tusi and Ibn-Nafi-the man who anticipated Harvey's theory of circulation of blood. No Sarton has yet chronicled the history of scientific creativity among the pre-Spanish Mayas and Aztecs, with their invention of the zero, of the calendars of the 'moon and Venus and of their diverse pharmacological discoveries, including quinine, but the outline of the story is the same-one of undoubted superiority to the Western contemporary correlates.

After 1350, however, the developing world loses out except for the occasional flash of scientific work, like that of Ulugh Beg-the grandson of Timurlane, in Samarkand in 1400 AD; or of Maharaja Jai Singh of Jaipur in 1720-who corrected the serious errors of the then Western tables of eclipses of the sun and the moon by as much as six minutes of arc. As it was, Jai Singh's techniques were surpassed soon after with the development of the telescope in Europe. As a contemporary Indian chronicler wrote: "With him on the funeral pyre, expired also all science in the East." And this brings us to this century when the cycle begun by Michael the Scot

turns full circle, and it is we in the developing world who turn to the Westwards for science. As Al-Kindi wrote 1100 years ago: "It is fitting then for us not to be ashamed to acknowledge and to assimilate it from whatever source it comes to us. For him who scales the truth there is nothing of higher value than truth itself; it never cheapens nor abases him."

Ladies and Gentlemen,

It is in the spirit of Al-Kindi that I start my lecture with a sincere expression of gratitude to the modern equivalents of the Universities of Toledo and Cordova, which I have been privileged to be associated with-Cambridge, Imperial College, and the Centre at Trieste.

I. FUNDAMENTAL PARTICLES, FUNDAMENTAL FORCES AND GAUGE UNIFICATION

The Nobel lectures this year are concerned with a set of ideas relevant to the gauge unification of the electromagnetic force with the weak nuclear force. These lectures coincide nearly with the 100th death-anniversary of Maxwell, with whom the first unification of forces (electric with the magnetic) matured and with whom gauge theories originated. They also nearly coincide with the 100th anniversary of the birth of Einstein-the man who gave us the vision of an ultimate unification of all forces.

The ideas of today started more than twenty years ago, as gleams in several theoretical eyes. They were brought to predictive maturity over a decade back. And they started to receive experimental confirmation some six years ago.

In some senses then, our story has a fairly long background in the past. In this lecture I wish to examine some of the theoretical gleams of today and ask the question if these may be the ideas to watch for maturity twenty years from now.

From time immemorial, man has desired to comprehend the complexity of nature in terms of as few elementary concepts as possible. Among his quests-in Feynman's words-has been the one for "wheels within wheels"-the task of natural philosophy being to discover the innermost wheels if any such exist. A second quest has concerned itself with the fundamental forces which make the wheels go round and enmesh with one another. The greatness of gauge ideas-of gauge field theories-is that they reduce these two quests to just one; elementary particles (described by relativistic quantum fields) are representations of certain charge operators, corresponding to gravitational mass, spin, flavour, colour, electric charge and the like, while the fundamental forces are the forces of attraction or repulsion between these same charges. A third quest seeks for a *unification* between the charges (and thus of the forces) by searching for a single entity, of which the various charges are components in the sense that they can be transformed one into the other.

But are all fundamental forces gauge forces? Can they be understood as such, in terms of charges-and their corresponding currents-only? And if they are, how many charges? What unified entity are the charges components of?

What is the nature of charge? Just as Einstein comprehended the nature of gravitational charge in terms of space-time curvature, can we comprehend the nature of the other charges-the nature of the entire unified set, *as a set,* in terms of something equally profound? This briefly is the dream, much reinforced by the verification of gauge theory predictions. But before I examine the new theoretical ideas on offer for the future in this particular context, I would like your indulgence to range over a one-man, purely subjective, perspective in respect of the developments of the last twenty years themselves. The point I wish to emphasize during this part of my talk was well made by G. P. Thomson in his 1937 Nobel Lecture. G. P. said ". . . The goddess of learning is fabled to have sprung full grown from the brain of Zeus, but it is seldom that a scientific conception is born in its final form, or owns a single parent. More often it is the product of a series of minds, each in turn modifying the ideas of those that came before, and providing material for those that come after."

II. THE EMERGENCE OF SPONTANEOUSLY BROKEN SU(2)xU(1) GAUGE THEORY

I started physics research thirty years ago as an experimental physicist in the Cavendish, experimenting with tritium-deuterium scattering. Soon I knew the craft of experimental physics was beyond me-it was the sublime quality of patience-patience in accumulating data, patience with recalcitrant equipment-which I sadly lacked. Reluctantly I turned my papers in, and started instead on quantum field theory with Nicholas Kemmer in the exciting department of P. A. M. Dirac.

The year 1949 was the culminating year of the Tomonaga-Schwinger-Feynman-Dyson reformulation of renormalized Maxwell-Dirac gauge theory, and its triumphant experimental vindication. A field theory must be renormalizable and be capable of being made free of infinities-first discussed by Waller-if perturbative calculations with it are to make any sense. More-a renormalizable theory, with no dimensional parameter in its interaction term, connotes *somehow* that the fields represent "structureless" elementary entities. With Paul Matthews, we started on an exploration of renormalizability of meson theories. Finding that renormalizability held only for spin-zero mesons and that these were the only mesons that empirically existed then, (pseudoscalar pions, invented by Kemmer, following Yukawa) one felt thrillingly euphoric that with the triplet of pions (considered as the carriers of the strong nuclear force between the proton-neutron doublet) one might resolve the dilemma of the origin of this particular force which is responsible for fusion and fission. By the same token, the so-called weak nuclear force-the force responsible for β-radioactivity (and described then by Fermi's non-renormalizable theory) had to be mediated by some unknown spin-zero mesons if it was to be renormalizable, If massive charged spin-one mesons were to mediate this interaction, the theory would be non-renormalizable, according to the ideas then.

Now this agreeably renormalizable spin-zero theory for the pion was a field theory, but not a gauge field theory. There was no conserved charge

which determined the pionic interaction. As is well known, shortly after the theory was elaborated, it was found wanting. The $(\frac{3}{2},\frac{3}{2})$ resonance Δ effectively killed it off as a fundamental theory; we were dealing with a complex dynamical system, not "structureless" in the held-theoretic sense.

For me, personally, the trek to gauge theories as candidates for fundamental physical theories started in earnest in September 1956-the year I heard at the Seattle Conference Professor Yang expound his and Professor Lee's ideas[1] on the possibility of the hitherto sacred principle of left-right symmetry, being violated in the realm of the *weak nuclearforce.* Lee and Yang had been led to consider abandoning left-right symmetry for weak nuclear interactions as a possible resolution of the (τ,θ) puzzle. I remember travelling back to London on an American Air Force (MATS) transport flight. Although I had been granted, for that night, the status of a Brigadier or a Field Marshal-I don't quite remember which-the plane was very uncomfortable; full of crying service-men's children-that is, the children were crying, not the servicemen. I could not sleep. I kept reflecting on why Nature should violate left-right symmetry in weak interactions. Now the hallmark of most weak interactions was the involvement in radioactivity phenomena of Pauli's neutrino. While crossing over the Atlantic, came back to me a deeply perceptive question about the neutrino which Professor Rudolf Peierls had asked when he was examining me for a Ph. D. a few years before. Peierls' question was: "The photon mass is zero because of Maxwell's principle of a gauge symmetry for electromagnetism; tell me, why is the neutrino mass zero?" I had then felt somewhat uncomfortable at Peierls. asking for a Ph. D. viva, a question of which he himselfsaid he did not know the answer. But during that comfortless night the answer came. The analogue for the neutrino, of the gauge symmetry for the photon existed; it had to do with the masslessness of the neutrino, with symmetry under the γ_5 transformation [2] (1ater christened "chiral symmetry"). The existence of this symmetry for the massless neutrino must imply a combination $(1+\gamma_5)$ or $(1-\gamma_5)$ for the neutrino interactions. Nature had the choice of an aesthetically satisfying but a left-right symmetry violating theory, with a neutrino which travels exactly with the velocity of light; or alternatively a theory where left-right symmetry is preserved, but the neutrino has a tiny mass-some ten thousand times smaller than the mass of the electron.

It appeared at that time clear to me what choice Nature must have made. Surely, left-right symmetry must be sacrificed in all neutrino interactions. I got off the plane the next morning, naturally very elated. I rushed to the Cavendish, worked out the Michel parameter and a few other consequences ofy, symmetry, rushed out again, got into a train to Birmingham where Peierls lived. To Peierls I presented my idea; he had asked the original question; could he approve of the answer? Peierls' reply was kind but firm. He said "I do not believe left-right symmetry is violated in weak nuclear forces at all. I would not touch such ideas with a pair of tongs." Thus rebuffed in Birmingham, like Zuleika Dobson, I wondered where I could go next and the obvious place was CERN in Geneva, with Pauli-the father of the 'neutrino-nearby in Zurich.

At that time CERN lived in a wooden hut just outside Geneva airport. Besides my friends, Prentki and d'Espagnat, the hut contained a gas ring on which was cooked the staple diet of CERN-Entrecôte à la creme. The hut also contained Professor Villars of MIT, who was visiting Pauli the same day in Zurich. I gave him my paper. He returned the next day with a message from the Oracle; "Give my regards to my friend Salam and tell him to think ofsomething better". This was discouraging, but I was compensated by Pauli's excessive kindness a few months later, when Mrs. Wu's[3], Lederman's[4] and Telegdi's[5] experiments were announced showing that left-right symmetry was indeed violated and ideas similar to mine about chiral symmetry were expressed independently by Landau[6] and Lee and Yang[7]. I received Pauli's first somewhat apologetic letter on 24 January 1957. Thinking that Pauli's spirit should by now be suitably crushed, I sent him two short notes[8] I had written in the meantime. These contained suggestions to extend chiral symmetry to electrons and muons, assuming that their masses were a consequence of what has come to be known as dynamical spontaneous symmetry breaking. With chiral symmetry for electrons, muons and neutrinos, the only mesons that could mediate weak decays of the muons would have to carry spin one. Reviving thus the notion of charged intermediate spin-one bosons, one could then postulate for these a type of gauge invariance which I called the "neutrino gauge". Pauli's reaction was swift and terrible. He wrote on 30th January 1957, then on 18 February and later on 11, 12 and 13 March: "I am reading (along the shores of Lake Zurich) in bright sunshine quietly your paper . ." "I am very much startled on the title of your paper 'Universal Fermi interaction'

For quite a while I have for myself the rule if a theoretician says *universal* it just means pure nonsense. This holds particularly in connection with the Fermi interaction, but otherwise too, and now you too, Brutus, my son, come with this word. ..." Earlier, on 30 January, he had written "There is a similarity between this type of gauge invariance and that which was published by Yang and Mills . . . In the latter, of course, no γ_5 was used in the exponent." and he gave me the full reference of Yang and Mills' paper; (Phys. Rev. 96, 191 (1954)). I quote from his letter: "However, there are dark points in your paper regarding the vector field B,. If the rest mass is infinite (or very large), how can this be compatible with the gauge transformation $B_\mu \rightarrow B_\mu - \partial_\mu \Lambda$?" and he concludes his letter with the remark: "Every reader will realize that you deliberately conceal here something and will ask you the same questions". Although he signed himself "With friendly regards", Pauli had forgotten his earlier penitence. He was clearly and rightly on the warpath.

Now the fact that I was using gauge ideas similar to the Yang-Mills (non-Abelian SU(2)-invariant) gauge theory was no news to me. This was because the Yang-Mills theory[9] (which married gauge ideas of Maxwell with the internal symmetry SU(2) of which the proton-neutron system constituted a doublet had been independently invented by a Ph. D. pupil of mine, Ronald Shaw,[10] at Cambridge at the same time as Yang and Mills had written. Shaw's work is relatively unknown; it remains buried in his Cambridge thesis. I must admit I was taken aback by 'Pauli's fierce prejudice against

universalism-against what we would today call unification of basic forces-but I did not take this too seriously. I felt this was a legacy of the exasperation which Pauli had always felt at Einstein's somewhat formalistic attempts at unifying gravity with electromagnetism-forces which in Pauli's phrase "cannot be joined-for God hath rent them asunder". But Pauli was absolutely right in accusing me of darkness about the problem of the masses of the Yang-Mills fields; one could not obtain a mass without wantonly destroying the gauge symmetry one had started with. And this was particularly serious in this context, because Yang and Mills had conjectured the desirable renormalizability of their theory with a proof which relied heavily and exceptionally on the masslessness of their spin-one intermediate mesons. The problem was to be solved only seven years later with the understanding of what is now known as the Higgs mechanism, but I will come back to this later.

Be that as it may, the point I wish to make from this exchange with Pauli is that already in early 1957, just after the first set of parity experiments, many ideas coming to fruition now, had started to become clear. These are:

1. First was the idea of chiral symmetry leading to a V-A theory. In those early days my humble suggestion [2], [8] of this was limited to neutrinos, electrons and muons only, while shortly after, that year, Sudarshan and Marshak,[11] Feynman and Gell-Mann,[12] and Sakurai[13] had the courage to postulate γ_5 symmetry for baryons as well as leptons, making this into a universal principle of physics.'

Concomitant with the (V-A) theory was the result that if weak interactions are mediated by intermediate mesons, these must curry spin one.

2. Second, was the idea of spontaneous breaking of chiral symmetry to generate electron and muon masses: though the price which those latter-day Shylocks, Nambu and Jona-Lasinio[14] and Goldstone[15] exacted for this (i.e. the appearance of massless scalars), was not yet appreciated.
3. And finally, though the use of a Yang-Mills-Shaw (non-Abelian) gauge theory for describing spin-one intermediate charged mesons was suggested already in 1957, the giving of masses to the intermediate bosons through spontaneous symmetry breaking, in a manner to preserve the renormalizability of the theory, was to be accomplished only during a long period of theoretical development between 1963 and 1971.

Once the Yang-Mills-Shaw ideas were accepted as relevant to the charged weak currents-to which the charged intermediate mesons were coupled in this theory-during 1957 and 1958 was raised the question of what was the third component of the SU(2) triplet, of which the charged weak currents were the two members. There were the two alternatives: the electroweak unification suggestion, where the electromagnetic current was assumed to be this third component; and the rival suggestion that the third component was a neutral current unconnected with electroweak unification. With hindsight, I shall

[1] Today we believe protons and neutrons are composites of quarks, so that γ_5 symmetry is now postulated for the elementary entities of today--the quarks.

call these the Klein[16] (1938) and the Kemmer[17] (1937) alternatives. The Klein suggestion, made in the context of a Kaluza-Klein five-dimensional space-time, is a real tour-de-force; it combined two hypothetical spin-one charged mesons with the photon in one multiplet, deducing from the compactification of the fifth dimension, a theory which looks like Yang-Mills-Shaw's. Klein intended his charged mesons for ***strong*** interactions, but if we read charged *weak* mesons for Klein's ***strong*** ones, one obtains the theory independently suggested by Schwinger[18] (1957), though Schwinger, unlike Klein, did not build in any non-Abelian gauge aspects. With just these non-Abelian Yang-Mills gauge aspects very much to the fore, the idea of uniting weak interactions with electromagnetism was developed by Glashow[19] and Ward and myself[20] in late 1958. The rival Kemmer suggestion of a global SU(2)-invariant triplet of weak charged and neutral currents was independently suggested by Bludrnan[21] (1958) in a gauge context and this is how matters stood till 1960.

To give you the flavour of, for example, the year 1960, there is a paper written that year of Ward and myself[22] with the statement: "Our basic postulate is that it should be possible to generate strong, weak and electromagnetic interaction terms with all their correct symmetry properties (as well as with clues regarding their relative strengths) by making local gauge transformations on the kinetic energy terms in the free Lagrangian for all particles. This is the statement of an ideal which, in this paper at least, is only very partially realized". I am not laying a claim that we were the only ones who were saying this, but I just wish to convey to you the temper of the physics of twenty years ago-qualitatively no different today from then. But what a quantitative difference the next twenty years made, first with new and far-reaching developments in theory-and then, thanks to CERN, Fermilab, Brookhaven, Argonne, Serpukhov and SLAG in testing it!

So far as theory itself is concerned, it was the next seven years between 1961-67 which were the crucial years of quantitative comprehension of the phenomenon of spontaneous symmetry breaking and the emergence of the SU(2) ***X*** U(1) theory in a form capable of being tested. The story is well known and Steve Weinberg has already spoken about it. So I will give the barest outline. First there was the realization that the two alternatives mentioned above a pure electromagnetic current versus a pure neutral current-Klein-Schwinger versus Kemmer-Bludman-were not alternatives; they were complementary. As was noted by Glashow[23] and independently by Ward and myself[24], both types of currents and the corresponding gauge particles (W', Z^0 and γ) were needed in order to build a theory that could simultaneously accommodate parity violation for weak and parity conservation for the electromagnetic phenomena. Second, there was the influential paper of Goldstone[25] in 1961 which, utilizing a non-gauge self-interaction between scalar particles, showed that the price of spontaneous breaking of a continuous internal symmetry was the appearance of zero 'mass scalars-a result foreshadowed earlier by Nambu. In giving a proof of this theorem[26] with Goldstone I collaborated with Steve Weinberg, who spent a year at Imperial College in London.

I would like to pay here a most sincerely felt tribute to him and to Sheldon Glashow for their warm and personal friendship.

I shall not dwell on the now well-known contributions of Anderson [27], Higgs[28], Brout & Englert[29], Guralnik, Hagen and Kibble[30] starting from 1963, which showed the way how spontaneous symmetry breaking using spin-zero fields could generate vector-meson masses, defeating Goldstone at the same time. This is the so-called Higgs mechanism.

The final steps towards the electroweak theory were taken by Weinberg[31] and myself [32] (with Kibble at Imperial College tutoring me about the Higgs phenomena). We were able to complete the present formulation of the spontaneously broken SU(2)xU(1) theory so far as leptonic weak interactions were concerned-with one parameter sin'8 describing all weak and electromagnetic phenomena and with one isodoublet Higgs multi let. An account of this development was given during the contribution [32] to the Nobel Symposium (organized by Nils Svartholm and chaired by Lamek Hulthén held at Gothenburg after some postponements, in early 1968). As is well known, we did not then, and still do not, have a prediction for the scalar Higgs mass.

Both Weinberg and I suspected that this theory was likely to be renormalizable.[2] Regarding spontaneously broken Yang-Mills-Shaw theories in general this had earlier been suggested by Englert, Brout and Thiry[29]. But this subject was not pursued seriously except at Veltman's school at Utrecht, where the proof of renormalizability was given by 't Hooft[33] in 1971. This was elaborated further by that remarkable physicist the late Benjamin Lee [34], working with Zinn Justin, and by 't Hooft and Veltman [35]. This followed on the earlier basic advances in Yang-Mills calculational technology by Feynman[36], DeWitt[37], Faddeev and Popov[38], Mandelstam[39], Fradkin and Tyutin[40] Boulware[41] Taylor[42], Slavnov[43], Strathdee[44] and Salam. In Coleman's eloquent phrase "'t Hooft's work turned the Weinberg-Salam frog into an enchanted prince". Just before had come the GIM (Glashow, Iliopoulos and Maiani) mechanism[45], emphasising that the existence of the fourth charmed quark (postulated earlier by several authors) was essential to the natural resolution of the dilemma posed by the absence of strangeness--violating currents. This tied in naturally with the understandin of the Steinberger-Schwinger-Rosenberg-Bell-Jackiw-Adler anomaly [46] and its removal for SU (2) x U (1) by the parallelism of four quarks and four leptons, pointed out by Bouchiat, Iliopoulos and Meyer and independently by Gross and Jackiw.[47]

[2] When I was discussing the final version of the SU(2) X U(1) theory and its possible renormalizability in Autumn 1967 during a post-doctoral course of lectures at Imperial College, Nino Zichichi from CERN happened to be present. I was delighted because Zichichi had been badgering me since 19.58 with persistent questioning of what theoretical avail his precise measurements on (g-2) for the muon as well as those of the muon lifetime were, when not only the magnitude of the electromagnetic corrections to weak decays was uncertain, but also conversely the effect of non-renormalizable weak interactions on "renormalized" electromagnetism was so unclear.

If one has kept a count, I have so far mentioned around fifty theoreticians. As a failed experimenter, I have always felt envious of the ambience of large experimental teams and it gives me the greatest pleasure to acknowledge the direct or the indirect contributions of the "series of minds" to the spontaneously broken SU (2) ***X*** U (1) gauge theory. My profoundest personal appreciation goes to my collaborators at Imperial College, Cambridge, and the Trieste Centre, John Ward, Paul Matthews, Jogesh Pati, John Strathdee, Tom Kibble and to Nicholas Kemmer.

In retrospect, what strikes me most about the early part of this story is how uninformed all of us were, not only of each other's work, but also of work done earlier. For example, only in 1972 did I learn of Kemmer's paper written at Imperial College in 1937.

Kemmer's argument essentially was that Fermi's weak theory was not globally SU(2) invariant and should be made so-though not for its own sake but as a prototype for strong interactions. Then this year I learnt that earlier, in 1936, Kemmer's Ph. D. supervisor, Gregor Wentzel[48], had introduced (the yet undiscovered) analogues of lepto-quarks, whose mediation could give rise to neutral currents after a Fierz reshuffle. And only this summer, Cecilia Jarlskog at Bergen rescued Oscar Klein's paper from the anonymity of the Proceedings of the International Institute of Intellectual Cooperation of Paris, and we learnt of his anticipation of a theory similar to Yang-Mills-Shaw's long before these authors. As I indicated before, the interesting point is that Klein was using his triplet, of two charged mesons plus the photon, not to describe weak interaction but for strong nuclear force unification with the electromagnetic-something our generation started on only in 1972-and not yet experimentally verified. Even in this recitation I am sure I have inadvertantly left off some names of those who have in some way contributed to SU (2) ***x*** U (1). Perhaps the moral is that not unless there is the prospect of quantitative verification, does a qualitative idea make its impress in physics.

And this brings me to experiment, and the year of the Gargamelle[49]. I still remember Paul Matthews and I getting off the train at Aix-en-Provence for the 1973 European Conference and foolishly deciding to walk with our rather heavy luggage to the student hostel where we were billeted. A car drove from behind us, stopped, and the driver leaned out. This was Musset whom I did not know well personally then. He peered out of the window and said: "Are you Salam?" I said "Yes". He said: "Get into the car. I have news for you. We have found neutral currents." I will not say whether I was more relieved for being given a lift because of our heavy luggage or for the discovery of neutral currents. At the Aix-en-Provence meeting that great and modest man, Lagarrigue, was also present and the atmosphere was that of a carnival-at least this is how it appeared to me. Steve Weinberg gave the rapporteur's talk with T. D. Lee as the chairman. T. D. was kind enough to ask me to comment after Weinberg finished. That summer Jogesh Pati and I had predicted proton decay within the context of what is now called grand unification and in the flush of this excitement I am afraid I ignored weak neutral currents as a subject which had already come to a successful conclusion, and concentrated on

speaking of the possible decays of the proton. I understand now that proton decay experiments are being planned in the United States by the Brookhaven, Irvine and Michigan and the Wisconsin-Harvard groups and also by a European collaboration to be mounted in the Mont Blanc Tunnel Garage No. 17. The later quantitative work on neutral currents at CERN, Fermilab., Brookhaven, Argonne and Serpukhov is, of course, history, but a special tribute is warranted to the beautiful SLAC-Yale-CERN experiment [50] of 1978 which exhibited the effective Z^0-photon interference in accordance with the predictions of the theory. This was foreshadowed by Barkov ***et al's*** experiments [51] at Novosibirsk in the USSR in their exploration of parity violation in the atomic potential for bismuth. There is the apocryphal story about Einstein, who was asked what he would have thought if experiment had not confirmed the light deflection predicted by him. Einstein is supposed to have said, "Madam, I would have thought the Lord has missed a most marvellous opportunity." I believe, however, that the following quote from Einstein's Herbert Spencer lecture of 1933 expresses his, my colleagues' and my own views more accurately. "Pure logical thinking cannot yield us any knowledge of the empirical world; all knowledge of reality starts from experience and ends in it." This is exactly how I feel about the Gargamelle-SLAC experience.

III. THE PRESENT AND ITS PROBLEMS

Thus far we have reviewed the last twenty years and the emergence of SU (2) ***X*** U (1) with the twin developments of a gauge theory of basic interactions, linked with internal symmetries, and of the spontaneous breaking of these symmetries. I shall first summarize the situation as we believe it to exist now and the immediate problems. Then we turn to the future.

1. To the level ofenergies explored, we believe that the following sets ofparticles are "structureless" (in a field-theoretic sense) and, at least to the level of energies explored hitherto, constitute the elementary entities of which all other objects are made.

		$SU_c(3)$ triplets	
Family I	quarks	$\begin{Bmatrix} u_R, u_Y, u_B \\ d_R, d_Y, d_B \end{Bmatrix}$	leptons $\begin{pmatrix} \nu_e \\ e \end{pmatrix}$ SU(2) doublets
Family II	quarks	$\begin{Bmatrix} c_R, c_Y, c_B \\ s_R, s_Y, s_B \end{Bmatrix}$	leptons $\begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}$,,
Family III	quarks	$\begin{Bmatrix} t_R, t_Y, t_B \\ b_R, b_Y, b_B \end{Bmatrix}$	leptons $\begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}$,,

Together with their antiparticles each family consists of 15 or 16 two-component fermions (15 or 16 depending on whether the neutrino is massless or not). The third family is still conjectural, since the top quark (t_R, t_Y, t_B) has not yet been discovered. Does this family really follow the pattern of the other two? Are there more families? Does the fact that the families are replicas of each other imply that Nature has discovered a dynamical stability about a system

of 15 (or 16) objects, and that by this token there is a more basic layer of structure underneath?[52]

2. Note that quarks come in three colours; Red (R), Yellow (Y) and Blue (B). Parallel with the electroweak SU(2) x U(l), a ***gauge*** field' theory (SU, (3)) of strong (quark) interactions (quantum chromodynamics, QCD)[53] has emerged which gauges the three colours. The indirect discovery of the (eight) gauge bosons associated with QCD (gluons), has already been surmised by the groups at DESY.[54]
3. All known baryons and mesons are singlets of colour $SU_c(3)$. This has led to a hypothesis that colour is always confined. One of the major unsolved problems of field theory is to determine if QCD-treated non-perturbatively-is capable of confining quarks and gluons.
4. In respect of the electroweak SU(2) XU(1), all known experiments on weak and electromagnetic phenomena below 100 GeV carried out to date agree with the theory which contains one theoretically undetermined parameter sin'8 = 0.230+0.009.[55] The predicted values of the associated gauge boson (W' and Z") masses are: $m_{W'} \approx 77\text{-}84$ GeV, $m_{Z} \approx 89\text{-}95$ GeV, for $0.25 \geqslant \sin\% \geqslant 0.21$.
5. Perhaps the most remarkable measurement in electroweak physics is that of the parameter $\rho = \left(\frac{m_W}{m_Z \cos\theta}\right)^2$ Currently this has been determined from the ratio of neutral to charged current cross-sections. The predicted value $\rho = 1$ for weak *iso-doublet Higgs* is to be compared with the experimental[4] p = 1.00+0.02.
6. Why does Nature favour the simplest suggestion in SU(2)xU(l) theory of the Higgs scalars being iso-doublet?[5] Is there just one physical Higgs?

[3] "To my mind the most striking feature of theoretical physics in the last thirty-six years is the fact that not a single new theoretical idea of a fundamental nature has been successful. The notions of relativistic quantum theory have in every instance proved stronger than the revolutionary ideas of a great number of talented physicists. We live in a dilapidated house and we seem to be unable to move out. The difference between this house and a prison is hardly noticeable" --Res Jost (1963) in Praise of Quantum Field Theory (Siena European Conference).

'The one-loop radiative corrections to ρ suggest that the maximum mass of leptons contributing top is less than 100 GeV.[56]

[5] To reduce the arbitrariness of the Higgs couplings and to motivate their iso-doublet character, one suggestion is to use supersymmetry. Supersymmetry is a Fermi-Bose symmetry, so that iso-doublet leptons like (ν_e, e) or (ν_μ, μ) in a super-symmetric theory must be accompanied in the same multiplet by iso-doublet Higgs.

Alternatively, one may identify the Higgs as composite fields associated with bound states of a yet new level of elementary particles and new (so-called techni-colour) forces (Dimopoulos & Susskind[57], Weinberg[58] and 't Hooft) of which, at present low energies, we have no cognisance and which may manifest themselves in the 1-100 TeV range. Unfortunately, both these ideas at first sight appear to introduce complexities, though in the context of a wider theory, which spans energy scales up to much higher masses, a satisfactory theory of the Higgs phenomena, incorporating these, may well emerge.

Of what mass? At present the Higgs interactions with leptons, quarks as well as their self-interactions are non-gauge interactions. For a three-family (6-quark) model, 21 out of the 26 parameters needed, are attributable to the Higgs interactions. Is there a basic principle, as compelling and as economical as the gauge principle, which embraces the Higgs sector? Alternatively, could the Higgs phenomenon itself be a manifestation of a dynamical breakdown of the gauge symmetry.'

7. Finally there is the problem of the families; is there a distinct SU(2) for the first, another for the second as well as a third SU(2), with spontaneous symmetry breaking such that the SU(2) apprehended by present experiment is a diagonal sum of these "family" SU(2)'s? To state this in another way, how far in energy does the e-μ universality (for example) extend? Are there more[59] Z^0 than just one, effectively differentially coupled to the e and the μ systems? (If there are, this will constitute mini-modifications of the theory, but not a drastic revolution of its basic ideas.)

In the next section 1 turn to a direct extrapolation of the ideas which went into the electroweak unification, so as to include strong interactions as well. Later I shall consider the more drastic alternatives which may be needed for the unification of all forces (including gravity)-ideas which have the promise of providing a deeper understanding of the charge concept. Regretfully, by the same token, 1 must also become more technical and obscure for the non-specialist. I apologize for this. The non-specialist may sample the flavour of the arguments in the next section (Sec. IV), ignoring the Appendices and then go on to Sec. V which is perhaps less technical.

IV. DIRECT EXTRAPOLATION FROM THE ELECTROWEAK TO THE ELECTRONUCLEAR

4.1 The three ideas

The three main ideas which have gone into the electronuclear-also called grand-unification of the electroweak with the *strong* nuclear force (and which date back to the period 1972-1974), are the following:

1. First: the psychological break (for us) of grouping quarks and leptons in the *same* multiplet of a unifying group G, suggested by Pati and myself in 1972[60]. The group G must contain $SU(2)xU(l)xSU_c(3)$; must be simple, if all quantum numbers (flavour, colour, lepton, quark and family numbers) are to be automatically quantized and the resulting gauge theory asymptotically free.
2. Second: an extension, proposed by Georgi and Glashow (1974)[61] which places not only (left-handed) quarks and leptons but also their antiparticles in the same multiplet of the unifying group.

Appendix I displays some examples of the unifying groups presently considered.

Now a gauge theory based on a "simple" (or with discrete symmetries, a "semi-simple") group G contains one basic gauge constant. This constant

would manifest itself physically above the "grand unification mass" M, exceeding all particle masses in the theory-these themselves being generated (if possible) hierarchially through a suitable spontaneous symmetry-breaking mechanism.

3. The third crucial development was by Georgi, Quinn and Weinberg (1974)[62] who showed how, using renormalization group ideas, one could relate the observed low-energy couplings $\alpha(\mu)$, and $\alpha_s(\mu)$ ($\mu \sim 100$ GeV) to the magnitude of the grand unifying mass M and the observed value of $\sin^2\theta(\mu)$; ($\tan\theta$ is the ratio of the U(1) to the SU(2) couplings).
4. If one extrapolates with Jowett[6], that nothing essentially new can possibly be discovered-i.e. one assumes that there are no new features, no new forces, or no new "types" of particles to be discovered, till we go beyond the grand unifying energy M-then the Georgi, Quinn, Weinberg method leads to a startling result: this featureless "plateau" with no "new physics" heights to be scaled stretches to fantastically high energies. More precisely, if $\sin^2\theta(\mu)$ is as large as 0.23, then the grand unifying mass M cannot be smaller than 1.3×10^{13} GeV.[63] (Compare with Planck mass $m_p \approx 1.2\times10^{19}$ GeV related to Newton's constant where gravity must come in.)[7] The result follows from the formula[63],[64]

$$\frac{11\alpha}{3\pi}\,\ell n\,\frac{M}{\mu} = \frac{\sin^2\theta(M)-\sin^2\theta(\mu)}{\cos^2\theta(M)}, \qquad \text{(I)}$$

if it is assumed that sin%(M)-the magnitude of $\sin^2\theta$ for energies of the order of the unifying mass M-equals 3/8 (see Appendix II).

This startling result will be examined more closely in Appendix II. I show there that it is very much a consequence of the assumption that the SU (2) X U(1) symmetry survives intact from the low regime energies μ right up to the grand unifying mass M. I will also show that there already is some experimental indication that this assumption is too strong, and that there may be likely peaks of new physics at energies of 10 TeV upwards (Appendix II).

"The universal urge to extrapolate from what we know to-day and to believe that nothing new can possibly be discovered, is well expressed in the following:

"I come first, My name is Jowett
I am the Master of this College,
Everything that is, I know it
If I don't, it isn't knowledge"

-The Balliol Masque.

'On account of the relative proximity of $M \approx 10^{13}$ GeV to m_p (and the hope of eventual unification with gravity), Planck mass m_p is now the accepted "natural" mass scale in Particle Physics. With this large mass as the input, the great unsolved problem of Grand Unification is the "natural" emergence of mass hierarchies (m_p, am,., a'm,...) or $m_p \exp(-c_n/\alpha)$, where c_n's are constants.

$\left(\frac{m_e}{m_P} \sim 10^{-22}.\right)$

4.2 Tests of electronuclear grand unification

The most characteristic prediction from the existence of the ELECTRONUCLEAR force is proton decay, first discussed in the context of grand unification at the Aix-en-Provence Conference (1973)[65]. For "semi-simple" unifying groups with multiplets containing quarks and leptons only, (but no antiquarks nor antileptons) the lepto-quark composites have masses (determined by renormalization group arguments), of the order of $\approx 10^5$- 10^6 GeV[66]. For such theories the characteristic proton decays (proceeding through exchanges of *three* lepto-quarks) conserve quark number+lepton number, i.e. $P = qqq \rightarrow \ell\ell\ell, \tau_P \sim 10^{29} - 10$"" years. On the contrary, for the "simple" unifying family groups like SU(5)[61] or SO(10)[67] (with multiplets containing antiquarks and antileptons) proton decay proceeds through an exchange of *one* lepto-quark into an antilepton (plus pions etc.) $(P \rightarrow \bar{\ell})$.

An intriguing possibility in this context is that investigated recently for the maximal unifying group SU(16)--the largest group to contain a 16-fold fermionic $(q, \ell, \bar{q}, \bar{\ell})$. This can permit four types of decay modes: $P \rightarrow 3\bar{\ell}$ as well as $P \rightarrow \bar{\ell}, P \rightarrow \ell$ (e.g. $P \rightarrow \ell^- + \pi^+ + \pi^+$) and $P \rightarrow 3\ell$ (e.g. $N \rightarrow 3\nu + \pi^0$, $P \rightarrow 2\nu + e^+ + \pi^0$), the relative magnitudes of these alternative decays being model-dependent on how precisely SU(16) breaks down to SU(3) x SU(2) x U(1). Quite clearly, it is the central fact of the existence of the proton decay for which the present generation of experiments must be designed, rather than for any specific type of decay modes.

Finally, grand unifying theories predict mass relations like:[68]

$$\frac{m_d}{m_e} = \frac{m_s}{m_\mu} = \frac{m_b}{m_\tau} \approx 2.8$$

for 6 (or at most 8) flavours *below the unification mass.* The important remark for proton decay and for mass relations of the above type as well as for an understanding of baryon excess[69] in the Universe[8], is that for the present *these are essentially characteristic of the fact of grand unification-rather than of specific models.*

"Yet each man kills the thing he loves" sang Oscar Wilde in his famous Ballad of Reading Gaol. Like generations of physicists before us, some in our generation also (through a direct extrapolation of the electroweak gauge methodology to the electronuclear)--and with faith in the assumption of

[8]The calculation of baryon excess in the Universe--arising from a combination of CP and baryon number violations-has recently been claimed to provide teleological arguments for grand unification. For example, Sanopoulos[70] has suggested that the "existence of human beings to measure the ratio n_B/n_γ (where n_B is the numbers of baryons and n_γ the numbers of photons in the Universe) necessarily imposes severe bounds on this quantity: i.e. $10^{-11} \approx (m_e/m_p)^{1/2} < n_B/n_\gamma < 10^{-4} (\approx 0(\alpha^2))$''. Of importance in deriving these constraints are the upper (and lower) bound on the numbers of flavours (≈ 6) deduced (1) from mass relations above, (2) from cosmological arguments which seek to limit the numbers of massless neutrinos, (3) from asymptotic freedom and (4) from numerous (one-loop) radiative calculations. It is clear that lack of accelerators as we move up in energy scale will force particle physics to reliance on teleology and cosmology (which in Landau's famous phrase is "often wrong, but never in doubt").

no "new physics", which leads to a grand unifying mass ~ 10^{13} GeV--are beginning to believe that the end of the problems of elementarity as well as of fundamental forces is nigh. They may be right, but before we are carried away by this prospect, it is worth stressing that even for the simplest grand unifying model (Georgi and Glashow's SU(5) with just two Higgs (a 5 and a 24)), the number of presently *ad hoc* parameters needed by the model is still unwholesomely large-22, to compare with 26 of the six-quark model based on the humble SU(2)xU(1)XSU$_c$(3). We cannot feel proud.

V. ELEMENTARITY: UNIFICATION WITH GRAVITY AND NATURE OF CHARGE

In some of the remaining parts of this lecture I shall be questioning two of the notions which have gone into the direct extrapolation of Sec. IV--first, do quarks and leptons represent the correct elementary" fields, which should appear in the matter Lagrangian, and which are structureless for renormalizaibility; second, could some of the presently considered gauge fields themselves be composite? This part of the lecture relies heavily on an address I was privileged to give at the European Physical Society meeting in Geneva in July this year.[64]

5.1 The quest for elementarity, prequarks (preons and pre-peons)

If quarks and leptons are elementary, we are dealing with 3 x 15 = 45 elementary entities. The "natural" group of which these constitute the fundamental representation is SU(45) with 2024 elementary gauge bosons. It is possible to reduce the size of this group to SU(11) for example (see Appendix I), with only 120 gauge bosons, but then the number of elementary fermions increases to 561, (of which presumably 3x15 = 45 objects are of low and the rest of Planckian mass). Is there any basic reason for one's instinctive revulsion when faced with these vast numbers of elementary fields.

The numbers by themselves would perhaps not matter so much. After all, Einstein in his description of gravity,[71] chose to work with 10 fields $(g_{\mu\nu}(x))$ rather than with just one (scalar field) as Nordstrom[72] had done before him. Einstein was not perturbed by the multiplicity he chose to introduce, since he relied on the sheet-anchor of a fundamental principle-(the equivalence principle)-which permitted him to relate the 10 fields for gravity $g_{\mu\nu}$ with the 10 components of the physically relevant quantity, the tensor $T_{\mu\nu}$ of energy and momentum. *Einstein knew that nature was not economical of structures:* only of principles of fundamental applicability. The question we must ask ourselves is this: have we yet discovered such principles in our quest for elementarity, to justify having fields with such large numbers of components as elementary.

[9] I would like to quote Feynman in a recent interview to the "Omni" magazine: "As long as it looks like the way things are built with wheels within wheels, then you are looking for the innermost wheel-but it might not be that way. in which case you are looking for whatever the hell it is you find!". In the same interview he remarks "a few years ago I was very sceptical about the gauge theories I was expecting mist. and now it looks like ridges and valleys after all."

Recall that quarks carry at least three charges (colour, flavour and a family number). Should one not, by now, entertain the notions of quarks (and possibly of leptons) as being composites of some more basic entities"' (PRE-QUARKS or PREENS), which each carry but *one* basic charge [52]. These ideas have been expressed before but they have become more compulsive now? with the growing multiplicity of quarks and leptons. Recall that it was similar ideas which led from the eight-fold of baryons to a triplet of (Sakatons and) quarks in the first place.

The preon notion is not new. In 1975, among others, Pati, Salam and Strathdee[52] introduced 4 chromons (the fourth colour corresponding to the lepton number) and 4 flavons, the basic group being SU(8)-of which the family group $SU_F(4)xSU_c(4)$ was but a subgroup. As an extension of these ideas, we now believe these preons carry magnetic charges and are bound together by very strong short-range forces, with quarks and leptons as their magnetically neutral composites[7 3] The important remark in this context is that in a theory containing *both* electric and magnetic generalized charges, the analogues of the well-known Dirac quantization condition [74] gives relations like $\frac{eg}{4\pi} = \frac{n}{2}$ for the strength of the two types of charges. Clearly, magnetic monopoles" of strength ±g and mass $\approx m_w/d \approx 10^4$ GeV, are likely to bind much more tightly than electric charges, yielding composites whose non-elementary nature will reveal itself only for very high energies. This appears to be the situation at least for leptons if they are composites.

In another form the preon idea has been revived this year by Curtwright and Freund[52],who motivated by ideas of extended supergravity (to be discussed in the next subsection). reintroduce an SU(8) of 3 chromons (R, Y, B), 2 flavons and 3 familons (horrible names). The family group SU(5) could be a subgroup of this SU(8). In the Curtwright-Freund scheme, the 3X15 = 45 fermions of SU(5)[61] can be found among the 8+28+56 of SU(8) (or alternatively the 3X 16 = 48 of SO(10) among the vectorial 56 fermions of SU(8)). (The next succession after the preon level may be the pre-preon level. It was suggested at the Geneva Conference [64] that with certain developments in field theory of composite fields it could be that just two-preons may suffice. But at this stage this is pure speculation.)

Before I conclude this section, I would like to make a prediction regarding the course of physics in the next decade, extrapolating from our past experience of the decades gone by:

[10] One must emphasise however that zero mass neutrinos are the hardest objects to conceive of as compositcs.

[11] According to 't Hooft's theorem. a monopoly corrosponding to the $SU_L(2)$ gauge symmetry is expected to possess a mass with the lower limit $\frac{m_W}{a}$. [75] [76] Xen if such monopolcs arc confined, their indirect effects must manifest themselves, if they exist. (Note that $\frac{m_W}{a}$ is very much a lower limit. For a grand unified theory like SU(5) for which the monopole mass is α^{-1} times the heavy lepto-quark mass.) The monopole force may be the techni-colour force of Footnote 5.

DECADE	1950—1960	1960—1970	1970—1980	1980→
Discovery in early part of the decade	The strange particles	The 8-fold way, Ω^-	Confirmation of neutral currents	W, Z, Proton decay
Expectation for the rest of the decade		SU(3) resonances		Grand Unification, Tribal Groups
Actual discovery		Hit the next level of elementarity with quarks		May hit the preon level, and composite structure of quarks

5.2 Post-Planck physics, supergravity and Einstein's dreams

I now turn to the problem of a deeper comprehension of the charge concept (the basis of gauging)-which, *in my humble view, is the real quest of particle physics.* Einstein, in the last thirty-live years of his life lived with two dreams: one was to unite gravity with matter (the photon)-he wished to see the "base wood" (as he put it) which makes up the stress tensor $T_{\mu\nu}$ on the right-hand side of his equation $R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -T_{\mu\nu}$ transmuted through this union, into the "marble" of gravity on the left-hand side. The second (and the complementary) dream was to use this unification to comprehend the nature of electric charge in terms of space-time geometry in the same manner as he had successfully comprehended the nature of gravitational charge in terms of space-time curvature.

In case some one imagines[12] that such deeper comprehension is irrelevant to quantitative physics, let me adduce the tests of Einstein's theory versus the proposed modifications to it (Brans-Dicke [77] for example). Recently (1976), the *strong* equivalence principle (i.e. the proposition that gravitational forces contribute equally to the inertial and the gravitational masses) was tested's to one part in 10^{12} (i.e. to the same accuracy as achieved in particle physics for (g-2),) through lunar-laser ranging measurements [78]. These measurements determined departures from Kepler equilibrium distances, of the moon, the earth and the sun to better than ±30 cms. and triumphantly vindicated Einstein.

There have been four major developments in realizing Einstein's dreams:

1. The Kaluza-Klein[79] miracle: An Einstein Lagrangian (scalar curvature) in live-dimensional space-time (where the fifth dimension is compactified in

[12]The following quotation from Einstein is relevant here. "We now realize, with special clarity, how much in error are those theorists who believe theory comes inductively from experience. Even the great Newton could not free himself from this error (Hypotheses "on lingo)." This quote is complementary to the quotation from Einstein at the end of Sec. II.

[13]The ***weak*** equivalence principle (the proposition that all but the gravitational force contribute equally to the inertial and the gravitational masses) was verified by Eötvös to $1 : 10^8$ and by Dicke and Braginsky and Panov to $1 : 10^{12}$.

the sense of all fields being explicitly independent of the fifth co-ordinate) precisely reproduces the *Einstein-Maxwell* theory in four dimensions, the $g_{\mu 5}$ (μ = 0, 1, 2, 3) components of the metric in five dimensions being identified with the Maxwell field A_μ. From this point of view, Maxwell's field is associated with the extra components of curvature implied by the (conceptual) existence of the fifth dimension.

2. The second development is the recent realization by Cremmer, Scherk, Englert, Brout, Minkowski and others that the compactification of the extra dimensions[80] -- (their curling up to sizes perhaps smaller than Planck length $\lesssim 10^{-33}$ cms. and the very high curvature associated with them)-- might arise through a spontaneous symmetry breaking (in the first 10^{-43} seconds) which reduced the higher dimensional space-time effectively to the four-dimensional that we apprehend directly.
3. So far we have considered Einstein's second dream, i.e. the unification of of electromagnetism (and presumably of other gauge forces) with gravity, giving a space-time significance to gauge charges as corresponding to extended curvature in extra bosonic dimensions. A full realization of the first dream (unification of spinor matter with gravity and with other gauge fields) had to await the development of supergravity[81], [82]--and an extension to extra fermionic dimensions. of superspace[83] (with extended torsion being brought into play in addition to curvature). I discuss this development later.
4. And finally there was the alternative suggestion by Wheeler[84] and Schemberg that electric charge may be associated with space-time topology-with worm-holes, with space-time Gruyère-cheesiness. This idea has recently been developed by Hawking[14] and his collaborators[85].

5.3 Extended supergravity, SU(8) preons and composite gauge fields

Thus far I have reviewed the developments in respect of Einstein's dreams as reported at the Stockholm Conference held in 1978 in this hall and organized by the Swedish Academy of Sciences.

A remarkable new development was reported during 1979 by Julia and Cremmer[87] which started with an attempt to use the ideas of Kaluza and Klein to formulate extended supergravity theory in a higher (compactified) spacetime-more precisely in eleven dimensions. This development links up, as we shall see, with preons and composite Fermi fields-and even more important-possibly with the notion of composite gauge fields.

Recall that simple supergravity[8l] is the gauge theory of super-symmetry[88] the gauge particles being the (helicity ±2) gravitons and

"The Einstein Langrangian allows large fluctuations of metric and topology on Planck-length scale. Hawking has surmised that the dominant contributions to the path integral of quantum gravity come from metrics which carry one unit of topology per Planck volume. On account of the intimate connection (de Rham, Atiyah-Singer)[86] of curvature with the measures of space-time topology (Euler number, Pontryagin number) the extended Kaluza-Klein and Wheeler-Hawking points of view may find consonance after all.

(helicity $\pm\frac{3}{2}$) gravitinos[15]. *Extendedsupergravity* gauges super-symmetry combined with SO(N) internal symmetry. For N = 8, the (tribal) supergravity multiplet consists of the following SO(8) families:[81], [87]

Helicity	
± 2	1
$\pm\frac{3}{2}$	8
± 1	28
$\pm\frac{1}{2}$	56
0	70

As is well known, SO(8) is too small to contain SU(2)xU(1)xSU$_c$(3). Thus this tribe has no place for $W^{\pm}$(though $Z^0$and γ are contained) and no places for μ or τ or the t quark.

This was the situation last year. This year, Cremmer and Julia [87] attempted to write down the N = 8 supergravity Langrangian explicitly, using an extension of the Kaluza-Klein ansatz which states that *extended supergravity* (with SO(8) internal symmetry) has the same Lagrangian in four space-time dimensions as *simple supergravity* in (compactified) eleven dimensions. This formal-and rather formidable ansatz-when carried through yielded a most agreeable bonus. *The supergravity Lagrangian possesses an unsuspected SU(8) "local" internal symmetry* although one started with an internal SO(8) only.

The tantalizing questions which now arise are the following.

1. Could this internal SU(8) be the symmetry group of the 8 preons (3 chromons, 2 flavons, 3 familons) introduced earlier?
2. When SU(8) is gauged, there should be 63 spin-one fields. The supergravity tribe contains only 28 spin-one fundamental objects which are not minimally coupled. Are the 63 fields of SU(8) to be identified with composite gauge fields made up of the 70 spin-zero objects of the form $V^{-1}\partial_\mu V$; Do these composites propagate, in analogy with the well-known recent result in CP^{n-1} theories[89], where a composite gauge field of this form propagates as a consequence of quantum effects (quantum completion)?

The entire development I have described-the unsuspected extension of SO(8) to SU(8) when extra compactified space-time dimensions are used-and the possible existence and quantum propagation of composite gauge fields-is of such crucial importance for the future prospects of gauge theories that one begins to wonder how much of the extrapolation which took SU(2) X U(1) X

[15] Supersymmetry algebra extends Poincaré group algebra by adjoining to it supersymmetric charges Q_α which transform bosons to fermions, $\{Q_\alpha, Q_\beta\} = (\gamma_\mu P_\mu)_{\alpha\beta}$. The currents which correspond to these charges (Q_α and P_μ) are $J_{\mu\alpha}$ and $T_{\mu\nu}$—these are essentially the currents which in gauged supersymmetry (i.e. supergravity) couple to the gravitino and the graviton, respectively.

$SU_c(3)$ into the electronuclear grand unified theories is likely to remain unaffected by these new ideas now unfolding.

But where in all this is the possibility to appeal directly to experiment? For grand unified theories, it was the proton decay. What is the analogue for supergravity? Perhaps the spin $\frac{3}{2}$ massive gravitino, picking its mass from a super-Higgs effect[90] provides the answer. Fayet[91] has shown that for a spontaneously broken globally supersymmetric weak theory the introduction of a local gravitational interaction leads to a super-Higgs effect. Assuming that supersymmetry breakdown is at mass scale m_w, the gravitino acquires a mass and an effective interaction. but of conventional weak rather than of the gravitational strength-an enhancement by a factor of 10^{34}. One may thus search for the gravitino among the neutral decay modes of J/ψ—the predicted rate being 10^{-3} - 10^{-5} times smaller than the observed rate for $J/\psi \rightarrow e^+e^-$. This will surely tax all the ingenuity of the great men (and women) at SLAC and DESY. Another effect suggested by Scherk[92] is antigravity-a cancellation of the attractive gravitational force with the force produced by spin-one gravi-photons which exist in all extended supergravity theories, Scherk shows that the Compton wave length of the gravi-photon is either smaller than 5 cms. or comprised between 10 and 850 metres in order that there is no conflict with what is presently known about the strength of the gravitational force.

Let me summarize: it is conceivable of course. that there is indeed a grand plateau--extending even to Planck energies. If so, the only eventual laboratory for particle physics will be the Early Universe, where we shall have to seek for the answers to the questions on the nature of charge. There may, however, be indications of a next level of structure around 10 TeV; there are also beautiful ideas (like, for example, of electric and magnetic monopole duality) which may manifest at energies of the order of $\alpha^{-1} m_W (= 10\ \mathrm{TeV})$. Whether even this level of structure will give us the final clues to the nature of charge, one cannot predict. All I can say is that I am for ever and continually being amazed at the depth revealed at each successive level we explore. I would like to conclude, as I did at the 1978 Stockholm Conference, with a prediction which J. R. Oppenheimer made more than twenty-five years ago and which has been fulfilled to-day in a manner he did not live to see. More than anything else, it expresses the faith for the future with which this greatest ofdecades in particle physics ends: “Physics will change even more If it is radical and unfamiliar

we think that the future will be only more radical and not less, only more strange and not more familiar, and that it will have its own new insights for the inquiring human spirit.”

J. R. Oppenheimer
Reith Lectures BBC 1953

APPENDIX I. EXAMPLES OF GRAND UNIFYING GROUPS

*Semi-simple groups**	*Multiplet*	*Exotic gauge particles*	*Proton decay*
(with left-right symmetry)	$G_L \to \begin{pmatrix} q \\ \ell \end{pmatrix}_L$, $G_R \to \begin{pmatrix} q \\ \ell \end{pmatrix}_R$	Lepto-quarks$\to(\bar{q}\ell)$	Lepto-quarks$\to$W + (Higgs) or
Example $[SU(6)_F \times SU(6)_c]_{L\to R}$	$G = G_L \times G_R$	Unifying mass $\approx 10^6$ GeV	Proton = qqq$\to \ell\ell\ell$
Simple groups		Diquarks$\to$(qq)	qq$\to\bar{q}\ell$ i.e.
Examples		Dileptons$\to(\ell\ell)$	Proton P = qqq$\to\bar{\ell}$
Family groups$\to$ SU(5)	or SO(10) $\left(G \to \begin{matrix} q \\ \ell \\ \bar{q} \\ \bar{\ell} \end{matrix}\right)_L$	Lepto-quarks$\to(\bar{q}\ell)$,	Also possible,
Tribal groups$\to$ SU(11)	SO(22)	$(q\ell)$	P$\to\ell$, P$\to 3\bar{\ell}$,
		Unifying mass $\approx 10^{13}-10^{15}$ GeV	P$\to 3\ell$

APPENDIX II

The following assumptions went into the derivation of the formula (I) in the text.

a) $SU_L(2) \times U_{L,R}(1)$ survives intact as the electroweak symmetry group from energies $\approx \mu$ right up to M. This intact survival implies that one eschews, for example, all suggestions that i) low-energy $SU_L(2)$ may be the diagonal sum of $SU_L^I(2), SU_L^{II}(2), SU_L^{III}(2)$, where I, II, III refer to the (three?) known families; ii) or that the $U_{L,R}(1)$ is a sum of pieces, where $U_R(1)$ may have differentially descended from a (V+A)-symmetric $SU_R(2)$ contained in G, or iii) that U(1) contains a piece from a four-colour symmetry $SU_c(4)$ (with lepton number as the fourth colour) and with $SU_c(4)$ breaking at an intermediate mass scale to $SU_c(3) \times U_c(1)$.

b) The second assumption which goes into the derivation of the formula above is that there are no unexpected heavy fundamental fermions, which might make $\sin^2\theta(M)$ differ from $\frac{3}{8}$—its value for the low mass fermions presently known to exist.'

* Grouping quarks (q) and leptons (ℓ) together, implies treating lepton number as the fourth colour, i.e. $SU_c(3)$ extends to $SU_c(4)$ (Pati and Salam)[93]. A "Tribal group, by definition, contains all known families in its basic representation. Favoured representations of Tribal SU(11) (Georgi)[94] and Tribal SO(22) (Gell-Man[95] et al.) contain 561 and 2048 fermions!

+ If one does not know G, one way to infer the parameter sin%(M) is from the formula:

$$\sin^2\theta(M) = \frac{\Sigma T^2_{3L}}{\Sigma Q^2}\left(= \frac{9\,N_q + 3\,N_\ell}{20\,N_q + 12\,N_\ell}\right)$$

Here N_q and N_ℓ are the numbers of the fundamental quark and lepton SU(2) doublets (assuming these are the only multiplets that exist). If we make the further assumption that $N_q = N_\ell$ (from the requirement of anomaly cancellation between quarks and leptons) we obtain sin%(M) = $\frac{3}{8}$. This assumption however is not compulsive; for example anomalies cancel also if (heavy) mirror fermions exist[98]. This is the case for $[SU(6)]^4$ for which $\sin^2\theta(M) = \frac{9}{28}$.

c) If these assumptions arc relaxed, for example, for the three family group $G = [SU_F(6) \times SU_c(6)]_{L\to R}$, where $\sin^2\theta(M) = \frac{9}{28}$, we find the grand unifying mass M tumbles down to 10" GeV.

d) The introduction of intermediate mass scales (for example, those connoting the breakdown of family universality, or of left-right symmetry, or of a breakdown of 4-colour $SU_c(4)$ down to $SU_c(3) \times U_c(1)$) will as a rule push the magnitude of the grand unifying mass M upwards[96] In order to secure a proton decay life, consonant with present empirical lower limits ($\sim 10^{30}$ years)[97] this is desirable anyway. (τ_{proton} for M ~ 10'" GeV is unacceptably low ~ 6×10^{23} years unless there arc 15 Higgs.) There is, from this point of view, an indication of there being in Particle Physics one or several intermediate mass scales which can be shown to start from around 10^4 GeV upwards. This *is the end result which I wished this Appendix to lead up to.*

REFERENCES

1. Lee, T. D. and Yang, C. N., Phys. Rev. 104, 254 (1956).
2. Abdus Salam, Nuovo Cimento 5, 299 (1957).
3. Wu, C. S., et al., Phys. Rev. 105, 1413 (1957).
4. Car-win, R., Lederman, L. and Weinrich, M., Phys. Rev. 105, 1415 (1957).
5. Friedman, J. I. and Telegdi, V. L., Phys. Rev. 105, 1681 (1957).
6. Landau, L., Nucl. Phys. 3, 127 (1957).
7. Lee, T. D. and Yang, C. N., Phys. Rev. 105, 1671 (1957).
8. Abdus Salam, Imperial College, London, preprint (1957). For reference, see Footnote 7, p. 89, of ***Theory*** of ***Weak*** Interactions ***in Particle Physics,*** by Marshak, R. E., Riazuddin and Ryan, C. P., (Wiley-Interscience, New York 1969), and W. Pauli's letters (CERN Archives).
9. Yang, C. N. and Mills, R. L., Phys. Rev. 96, 191 (1954).
10. Shaw, R., "The problem of particle types and other contributions to the theory of elementary particles", Cambridge Ph. D. Thesis (1955), unpublished.
11. Marshak, R. E. and Sudarshan, E. C. G., Proc. Padua-Venice Conference on Mesons and Recently Discovered Particles (1957), and Phys. Rev. 109, 1860 (1958). The idea of a universal Fermi interaction for (P,N), (υ_e, e) and (Υ_μ, μ) doublets goes back to Tiomno, J. and Wheeler, J. A., Rev. Mod. Phys. ***21,*** 144 (1949); ***21,*** I53 (1949) and by Yang, C. N. and Tiomno, J., Phys. Rev. 75, 495 (1950). Tiomno, J., considered γ_5 transformations of Fermi fields linked with mass reversal in 11 Nuovo Cimento ***1, 226 (1956).***
12. Feynman, R. P. and Gell-Mann, M., Phys. Rev. 109, 193 (1958).
13. Sakurai, J. J., Nuovo Cimento 7, 1306 (1958).
14. Nambu, Y. and Jona-Lasinio, G., Phys. Rev. 122, 345 (1961).
15. Nambu, Y., Phys. Rev. Letters 4, 380 (1960); Goldstone, J., Nuovo Cimento 19, 154 (1961).
16. Klein, O., "On the theory of charged fields", Proceedings of the Conference organized by International Institute of Intellectual Cooperation, Paris (1939).
17. Kemmer, N., Phys. Rev. 52, 906 (1937).
18. Schwinger, J., Ann. Phys. (N. Y.)2, 407 (1957).
19. Glashow, S. L., Nucl. Phys. ***10,*** 107 (1959).
20. Abdus Salam and Ward, J. C., Nuovo Cimento II, 568 (1959).
21. Bludman, S., Nuovo Cimento 9, 433 (1958).

22. Abdus Salam and Ward, J. C., Nuovo Cimento 19, 165 (1961).
23. Glashow, S. L., Nucl. Phys. 22, 579 (1961).
24. Abdus Salam and Ward, J. C., Phys. Letters 13, 168 (1964).
25. Goldstone, J., see Ref. 15.
26. Goldstone, J.. Abdus Salam and Weinberg, S., Phys. Rev. 127, 965 (1962).
27. Anderson, P. W., Phys. Rev. 130, 439 (1963).
28. Higgs, P. W., Phys. Letters 12, 132 (1964); Phys. Rev. Letters 23, 508 (1964); Phys. Rev. 245, 1156 (1966).
29. Englert, F. and Brout, R., Phys. Rev. Letters 23, 321 (1964);
Englert, F., Brout, R. and Thiry, M. F., Nuovo Cimento 48, 244 (1966).
30. Guralnik, G. S., Hagen, C. R. and Kibble, T. W. B., Phys. Rev. Letters 13, 585 (1964); Kibble, T. W. B., Phys. Rev. ***155,*** 1554 (1967).
31. Weinberg, S., Phys. Rev. Letters 27, 1264 (1967).
32. Abdus Salam, Proceedings of the 8th Nobel Symposium, Ed. Svartholm, N., (Almqvist and Wiksell, Stockholm 1968).
33. 't Hooft, G., Nucl. Phys. B33, 173 (1971); ibid. B35, 167 (1971).
34. Lee, B. W., Phys. Rev. D5, 823 (1972); Lee, B. W. and Zinn-Justin, J., Phys. Rev. D5, 3137 (1972);ibid. D7, 1049 (1973).
35. 't Hooft, G. and Veltman, M., Nucl. Phys. B44, 189 (1972); ibid. B50, 318 (1972). An important development in this context was the invention of the dimensional regularization technique by Bollini, C. and Giambiagi, J., Nuovo CimentoEl2, 20 (1972);
Ashmore, J., Nuovo Cimento Letters 4, 289 (1972) and by 't Hooft, G. and Veltman, M.
36. Feynman, R. P., Acta Phys. Polonica 24, 297 (1963).
37. Dewitt, B. S., Phys. Rev. 162, 1195 and 1239 (1967).
38. Faddeev, L. D. and Popov, V. N., Phys. Letters 258, 29 (1967).
39. Mandelstam, S., Phys. Rev. 175, 1588 and 1604 (1968).
40. Fradkin, E. S. and Tyutin, I. V., Phys. Rev. D2, 2841 (1970).
41. Boulware, D. G., Ann. Phys. (N.Y.)56, 140 (1970).
42. Taylor, J. C., Nucl. Phys. B33, 436 (1971).
43. Slavnov, A., Theor. Math. Phys. 10, 99 (1972).
44. Abdus Salam and Strathdee, J., Phys. Rev. ***D2, 2869 (1970).***
45. Glashow, S., Iliopoulos, J. and Maiani, L., Phys. Rev. ***D2,*** 1285 (1970).
46. For a review, see Jackiw, R., in ***Lectures on Current Algebra and its Applications,*** by Treiman, S. B., Jackiw, R. and Gross, D. J., (Princeton Univ. Press, 1972).
47. Bouchiat, C., Iliopoulos, J. and Meyer, Ph., Phys. Letters ***38B,*** 519 (1972); Gross, D. J. and Jackiw, R., Phys. Rev. ***D6, 477 (1972).***
48. Wentzel, G., Helv. Phys. Acta 10, 108 (1937).
49. Hasert, F. J., et al., Phys. Letters 46B, 138 (1973).
50. Taylor, R. E., Proceedings of the 19th International Conference on High Energy Physics, Tokyo, Physical Society of Japan, 1979, p. 422.
51. Barkov, L. M., Proceedings of the 19th International Conference on High Energy Physics, Tokyo, Physical Society of Japan, 1979, p. 425.
52. Pati, J. C. and Abdus Salam, ICTP, Trieste, IC/75/106, Palermo Conference, June 1975; Pati, J, C., Abdus Salam and Strathdee, J., Phys. Letters 598, 265 (1975); Harari, H., Phys. Letters ***86B,*** 83 (1979); Schupe, M., ***ibid.*** 86B, 87 (1979); Curtwright, T. L. and Freund, P. G.O., Enrico Fermi Inst. preprint EFI 79/25, University of Chicago, April 1979.
53. Pati, J. C. and Abdus Salam, see the review by Bjorken, J. D., Proceedings of the 16th International Conference on High Energy Physics, Chicago-Batavia, 1972, Vol. 2, p. 304; Fritsch, H. and Gell-Mann, M., ***ibid.*** p. 135; Fritzsch, H., Gell-Mann, M. and Leutwyler, H., Phys. Letters 478, 365 (1973);
Weinberg, S., Phys. Rev. Letters 31, 494 (1973); Phys. Rev. ***D8, 4482*** (1973); Gross, D. J. and Wilczek, F., Phys. Rev. ***D8,*** 3633 (1973); For a review see Marciano, W. and Pagels, H., Phys. Rep. 36C, 137 (1978).
54. Tasso Collaboration, Brandelik et al., Phys. Letters 868, 243 (1979); Mark-J. Collaboration, Barber et al., Phys. Rev. Letters 43, 830 (1979); See also reports of the Jade, Mark-J,

Pluto and Tasso Collaborations to the International Symposium on Lepton and Photon Interactions at High Energies, Fermilab, August 1979.
55. Winter, K., Proceedings of the International Symposium on Lepton and Photon Interactions at High Energies, Fermilab, August 1979.
56. Ellis, J., Proceedings of the "Neutrino-79" International Conference on Neutrinos, Weak Interactions and Cosmology, Bergen, June 1979.
57. Dimopoulos, S. and Susskind, L., Nucl. Phys B155, 237 (1979).
58. Weinberg, S., Phys. Rev. ***D19, 1277 (1979).***
59. Pati, J. C. and Abdus Salam, Phys. Rev. ***D10, 275 (1974);***
Mohapatra, R. N. and Pati. J. ***C.,*** Phys. Rev. ***D11, 566,2558*** (1975);
Elias, V., Pati, J. C. and Abdus Salam, Phys. Letters ***73B, 450*** (1978);
Pati, J. C. and Rajpoot, S., Phys. Letters 798, 65 (1978).
60. See Pati, J. C. and Abdus Salam, Ref. 53 above and Pati, J. C. and Abdus Salam, Phys. Rev. ***D8,*** 1240 (1973).
61. Georgi, H. and Glashow, S. L., Phys. Rev. Letters 32, 438 (1974).
62. Georgi, H., Quinn, H. R. and Weinberg, S., Phys. Rev. Letters 33, 451 (1974).
63. Marciano, W. J., Phys. Rev. ***020, 274*** (1979).
64. See Abdus Salam, Proceedings of the European Physical Society Conference, Geneva, August 1979, ICTP, Trieste, preprint IC/79/142, with references to H. Harari's work.
65. Pati, J. C. and Abdus Salam, Phys, Rev. Letters 31, 661 (1973).
66. Elias, V., Pati, J. C. and Abdus Salam, Phys. Rev. Letters 40, 920 (1978);
Rajpoot, S. and Elias, V., ICTP, Trieste, preprint IC/78/159.
67. Fritzsch, H. and Minkowski, P., Ann. Phys. (N. Y.) 93, ***193*** (1975); Nucl. Phys. B103, 61 (1976);
Georgi, H., Particles and Fields (APS/OPF Williamsburg), Ed. Carlson, C. E., AIP, New York, 1975, p. 575;
Georgi, H. and Nanopoulos, D. V., Phys. Letters 82B, 392 (1979).
68. Buras, A., Ellis, J., Gaillard, M. K. and Nanopoulos, D. V., Nucl. Phys. ***Bl35, 66*** (1978).
69. Yoshimura, M., Phys. Rev. Letters 41, 381 (1978);
Dimopoulos, S. and Susskind, L., Phys. Rev. ***D18, 4500 (1978);***
Toussaint, B., Treiman, S. B., Wilczek, F. and Zee, A., Phys. Rev. ***D19,*** 1036 (1979);
Ellis J., Gaillard, M. K. and Nanopoulos, D. V., Phys. Letters 8OB, 360 (1979);
Erratum 82B, 464 (1979);
Weinberg, S., Phys. Rev. Letters 42, 850 (1979); Nanopoulos, D. V. and Weinberg, S., Harvard University preprint HUTP-79/A023.
70. Nanopoulos, D. V., CERN preprint TH2737, September 1979.
71. Einstein, A., Annalen der Phys. 49, 769 (1916). For an English translation, see ***The Principle of Relativity*** (Methuen, 1923, reprinted by Dover Publications), p. 35.
72. Nordstrom, G., Phys. Z. ***13,*** 1126 (1912); Ann. Phys. (Leipzig) ***40, 856 (1913); ibid. 42, 533 (1913);ibid.*** 43, 1101 (1914); Phys. Z. 15, 375 (1914); Ann. Acad. Sci. Fenn.57 (1914, 1915);
See also Einstein, A., Ann. Phys. Leipzig38, 355, 433 (1912).
73. Pati, J. C. and Abdus Salam, in preparation.
74. Dirac, P. A. M., Proc. Roy. Soc. (London)A133, 60 (1931).
75. 't Hooft, G., Nucl. Phys. 879, 276 (1974).
76. Polyakov, A. M., JETP Letters 20, 194 (1974).
77. Brans, C. H. and Dicke, R. H., Phys. Rev. ***124, 925 (1961).***
78. Williams, J. G., ***et al.,*** Phys. Rev. Letters 36, 551 (1976);
Shapiro, I. I., et al., Phys. Rev. Letters 36, 555 (1976);
For a discussion, see Abdus Salam, in ***Physics and Cmtemporary*** Need.s Ed. Riazuddin (Plenum Publishing Corp., 1977), p. 301.
79. Kaluza, Th., Sitzungsber. Preuss. Akad. Wiss. p. 966 (1921);
Klein, O.. Z. Phys. 37, 895 (1926).
80. Cremmer, E. and Scherk, J., Nucl. Phys. 8103, 399 (1976); ***ibid. B108, 409 (1976); ibid.*** B118, 61 (1976);
Minkowski, P., Univ. of Berne preprint, October 1977.

81. Freedman, D. Z.. van Nieuwenhuizen, P. and Ferrara, S., Phys. Rev, D13, 3214 (1976); Deser, S. and Zumino, B., Phys. Letters 62B, 335 (1976);
For a review and comprehensive list of references, see D. Z. Freedman's presentation to the 19th International Conference on High Energy Physics, Tokyo, Physical Society of Japan, 1979.
82. Amowitt, R., Nath, P. and Zumino, B., Phys. Letters 56B, 81 (1975); Zumino, B., in Proceedings of the Conference ***on*** Gauge Theories and Modern Field Theory, Northeastern University, September 1975, Eds. Arnowitt, R. and Nath, P., (MIT Press);
Wess, J. and Zumino, B., Phys. Letters 66B, 361 (1977);
Akulov, V. P., Volkov, D. V. and Soroka, V. A., JETP Letters22, 187 (1975);
Brink, L., Gell-Mann, M., Ramond, P. and Schwarz, J. H., Phys. Letters 748, 336 (1978);
Taylor, J. G., King's College, London, preprint, 1977 (unpublished);
Siegel, W., Harvard University preprint HUTP-77/A068, 1977 (unpublished);
Ogievetsky, V. and Sokatchev, E., Phys. Letters ***79B, 222*** (1978);
Chamseddine, A. H. and West, P. C., Nucl. Phys. ***B129, 39*** (1977);
MacDowell, S. W. and Mansouri, F., Phys. Rev. Letters 38, 739 (1977).
83. Abdus Salam and Strathdee, J., Nucl. Phys. ***B79, 477*** (1974).
84. Fuller, R. W. and Wheeler, J. A., Phys. Rev. 128, 919 (1962);
Wheeler, J, A., in ***Relativity*** Groups and ***Topology,*** Proceedings of the Les Houches Summer School, 1963, Eds. Dewitt, B. S. and Dewitt, C. M., (Gordon and Breach, New York 1964).
85. Hawking, S. W., in General ***Relativity: An Einstein Centenary Survey*** (Cambridge University Press, 1979);
See also "Euclidean quantum gravity", DAMTP, Univ. of Cambridge preprint, 1979;
Gibbons, G. W., Hawking, S. W. and Perry, M. J., Nucl. Phys. ***B138,*** 141 (1978);
Hawking, S. W., Phys. Rev. ***D18, 1747*** (1978).
86. Atiyah, M. F. and Singer, 1. M., Bull. Am. Math. Soc. 69, 422 (1963).
87. Cremmer, E., Julia, B. and Scherk, J., Phys. Letters 768, 409 (1978); Cremmer, E. and Julia, B., Phys. Letters 80B, 48 (1978); Ecole Normale Superieure preprint, LPTENS 79/6, March 1979;
See also Julia, B., in Proceedings of the Second Marcel Grossmann Meeting, Trieste, July 1979 (in preparation).
88. Gol'fand, Yu. A. and Likhtman, E. P., JETP Letters ***13, 323*** (1971);
Volkov, D. V. and Akulov, V. P., JETP Letters ***16, 438*** (1972);
Wess, J. and Zumino, B., Nucl. Phys. 870, 39 (1974);
Abdus Salam and Strathdee, J., Nucl. Phys. 879, 477 (1974); ***ibid. B80, 499*** (1974); Phys. Letters ***51B, 353*** (1974);
For a review, see Abdus Salam and Strathdee, J., Fortschr. Phys. 26, 57 (1978).
89. D'Adda, A., Lüscher, M. and Di Vecchia, P., Nucl. Phys. 8146, 63 (1978).
90. Cremmer, E., et al., Nucl. Phys. ***B147, 105*** (1979);
See also Ferrara, S., in Proceedings of the Second Marcel Grossmann Meeting, Trieste, July 1979 (in preparation), and references therein.
91. Fayet, P., Phys. Letters 7OB, 461 (1977); ibid. ***84B, 421*** (1979).
92. Scherk, J., Ecole Normale Superieure preprint, LPTENS 79/17, September 1979.
93. Pati, J. C. and Abdus Salam, Phys. Rev. ***D10, 275*** (1974).
94. Georgi, H., Harvard University Report No. HUTP-29/AO13 (1979).
95. Gell-Mann, M., (unpublished).
96. See Ref. 64 above and also Shafi, Q. and Wetterich, C., Phys. Letters 85B, 52 (1979).
97. Learned, J., Reines, F. and Soni, A., Phys. Letters 43, 907 (1979).
98. Pati, J. C., Abdus Salam and Strathdee, J., Nuovo Cimento 26A, 72 (1975);
Pati, J. C. and Abdus Salam, Phys. Rev. ***D11,*** 1137, 1149 (1975);
Pati, J. C., invited talk, Proceedings Second Orbis Scientiae, Coral Gables, Florida, 1975, Eds. Perlmutter, A. and Widmayer, S., p. 253.

Steven Weinberg

STEVEN WEINBERG

I was born in 1933 in New York City to Frederick and Eva Weinberg. My early inclination toward science received encouragement from my father, and by the time I was 15 or 16 my interests had focused on theoretical physics.

I received my undergraduate degree from Cornell in 1954, and then went for a year of graduate study to the Institute for Theoretical Physics in Copenhagen (now the Niels Bohr Institute). There, with the help of David Frisch and Gunnar Källen. I began to do research in physics. I then returned to the U.S. to complete my graduate studies at Princeton. My Ph.D thesis, with Sam Treiman as adviser, was on the application of renormalization theory to the effects of strong interactions in weak interaction processes.

After receiving my Ph.D. in 1957, I worked at Columbia and then from 1959 to 1966 at Berkeley. My research during this period was on a wide variety of topics - high energy behavior of Feynman graphs, second-class weak interaction currents, broken symmetries, scattering theory, muon physics, etc. - topics chosen in many cases because I was trying to teach myself some area of physics. My active interest in astrophysics dates from 1961-62; I wrote some papers on the cosmic population of neutrinos and then began to write a book, *Gravitation and* Cosmology, which was eventually completed in 197 1. Late in 1965 I began my work on current algebra and the application to the strong interactions of the idea of spontaneous symmetry breaking.

From 1966 to 1969, on leave from Berkeley, I was Loeb Lecturer at Harvard and then visiting professor at M.I.T. In 1969 I accepted a professorship in the Physics Department at M.I.T., then chaired by Viki Weisskopf. It was while I was a visitor to M.I.T. in 1967 that my work on broken symmetries, current algebra, and renormalization theory turned in the direction of the unification of weak and electromagnetic interactions. In 1973, when Julian Schwinger left Harvard, I was offered and accepted his chair there as Higgins Professor of Physics, together with an appointment as Senior Scientist at the Smithsonian Astrophysical Observatory.

My work during the 1970's has been mainly concerned with the implications of the unified theory of weak and electromagnetic interactions, with the development of the related theory of strong interactions known as quantum chromodynamics, and with steps toward the unification of all interactions.

In 1982 I moved to the physics and astronomy departments of the University of Texas at Austin, as Josey Regental Professor of Science.

I met my wife Louise when we were undergraduates at Cornell, and we were married in 1954. She is now a professor of law. Our daughter Elizabeth was born in Berkeley in 1963.

Awards and Honors

Honorary Doctor of Science degrees, University of Chicago, Knox College, City University of New York, University of Rochester, Yale University
American Academy of Arts and Sciences, elected 1968
National Academy of Sciences, elected 1972
J. R. Oppenheimer Prize, 1973
Richtmeyer Lecturer of Am. Ass'n. of Physics Teachers, 1974
Scott Lecturer, Cavendish Laboratory, 1975
Dannie Heineman Prize for Mathematical Physics, 1977
Silliman Lecturer, Yale University, 1977
Am. Inst. of Physics - U.S. Steel Foundation Science Writing Award, 1977, for authorship of *The First Three Minutes* (1977)
Lauritsen Lecturer, Cal. Tech., 1979
Bethe Lecturer, Cornell Univ., 1979
Elliott Cresson Medal (Franklin Institute), 1979
Nobel Prize in Physics, 1979

Awards and Honors since 1979

Honorary Doctoral degrees, Clark University, City University of New York, Dartmouth College, Weizmann Institute, Clark University, Washington College, Columbia University
Elected to American Philosophical Society, Royal Society of London (Foreign Honorary Member), Philosophical Society of Texas
Henry Lecturer, Princeton University, 1981
Cherwell-Simon Lecturer, University of Oxford, 1983
Bampton Lecturer, Columbia University, 1983
Einstein Lecturer, Israel Academy of Arts & Sciences, 1984
McDermott Lecturer, University of Dallas, 1985
Hilldale Lecturer, University of Wisconsin, 1985
Clark Lecturer, University of Texas at Dallas, 1986
Brickweede Lecturer, Johns Hopkins University, 1986
Dirac Lecturer, University of Cambridge, 1986
Klein Lecturer, University of Stockholm, 1989
James Madison Medal of Princeton University, 1991
National Medal of Science, 1991

CONCEPTUAL FOUNDATIONS OF THE UNIFIED THEORY OF WEAK AND ELECTROMAGNETIC INTERACTIONS

Nobel Lecture, December 8, 1979
by STEVEN WEINBERG
Lyman Laboratory of Physics Harvard University and Harvard-Smithsonian Center for Astrophysics Cambridge, Mass., USA.

Our job in physics is to see things simply, to understand a great many complicated phenomena in a unified way, in terms of a few simple principles. At times, our efforts are illuminated by a brilliant experiment, such as the 1973 discovery of neutral current neutrino reactions. But even in the dark times between experimental breakthroughs, there always continues a steady evolution of theoretical ideas, leading almost imperceptibly to changes in previous beliefs. In this talk, I want to discuss the development of two lines of thought in theoretical physics. One of them is the slow growth in our understanding of symmetry, and in particular, broken or hidden symmetry. The other is the old struggle to come to terms with the infinities in quantum field theories. To a remarkable degree, our present detailed theories of elementary particle interactions can be understood deductively, as consequences of symmetry principles and of a principle of renormalizability which is invoked to deal with the infinities. I will also briefly describe how the convergence of these lines of thought led to my own work on the unification of weak and electromagnetic interactions. For the most part, my talk will center on my own gradual education in these matters, because that is one subject on which I can speak with some confidence. With rather less confidence, I will also try to look ahead, and suggest what role these lines of thought may play in the physics of the future.

Symmetry principles made their appearance in twentieth century physics in 1905 with Einstein's identification of the invariance group of space and time. With this as a precedent, symmetries took on a character in physicists' minds as a priori principles of universal validity, expressions of the simplicity of nature at its deepest level. So it was painfully difficult in the 1930's to realize that there are internal symmetries, such as isospin conservation, [1] having nothing to do with space and time, symmetries which are far from self-evident, and that only govern what are now called the strong interactions. The 1950's saw the discovery of another internal symmetry - the conservation of strangeness [2] - which is not obeyed by the weak interactions, and even one of the supposedly sacred symmetries of space-time - parity - was also found to be violated by weak interactions. [3] Instead of moving toward unity, physicists were learning that different

interactions are apparently governed by quite different symmetries. Matters became yet more confusing with the recognition in the early 1960's of a symmetry group - the "eightfold way" - which is not even an exact symmetry of the strong interactions. [4]

These are all "global" symmetries, for which the symmetry transformations do not depend on position in space and time. It had been recognized [5] in the 1920's that quantum electrodynamics has another symmetry of a far more powerful kind, a "local" symmetry under transformations in which the electron field suffers a phase change that can vary freely from point to point in space-time, and the electromagnetic vector potential undergoes a corresponding gauge transformation. Today this would be called a U(1) gauge symmetry, because a simple phase change can be thought of as multiplication by a 1 x 1 unitary matrix. The extension to more complicated groups was made by Yang and Mills [6] in 1954 in a seminal paper in which they showed how to construct an SU(2) gauge theory of strong interactions. (The name "SU(2)" means that the group of symmetry transformations consists of 2 x 2 unitary matrices that are "special," in that they have determinant unity). But here again it seemed that the symmetry if real at all would have to be approximate, because at least on a naive level gauge invariance requires that vector bosons like the photon would have to be massless, and it seemed obvious that the strong interactions are not mediated by massless particles. The old question remained: if symmetry principles are an expression of the simplicity of nature at its deepest level, then how can there be such a thing as an approximate symmetry? Is nature only approximately simple?

Some time in 1960 or early 1961, I learned of an idea which had originated earlier in solid state physics and had been brought into particle physics by those like Heisenberg, Nambu, and Goldstone, who had worked in both areas. It was the idea of "broken symmetry," that the Hamiltonian and commutation relations of a quantum theory could possess an exact symmetry, and that the physical states might nevertheless not provide neat representations of the symmetry. In particular, a symmetry of the Hamiltonian might turn out to be not a symmetry of the vacuum.

, As theorists sometimes do, I fell in love with this idea. But as often happens with love affairs, at first I was rather confused about its implications. I thought (as turned out, wrongly) that the approximate symmetries - parity, isospin, strangeness, the eight-fold way - might really be exact *a priori* symmetry principles, and that the observed violations of these symmetries might somehow be brought about by spontaneous symmetry breaking. It was therefore rather disturbing for me to hear of a result of Goldstone, [7] that in at least one simple case the spontaneous breakdown of a continuous symmetry like isospin would necessarily entail the existence of a massless spin zero particle - what would today be called a "Goldstone boson." It seemed obvious that there could not exist any new type of massless particle of this sort which would not already have been discovered.

I had long discussions of this problems with Goldstone at Madison in the summer of 1961, and then with Salam while I was his guest at Imperial College in 196 l-62. The three of us soon were able to show that Goldstone bosons must in fact occur whenever a symmetry like isospin or strangeness is spontaneously broken, and that their masses then remain zero to all orders of perturbation theory. I remember being so discouraged by these zero masses that when we wrote our joint paper on the subject, [8] I added an epigraph to the paper to underscore the futility of supposing that anything could be explained in terms of a non-invariant vacuum state: it was Lear's retort to Cordelia, "Nothing will come of nothing: speak again." Of course, The Physical Review protected the purity of the physics literature, and removed the quote. Considering the future of the non-invariant vacuum in theoretical physics, it was just as well.

There was actually an exception to this proof, pointed out soon afterwards by Higgs, Kibble, and others. [9] They showed that if the broken symmetry is a local, gauge symmetry, like electromagnetic gauge invariance, then although the Goldstone bosons exist formally, and are in some sense real, they can be eliminated by a gauge transformation, so that they do not appear as physical particles. The missing Goldstone bosons appear instead as helicity zero states of the vector particles, which thereby acquire a mass.

I think that at the time physicists who heard about this exception generally regarded it as a technicality. This may have been because of a new development in theoretical physics which suddenly seemed to change the role of Goldstone bosons from that of unwanted intruders to that of welcome friends.

In 1964 Adler and Weisberger [10] independently derived sum rules which gave the ratio g_A/g_V of axial-vector to vector coupling constants in beta decay in terms of pion-nucleon cross sections. One way of looking at their calculation, (perhaps the most common way at the time) was as an analogue to the old dipole sum rule in atomic physics: a complete set of hadronic states is inserted in the commutation relations of the axial vector currents. This is the approach memorialized in the name of "current algebra." [11] But there was another way of looking at the Adler-Weisberger sum rule. One could suppose that the strong interactions have an approximate symmetry, based on the group SU(2) x SU(2), and that this symmetry is spontaneously broken, giving rise among other things to the nucleon masses. The pion is then identified as (approximately) a Goldstone boson, with small non-zero mass, an idea that goes back to Nambu. [12] Although the SU(2) X SU(2) symmetry is spontaneously broken, it still has a great deal of predictive power, but its predictions take the form of approximate formulas, which give the matrix elements for low energy pionic reactions. In this approach, the Adler-Weisberger sum rule is obtained by using the predicted pion nucleon scattering lengths in conjunction with a well-known sum rule [13], which years earlier had been derived from the dispersion relations for pion-nucleon scattering.

In these calculations one is really using not only the fact that the strong interactions have a spontaneously broken approximate SU(2) X SU(2) symmetry, but also that the currents of this symmetry group are, up to an overall constant, to be identified with the vector and axial vector currents of beta decay. (With this assumption g_A/g_V gets into the picture through the Goldberger-Treiman relation, [14] which gives g_A/g_V in terms of the pion decay constant and the pion nucleon coupling.) Here, in this relation between the currents of the symmetries of the strong interactions and the physical currents of beta decay, there was a tantalizing hint of a deep connection between the weak interactions and the strong interactions. But this connection was not really understood for almost a decade.

I spent the years 1965-67 happily developing the implications of spontaneous symmetry breaking for the strong interactions. [15] It was this work that led to my 1967 paper on weak and electromagnetic unification. But before I come to that I have to go back in history and pick up one other line of though, having to do with the problem of infinities in quantum field theory.

I believe that it was Oppenheimer and Waller in 1930 [16] who independently first noted that quantum field theory when pushed beyond the lowest approximation yields ultraviolet divergent results for radiative self energies. Professor Waller told me last night that when he described this result to Pauli, Pauli did not believe it. It must have seemed that these infinities would be a disaster for the quantum field theory that had just been developed by Heisenberg and Pauli in 1929-30. And indeed, these infinites did lead to a sense of discouragement about quantum field theory, and many attempts were made in the 1930's and early 1940's to find alternatives. The problem was solved (at least for quantum electrodynamics) after the war, by Feynman, Schwinger, and Tomonaga [17] and Dyson [19]. It was found that all infinities disappear if one identifies the observed finite values of the electron mass and charge, not with the parameters m and e appearing in the Lagrangian, but with the electron mass and charge that are *calculated* from m and e, when one takes into account the fact that the electron and photon are always surrounded with clouds of virtual photons and electron-positron pairs [18]. Suddenly all sorts of calculations became possible, and gave results in spectacular agreement with experiment.

But even after this success, opinions differed as to the significance of the ultraviolet divergences in quantum field theory. Many thought-and some still do think-that what had been done was just to sweep the real problems under the rug. And it soon became clear that there was only a limited class of so-called "renormalizable" theories in which the infinities could be eliminated by absorbing them into a redefinition, or a "renormalization," of a finite number of physical parameters. (Roughly speaking, in renormalizable theories no coupling constants can have the dimensions of negative powers of mass. But every time we add a field or a space-time derivative to an interaction, we reduce the dimensionality of the associated coupling

constant. So only a few simple types of interaction can be renormalizable.) In particular, the existing Fermi theory of weak interactions clearly was not renormalizable. (The Fermi coupling constant has the dimensions of $[\text{mass}]^{-2}$.) The sense of discouragement about quantum field theory persisted into the 1950's and 1960's.

I learned about renormalization theory as a graduate student, mostly by reading Dyson's papers. [19] From the beginning it seemed to me to be a wonderful thing that very few quantum field theories are renormalizable. Limitations of this sort are, after all, what *we* most *want,* not mathematical methods which can make sense of an infinite variety of physically irrelevant theories, but methods which carry constraints, because these constraints may point the way toward the one true theory. In particular, I was impressed by the fact that quantum electrodynamics could in a sense be *de-rived* from symmetry principles and the constraints of renormalizability; the only Lorentz invariant and gauge invariant renormalizable Lagrangian for photons and electrons is precisely the orginal Dirac Lagrangian of QED. Of course, that is not the way Dirac came to his theory. He had the benefit of the information gleaned in centuries of experimentation on electromagnetism, and in order to fix the final form of his theory he relied on ideas of simplicity (specifically, on what is sometimes called minimal electromagnetic coupling). But we have to look ahead, to try to make theories of phenomena which have not been so well studied experimentally, and we may not be able to trust purely formal ideas of simplicity. I thought that renormalizability might be the key criterion, which also in a more general context would impose a precise kind of simplicity on our theories and help us to pick out the one true physical theory out of the infinite variety of conceivable quantum field theories. As I will explain later, I would say this a bit differently today, but I am more convinced than ever that the use of renormalizability as a constraint on our theories of the observed interactions is a good strategy. Filled with enthusiasm for renormalization theory, I wrote my Ph.D. thesis under Sam Treiman in 1957 on the use of a limited version of renormalizability to set constraints on the weak interactions, [20] and a little later I worked out a rather tough little theorem [21] which completed the proof by Dyson [19] and Salam [22] that ultraviolet divergences really do cancel out to all orders in nominally renormalizable theories. But none of this seemed to help with the important problem, of how to make a renormalizable theory of weak interactions.

Now, back to 1967. I had been considering the implications of the broken SU(2) x SU(2) symmetry of the strong interactions, and I thought of trying out the idea that perhaps the SU(2) x SU(2) symmetry was a "local," not merely a "global," symmetry. That is, the strong interactions might be described by something like a Yang-Mills theory, but in addition to the vector ϱ mesons of the Yang-Mills theory, there would also be axial vector Al mesons. To give the ϱ meson a mass, it was necessary to insert a common ϱ and Al mass term in the Lagrangian, and the spontaneous

breakdown of the SU(2) x SU(2) symmetry would then split the ϱ and Al by something like the Higgs mechanism, but since the theory would not be gauge invariant the pions would remain as physical Goldstone bosons. This theory gave an intriguing result, that the Al/ϱ mass ratio should be $\sqrt{2}$, and in trying to understand this result without relying on perturbation theory, I discovered certain sum rules, the "spectral function sum rules," [23] which turned out to have variety of other uses. But the SU(2) x SU(2) theory was not gauge invariant, and hence it could not be renormalizable, [24] so I was not too enthusiastic about it. [25] Of course, if I did not insert the ϱ-Al mass term in the Lagrangian, then the theory would be gauge invariant and renormalizable, and the Al would be massive. But then there would be no pions and the ϱ mesons would be massless, in obvious contradiction (to say the least) with observation.

At some point in the fall of 1967, I think while driving to my office at M.I.T., it occurred to me that I had been applying the right ideas to the wrong problem. It is not the ϱ mesons that is massless: it is the photon. And its partner is not the Al, but the massive intermediate boson, which since the time of Yukawa had been suspected to be the mediator of the weak interactions. The weak and electromagnetic interactions could then be described [26] in a unified way in terms of an exact but spontaneously broken gauge symmetry. [Of course, not necessarily SU(2) X SU(2)]. And this theory would be renormalizable like quantum electrodynamics because it is gauge invariant like quantum electrodynamics.

It was not difficult to develop a concrete model which embodied these ideas. I had little confidence then in my understanding of strong interactions, so I decided to concentrate on leptons. There are two left-handed electron-type leptons, the ν_{eL} and e_L, and one right-handed electron-type lepton, the e_R, so I started with the group U(2) $\times$ U(1): all unitary 2 x 2 matrices acting on the left-handed e-type leptons, together with all unitary 1 X 1 matrices acting on the right-handed e-type lepton. Breaking up U(2) into unimodular transformations and phase transformations, one could say that the group was SU(2) X U(1) X U(1). But then one of the U(l)'s could be identified with ordinary lepton number, and since lepton number appears to be conserved and there is no massless vector particle coupled to it, I decided to exclude it from the group. This left the four-parameter group SU(2) x U(1). The spontaneous breakdown of SU(2) x U(1) to the U(1) of ordinary electromagnetic gauge invariance would give masses to three of the four vector gauge bosons: the charged bosons $W^{\pm}$, and a neutral boson that I called the Z^0. The fourth boson would automatically remain massless, and could be identified as the photon. Knowing the strength of the ordinary charged current weak interactions like beta decay which are mediated by $W^{\pm}$, the mass of the $W^{\pm}$ was then determined as about 40 $GeV/\sin\theta$, where θ is the γ-Z^0 mixing angle.

To go further, one had to make some hypothesis about the mechanism for the breakdown of SU (2) x U (1). The only kind of field in a renormalizable SU(2) X U(1) theory whose vacuum expectation values could give the

electron a mass is a spin zero SU(2) doublet (Φ^+, Φ^0), so for simplicity I assumed that these were the only scalar fields in the theory. The mass of the Z^0 was then determined as about 80 GeV/sin 2Θ. This fixed the strength of the neutral current weak interactions. Indeed, just as in QED, once one decides on the menu of fields in the theory all details of the theory are completely determined by symmetry principles and renormalizability, with just a few free parameters: the lepton charge and masses, the Fermi coupling constant of beta decay, the mixing angle Θ, and the mass of the scalar particle. (It was of crucial importance to impose the constraint of renormalizability; otherwise weak interactions would receive contributions from SU(2)xU(I)-invariant four-fermion couplings as well as from vector boson exchange, and the theory would lose most of its predictive power.) The naturalness of the whole theory is well demonstrated by the fact that much the same theory was independently developed [27] by Salam in 1968.

The next question now was renormalizability. The Feynman rules for Yang-Mills theories with unbroken gauge symmetries had been worked out [28] by deWitt, Faddeev and Popov and others, and it was known that such theories are renormalizable. But in 1967 I did not know how to prove that this renormalizability was not spoiled by the spontaneous symmetry breaking. I worked on the problem on and off for several years, partly in collaboration with students, [29] but I made little progress. With hindsight, my main difficulty was that in quantizing the vector fields I adopted a gauge now known as the unitarity gauge [30]: this gauge has several wonderful advantages, it exhibits the true particle spectrum of the theory, but it has the disadvantage of making renormalizability totally obscure.

Finally, in 1971 ‘t Hooft [31] showed in a beautiful paper how the problem could be solved. He invented a gauge, like the “Feynman gauge” in QED, in which the Feynman rules manifestly lead to only a finite number of types of ultraviolet divergence. It was also necessary to show that these infinities satisfied essentially the same constraints as the Lagrangian itself, so that they could be absorbed into a redefinition of the parameters of the theory. (This was plausible, but not easy to prove, because a gauge invariant theory can be quantized only after one has picked a specific gauge, so it is not obvious that the ultraviolet divergences satisfy the same gauge invariance constraints as the Lagrangian itself.) The proof was subsequently completed [32] by Lee and Zinn-Justin and by ‘t Hooft and Veltman. More recently, Becchi, Rouet and Stora [33] have invented an ingenious method for carrying out this sort of proof, by using a global supersymmetry of gauge theories which is preserved even when we choose a specific gauge.

I have to admit that when I first saw ‘t Hooft’s paper in 197 1, I was not convinced that he had found the way to’ prove renormalizability. The trouble was not with ‘t Hooft, but with me: I was simply not familiar enough with the path integral formalism on which ‘t Hooft’s work was based, and I wanted to see a derivation of the Feynman rules in ‘t Hooft’s

gauge from canonical quantization. That was soon supplied (for a limited class of gauge theories) by a paper of Ben Lee, [34] and after Lee's paper I was ready to regard the renormalizability of the unified theory as essentially proved.

By this time, many theoretical physicists were becoming convinced of the general approach that Salam and I had adopted: that is, the weak and electromagnetic interactions are governed by some group of exact local gauge symmetries; this group is spontaneously broken to U(l), giving mass to all the vector bosons except the photon; and the theory is renormalizable. What was not so clear was that our specific simple model was the one chosen by nature. That, of course, was a matter for experiment to decide.

It was obvious even back in 1967 that the best way to test the theory would be by searching for neutral current weak interactions, mediated by the neutral intermediate vector boson, the Z^0. Of course, the possibility of neutral currents was nothing new. There had been speculations [35] about possible neutral currents as far back as 1937 by Gamow and Teller, Kemmer, and Wentzel, and again in 1958 by Bludman and Leite-Lopes. Attempts at a unified weak and electromagnetic theory had been made [36] by Glashow and Salam and Ward in the early 1960's, and these had neutral currents with many of the features that Salam and I encountered in developing the 1967-68 theory. But since one of the predictions of our theory was a value for the mass of the Z^0, it made a definite prediction of the strength of the neutral currents. More important, now we had a comprehensive quantum field theory of the weak and electromagnetic interactions that was physically and mathematically satisfactory in the same sense as was quantum electrodynamics-a theory that treated photons and intermediate vector bosons on the same footing, that was based on an exact symmetry principle, and that allowed one to carry calculations to any desired degree of accuracy. To test this theory, it had now become urgent to settle the question of the existence of the neutral currents.

Late in 1971, I carried out a study of the experimental possibilites. [37] The results were striking. Previous experiments had set upper bounds on the rates of neutral current processes which were rather low, and many people had received the impression that neutral currents were pretty well ruled out, but I found that in fact the 1967-68 theory *predicted* quite low rates, low enough in fact to have escaped clear detection up to that time. For instance, experiments [38] a few years earlier had found an upper bound of 0.12 ± 0.06 on the ratio of a neutral current process, the elastic scattering of muon neutrinos by protons, to the corresponding charged current process, in which a muon is produced. I found a predicted ratio of 0.15 to 0.25, depending on the value of the Z^0-γ mixing angle θ. So there was every reason to look a little harder.

As everyone knows, neutral currents were finally discovered [39] in 1973. There followed years of careful experimental study on the detailed properties of the neutral currents. It would take me too far from my subject to survey these experiments, [40] so I will just say that they have

confirmed the 1967-68 theory with steadily improving precision for neutrino-nucleon and neutrino-electron neutral current reactions, and since the remarkable SLAC-Yale experiment [41] last year, for the electron-nucleon neutral current as well.

This is all very nice. But I must say that I would not have been too disturbed if it had turned out that the correct theory was based on some other spontaneously broken gauge group, with very different neutral currents. One possibility was a clever SU(2) theory proposed in 1972 by Georgi and Glashow, [42] which has no neutral currents at all. The important thing to me was the idea of an exact spontaneously broken gauge symmetry, which connects the weak and electromagnetic interactions, and allows these interactions to be renormalizable. Of this I was convinced, if only because it fitted my conception of the way that nature ought to be.

There were two other relevant theoretical developments in the early 1970's, before the discovery of neutral currents, that I must mention here. One is the important work of Glashow, Iliopoulos, and Maiani on the charmed quark. [43] Their work provided a solution to what otherwise would have been a serious problem, that of neutral strangeness changing currents. I leave this topic for Professor Glashow's talk. The other theoretical development has to do specifically with the strong interactions, but it will take us back to one of the themes of my talk, the theme of symmetry.

In 1973, Politzer and Gross and Wilczek discovered [44] a remarkable property of Yang-Mills theories which they called "asymptotic freedom" - the effective coupling constant [45] decreases to zero as the characteristic energy of a process goes to infinity. It seemed that this might explain the experimental fact that the nucleon behaves in high energy deep inelastic electron scattering as if it consists of essentially free quarks. [46] But there was a problem. In order to give masses to the vector bosons in a gauge theory of strong interactions one would want to include strongly interacting scalar fields, and these would generally destroy asymptotic freedom. Another difficulty, one that particularly bothered me, was that in a unified theory of weak and electromagnetic interactions the fundamental weak coupling is of the same order as the electronic charge, e, so the effects of virtual intermediate vector bosons would introduce much too large violations of parity and strangeness conservation, of order $1/137$, into the strong interactions of the scalars with each other and with the quarks. [47] At some point in the spring of 1973 it occurred to me (and independently to Gross and Wilczek) that one could do away with strongly interacting scalar fields altogether, allowing the strong interaction gauge symmetry to remain unbroken so that the vector bosons, or "gluons", are massless, and relying on the increase of the strong forces with increasing distance to explain why quarks as well as the massless gluons are not seen in the laboratory. [48] Assuming no strongly interacting scalars, three "colors" of quarks (as indicated by earlier work of several authors [49]), and an SU(3) gauge group, one then had a specific theory of strong interactions, the theory now generally known as quantum chromodynamics.

Experiments since then have increasingly confirmed QCD as the correct theory of strong interactions. What concerns me here, though, is its impact on our understanding of symmetry principles. Once again, the constraints of gauge invariance and renormalizability proved enormously powerful. These constraints force the Lagrangian to be so simple, that the strong interactions in QCD must conserve strangeness, charge conjugation, and (apart from problems [50] having to do with instantons) parity. One does not have to assume these symmetries as *a* priori principles; there is simply no way that the Lagrangian can be complicated enough to violate them. With one additional assumption, that the u and d quarks have relatively small masses, the strong interactions must also satisfy the approximate SU(2) X SU(2) symmetry of current algebra, which when spontaneously broken leaves us with isospin. If the s quark mass is also not too large, then one gets the whole eight-fold way as an approximate symmetry of the strong interactions. And the breaking of the SU(3)xSU(3) symmetry by quark masses has just the (3,3)+(3,3) form required to account for the pion-pion scattering lengths [15] and Gell-Mann-Okubo mass formulas. Furthermore, with weak and electromagnetic interactions also described by a gauge theory, the weak currents are necessarily just the currents associated with these strong interaction symmetries. In other words, pretty much the whole pattern of approximate symmetries of strong, weak, and electromagnetic interactions that puzzled us so much in the 1950's and 1960's now stands explained as a simple consequence of strong, weak, and electromagnetic gauge invariance, plus renormalizability. Internal symmetry is now at the point where space-time symmetry was in Einstein's day. All the approximate internal symmetries are explained dynamically. On a fundamental level, there are no approximate or partial symmetries; there are only exact symmetries which govern all interactions.

I now want to look ahead a bit, and comment on the possible future development of the ideas of symmetry and renormalizability.

We are still confronted with the question whether the scalar particles that are responsible for the spontaneous breakdown of the electroweak gauge symmetry SU(2) X U(1) are really elementary. If they are, then spin zero semi-weakly decaying "Higgs bosons" should be found at energies comparable with those needed to produce the intermediate vector bosons. On the other hand, it may be that the scalars are composites. [51] The Higgs bosons would then be indistinct broad states at very high mass, analogous to the possible s-wave enhancement in π-π scattering. There would probably also exist lighter, more slowly decaying, scalar particles of a rather different type, known as pseudo-Goldstone bosons. [52] And there would have to exist a new class of "extra strong" interactions [53] to provide the binding force, extra strong in the sense that asymptotic freedom sets in not at a few hundred MeV, as in QCD, but at a few hundred GeV. This "extra strong" force would be felt by new families of fermions, and would give these fermions masses of the order of several hundred GeV. We shall see.

Of the four (now three) types of interactions, only gravity has resisted incorporation into a renormalizable quantum field theory. This may just mean that we are not being clever enough in our mathematical treatment of general relativity. But there is another possibility that seems to me quite plausible. The constant of gravity defines a unit of energy known as the Planck energy, about 10^{19} GeV. This is the energy at which gravitation becomes effectively a strong interaction, so that at this energy one can no longer ignore its ultraviolet divergences. It may be that there is a whole world of new physics with unsuspected degrees of freedom at these enormous energies, and that general relativity does not provide an adequate framework for understanding the physics of these superhigh energy degrees of freedom. When we explore gravitation or other ordinary phenomena, with particle masses and energies no greater than a TeV or so, we may be learning only about an "effective" field theory; that is, one in which superheavy degrees of freedom do not explicitly appear, but the coupling parameters implicitly represent sums over these hidden degrees of freedom.

To see if this makes sense, let us suppose it is true, and ask what kinds of interactions we would expect on this basis to find at ordinary energy. By "integrating out" the superhigh energy degrees of freedom in a fundamental theory, we generally encounter a very complicated effective field theory - so complicated, in fact, that it contains *all* interactions allowed by symmetry principles. But where dimensional analysis tells us that a coupling constant is a certain power of some mass, that mass is likely to be a typical superheavy mass, such as 10^{19} GeV. The infinite variety of nonrenormalizable interactions in the effective theory have coupling constants with the dimensionality of negative powers of mass, so their effects are suppressed at ordinary energies by powers of energy divided by superheavy masses. Thus the only interactions that we can detect at ordinary energies are those that are renormalizable in the usual sense, plus any nonrenormalizable interactions that produce effects which, although tiny, are somehow exotic enough to be seen.

One way that a very weak interaction could be detected is for it to be coherent and of long range, so that it can add up and have macroscopic effects. It has been shown [54] that the only particles whose exchange could produce such forces are massless particles of spin 0, 1, or 2. And furthermore, Lorentz's invariance alone is enough to show that the long-range interactions produced by any particle of mass zero and spin 2 must be governed by general relativity. [55] Thus from this point of view we should not be too surprised that gravitation is the only interaction discovered so far that does not seem to be described by a renormalizable field theory - it is almost the only superweak interaction that *could* have been detected. And we should not be surprised to find that gravity is well described by general relativity at macroscopic scales, even if we do not think that general relativity applies at 10^{19} GeV.

Non-renormalizable effective interactions may also be detected if they violate otherwise exact conservation laws. The leading candidates for violation are baryon and lepton conservation. It is a remarkable consequence of the SU(3) and SU(2) x U(1) gauge symmetries of strong, weak, and electromagnetic interactions, that all renormalizable interactions among known particles automatically conserve baryon and lepton number. Thus, the fact that ordinary matter seems pretty stable, that proton decay has not been seen, should not lead us to the conclusion that baryon and lepton conservation are fundamental conservation laws. To the accuracy with which they have been verified, baryon and lepton conservation can be explained as dynamical consequences of other symmetries, in the same way that strangeness conservation has been explained within QCD. But superheavy particles may exist, and these particles may have unusual SU(3) or SU(2) x SU(1) transformation properties, and in this case, there is no reason why their interactions should conserve baryon or lepton number. I doubt that they would. Indeed, the fact that the universe seems to contain an excess of baryons over antibaryons should lead us to suspect that baryon nonconserving processes have actually occurred. If effects of a tiny nonconservation of baryon or lepton number such as proton decay or neutrino masses are discovered experimentally, we will then be left with gauge symmetries as the only true internal symmetries of nature, a conclusion that I would regard as most satisfactory.

The idea of a new scale of superheavy masses has arisen in another way. [56] If any sort of "grand unification" of strong and electroweak gauge couplings is to be possible, then one would expect all of the SU(3) and SU(2) x U(1) gauge coupling constants to be of comparable magnitude. (In particular, if SU(3) and SU(2) x U(1) are subgroups of a larger simple group, then the ratios of the squared couplings are fixed as rational numbers of order unity.[57]) But this appears in contradiction with the obvious fact that the strong interactions are stronger than the weak and electromagnetic interactions. In 1974 Georgi, Quinn and I suggested that the grand unification scale, at which the couplings are comparable, is at an enormous energy, and that the reason that the strong coupling is so much larger than the electroweak couplings at ordinary energies is that QCD is asymptotically free, so that its effective coupling constant rises slowly as the energy drops from the grand unification scale to ordinary values. The change of the strong couplings is very slow (like $1/\sqrt{\ln E}$) so the grand unification scale must be enormous. We found that for a fairly large class of theories the grand unification scale comes out to be in the neighborhood of 10^{16} GeV, an energy not all that different from the Planck energy of 10^{19} GeV. The nucleon lifetime is very difficult to estimate accurately, but we gave a representative value of 10^{32} years, which may be accessible experimentally in a few years. (These estimates have been improved in more detailed calculations by several authors.) [58] We also calculated a value for the mixing parameter $\sin^2\Theta$ of about 0.2, not far from the present experimental value[40] of 0.23±0.01. It will be an important task for future

experiments on neutral currents to improve the precision with which $\sin^2\Theta$ is known, to see if it really agrees with this prediction.

In a grand unified theory, in order for elementary scalar particles to be available to produce the spontaneous breakdown of the electroweak gauge symmetry at a few hundred GeV, it is necessary for such particles to escape getting superlarge masses from the spontaneous breakdown of the grand unified gauge group. There is nothing impossible in this, but I have not been able to think of any reason why it should happen. (The problem may be related to the old mystery of why quantum corrections do not produce an enormous cosmological constant; in both cases, one is concerned with an anomalously small "super-renormalizable" term in the effective Lagrangian which has to be adjusted to be zero. In the case of the cosmological constant, the adjustment must be precise to some fifty decimal places.) With elementary scalars of small or zero bare mass, enormous ratios of symmetry breaking scales can arise quite naturally [59]. On the other hand, if there are no elementary scalars which escape getting superlarge masses from the breakdown of the grand unified gauge group, then as I have already mentioned, there must be extra strong forces to bind the composite Goldstone and Higgs bosons that are associated with the spontaneous breakdown of SU(2) x U(1). Such forces can occur rather naturally in grand unified theories. To take one example, suppose that the grand gauge group breaks, not into SU(3) x SU(2) x U(l), but into SU(4) x SU(3) x SU(2) x U(1). Since SU(4) is a bigger group than SU(3), its coupling constant rises with decreasing energy more rapidly than the QCD coupling, so the SU(4) force becomes strong at a much higher energy than the few hundred MeV at which the QCD force becomes strong. Ordinary quarks and leptons would be neutral under SU(4), so they would not feel this force, but other fermions might carry SU(4) quantum numbers, and so get rather large masses. One can even imagine a sequence of increasingly large subgroups of the grand gauge group, which would fill in the vast energy range up to 10^{15} or 10^{19} GeV with particle masses that are produced by these successively stronger interactions.

If there are elementary scalars whose vacuum expectation values are responsible for the masses of ordinary quarks and leptons, then these masses can be affected in order α by radiative corrections involving the superheavy vector bosons of the grand gauge group, and it will probably be impossible to explain the value of quantities like m_e/m_μ without a complete grand unified theory. On the other hand, if there are no such elementary scalars, then almost all the details of the grand unified theory are forgotten by the effective field theory that describes physics at ordinary energies, and it ought to be possible to calculate quark and lepton masses purely in terms of processes at accessible energies. Unfortunately, no one so far has been able to see how in this way anything resembling the observed pattern of masses could arise. [60]

Putting aside all these uncertainties, suppose that there is a truly fundamental theory, characterized by an energy scale of order 10^{16} to 10^{19} GeV,

at which strong, electroweak, and gravitational interactions are all united. It might be a conventional renormalizable quantum field theory, but at the moment, if we include gravity, we do not see how this is possible. (I leave the topic of supersymmetry and supergravity for Professor Salam's talk.) But if it is not renormalizable, what then determines the infinite set of coupling constants that are needed to absorb all the ultraviolet divergences of the theory?

I think the answer must lie in the fact that the quantum field theory, which was born just fifty years ago from the marriage of quantum mechanics with relativity, is a beautiful but not very robust child. As Landau and Kallen recognized long ago, quantum field theory at superhigh energies is susceptible to all sorts of diseases--tachyons, ghosts, etc.-and it needs special medicine to survive. One way that a quantum field theory can avoid these diseases is to be renormalizable and asymptotically free, but there are other possibilities. For instance, even an infinite set of coupling constants may approach a non-zero fixed point as the energy at which they are measured goes to infinity. However, to require this behavior generally imposes so many constraints on the couplings that there are only a finite number of free parameters left[6 1] -just as for theories that are renormalizable in the usual sense. Thus, one way or another, I think that quantum field theory is going to go on being very stubborn, refusing to allow us to describe all but a small number of possible worlds, among which, we hope, is ours.

I suppose that I tend to be optimistic about the future of physics. And nothing makes me more optimistic than the discovery of broken symmetries. In the seventh book of the *Republic,* Plato describes prisoners who are chained in a cave and can see only shadows that things outside cast on the cave wall. When released from the cave at first their eyes hurt, and for a while they think that the shadows they saw in the cave are more real than the objects they now see. But eventually their vision clears, and they can understand how beautiful the real world is. We are in such a cave, imprisoned by the limitations on the sorts of experiments we can do. In particular, we can study matter only at relatively low temperatures, where symmetries are likely to be spontaneously broken, so that nature does not appear very simple or unified. We have not been able to get out of this cave, but by looking long and hard at the shadows on the cave wall, we can at least make out the shapes of symmetries, which though broken, are exact principles governing all phenomena, expressions of the beauty of the world outside.

***

It has only been possible here to give references to a very small part of the literature on the subjects discussed in this talk. Additional references can be found in the following reviews:.

Abers, E.S. and Lee, B.W., Gauge *Theories* (Physics Reports 9C, No. 1, 1973).

Marciano, W. and Pagels, H., *Quantum Chromodynamics* (Physics Reports *36C, No. 3,* 1978).

Taylor, J.C., *Gauge Theories of Weak Interactions* (Cambridge Univ. Press, 1976).

REFERENCES

1. Tuve, M. A., Heydenberg, N. and Hafstad, L. R. Phys. Rev. 50, 806 (1936); Breit, G., Condon, E. V. and Present, R. D. Phys. Rev. 50, 825 (1936); Breit, G. and Feenberg, E. Phys. Rev. 50, 850 (1936).
2. Gell-Mann, M. Phys. Rev. 92, 833 (1953); Nakano. T. and Nishijima, K. Prog. Theor. Phys. 10, 581 (1955).
3. Lee, T. D. and Yang, C. N. Phys. Rev. 104, 254 (1956); Wu. C. S. et.al. Phys. Rev. 105, 1413 (1957); Garwin, R., Lederman, L. and Weinrich, M. Phys. Rev. 105, 1415 (1957); Friedman, J. I. and Telegdi V. L. Phys. Rev. 105, 1681 (1957).
4. Gell-Mann, M. Cal. Tech. Synchotron Laboratory Report CTSL-20 (1961). unpublished; Ne'eman, Y. Nucl. Phys. 26, 222 (1961).
5. Fock, V. Z. f. Physik 39, 226 (1927); Weyl, H. Z. f. Physik 56, 330 (1929). The name "gauge invariance" is based on an analogy with the earlier speculations of Weyl, H. in Raum, Zeit, ***Materie,*** 3rd edn, (Springer, 1920). Also see London, F. Z. f. Physik 42, 375 (1927). (This history has been reviewed by Yang, C. N. in a talk at City College, (1977).)
6. Yang, C. N. and Mills, R. L. Phys. Rev. 96, 191 (1954).
7. Goldstone, J. Nuovo Cimento ***19,*** 154 (1961).
8. Goldstone, J., Salam, A. and Weinberg, S. Phys. Rev. 127, 965 (1962).
9. Higgs, P. W. Phys. Lett. 12, 132 (1964); 13, 508 (1964); Phys. Rev. ***145,*** 1156 (1966); Kibble, T. W. B. Phys. Rev. ***155,*** 1554 (1967); Guralnik, G. S., Hagen, C. R. and Kibble, T. W. B. Phys. Rev. Lett. 13, 585 (1964); Englert, F. and Brout, R. Phys. Rev. Lett. 13, 32 1 (1964); Also see Anderson, P. W. Phys. Rev. 130, 439 (1963).
10. Adler, S. L. Phys. Rev. Lett. ***14,*** 1051 (1965); Phys Rev. 140, B736 (1965); Weisberger, W. I. Phys. Rev. Lett. 14, 1047 (1965); Phys Rev. 143, 1302 (1966).
11. Gell-Mann, M. Physics I, 63 (1964).
12. Nambu, Y. and Jona-Lasinio, G. Phys. Rev. ***122,*** 345 (1961); ***124,*** 246 (1961); Nambu, Y, and Lurie, D. Phys. Rev. ***125,*** 1429 (1962); Nambu. Y. and Shrauner, E. Phys. Rev. ***128,*** 862 (1962); Also see Gell-Mann, M. and Levy, M., Nuovo Cimento 16, 705 (1960).
13. Goldberger, M. L., Miyazawa, H. and Oehme, R. Phys Rev. 99, 986 (1955).
14. Goldberger, M. L., and Treiman, S. B. Phys. Rev. ***111, 354*** (1958).
15 .Weinberg, S. Phys. Rev. Lett. 16, 879 (1966); 17, 336 (1966); ***17,*** 616 (1966); 18, 188 (1967); Phys Rev. 166, 1568 (1967).
16. Oppenheimer, J, R. Phys. Rev. 35, 461 (1930); Waller, I. Z. Phys. 59, 168 (1930); ibid., ***62, 673*** (1930).
17. Feynman, R. P. Rev. Mod. Phys. 20, 367 (1948); Phys. Rev. 74, 939, 1430 (1948); 76, 749, 769 (1949); 80, 440 (1950); Schwinger, J. Phys. Rev. 73, 146 (1948); 74, 1439 (1948); 75, 651 (1949); 76, 790 (1949); 82, 664, 914 (1951);91, 713 (1953); Proc. Nat. Acad. Sci.37, 452 (1951); Tomonaga, S. Progr. Theor. Phys. (Japan) I, 27 (1946); Koba, Z., Tati, T. and Tomonaga, S. ibid. 2, 101 (1947); Kanazawa, S. and Tomonaga, S. ibid. 3, 276 (1948); Koba, Z. and Tomonaga, S. ibid 3, 290 (1948).
18. There had been earlier suggestions that infinities could be eliminated from quantum field theories in this way, by Weisskopf, V. F. Kong. Dansk. Vid. Sel. Mat.-Fys. Medd. ***15*** (6) 1936, especially p. 34 and pp. 5-6; Kramers,.H. (unpublished).
19. Dyson, F. J. Phys. Rev. 75, 486, 1736 (1949).
20. Weinberg, S. Phys. Rev. 106, 1301 (1957).
21. Weinberg, S. Phys. Rev. 118, 838 (1960).
22. Salam, A. Phys. Rev. 82, 217 (1951); 84, 426 (1951).

23. Weinberg, S. Phys. Rev. Lett. ***18, 507 (1967).***
24. For the non-renormalizability of theories with intrinsically broken gauge symmetries, see Komar, A. and Salam, A. Nucl. Phys. 21, 624 (1960); Umezawa, H. and Kamefuchi, S. Nucl. Phys. 23, 399 (1961); Kamefuchi, S., O'Raifeartaigh, L. and Salam, A. Nucl. Phys. 28, 529 (1961); Salam, A. Phys. Rev. 127, 331 (1962); Veltman, M. Nucl. Phys. B7, 637 (1968); B21, 288 (1970); Boulware, D. Ann. Phys. (N, Y,)56, 140 (1970).
25. This work was briefly reported in reference 23, footnote 7.
26. Weinberg, S. Phys. Rev. Lett. ***19,*** 1264 (1967).
27. Salam, A. In ***Elementary Particle Physics*** (Nobel Symposium No. 8), ed. by Svartholm, N. (Almqvist and Wiksell, Stockholm, 1968), p. 367.
28. deWitt, B. Phys. Rev. Lett. 12, 742 (1964); Phys. Rev. 162, 1195 (1967); Faddeev L. D., and Popov, V. N. Phys. Lett. B25, 29 (1967); Also see Feynman, R. P. Acta. Phys. Pol. 24, 697 (1963); Mandelstam, S. Phys. Rev. 175, 1580, 1604 (1968).
29. See Stuller, I.. M. I. T., Thesis, Ph. D. (1971), unpublished.
30. My work with the unitarity gauge was reported in Weinberg, S. Phys. Rev. Lett. 27, 1688 (1971), and described in more detail in Weinberg, S. Phys. Rev. D7, 1068 (1973).
31. 't Hooft, G Nucl. Phys. B35, 167 (1971).
32. Lee, B. W. and Zinn-Justin, J. Phys. Rev. D5, 3121, 3137, 3155 (1972); 't Hooft, G. and Veltman, M. Nucl. Phys. 844, 189 (1972), B50, 318 (1972). There still remained the problem of possible Adler-Bell-Jackiw anomalies, but these nicely cancelled; see D. J. Gross and R. Jackiw, Phys. Rev. D6, 477 (1972) and C. Bouchiat, J. lliopoulos, and Ph. Meyer, Phys. Lett. 388, 519 (1972).
33. Beechi, C., Rouet, A. and Stora R. Comm. Math. Phys. 42, 127 (1975).
34. Lee, B. W. Phys. Rev. ***D5, 823*** (1972).
35. Gamow, G. and Teller, E. Phys. Rev. 51, 288 (1937); Kemmer, N. Phys. Rev. 52, 906 (1937); Wentrel, G. Helv. Phys. Acta. 10, 108 (1937); Bludman, S. Nuovo Cimento 9, 433 (1958); Leite-Lopes, J. Nucl. Phys. 8, 234 (1958).
36. Glashow, S. L. Nucl. Phys. 22, 519 (1961); Salam, A. and Ward, J. C. Phys. Lett. 13, 168 (1964).
37. Weinberg, S. Phys. Rev. 5, 1412 (1972).
38. Cundy, D. C. et al., Phys. Lett. 31B, 478 (1970).
39. The first published discovery of neutral currents was at the Gargamelle Bubble Chamber at CERN: Hasert, F. J. et al., Phys. Lett. 468, 121, 138 (1973). Also see Musset, P. Jour. de Physique 11 /12 T34 (1973). Muonless events were seen at about the same time by the HPWF group at Fermilab, but when publication of their paper was delayed, they took the opportunity to rebuild their detector, and then did not at first find the same neutral current signal. The HPWF group published evidence for neutral currents in Benvenuti, A. et al., Phys. Rev. Lett. 52, 800 (1974).
40. For a survey of the data see Baltay, ***C. Proceedings*** *of* ***the 19th*** International Conference on ***High Energy Physics,*** Tokyo, 1978. For theoretical analyses, see Abbott, L. F. and Barnett, R. M. Phys. Rev. D19, 3230 (1979); Langacker, P., Kim, J. E., Levine, M., Williams, H. H. and Sidhu, D. P. Neutrino Conference '79; and earlier references cited therein.
41. Prescott, C. Y. et.al., Phys. Lett. 778, 347 (1978).
42. Glashow, S. L. and Georgi, H. L. Phys. Rev. Lett. 28, 1494 (1972). Also see Schwinger, J. Annals of Physics (N. Y.)2, 407 (1957).
43. Glashow, S. L., Iliopoulos, J. and Maiani, L. Phys. Rev. D2, 1285 (1970). This paper was cited in ref. 37 as providing a possible solution to the problem of strangeness changing neutral currents. However, at that time I was skeptical about the quark model, so in the calculations of ref. 37 baryons were incorporated in the theory by taking the protons and neutrons to form an SU(2) doublet, with strange particles simply ignored.
44. Politzer, H. D. Phys. Rev. Lett. 30, 1346 (1973); Gross, D. J. and Wilczek, F. Phys. Rev. Lett. 30, 1343 (1973).
45. Energy dependent effective couping constants were introduced by Gell-Mann, M. and Low, F. E. Phys. Rev. 95, 1300 (1954).
46. Bloom, E. D. et.al., Phys. Rev. Lett. 23, 930 (1969); Breidenbach, M. et.al., Phys. Rev. Lett. 23, 935 (1969).

47. Weinberg, S. Phys. Rev. ***D8, 605 (1973).***
48. Gross, D. J. and Wilczek, F. Phys. Rev. ***D8,*** 3633 (1973); Weinberg, S. Phys. Rev. Lett. 31, 494 (1973). A similar idea had been proposed before the discovery of asymptotic freedom by Fritzsch, H., Gell-Mann, M. and Leutwyler, H. Phys. Lett. 478, 365 (1973).
49. Greenberg, 0. W. Phys. Rev. Lett. 13, 598 (1964); Han, M. Y. and Nambu, Y. Phys. Rev. 139, B1006 (1965); Bardeen, W. A., Fritzsch, H. and Gell-Mann, M. in Scale and Conformu1 Symmetry in ***Hadron Physics,*** ed. by Gatto, R. (Wiley, 1973), p. 139; etc.
50. 't Hooft, G. Phys. Rev. Lett. 37, 8 (1976).
51. Such "dynamical" mechanisms for spontaneous symmetry breaking were first discussed by Nambu, Y. and Jona-Lasinio, G. Phys. Rev. 122, 345 (1961); Schwinger, J. Phys. Rev. 125, 397 (1962); ***128,*** 2425 (1962); and in the context of modern gauge theories by Jackiw, R. and Johnson, K. Phys. Rev. ***D8,*** 2386 (1973); Cornwall, J. M. and Norton, R. E. Phys. Rev. ***D8,*** 3338 (1973). The implications of dynamical symmetry breaking have been considered by Weinberg, S. Phys. Rev. ***D13,*** 974 (1976); ***D19,*** 1277 (1979); Susskind, L. Phys. Rev. ***D20,*** 2619 (1979).
52. Weinberg, S. ref 51, The possibility of pseudo-Goldstone bosons was originally noted in a different context by Weinberg, S. Phys. Rev. Lett. 29, 1698 (1972).
53. Weinberg, S. ref. 51. Models involving such interactions have also been discussed by Susskind, L. ref. 51.
54. Weinberg, S. Phys. Rev. ***135,*** B1049 (1964).
55. Weinberg. S. Phys. Lett. ***9, 357*** (1964); Phys. Rev. ***8138, 988*** (1965); ***Lectures in Particles*** and Field ***Theory,*** ed. by Deser, S. and Ford, K. (Prentice-Hall, 1965), p. 988; and ref. 54. The program of deriving general relativity from quantum mechanics and special relativity was completed by Boulware, D. and Deser, S. Ann. Phys. 89, 173 (1975). I understand that similar ideas were developed by Feynman, R. in unpublished lectures at Cal. Tech.
56. Georgi, H., Quinn, H. and Weinberg, S. Phys. Rev. Lett. 33, 45 1 (1974).
57. An example of a simple gauge group for weak and electromagnetic interactions (for which $\sin^2\theta=\frac{1}{4}$) was given by S. Weinberg, Phys. Rev. ***D5,*** 1962 (1972). There are a number of specific models of weak, electromagnetic, and strong interactions based on simple gauge groups, including those of Pati, J. C. and Salam, A. Phys. Rev. ***D10, 275*** (1974); Georgi, H. and Glashow, S. L. Phys. Rev. Lett. 32, 438 (1974); Georgi, H. in ***Particles and Fields*** (American Institute of Physics, 1975); Fritzsch, H. and Minkowski, P. Ann. Phys. 93, 193 (1975); Georgi, H. and Nanopoulos, D. V. Phys. Lett. ***82B, 392*** (1979); Gürsey, F. Ramond, P. and Sikivie, P. Phys. Lett. ***B60,*** 177 (1975); Gürsey, F. and Sikivie, P. Phys. Rev. Lett. 36, 775 (1976); Ramond, P. Nucl. Phys, ***B110,*** 214 (1976); etc; all these violate baryon and lepton conservation, because they have quarks and leptons in the same multiplet; see Pati, J. C. and Salam, A. Phys. Rev. Lett. 31, 661 (1973); Phys. Rev. D8, 1240 (1973).
58. Buras, A., Ellis, J., Gaillard, M. K. and Nanopoulos, D. V. Nucl. Phys. ***B135, 66*** (1978); Ross, D. Nucl. Phys. ***B140,*** 1 (1978); Marciano, W. J. Phys. Rev. ***D20, 274*** (1979); 'Goldman, T. and Ross, D. CALT 68-704, to be published; Jarlskog, C. and Yndurain, F. J. CERN preprint, to be published. Machacek, M. Harvard preprint HUTP-79/AO21, to be published in Nuclear Physics; Weinberg, S. paper in preparation. The phenomenonology of nucleon decay has been discussed in general terms by Weinberg, S. Phys. Rev. Lett. 43, 1566 (1979); Wilczek, F. and Zee, A. Phys. Rev. Lett. 43, 1571 (1979).
59. Gildener, E. and Weinberg, S. Phys. Rev. ***D13, 3333*** (1976); Weinberg, S. Phys. Letters ***82B,*** 387 (1979). In general there should exist at least one scalar particle with physical mass of order 10 GeV. The spontaneous symmetry breaking in models with zero bare scalar mass was first considered by Coleman, S. and Weinberg, E., Phys. Rev. D 7, 1888 (1973).
60. This problem has been studied recently by Dimopoulos, S. and Susskind, L. Nucl. Phys. ***B155, 237*** (1979); Eichten, E. and Lane, K. Physics Letters, to be published; Weinberg, S. unpublished.
61. Weinberg, S. in General ***Relativity - An Einstein Centenary*** Survey, ed. by Hawking, S. W. and Israel, W. (Cambridge Univ. Press, 1979), Chapter-16.

Physics 1980

JAMES W CRONIN and VAL L FITCH

for the discovery of violations of fundamental symmetry principles in the decay of neutral K-mesons

THE NOBEL PRIZE FOR PHYSICS

Speech by Professor GÖSTA EKSPONG of the Royal Academy of Sciences. Translation from the Swedish text.

Your Majesty, Your Royal Highnesses, Ladies and Gentlemen.

By decision of the Royal Swedish Academy of Sciences, this year's Nobel Prize for Physics has been awarded Professor James Cronin and Professor Val Fitch for their discovery in a joint experiment of violations of fundamental symmetry principles. The experiment was carried out in 1964 at Brookhaven National Laboratory in the United States of America and was concerned with a forbidden decay of a certain type of elementary particles, named the neutral K-meson.

Suppose the TV-news suddenly reported one evening that visitors from outer space were planning to land on Earth; that the space travellers have radioed a demand for immediate information about the composition of the Earth. Does it consist of Matter or Antimatter? The answer to this question is one of lift and death. The two kinds of matter are known to annihilate each other atom by atom. The space travelers claim, furthermore, that the nature of their own kind of matter was determined before leaving. What they now want to know is, whether the same tests have been made on Earth. Thanks to Cronin's and Fitch's discovery it is now possible to give them a clear-cut answer, so they can avoid a disastrous landing. Let us now leave the world of science fiction, remembering, however, what a fortunate circumstance it was that no space visits occurred before 1964.

Symmetries are science's lodestars and symmetry principles act as guiding rules to help us discover the mathematical laws of Nature. Three mirror symmetries arc of immediate interest in relation to the prize-winning discovery. One of them is ordinary mirror reflection, which corresponds to switching left and right. The other two symmetries of interest concern reflection of time and of charge, which implies switching forward and backward movements and switching matter and antimatter, respectively. In the latter case it is positive and negative electric charges that are switched.

The beauty of spatial symmetries is well known in the realm of art and architecture, from the ornamental arabesques of the old Alhambra to the recent intricate woodcuts signed by Escher, from the palace of the Doges in Venice to the Town Hall in Stockholm. A master such as Johann Sebastian Bach has created music with ingenious symmetries, generated both by reflection in space of the theme and by reflection in time when the theme is played backwards. The laws of physics resemble a canon by Bach. They arc symmetric in space and time. They do not distinguish between left and right, nor between forward and backward movements. For a long time everyone thought it had to be like that. A remarkable exception exists, however, in the law governing radioactive

decay, which violates the left-right symmetry. Lee and Yang were awarded the Nobel prize for physics in 1957 for this revolutionary discovery.

The third mirror symmetry is not present in art. The laws for electric and magnetic phenomena contain a complete symmetry between the two kinds of electric charge. The discoveries of antimatter with plus and minus charges in exchanged roles are among the most profound of the last half-century. Nowadays microscopic amounts of antiparticles are produced with relative ease in such special laboratories as Brookhaven National Laboratory in the U.S. or CERN in Europe.

Cronin and Fitch elected to carry out tests to find out whether a certain decay of K-mesons occurred, in spite of being forbidden by symmetry. Their research team found that two out of a thousand K-mesons did in fact decay in the forbidden manner. This means that some law of Nature now must be changed or a new law invoked. In what way does this discovery concern antimatter? As early as 1955 Gell-Mann and Pais had analyzed the neutral K-mesons and found that they are strange, indeed unique in their ambivalence with respect both to matter and antimatter. If perfect symmetry were to prevail, a decaying K-meson would have to be antimatter in exactly half the cases and in the other half, matter. Lee's and Yang's Chinese revolution did not change the conclusions, but new arguments were required. Cronin and Fitch interpreted the results of their experiment as a small but clear lack of symmetry. Their conclusion has been confirmed in a long series of other experiments. The new symmetry violation constitutes the basic prerequisite for the claim that a definite answer can be relayed to our visitors from outer space.

The discovery also implied consequences for time reflection. At least one theme is played more slowly backwards than forwards by Nature.

Artists nearly always introduce symmetry breaking elements into their works. Perhaps, the laws of nature, too, are in the deepest sense works of art. Violations of perfect symmetry open roads to new insights, or in the words of a poet:

"A knot there is in th'entendrill'd arabesque
No mortal eye but mine has ever seen".

Professor Fitch, Professor Cronin,

The scientific world was shocked when you first announced your discovery. Nobody, absolutely nobody, had anticipated anything like it. You had pursued your experiment with skill and determination and found the impossible to be possible.

On behalf of the Royal Swedish Academy of Sciences I have the pleasure and the honour of extending to you our warmest congratulations. I now invite you to receive your Prizes from the hands of His Majesty the King.

James W Cronin

JAMES W. CRONIN

I was born on September 29, 1931 in Chicago, Illinois, while my father, James Farley Cronin, was a graduate student at the University of Chicago. He was a student of classical languages. My mother, Dorothy Watson, had met my father in a Greek class at Northwestern University. After a brief stay at a small school in Alabama, my father became Professor of Latin and Greek at Southern Methodist University in Dallas, Texas, in September 1939. My primary and secondary education was provided by the Highland Park Public School System. I received my undergraduate degree from Southern Methodist University with a major in physics and mathematics in 195 1. In high school my natural interest in science was encouraged by an excellent physics teacher, Mr. Charles H. Marshall. He stressed analytical methods as applied to simple physical systems as well as practical experimental problems.

My real education began when I entered the University of Chicago in September 1951 as a graduate student. I was fortunate to have among my classroom teachers, Enrico Fermi, Maria Mayer, Edward Teller, Gregor Wentzel, Val Telegdi, Marvin Goldbergcr and Murray Gell-Mann. I did a thesis in experimental nuclear physics under the direction of Samuel K. Allison. While at Chicago my interest in the new field of particle physics was stimulated by a course given by Gell-Mann, who was developing his ideas about Strangeness at the time.

It was also at the University of Chicago that I met my future wife, Annette Martin, in the summer of 1953. It was a wonderful, happy summer; I had passed my Ph.D. qualifying exams the previous winter, and I realized that I had met my lifetime companion. We were married in September 1954. The stable point in my life became our home. On even the worst days, when nothing was working at the lab, I knew that at home 1 would find warmth, peace, companionship, and encouragement. As a consequence, the next day would surely be better. Annette, with great patience and good spirit, tolerated my many long absences when experiments were carried out at distant laboratories.

After receiving my Ph.D. in 1955 I had the opportunity to join the group of Rodney Cool and Oreste Piccioni who were working at the Brookhaven Cosmotron, a newly completed 3 GeV accelerator. That period was an exciting time in physics. The famous τ—θ puzzle led to the prediction of parity violation and the experimental demonstration of its violation. The long-lived K meson was discovered at Brookhaven.

When the violation of parity was discovered I began a series of electronic experiments to investigate parity violation in hyperon decays. In early 1958 the

Cosmotron suffered a severe magnet failure. As a consequence, we moved our experiment to the Berkeley Bevatron. Here I had the good fortune to meet William Wenzel and Bruce Cork. These physicists had a great influence on me. From their example I learned not to be intimidated by complex pieces of apparatus.

While at Brookhaven I met Val Fitch who was responsible for my coming to Princeton University in the fall of 1958. At Princeton all the work in particle physics was supported through a contract with the Office of Naval Research. The Director of the Laboratory, George Reynolds, was most supportive of my efforts to work independently. There followed for ten years a glorious time for research. I was much involved in the development of the spark chamber as a practical research tool. During this period, with a series of excellent students, we further studied hyperon decays. Then we joined with Val Fitch to study neutral K meson decays which led to the discovery of CP violation.

Following the discovery in the summer of 1964, I spent a year in France working at the Centre d'Etudes Nucléaires at Saclay with Rene Turlay. In addition to the research, I enjoyed learning French and assimilating the culture of another country. One of the greatest joys in my life was giving a lecture in French at the Collège de France.

On returning to Princeton in 1965, I began with students a series of experiments to study the neutral CP violating decay modes of the long lived neutral K meson. These experiments lasted until 1971. In 1971 I returned to the University of Chicago as Professor of Physics. The fact that the new Fermilab 400 GeV Accelerator was being built near Chicago made this move an attractive one. At Fermilab, with younger associates and students, I carried out experiments on the production of particles at high transverse momentum, and on the production of direct leptons. At present with my colleague at Chicago, Bruce Winstein, I am preparing to study with much greater accuracy some of the CP violating parameters of the neutral K meson.

I now live in Chicago near the campus with my wife Annette, and son Daniel. My oldest daughter Cathryn lives and works in New York City. My daughter Emily attends the University of Minnesota. My mother remained in Dallas, Texas, after the death of my father in 1959. For recreation we have a cabin in the woods in Wisconsin which we visit year-round. In the summer we spend some time in Aspen, Colorado. Our whole family assembles in Chicago at Christmas and usually in Aspen in the summer.

Education
B.S., Southern Methodist University, 1951
M.S., University of Chicago, 1953
Ph.D., (Physics) University of Chicago, 1955

Career
National Science Foundation Fellow, 1952-1955
Assistant Physicist, Brookhaven National Laboratory, 1955-1958

Assistant Professor of Physics, Princeton University, 1958-1962
Associate Professor of Physics, Princeton University, 1962-1964
Professor of Physics, Princeton University, 1964-1971
University Professor of Physics, University of Chicago, 1971-

Member
American Academy of Arts and Sciences
American Physical Society
National Academy of Sciences

Recipient
Research Corporation Award, 1968
John Price Wetherill Medal of the Franklin Institute, 1975
Ernest 0. Lawrence Award, 1977

CP SYMMETRY VIOLATION - THE SEARCH FOR ITS ORIGIN

Nobel lecture, 8 December, 1980

by

JAMES W. CRONIN

The University of Chicago, Illinois 60637, USA

The greatest pleasure a scientist can experience is to encounter an unexpected discovery. I am always astonished when a simple apparatus, designed to ask the right question of nature, receives a clear response. Our experiment, carried out with James Christenson, Val Fitch and Renk Turlay, gave convincing evidence that the long-lived neutral K meson (K_L) decayed into two charged pions, a decay mode forbidden by CP symmetry. The forbidden decay mode was found to be a small fraction (2.0±0.4) $\times 10^{-3}$ of all charged decay modes. Professor Fitch has described our discovery of CP symmetry violation. He has discussed how it was preceded by brilliant theoretical insights and incisive experiments with K mesons. My lecture will review the knowledge that we have obtained about CP violation since its discovery.' The discovery triggered an intense international experimental effort. It also provoked many theoretical speculations which in turn stimulated a variety of experiments.

At present there is no satisfactory theoretical understanding of CP violation. Such understanding as we do have has come entirely from experimental studies. These studies have extended beyond the high energy accelerator laboratories into nuclear physics laboratories and research reactor laboratories. The experiments which have sought to elucidate the tiny effect have involved both ingenuity and painstaking attention to detail.

Upon learning of the discovery in 1964, the natural reaction of our colleagues was to ask what was wrong with the experiment. Or, if they were convinced of the correctness of the measurements, they asked how could the effect be explained while still retaining CP symmetry. I remember vividly a special session organized at the 1964 International Conference on High Energy Physics at Dubna in the Soviet Union. There, for an afternoon, I had to defend our experiment before a large group of physicists who wanted to know every detail of the experiment-more details than could have been given in the formal conference session.

As the session neared a close, one of my Soviet colleagues suggested that, perhaps, the effect was due to regeneration of short-lived K mesons (K_s) in a fly unfortunately trapped in the helium bag. We did a quick “back of the envelope” estimate of the density of the fly necessary to produce the effect. The density required was far in excess of uranium.

More serious questions were raised at this session and by many other

physicists who had thought deeply about our result. While we were confident that the experiment had been correctly carried out and interpreted, many sought reassurance through confirmation of the experiment by other groups. This confirmation came quickly from experiments at the Rutherford Laboratory[2] in England, and at CERN[3] in Geneva, Switzerland.

Another important issue was raised. In the original experiment, the decay to two pions was inferred kinematically, but no proof was given that these pions were identical to the ordinary pions or that the decay was not accompanied by a third light particle emitted at a very low energy. The direct proof that the effect was indeed a violation of CP symmetry was the demonstration of interference between the decay of the long-lived and short-lived K meson to two charged pions. This interference was first demonstrated in a simple and elegant experiment by my colleague Val Fitch with Roth, Russ and Vernon.'

Their experiment compared the rate of decay of a K_L, beam into two charged pions in vacuum and in the presence of a diffuse beryllium regenerator. The density of the regenerator was adjusted so that the regeneration amplitude A, was equal to the CP violating amplitude η_{+-}. These amplitudes are defined by

$$\eta_{+-} = \frac{\text{amplitude } (K_L \to \pi^+\pi^-)}{\text{amplitude } (K_S \to \pi^+\pi^-)}$$

and

$$A_r = i\pi N\Lambda \left(\frac{f-\bar{f}}{k}\right)\left(i\delta+\frac{1}{2}\right)^{-1}.$$

The yield of $K_L \to \pi^+\pi^-$ in the presence of the regenerator is proportional to

$$|A_r+\eta_{+-}|^2.$$

In the expression for A_r, δ is given by $(M_S—M_L)/\Gamma_S$ where M_s and M_L, are the K_s and K_L, masses, and Γ_S the decay rate of the K_s meson, A is the mean decay length of the K_s meson, k is the wave number of the incident K_L, beam and f and f are the forward scattering amplitudes for K and K, respectively on the nuclei of the regenerator. The regeneration amplitude is proportional to N, the number density of the material. The quantity (f-f)/k was determined in an auxiliary experiment with a dense regenerator. Then a regenerator of appropriate density was constructed using the formula for A,.' The actual regenerator was constructed of 0.5 mm sheets separated by 1 cm. Such an arrangement behaves as a homogeneous regenerator of 1/20) normal density if the separation of the sheets is small compared to the quantity $\delta\Lambda$.

In the earliest experiment Fitch and his colleagues found that with $|A_r|$ chosen to be equal to $|\eta_{+-}|$ the rate of $\pi^+\pi^-$ decays was about *four times* the rate without the regenerator. This result showed not only that there was interference, but also that the interference was fully constructive. Complete analysis of this experiment reported subsequently" gave the $\pi^+\pi^-$ yield as a function of density as shown in Fig 1. The quantity *a* in the figure is the relative phase between the regeneration amplitude and the CP violating amplitude.

The result of this experiment also permits the experimental distinction

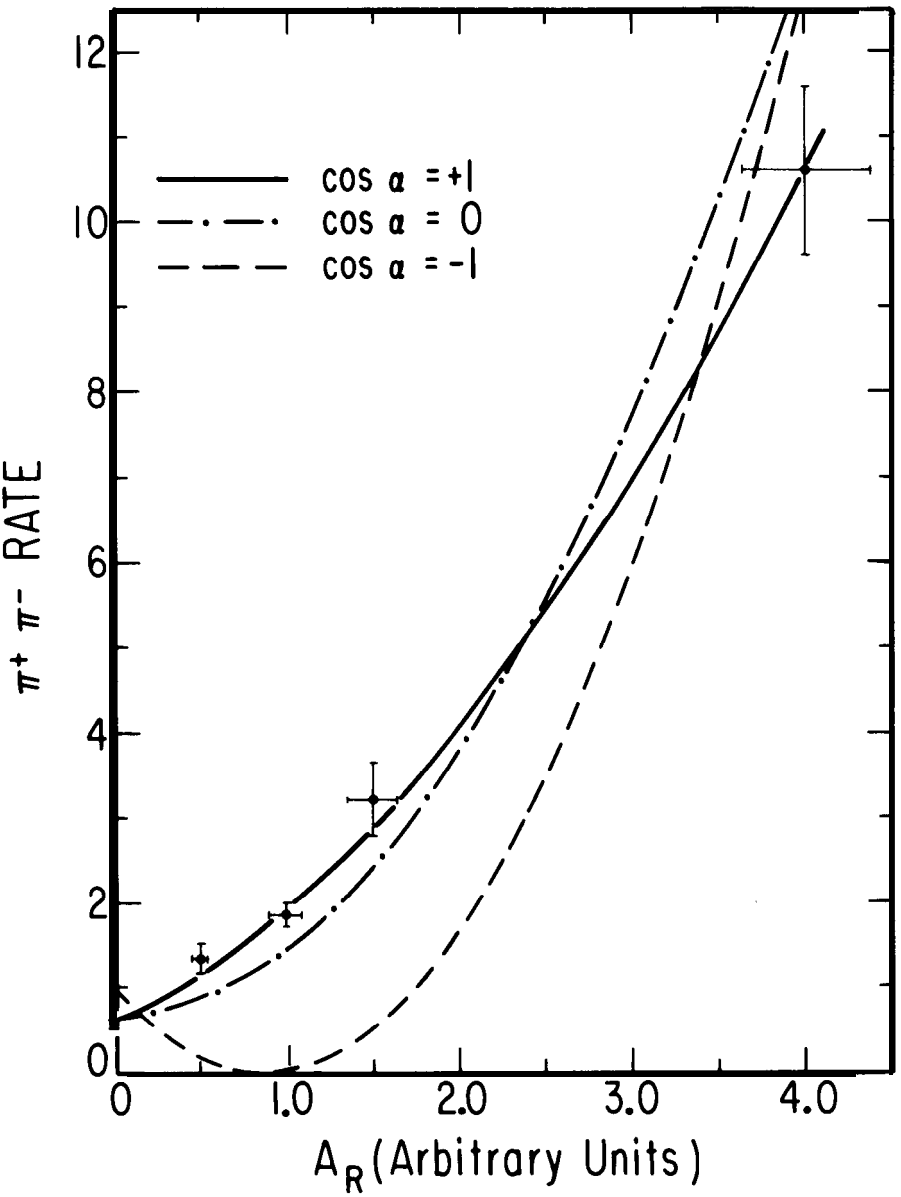


Fig. I. Yield of $\pi^+\pi^-$ events as a function of the diffuse regenerator amplitude. The three curves correspond to the three stated values of the phase between the regeneration amplitude A_r and the CP violating amplitude η_{+-}.

between a world composed of matter and a world composed of antimatter.' Imagine that this experiment were performed in the antiworld. The only difference would be that the regenerator material would be antimatter. If we assume C invariance for the strong interactions, the forward scattering amplitudes for K and K would be interchanged so that A, would have the opposite sign. Thus, in the antiworld an investigator performing the interference experiment would observe destructive interference similar to the dashed curve of Fig 1, an unmistakable difference from the result found in our world. The

interference experiment of Fitch and collaborators eliminated alternate explanations of the $K_L \rightarrow \pi^+\pi^-$ decay, since the effect was of such a nature that an experiment distinguishing a world of matter and antimatter was possible.

It was also suggested that the effect might be due to a long range vector field of cosmological origin.' Such a source of the effect would lead to a decay rate for $K_L \rightarrow \pi^+\pi^-$ which would be proportional to the square of the K_L energy in the laboratory. Our original experiment was carried out at a mean K_L energy of 1.1 GeV. The confirming experiments at the Rutherford Laboratory and CERN were carried out at mean K_L. energies of 3.1 and 10.7 GeV, respectively. Since the three experiments found the same branching ratio for $K_L \rightarrow \pi^+\pi^-$, the possibility of a long range vector field was eliminated.

Before continuing, it is necessary to state some of the phenomenology which describes the CP violation in the neutral K system. The basic notation was introduced by Wu and Yang." For this discussion CPT conservation is assumed. Later we shall refer to the evidence from K-meson decays which show that all data are consistent with a corresponding T violation. Any CPT violation is consistent with zero within the present sensitivity of the measurements.

There are two basic complex parameters which are required to discuss CP violation as observed in the two pion decays of K_L mesons. The first quantity ε is a measure of the CP impurity in the eigenstates $|K_S>$ and $|K_L>$. These eigenstates are given by

$$|K_S> = \frac{1}{\sqrt{2}\sqrt{1+|\varepsilon|^2}}[(1+\varepsilon)|K>+(1-\varepsilon)|\bar{K}>],$$

and

$$|K_L> = \frac{1}{\sqrt{2}\sqrt{1+|\varepsilon|^2}}[(1+\varepsilon)|K>-(1-\varepsilon)|\bar{K}>].$$

The quantity ε can be expressed in terms of the elements of the mass and decay matrices which couple and control the time evolution of the $|K>$ and $|\bar{K}>$ states. It is given by

$$\varepsilon = \frac{-\mathrm{Im}M_{12}+i\mathrm{Im}\Gamma_{12}/2}{i(M_S-M_L)+(\Gamma_S-\Gamma_L)/2}$$

Limits on the size of $\mathrm{Im}\Gamma_{12}$ can be obtained from the observed decay rates of K_s and K_L, to the various decay modes. If $\mathrm{Im}\Gamma_{12}$ were zero, then the phase of ε would be determined by the denominator which is just the difference in eigenvalues of the matrix which couples K and K. These quantities have been experimentally measured and give arg $\varepsilon \sim 45$".

The second quantity ε' is defined by

$$\varepsilon' = \frac{i}{\sqrt{2}}\,\mathrm{Im}\left(\frac{A_2}{A_0}\right) e^{i(\delta_2-\delta_0)}.$$

Here A_0 and A_2 are respectively the amplitudes for a K meson to decay to standing wave states of two pions in the isotopic spin 0 and 2 states, respectively. Time reversal symmetry demands that A_0 and A_2 be relatively real.[10] The quantities δ_0 and δ_2 are the s-wave $\pi\pi$ scattering phase shifts for the states

I = 0 and I = 2, respectively. The parameters ε and ε' are related to observable quantities defined by

$$|\eta_{+-}|\, e^{i\phi_{+-}} = \frac{\text{amp}\ (K_L \to \pi^+\pi^-)}{\text{amp}\ (K_S \to \pi^+\pi^-)},$$

$$|\eta_{00}|\, e^{i\phi_{00}} = \frac{\text{amp}\ (K_L \to \pi^0\pi^0)}{\text{amp}\ (K_S \to \pi^0\pi^0)},$$

and

$$\delta_\ell = \frac{\Gamma(K_L \to \pi^-\ell^+\nu_\ell) - \Gamma(K_L \to \pi^+\ell^-\bar{\nu}_\ell)}{\Gamma(K_L \to \pi^-\ell^+\nu_\ell) + \Gamma(K_L \to \pi^+\ell^-\bar{\nu}_\ell)}$$

These experimentally measured quantities are related to ε and ε' by the following expressions:[11]

$$\eta_{+-} = \varepsilon + \varepsilon'$$
$$\eta_{00} = \varepsilon - 2\varepsilon'$$
$$\delta_\ell = 2\,\mathrm{Re}\varepsilon.$$

The magnitude and phase of the quantity η_{+-} have been most precisely measured by studying the time dependence of $\pi^+\pi^-$ decays from a K beam which was prepared as a mixture of K_S and K_L. This experimental technique was suggested by Whatley,[12] long before the discovery of CP violation. If we let ρ be the amplitude for K_S at t = 0, relative to the K_L amplitude, then the time dependence of $\pi^+\pi^-$ decays will be given by'"

$$N_{+-}(t) = |\rho \exp[(-i\Delta M - \Gamma_S/2)t] + \eta_{+-}|^2$$

The initial amplitude for the K_S component can be prepared by two different methods. In the first method we pass a K_L, beam through a regenerator. Then ρ is the regeneration amplitude. Here the interference term is $2|\rho|\,|\eta_{+-}|\, e^{-\Gamma_S t/2} \cos(-\Delta M t + \phi_\rho - \phi_{+-})$. In the second method we produce a beam which is pure K (or $\bar{K}$) at t = 0. In practice protons of ≈20 GeV produce at small angles about three times as many K as $\bar{K}$. The K dilution is a detail which need not be of concern here. In this case ρ = + 1, and the interference term is $2|\eta_{+-}|e^{-\Gamma_S t/2}\cos(-\Delta M t - \phi_{+-})$.

The important CP parameters are $|\eta_{+-}|$ and ϕ_{+-}. We see, however, that a knowledge of the auxiliary parameters Γ_S and AM is also required. In the first method one measures $\phi_{+-} - \phi_\rho$ and one must also have a technique to independently measure ϕ_ρ. In both cases the $\pi^+\pi^-$ yield is most sensitive to the interference term when the two interfering amplitudes are of the same size. For the second method we require observation at 12 K_S lifetimes. (We want $e^{-\Gamma_S t/2} \approx |\eta_{+-}| \approx 2\times10^{-3}$.) As a consequence, a small error in AM can lead to a large uncertainty on ϕ_{+-}, and, more importantly, a systematic error in AM can lead to an incorrect value for ϕ_{+-}. A one percent error in AM corresponds to an error in ϕ_{+-} of about 3". The measurement of AM with satisfactory precision has required an effort as formidable as the interference experiments themselves."

Time and space do not permit a survey which does justice to the many

groups at CERN, Brookhaven, Argonne, and SLAC who have made the meticulous measurements which have led to the following parameters:[15]

$$\eta_{+-} = [(2.27 \pm 0.02) \times 10^{-3}] \exp [i(44.7° \pm 1.2°)],$$
$$\Delta M = M_S - M_L = -(0.535 \pm 0.002) \times 10^{10}/\text{sec},$$
$$\Gamma_S = (1.121 \pm 0.003) \times 10^{10}/\text{sec}.$$

As an example of the quality of the measurements mentioned above, Fig 2 shows a time distribution of $\pi^+\pi^-$ decays following the passage of a K_L beam of 4 to 10 GeV/c momentum through an 81 cm thick carbon regenerator.[16]

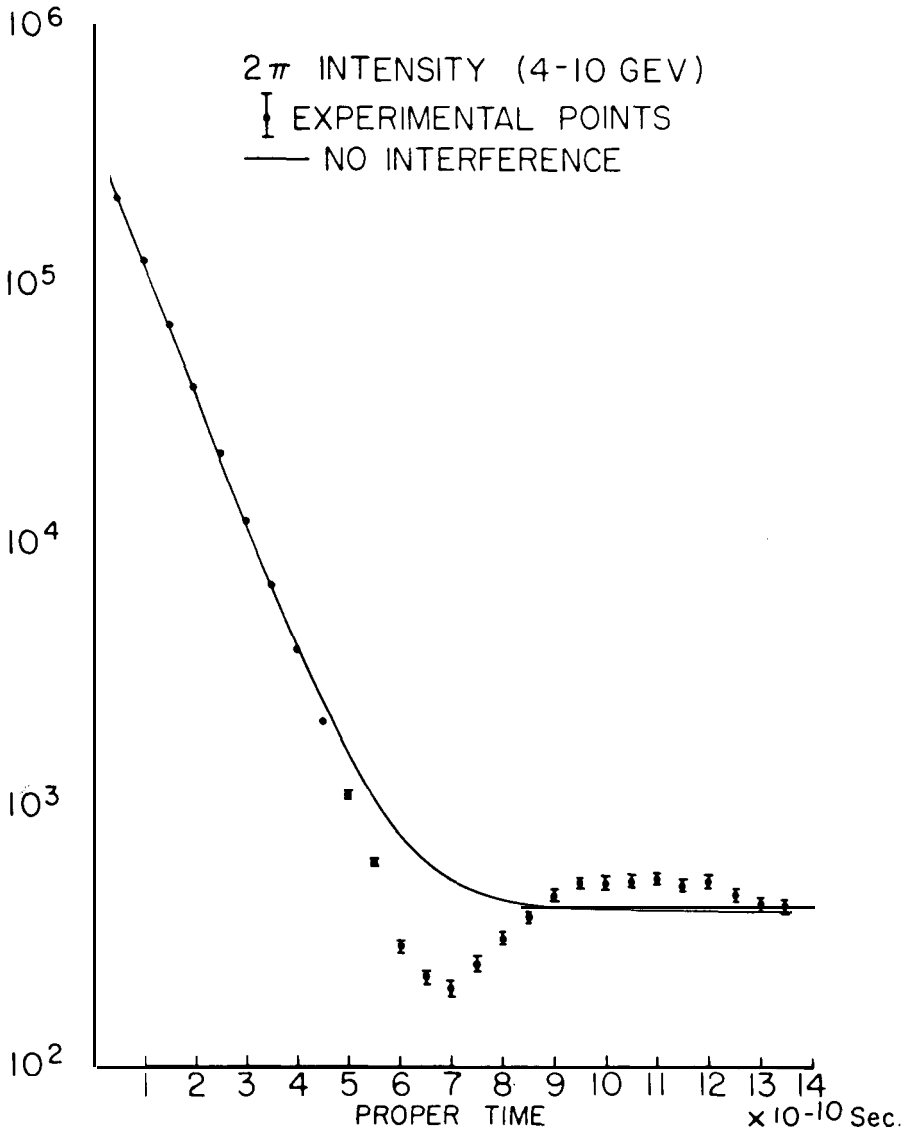


Fig. 2. Yield of $\pi^+\pi^-$ events as a function of proper time downstream from an 81 cm carbon regenerator placed in a K_L beam.

The destructive interference is clearly seen. If the experiment were carried out with a regenerator of anticarbon, then constructive interference would have been observed.

Measurements of the charge asymmetry δ_ℓ for K_L decays began in 1966. This asymmetry is found in the abundant semileptonic decay modes $K_L \to \pi^\pm \ell^\mp \nu$, where ℓ is either an electron or muon. It basically measures the difference in amplitude of K and $\bar{K}$ in the eigenstate of the K_L. It does so by virtue of the AS = AQ rule, which states that all semileptonic decays have the change in charge of the hadron equal the change in strangeness. Thus, K mesons decay to $\pi^- \ell^+ \nu$ and $\bar{K}$ mesons decay to $\pi^+ \ell^- \bar{\nu}$. The validity of the AS = AQ rule was in doubt for many years, but it has finally been established that the AQ = -AS transitions are no more than about 2% of the AQ = +AS transitions.[17] The size of the charge asymmetry expected is $\sim\sqrt{2}\,|\eta_{+-}| \approx 3\text{X}\ 10^{-3}$. Millions of events are required to measure δ_ℓ accurately, and excellent control of the symmetry of the apparatus and understanding of charge dependent biases are needed to reduce systematic errors.

Again, we must omit a detailed review of all asymmetry measurements. These have been carried out at CERN, Brookhaven, and SLAC. The net result of these measurements gives'"

and
$$\delta_e = (3.33 \pm 0.14) \times 10^{-3}$$
$$\delta_\mu = (3.19 \pm 0.24) \times 10^{-3}.$$

We expect these two asymmetries to be equal since they both are a measure of $2\,\mathrm{Re}\varepsilon$. These asymmetries are measured for a pure K_L beam. For a beam which is pure K at t = 0 the charge asymmetry shows a strong oscillation term with angular frequency AM. Figure 3 shows the time dependence of the

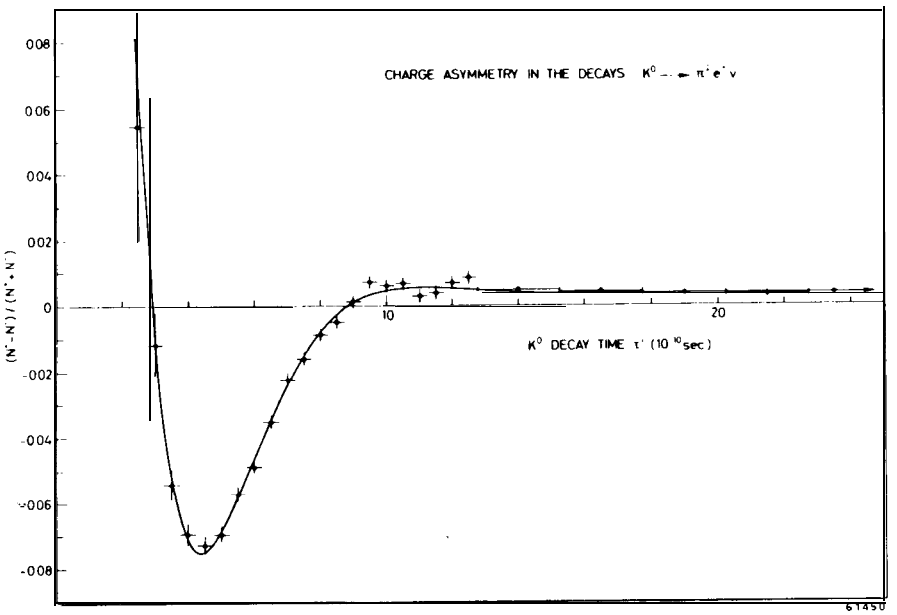


Fig. 3. Time dependence of the charge asymmetry of semileptonic decays.

charge asymmetry taken from the thesis of V. Lüth.[18] The small residual charge asymmetry of the K_L decays after the oscillations have died out is clearly resolved.

The charge asymmetry is a manifest violation of CP, and as such also permits an experimental distinction between a world and an antiworld. In our world we find that the positrons in the decay are slightly in excess. The positrons are leptons which have the same charge as our atomic nuclei. In the antiworld the experimenter will find that the excess leptons have opposite charge to his atomic nuclei; hence, he would report a different result for the same experiment.

Simple examination of the relations between the experimentally measurable parameters and the complex quantities ε and ε' show that measurements of $|\eta_{oo}|$ and ϕ_{oo} are essential to finding ε and ε'. The path to reliable results for $|\eta_{oo}|$ and ϕ_{oo} has been torturous. This statement is based on personal experience; six years of my professional life have been spent on the measurement of $|\eta_{oo}|$.

Measurement of the parameters associated with $K_L \to \pi^\circ\pi^\circ$ is complicated by the fact that each π° decays rapidly (10^{-16} sec) into two photons. For typical K_L beams used in these experiments the photon energies are in the range of 0.25 to 5 GeV. It is difficult to measure accurately the direction and energy of such photons. In addition to that difficulty, the CP conserving decay $K_L \to 3\pi^\circ$ occurs at a rate which is about 200 times as frequent, and presents a severe background.

Early results suggested that $|\eta_{oo}|$ was about twice $|\eta_{+-}|$ with the consequence that ε' was a large number. By 1968 however, an improved experiment using

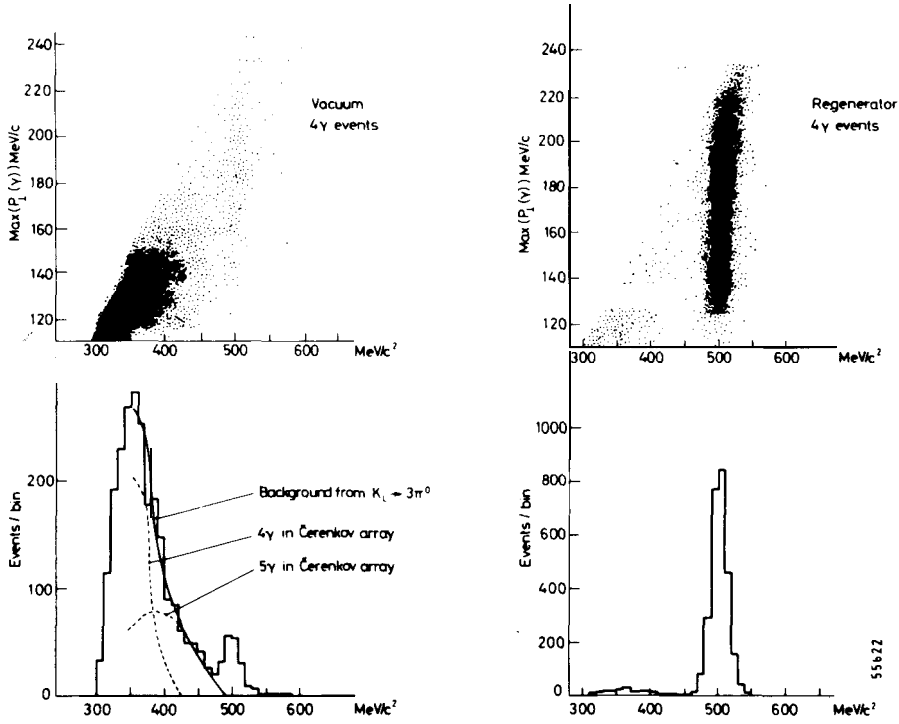


Fig. 4. Distributions of reconstructed $K_L \to \pi^\circ\pi^\circ$ events, and regenerated $K_S \to \pi^\circ\pi^\circ$ events

spark chambers'" and a painstaking heavy-liquid bubble chamber experiment from CERN[20] showed that $|\eta_{00}|$ was rather close in value to $|\eta_{+-}|$. Figure 4 shows the results from the most accurate measurement of $|\eta_{00}|/|\eta_{+-}|$.[21] Shown are reconstructed events from free K_L, decays as well as a sample of $K_S \rightarrow \pi^\circ\pi^\circ$ from a regenerator used to determine the resolution of the apparatus. The serious background from the $3\pi^\circ$ decays is clearly seen. The result $|\eta_{00}|/|\eta_{+-}| = 1.00\pm0.06$ is based on only 167 events. The equality of $|\eta_{00}|$ and $|\eta_{+-}|$ means that the ratio of charged 2π decays to neutral 2π decays is the same for CP violating K_L, decays as for CP conserving K_S decays. This result implies that ε' is very small providing ϕ_{00} is close to ϕ_{+-}.

The $K_L \rightarrow \pi^\circ\pi^\circ$ events cannot be collected at the rate of the $\pi^+\pi^-$ decays, nor can they be separated so cleanly from backgrounds. As a consequence, the precision with which we know the parameters $|\eta_{00}|$ and ϕ_{00} is much less than the charged parameters. A weighted average of all the data presently available gives"

$$|\eta_{00}|/|\eta_{+-}| = 1.02\pm0.04,$$

and

$$\phi_{00}-\phi_{+-} = 10^\circ\pm6^\circ.$$

The results are quoted with reference to the charged decay mode parameters because the most accurate experiments have measured the quantity $|\eta_{00}|/|\eta_{+-}|$ directly. The result for ϕ_{00} is principally due to a recent experiment by J. Christenson et al."

The phase of the quantity ε' is given by the angle $\pi+\delta_2-\delta_0$. Information concerning the pion-pion scattering phase shifts comes from several sources." A compendium of these sources gives $\delta_2-\delta_0 = -45^\circ\pm10^\circ$. The phase of ε is naturally related to $\phi_n \equiv \arg\left([i(M_S-M_L)+(\Gamma_S-\Gamma_L)/2]^{-1}\right) = 43.7^\circ\pm0.2^\circ$. This is the phase ε would have if there were no contributions from $\mathrm{Im}\Gamma_{12}$. The measured phase of $\eta_{+-}(44.7^\circ\pm1.2^\circ)$ is within measurement precision equal to ϕ_n.

The measured parameters are plotted on the complex plane in Fig 5a. The size of the box for η_{+-} and η_{00} and the width of the bar for δ_ℓ correspond to one standard deviation. The derived quantities ε and ε' are plotted in Fig 5b. Boxes corresponding to both one and two standard deviations are shown. Also plotted is the constraint coming from the π—π scattering phase shifts which defines the phase of ε' to be $45^\circ\pm10^\circ$. With this constraint we find that ε, ε', η_{00} and η_{+-} lie nearly on a common line. There is a mild disagreement between the π—π phase shift constraint and the result of Christenson et al. for ϕ_{00}.

A more general analysis of the neutral K system which includes the possibility of violation of CPT with T conservation as well as CP violation with CPT conservation has been given by Bell and Steinberger.[24] The analysis does depend on the assumption of unitarity which requires that the M and Γ matrices remain Hermitian. The Bell-Steinberger analysis has been applied to the data with the conclusion that while a small CPT violation is possible, the predominant effect is one of CP violation. All experiments are consistent with exact CPT

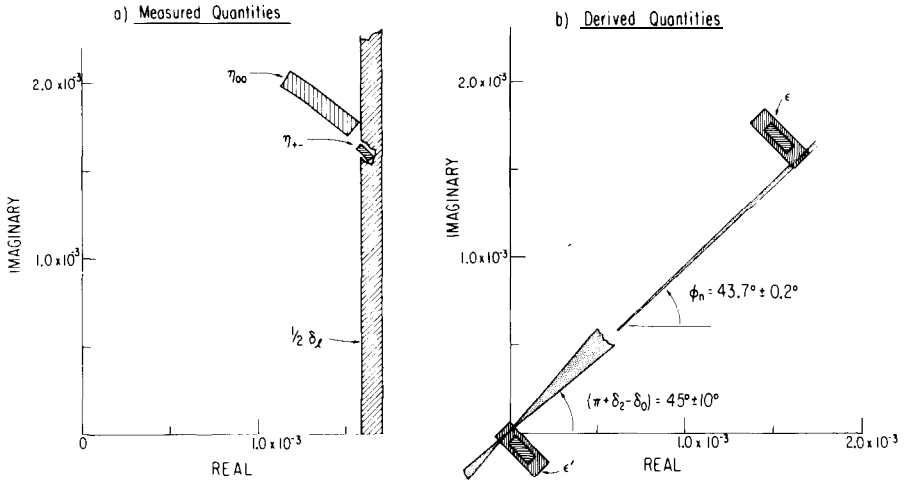


Fig. 5. Summary of CP violating parameters in the neutral K system
(a) Measured quantities.
(b) Derived quantities.

conservation,[25] and, hence, imply a violation of time reversal symmetry. The conservation or non-conservation of CPT remains, however, a question that must continue to be addressed by experiment. A briefdiscussion of the unitarity analysis is given in an appendix.

The essential point of this analysis rests on the measurement of the phase of η_{+-}. Limits on the contribution of $\mathrm{Im}\Gamma_{12}$ can be estimated from measured decay rates to all modes of decay of the neutral K mesons. The absence, within present experimental limits, of CP violation in the decay modes other than the 2π modes limits the contribution of $\mathrm{Im}\Gamma_{12}$ to ε to be $\leqslant 0.3\times10^{-3}$, a value small compared to $|\eta_{+-}|$. Thus the phase of ε and hence η_{+-} is expected to be close to ϕ_{n}. We can examine the other extreme, namely, that CP and CPT symmetry are both violated while time reversal symmetry remains valid. Under these conditions we would find the natural phase ϕ_{n} to be $\sim 135°$, and would expect ϕ_{+-} to be close to 135". The fact that this is not the case is the essence of the argument that CPT is not violated.

We note that the natural phase depends on the sign of the mass difference. We have assumed $\mathrm{AM} = (\mathrm{M_S}-\mathrm{M_L}) < 0$. If the sign of the mass difference were opposite, we would expect the phase of ε to be equal to 135" or -45" for CP violation with CPT symmetry. The phase of ε' would remain the same, however, since it does not depend on AM in any way. Thus, the conclusion that the phase of ε and ε' are approximately the same is a consequence of the fact that the long lived K is heavier than the short lived K. The sign of the mass difference has been measured by several groups with complete agreement.[26]

Independent of any particular theory, we would expect results which are similar to those observed. The constraint of unitarity and $\pi\pi$ scattering phase shifts force $\phi_{00} \approx \phi_{+-}$ for $\varepsilon' \ll \varepsilon$. Under these circumstances, a measurement of the ratio $(|\eta_{00}|/|\eta_{+-}|)^2$ is a direct measurement of the quantity ε' by means of the relation $\varepsilon'/\varepsilon \approx [1-(|\eta_{00}|/|\eta_{+-}|)^2]/6$. Applying this relation to the present

data we have ε'/ε = -0.007±0.013. New experiments at the Fermilab and at Brookhaven will attempt to increase the sensitivity of the measurement by a factor 10.

As we have shown, detailed analysis of the CP violation in the neutral K meson system leads to the conclusion that time reversal is also violated. Table I gives a representative set of experiments which have searched for T violation, CP violation, and C violation (in non-weak interactions). None of these experiments has led to a positive result. Many of the experiments are approaching a sensitivity for the violation of 10^{-3}, but few have attained this value. A strength of 10^{-3} in amplitude or relative phase is what we might expect for the CP violation based on the results of K-decay. For experiments involving decays with electromagnetic interactions in the final states, an apparent T-violation effect is usually expected at the 10^{-3} level. An example of this is the result for the ^{191}Ir decay in which a significant effect is found, but it is of the size expected on the basis of the final state electromagnetic interaction.

Table I. Searches for CP, 'I', and C Violation

Measurement	Result	Test	Ref
$\frac{\Gamma(K^+\to\pi^+\pi^+\pi^-)-\Gamma(K^-\to\pi^-\pi^-\pi^+)}{\text{average}}$	$(0.8\pm1.2)\times10^{-3}$	CP	37
$\frac{\Gamma(K^+\to\pi^+\pi^\circ\pi^\circ)-\Gamma(K^-\to\pi^-\pi^\circ\pi^\circ)}{\text{average}}$	$(0.8\pm5.8)\times10^{-3}$	CP	38
$\frac{a_{\tau^+}-a_{\tau^-}}{\text{average}}$, where a_τ is the slope of the odd pion in the $K^\pm\to\pi^\pm\pi^\pm\pi^\mp$ Dalitz plot	$(-7.0\pm5.3)\times10^{-3}$	CP	37
Muon polarization transverse to decay plane in $K_l\to\pi^-\mu^+\nu_\mu$	$(2.1\pm4.8)\times10^{-3}$	T	39
Coefficient of T odd correlation $\langle\vec{J}\cdot\vec{P}_e\times\vec{P}_\nu\rangle$ in the β-decay of polarized ^{19}Ne	$(-0.5\pm1.0)\times10^{-3}$	T	40
Coefficient of T odd correlation $\langle\vec{\sigma}_n\cdot\vec{P}_e\times\vec{P}_\nu\rangle$ in the β-decay of the neutron	$(-1.1\pm1.7)\times10^{-3}$	T	41
Asymmetry in distribution of $(T_{\pi^+}-T_{\pi^-})$ in the decay of $\eta\to\pi^+\pi^-\pi^\circ$	$(1.2\pm1.7)\times10^{-3}$	C	42
Electric dipole moment of the neutron	$(0.4\pm1.5)\times10^{-24}$ e-cm $(0.4\pm0.75)\times10^{-24}$ e-cm	T	43 44
Angular correlation in γ decay of polarized iridium, $^{191}Ir^*\to{}^{191}Ir+\gamma$. Measure phase angle between E_2 and M_1 decay amplitudes. Result expected on basis of electromagnetic interaction in final state	$(4.7\pm0.3)\times10^{-3}$ 4.3×10^{-3}	T	45 46
Detailed balance in nuclear rections, e.g., $^{24}Mg+\alpha\rightleftarrows{}^{27}A\ell+p$ Measure: $\frac{\text{amplitude T violating}}{\text{amplitude T conserving}}$	$\leq 3\times10^{-3}$	T	47

Among the many measurements listed in Table I, we would like to single out the electric dipole moment of the neutron. The first measurement of this quantity was made in 1950 by Purcell, Ramsey and Smith" with the avowed purpose of testing the assumptions on which one presumed the electric dipole moment would be zero. Today, outside of the K-system, the search for an electric dipole moment of the neutron is the most promising approach to the detection of T violation. At present the upper limit is ~ 10^{-24} e-cm. New experiments using ultra-cold neutrons give promise of an increase in intensity by 100-fold within the next several years. The significance of a negative result for the electric dipole moment, or for any of the measurements in Table I, is difficult to assess without a theory of CP violation."

Up to now our discussion has been entirely experimental. In the analysis of the CP violation in the neutral K system general principles of quantum mechanics have been used. The manifest charge asymmetry of the K_L semi-leptonic decays requires no assumptions at all for its interpretation. The literature abounds with theoretical speculations about CP violation. One of these speculations by Wolfenstein[29] is frequently referred to. He hypothesizes a direct $\Delta S = 2$ superweak interaction which is constructed to produce a CP violation. This direct interaction interferes with the second order weak interaction to produce the CP-violating AS = 2 coupling between K and K. Since the hypothesized superweak transition is first order, it need have only ~ 10^{-7} of the strength of the normal weak interaction. As such the only observable consequence is a CP violation in $K \rightarrow 2\pi$ decay characterized by a single number, the value of ImM_{12} in the mass matrix.

At present the data are in agreement with this hypothesis, which leads to predictions that $|\eta_{00}| = |\eta_{+-}|$, and $\phi_{00} = \phi_{+-} = \phi_n$. However, the relation $\phi_{00} = \phi_{+-} = \phi_n$ to a good approximation follows from the constraints of unitarity and the π—π scattering phase shifts with no further assumptions. On the other hand, the relation $|\eta_{00}| = |\eta_{+-}|$ has not been tested to very high accuracy, especially considering the difficulty of experiments which attempt to measure the properties of $K_L \rightarrow \pi^\circ\pi^\circ$. These experiments are more prone to systematic errors, and in truth $|\eta_{00}|$ and $|\eta_{+-}|$ could differ considerably more than appears to be allowed by the experiments. Thus, while the superweak hypothesis is in agreement with the present data, the data by no means make a compelling case for the superweak hypothesis.

In 1973, Kobayashi and Maskawa[30] in a remarkable paper pointed out that with the (then) current understanding of weak interactions, CP violation could be accommodated only if there were three or more pairs of strongly interacting quarks. The paper was remarkable because at that time only three quarks were known to exist experimentally. Since then, strong evidence has been accumulated to support the existence of a charmed quark and a fifth bottom quark. It is presumed that the sixth quark, top, will be eventually found. With six quarks the weak hadronic current involving quarks can be characterized by three Cabibbo-angles, and a phase δ. This phase, if non-zero, would imply a CP violation in the weak interaction.

In principle, the magnitude of this phase δ which appears in the weak

currents of quarks can be related to the CP violation observed in the laboratory. Unfortunately, all the experimental investigations are carried out with hadrons, which are presumed to be structures of bound quarks, while the parameter one wants to establish, δ, is expressed in terms of interactions between free quarks. The theoretical "engineering" required to relate the free quark properties to bound quark properties is difficult and, as a consequence, is not well developed. A balanced and sober view of this problem is given in a paper by Guberina and Peccei.[31] Even if the CP violation has its origin in the weak currents, it is not clear whether the experimental consequences with respect to K decay can be distinguished from the superweak hypothesis. If we are successful in establishing the fact that CP violation is the result of a phase in the weak currents between quarks, we will still have to understand why it has the particular value we find.

There are, however, on the horizon new systems which have some promise to give additional information about CP violation. These are the new neutral mesons, $D^\circ, B^\circ, B_s^\circ$, (composed of $c\bar{u}$, bd, and $b\bar{s}$ quarks), and their antiparticles $\bar{D}^\circ, \bar{B}^\circ, \bar{B}_s^\circ$. These mesons have the same general properties as K mesons. They are neutral particles that, with respect to strong interactions, are distinct from their own antiparticles, and yet are coupled to them by common weak decay modes. While we may not expect any stronger CP impurities on the eigenstates (the parameter analogous to ε), we might expect stronger effects in the decay amplitudes (the parameter analogous to E'). We might expect this since the CP violation comes about through the weak interactions of the heavy quarks, c, b, t, which participate only virtually in K decay, but can be more influential in heavy neutral meson decay. At present, D mesons can be made rather copiously at the e^+e^- storage ring SPEAR at SLAG,[32] and B mesons are beginning to be produced at the e^+e^- storage ring CESR at Cornell.[33]

It is conceivable that the effect of CP violation may become stronger with energy. Soon collisions of protons with antiprotons will be observed at CERN with a total center of mass energy greater than 500 GeV. It will be most interesting to look for C violations in the spectra of particles produced in those collisions. Also, improvements in technology of detectors over the next several decades may permit sensitive searches for time reversal violating observables in high energy neutrino interactions.

Recently, much attention has been given to the role that CP violation may play in the early stages of the evolution of the universe.[34] A mechanism has been proposed with CP violation as one ingredient which leads from matter--antimatter symmetry in the early universe to the small excess of matter observed in the universe at the present time. The first published account of this mechanism, of which I am aware, was made by Sakharov[35] in 1967. He explicitly stated the three ingredients which form the foundation of the mechanism as it is presently discussed. These ingredients are: (1) baryon instability, (2) CP violation, and (3) appropriate lack of thermal equilibrium. The recent intense interest in this problem has risen because baryon instability is a natural consequence of the present ideas of unification of the strong interactions with the successfully unified electromagnetic and weak interactions. This latter unifica-

tion was discussed in the 1979 Nobel lectures of Glashow, Salam, and Weinberg.“’

A very oversimplified explanation of the process which leads to a net baryon number can be given with the aid of Fig 6a. Quarks and leptons are linked by a very heavy boson X and its antiparticle $\bar{X}$. While the total decay rates

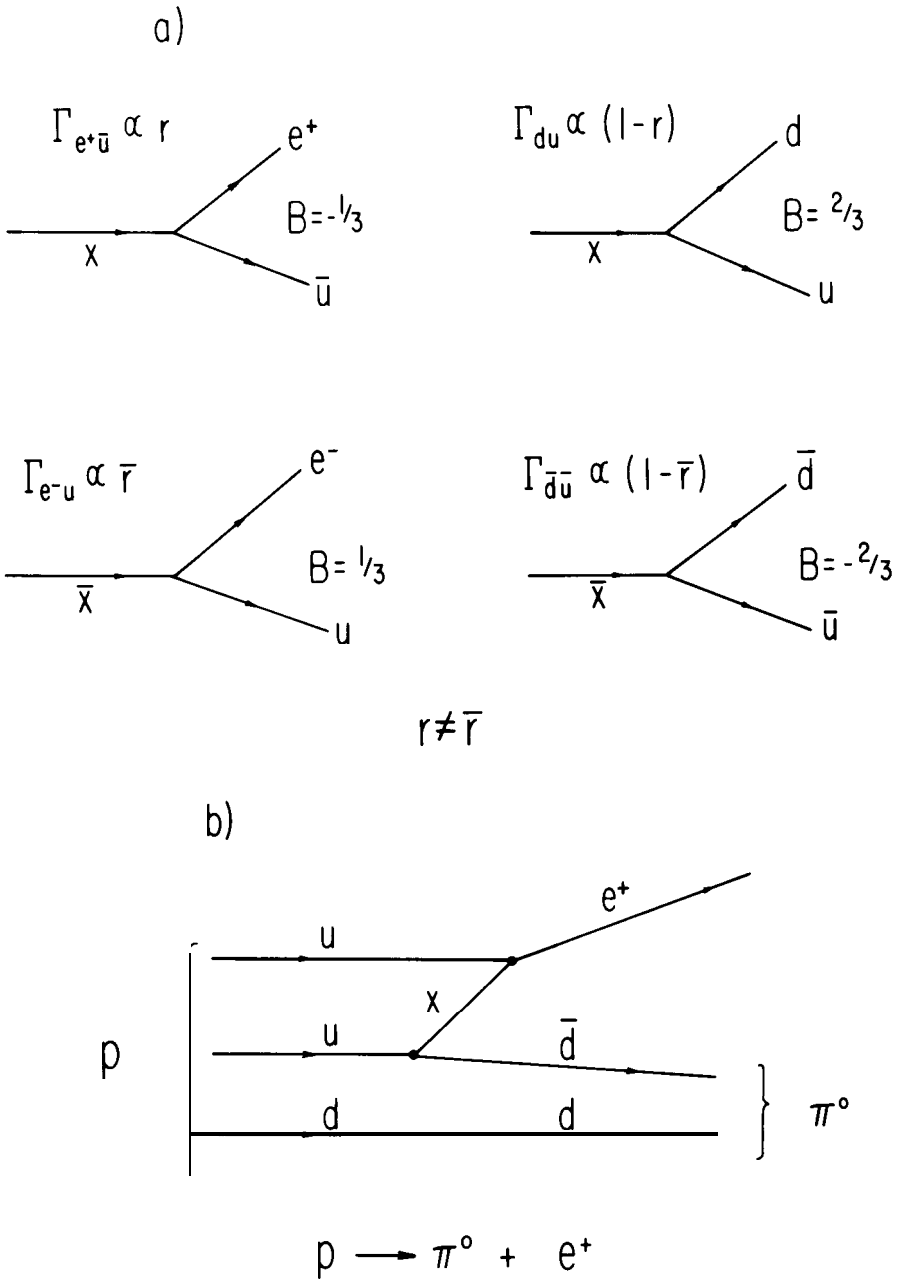


Fig. 6. (a) Simplified diagrams of baryon number non-conserving X boson decays. (h) A proton decay mediated by an X boson.

of X and $\bar{X}$ may be equal, with CP violation the fractional partial rates r and r to B = -3^{1} and B = + $\frac{1}{3}$ decay channels of X and $\bar{X}$, respectively, can differ. At an early stage where the temperature is large compared to the mass of X, the density of X and X may be equal. On decay, however, the net evolution of baryon number is proportional to (r--Y). The excess can be quite small since the ratio of baryons to photons today is ~ 10^{-9}. Figure 6b shows how such an X boson can mediate the decay p + e++n'. If nucleon decay is discovered it will give a strong support to these present speculations.

Whether the CP violation that we observe today is a "fossil remain" of these conjectured events in the early universe is a question that cannot be answered at present. That is to say, does the CP violation we observe today provide supporting evidence for these speculations? We simply do not know enough about CP violation. Our experimental knowledge is limited to its observation in only one extraordinarily sensitive system that nature has provided us. We need to know the theoretical basis for CP violation and we need to know how to reliably extrapolate the behavior of CP violation to the very high energies involved.

At present our experimental understanding of CP violation can be summarized by the statement of a single number. If we state that the mass matrix which couples K and K has an imaginary off-diagonal term given by

$$\mathrm{Im}M_{12} = -1.16\times10^{-8}\,\mathrm{eV},$$

then all the experimental results related to CP violation can be accounted for. If this is all the information nature is willing to provide about CP violation it is going to be difficult to understand its origin. I have emphasized, however, that despite the enormous experimental effort, punctuated by some experiments of exceptional beauty, we have not reached a level of sensitivity for which a single parameter description should either surprise or discourage us.

We must continually remind ourselves that the CP violation, however small, is a very real effect. It has been used almost routinely as a calibration signal in several high energy physics experiments. But more importantly, the effect is telling us that there is a fundamental asymmetry between matter and antimatter, and it is also telling us that at some tiny level interactions will show an asymmetry under the reversal of time. We must continue to seek the origin of the CP symmetry violation by all means at our disposal. We know that improvements in detector technology and quality ofaccelerators will permit even more sensitive experiments in the coming decades. We are hopeful then, that at some epoch, perhaps distant, this cryptic message from nature will be deciphered.

APPENDIX

The evolution of a neutral K system characterized by time dependent amplitudes a and Z for the IK> and IK> components, respectively, is given by

$$-\frac{d}{dt}\begin{pmatrix} a \\ \bar{a} \end{pmatrix} = \left(iM+\frac{1}{2}\Gamma\right)\begin{pmatrix} a \\ \bar{a} \end{pmatrix},$$

where M and Γ are each Hermitian matrices, and t is the time measured in the rest system of the K meson. Expressed in terms of their elements the matrices are

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{12}^* & M_{22} \end{pmatrix} \text{ and } \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{12}^* & \Gamma_{22} \end{pmatrix}.$$

The matrix $iM+\frac{1}{2}\Gamma$ has eigenvalues $\gamma_S = iM_S+\frac{1}{2}\Gamma_S$ and $\gamma_L = iM_L+\frac{1}{2}\Gamma_L$. We define small parameters $\varepsilon = (-\mathrm{Im}M_{12}+i\mathrm{Im}\Gamma_{12}/2)/(\gamma_S-\gamma_L)$ and $A = [i(M_{11}-M_{22})+(\Gamma_{11}-\Gamma_{22})/2]/[2(\gamma_S-\gamma_L)]$. We can then express the eigenvectors as

$$|K_S> = \frac{1}{\sqrt{2}}\frac{1}{\sqrt{1+|\varepsilon+\Delta|^2}}\left[(1+\varepsilon+\Delta)\,|K>+(1-\varepsilon-\Delta)\,|\bar{K}>\right]$$

and

$$|K_L> = \frac{1}{\sqrt{2}}\frac{1}{\sqrt{1+|\varepsilon-\Delta|^2}}\left[(1+\varepsilon-\Delta)\,|K>-(1-\varepsilon+\Delta)\,|\bar{K}>\right].$$

The parameter ε represents a CP violation with T non-conservation. The parameter A represents a CP violation with CPT non-conservation.

If we form a state $|K(t)>$ which is an arbitrary superposition of $|K_S>$ and $|K_L>$ with amplitudes a_S and a_L at t = 0, we can compute its norm $<K(t)|K(t)>$ as a function of time. At t = 0 by conservation of probability we have the relation.

$$-\frac{d}{dt}<K(t)|K(t)>\Big|_{t=0} = \sum_f |a_S\,\mathrm{amp}(K_S\to f)+a_L\,\mathrm{amp}(K_L\to f)|^2,$$

where f represents the set of final states. Explicit evaluation of the expression gives

$$[-i(M_S-M_L)+(\Gamma_S+\Gamma_L)/2]<K_S|K_L> = \sum_f (\mathrm{amp}(K_S\to f))^*\,(\mathrm{amp}(K_L\to f))$$

A number of definitions and a particular phase convention are used. We define $\tilde{\Delta} = \Delta-(A_0-\bar{A}_0)/(A_0+\bar{A}_0)$ where A_0 and $\bar{A}_0$ are the standing wave amplitudes for K and $\bar{K}$, respectively, to decay to the I = 0 state of two pions. A_0 and $\bar{A}_0$ are chosen real and define the phase convention used in the analysis. From the experimental parameters we define $\varepsilon_0 = \frac{2}{3}\eta_{+-}+\frac{1}{3}\eta_{00}$ and $\varepsilon_2 = \frac{\sqrt{2}}{3}(\eta_{+-}-\eta_{00})$, and $\alpha(f) = (1/\Gamma_S)\,(\mathrm{amp}(K_S\to f))^*\,(\mathrm{amp}(K_L\to f))$. With these definitions we find to a good approximation that

$$\left[-i\Delta M/\Gamma_S+\frac{1}{2}\right]\left[2\mathrm{Re}\varepsilon-2i\mathrm{Im}\tilde{\Delta}\right] = \varepsilon_0+\sum_f \alpha(f) \qquad (1)$$

and

$$\varepsilon-\tilde{\Delta} = \varepsilon_0 \qquad (2)$$

The sum over f, which now excludes the I = 0 $\pi\pi$ state, consists of the following terms:

$$\alpha(\pi\pi, I = 2) = A_2/A_0 e^{i(\delta_2-\delta_0)}\varepsilon_2^*,$$

$$\alpha(\pi^+\pi^-\pi^\circ) = (\Gamma(K_L \to \pi^-\pi^+\pi^\circ)/\Gamma_S)\eta^*_{+-\circ},$$

$$\alpha(\pi^\circ\pi^\circ\pi^\circ) = (\Gamma(K_L \to \pi^\circ\pi^\circ\pi^\circ)/\Gamma_S)\eta^*_{\circ\circ\circ},$$

$$\alpha(\pi e\nu) = (\Gamma(K_L \to \pi e\nu)/\Gamma_S)2i\mathrm{Im}x_e,$$

and $$\alpha(\pi\mu\nu) = (\Gamma(K_L \to \pi\mu\nu)/\Gamma_S)2i\mathrm{Im}x_\mu,$$

where $\eta_{+-\circ} = \mathrm{amp}\,(K_S \to \pi^+\pi^-\pi^\circ)/\mathrm{amp}\,(K_L \to \pi^+\pi^-\pi^\circ)$, $\eta_{\circ\circ\circ} = \mathrm{amp}\,(K_S \to \pi^\circ\pi^\circ\pi^\circ)/\mathrm{amp}\,(K_L \to \pi^\circ\pi^\circ\pi^\circ)$, and x_ℓ is the ratio, $\mathrm{amp}(\Delta Q = -\Delta S)/\mathrm{amp}(\Delta Q = \Delta S)$, for $K \to \pi\ell\nu_\ell$. The quantities $\eta_{+-\circ}$ and $\eta_{\circ\circ\circ}$ are CP violating ratios. (The final state $\pi^+\pi^-\pi^\circ$ can be CP even or odd. Here we refer only to the odd state.) The measurements of $\eta_{\circ\circ\circ}$ and $\eta_{+-\circ}$ are not at present very accurate and are consistent with zero. If we use the experimental limits which exist,[15] we find

$$\mathrm{Re}\alpha = \mathrm{Re}\sum_f \alpha(f) = (0.14\pm0.19)\times10^{-3}$$

and $$\mathrm{Im}\alpha = \mathrm{Im}\sum_f \alpha(f) = (-0.19\pm0.25)\times10^{-3}.$$

The equations (1) and (2) take a very simple form if we resolve the components of ε and $\tilde{\Delta}$ parallel and perpendicular to the direction which makes an angle ϕ_n with the real axis, where

$$\phi_n = \tan^{-1}\left[-\frac{2(M_S-M_L)}{(\Gamma_S-\Gamma_L)}\right].$$

We then find $$\varepsilon_\parallel = \varepsilon_{o\parallel} + \cos\phi_n\,\mathrm{Re}\alpha,$$

$$\varepsilon_\perp = -\cos\phi_n\,\mathrm{Im}\alpha,$$

$$\tilde{\Delta}_\parallel = \cos\phi_n\,\mathrm{Re}\alpha,$$

and $$\tilde{\Delta}_\perp = -\varepsilon_{o\perp} - \cos\phi_n\,\mathrm{Im}\alpha.$$

The experimental values of $\varepsilon_{o\parallel}$ and $\varepsilon_{o\perp}$ are, respectively, $(2.27\pm0.03)\times10^{-3}$ and $(0.16\pm0.09)\times10^{-3}$. We then find

$$\varepsilon_\parallel = (2.37\pm0.19)\times10^{-3},$$

$$\varepsilon_\perp = (0.14\pm0.18)\times10^{-3},$$

$$\tilde{\Delta}_\parallel = (0.10\pm0.14)\times10^{-3},$$

and $$\tilde{\Delta}_\perp = (-0.02\pm0.20)\times10^{-3}.$$

Within the present experimental limits, we find that all the measurements are consistent with T violation and CPT conservation. In particular, we see the limit on $\varepsilon_\perp$ is very small so that we cannot expect ϕ_{+-} and ϕ_{oo} to differ greatly from ϕ_n. Further, if the values of $\eta_{\circ\circ\circ}, \eta_{+-}, x_e$, and x_μ were $< 10^{-2}$, then

we would find $|\varepsilon_\perp| \leqslant 10^{-5}$. Such an expectation is reasonable if the strength of the CP violation is roughly the same in all modes.

ACKNOWLEDGEMENTS

I would like to thank Professors S. Chandrasekhar, R. Oehme, R. G. Sachs and B. Winstein for their advice and critical comments concerning this lecture. I would also like to thank Professors V. Telegdi, S. Treiman, and L. Wolfenstein for many valuable discussions concerning CP violation over the years.

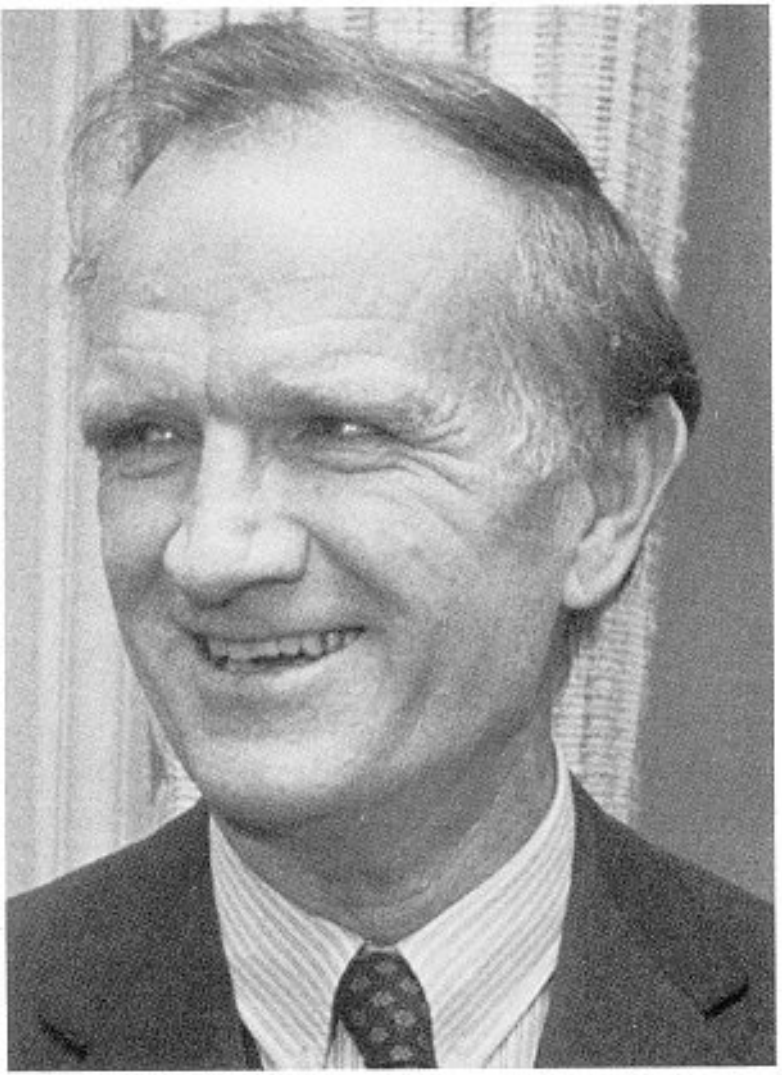
REFERENCES

1. This lecture cannot cover all the important details, refer to all the important work concerning CP violation, or do justice to the work that is referred to in the text. For a more complete discussion, the reader is referred to the most recent review by K. Kleinknecht, ***Annual Reviews*** of Nuclear ***Science,*** Vol. 26, I, 1976. Also, a good perspective of the progress in the field can he found by reading the appropriate sections of the Proceedings of the biannual International Conferences on High Energy Physics, 1964-1974.
2. Galhraith, W. et al., Phys. Rev. Letters 14, 383 (1965).
3. de Bouard, X. et al., Phys. Letters 15, 58 (1965).
4. Fitch, V. L. et al., Phys. Rev. Letters 15, 73 (1965).
5. The value of $A_\perp$ depends on the K_S-K_L mass difference $|\delta|$ which was, at the time, measured to he 0.5±0.1 by J. H. Christenson et al.. Phys. Rev. ***140B, 74*** (1965).
6. Fitch, V. L. et al., Phys. Rev. 164, 1711 (1967).
7. This argument was presented in the literature by Wattenberg, A. and Sakurai, J., Phys. Rev. 161, 1449 (1967).
8. Bell, J. S. and Perring, J. K., Phys. Rev. Letters 13, 348 (1964); Bernstein. J., Cabibbo, N. and Lee, T. D., Phys. Letters 12, 146 (1964).
9. Wu, T. T. and Yang, C. N., Phys. Rev. Letters 13, 380 (1964); the phenomenology was first discussed by Lee, Oehme and Yang, Phys. Rev. 106, 340 (1957).
10. Wigner, E. P., Göttinger Nachrichten 31, 546 (1932); see also paper of Lee, Oehme, and Yang in Ref. 9.
11. Some approximations have been made. The first two expressions should read: $\eta_{+-} = \varepsilon + \varepsilon'/(1+\omega)$ and $\eta_{00} = \varepsilon - 2\varepsilon'/(1-2\omega)$ where $\omega = 1/\sqrt{2}\ \mathrm{Re}(A_2/A_0)e^{i(\delta_2-\delta_0)}$. The magnitude of $\omega \approx 0.05$, so that its effect is not large. The charge asymmetry δ_L should he given by $\delta_L = 2\frac{(1-|x|^2)}{|1-x|^2}\ \mathrm{Re}\varepsilon$, where x is the ratio of the ΔQ = -AS amplitude to the $\Delta Q = +\Delta S$ amplitude in the semi-leptonic decay. Evidence strongly favors $x \approx 0$; see Ref. 17.
12. Whatley, M. C., Phys. Rev. Letters 9, 317 (1962).
13. Here $\Delta M = M_S - M_L$ and we have neglected Γ_L compared to Γ_S. The interference experiments are always performed over a time scale such that $t \ll \frac{1}{\Gamma_L}$ so that the decay of the K_L amplitude is negligible.
14. Precise measurements of AM have been reported by: Aronson, S. H. et al.. Phys. Rev. Letters 25, 1057 (1970); Cullen, M. et al.. Phys. Letters 32B, 523 (1970); Carnegie, R. K. et al., Phys. Rev. D4, I (1971); Geweniger, C. et al., Phys. Letters ***52B,*** 108 (1974).
15. The data for this compilation are most readily available from the Particle Data Group, Barash-Schmidt, N. et al., Rev. Mod. Phys. 52, S1 (1980).

16. Figure taken from thesis of Modis, T., Columbia University (1973), (unpublished); a published version of this work is given by Carithers, W. et al., Phys. Rev. Letters 34, 1244 (1975).
17. Niebergall, F. et al., Phys. Letters 49B, 103 (1974).
18. Figure taken from thesis of Lüth, V., Heidelberg University (1974), (unpublished); a published version of this work can be found in Gjesdal, S. et al., Phys. Letters 52B, 113 (1974).
19. Banner, M. et al., Phys. Rev. Letters 21, 1103 (1968).
20. Budagov, I. A. et al., Phys. Letters 28B, 215 (1968).
21. Holder, M. et al., Phys. Letters ***40B,*** 141 (1972).
22. Christenson, J. H. et al., Phys. Rev. Letters 43, 1209 (1979).
23. From extrapolation of the phase shift analysis of Ke_4 decays, one finds $\delta_2-\delta_0 = 36°\pm10°$, Rosselet, J. et al., Phys. Rev. ***D15,*** 574 (1977). From analysis of K^+ and K_S decays to p^+p^- and $\pi^0\pi^0$, one finds $\delta_2-\delta_0 = \pm(53°\pm6°)$, Abbud, Lee, and Yang, Phys. Rev. Letters, 18, 980 (1967); Particle Data Group, Barash-Schmidt, N. et al., Rev. Mod. Phys. 52, S1 (1980). From analysis of pion production by pions, one finds $\delta_2-\delta_0 = 40°\pm6°$, see for example, Baton, J. et al., Phys. Letters 33B, 528 (1970); Estabrooks, P. and Martin, A. D., Nucl. Phys. B79, 301 (1974).
24. Bell, J. S. and Steinberger, j., Proceedings of the Oxford International Conference on Elementary Particles, 1965, edited by Walsh, T. et al., (Rutherford Laboratory, Chilton, England 1966). An analysis with a similar purpose has also been given by Sachs, R. G., Progr. Theor. Phys. (Japan) 54, 809 (1975). References to the literature concerning the analysis of the neutral K meson decay data into a T conserving part and T violating part are given by Sachs.
25. Schubert, K. R. et al., Phys. Letters ***3lB, 662 (1970).***
26. Canter, J. et al., Phys. Rev. Letters 17, 942 (1966); Meisner, G. W. et al., Phys. Rev. Letters 17, 492 (1966); Mehlhop, W. A. W. et al., Phys. Rev. 172, 1613 (1968). The last reference uses a highly innovative technique, and will give pleasure to those who take the time to read it.
27. Purcell, E. M. and Ramsey, N. F., Phys. Rev. 78, 807 (1950); also, Smith, J. H. Ph.D. thesis Harvard University (1951) (unpublished).
28. Weinberg has suggested a mechanism whereby the CP violation is due to Higgs mesons. The suggestion is attractive because the CP violation can be maximal and a neutron electric dipole moment of ~ 10-24 might be expected, see Weinberg, S., Phys. Rev. Letters 37, 657 (1976).
29. Wolfenstein, L., Phys. Rev. Letters ***13, 569 (1964).***
30. Kobayashi, M. and Maskawa, K., Progr. Theor. Phys. 49,652 (1973).
31. Guberina, B. and Peccei, R. D., Nucl. Phys. ***B163, 289*** (1980).
32. Lüth, V., Proceedings of the 1979 International Symposium on Lepton and Photon Interactions at High Energies, 1979, edited by Kirk, T. B. W., p. 83 (Fermilab, Batavia, IL, USA, 1980).
33. Andrews, D. et al., Phys. Rev. Letters 4.5, 219 (1980); Finocchiaro, G. et al., Phys. Rev. Letters 45, 222 (1980).
34. Yoshimura, .M., Phys. Rev. Letters 41,381 (1978); Dimopoulos, S. and Susskind, I,., Phys. Rev. D18, 4500 (1978); Toussaint, B. et al., Phys. Rev. ***D19, 1036 (1979);*** Ellis, J., Gaillard, M. K. and Nanopoulos, D. V., Phys. Letters ***80B, 360 (1979);*** Weinberg, S., Phys. Rev. Letters 42,850 (1979).
35. Sakharov, A. D., ZhETF Pis'ma 5, 32 (1967); English translation JETP Letters 5, 24 (1967).
36. Glashow, S. L., Rev. Mod. Phys. 52, 539 (1980); Salam, A., ibid. 52, 525 (1980); Weinberg, S., ibid, 52, 515 (1980).
37. Ford, W. T. et al., Phys. Rev. Letters 2.5, 1370 (1970).
38. Smith, K. M. et al., Nucl. Phys. ***B60, 411*** (1970).
39. Schmidt, M. et al., Phys. Rev. Letters 43, 556 (1979).
40. Baltrusaitis, R. M. and Calaprice, F. P., Phys. Rev. Letters 38, 464 (1977).
41. Steinberg, R. I. et al., Phys. Rev. Letters 33, 41 (1974).
42. Layter, J. G. et al., Phys. Rev. Letters 29, 316 (1972); Jane, M. R. et al., Phys. Letters ***48B,*** 260 (1974).
43. Dress, W. B. et al., Phys. ***Rev. D15, 9 (1977).***
44. Altarev, I. S. et al., "Search for an Electric Dipole Moment by Means of Ultra-cold Neutrons," p. 541, Proceedings of the Third International Symposium on Neutron Capture Gamma-Ray

Spectroscopy and Related Topics, 1978, Brookhaven National Laboratory, Upton, NY and State University of New York, Stony Brook, NY, ed. By Chrien, Robert E. and Kane, Walter R., (Plenum Press).

45. Gimlett, J. L., Phys. Rev. Letters 42, 3.54 (1979).
46. Davis, B. R. et al., Phys. Rev. C22, 1233 (1980).
47. Weitkamp, W. G. et al., Phys. Rev. 165, 1233 (1968); von Witsch, W. et al., Phys. Rev. 169, 923 (1968); Thornton, S. T. et al., Phys. Rev. C3, 1065 (1971); Driller, M. et al., Nucl. Phys. ***A317 300*** (1979).



Val L. Fitch

VAL LOGSDON FITCH

I was born the youngest of three children, on a cattle ranch in Cherry County, Nebraska, not far from the South Dakota border, on March 10, 1923. This is a very sparsely populated part of the United States and remote from any center of population. It seems incredible by modern standards that by the age of 20 my father, Fred Fitch, had acquired a ranch of more than 4 square miles and had persuaded a local school teacher, Frances Logsdon, to marry and join him in living there. They moved to the ranch just 20 years after the battle of Wounded Knee, which occurred about 40 miles northwest. I mention this because our living close to their reservation made the Sioux Indians very much a part of our environment. My father, while not fluent, spoke their language. They recognized his friendly interest on their behalf by making him an honorary chief.

Not long after my birth my father was badly injured when a horse he was riding fell with him. He subsequently had to give up the physically strenuous activity associated with running a ranch and raising cattle. The family moved to Gordon, Nebraska, a town about 25 miles away, where my father entered the insurance business. All of my formal schooling through high school was in the public schools of Gordon. During this period my parents retained ownership of the ranch but the operation was largely left to others. E. B. White has defined farming as 10 % agriculture and 90 % fixing something that has gotten broken. My memories of ranching are primarily not the romantic ones of rounding up and branding cattle but rather of oiling windmills and fixing fences.

Probably the most significant occurrence in my education came when, as a soldier in the U.S. Army in WWII, I was sent to Los Alamos, New Mexico, to work on the Manhattan Project. The work I did there under the direction of Ernest Titterton, a member of the British Mission, was highly stimulating. The laboratory was small and even as a technician garbed in a military fatigue uniform I had the opportunity to meet and see at work many of the great figures in physics: Fermi, Bohr, Chadwick, Rabi, Tolman. I have recorded some of the experiences from those days in a chapter in *All in Our Time,* a book edited by Jane Wilson and published by the Bulletin of Atomic Scientists. I spent 3 years at Los Alamos and in that period learned well the techniques of experimental physics. I observed that the most accomplished experimentalists were also the ones who knew most about electronics and electronic techniques were the first I learned. But mainly I learned, in approaching the measurement of new phenomena, not just to consider using existing apparatus but to allow the mind to wander freely and invent new ways of doing the job.

Robert Bather, the leader of the physics division in which I worked, offered

me a graduate assistantship at Cornell after the war but I still had to finish the work for an undergraduate degree. This I did at McGill University. And then another opportunity for graduate work came from Columbia and I ended up there working with Jim Rainwater for my Ph. D. thesis. One day in his office, which he shared at the time with Aage Bohr, he handed me a preprint of a paper by John Wheeler devoted to μ-mesic atoms. This paper emphasized, in the case of the heavier nuclei, the extreme sensitivity of the 1s level to the size of the nucleus. Even though the radiation from these **atoms** had never been observed, these atomic systems might be a good thesis topic. At this same time a convergence of technical developments took place. The Columbia Nevis cyclotron was just coming into operation. The beams of -measons from the cyclotron contained an admixture of μ-measons which came from the decay of the n's and which could be separated by range. Sodium iodide with thallium activation had just been shown by Hofstadter to be an excellent scintillation counter and energy spectrometer for γ rays. And there were new phototubes just being produced by RCA which were suitable matches to sodium iodide crystals to convert the scintillations to electrical signals. The other essential ingredient to make a y-ray spectrometer was a multichannel pulse height analyzer which, utilizing my Los Alamos experience, I designed and built with the aid of a technician. The net result of all the effort for my thesis was the pioneering work **on** μ-mesic atoms. It is of interest to note that we came very close to missing the observation of the y-rays completely. Wheeler had calculated the 2p-1s transition energy in Pb, using the then accepted nuclear radius 1.4 $A^{1/3}$ fermi, to be around 4.5 MeV. Correspondingly, we had set our spectrometer to look in that energy region. After several frustrating days, Rainwater suggested we broaden the range and then the peak appeared-not at 4.5 MeV but at 6 MeV! The nucleus was substantially smaller than had been deduced from other effects. Shortly afterwards Hofstadter got the same results from his electron scattering experiments. While the μ-mesic atom measurements give the rms radius of the nucleus with extreme accuracy the electron scattering results have the advantage of yielding many moments to the charge distribution. Now the best information is obtained by combining the results from both μ-mesic atoms and electron scattering

Subsequently, in making precise y-ray measurements to obtain a better mass value for the p-meson, we found that substantial corrections for the vacuum polarization were required to get agreement with independent mass determinations. While the vacuum polarization is about 2 % of the Lamb shift in hydrogen it is the very dominant electrodynamic correction in μ-mesic **atoms.**

My interest then shifted to the strange particles and K mesons but I had learned from my work at Columbia the delights of unexpected results and the challenge they present in understanding nature. I took a position at Princeton where, most often working with a few graduate students, I spent the next 20 years studying K-mesons. The ultimate in unexpected results was that which was recognized by the Nobel Foundation in 1980, the discovery of CP-violation.

At any one time there is a natural tendency among physicists to believe that

we already know the essential ingredients of a comprehensive theory. But each time a new frontier of observation is broached we inevitably discover new phenomena which force us to modify substantially our previous conceptions. I believe this process to be unending, that the delights and challenges of unexpected discovery will continue always.

It is highly improbable, a priori, to begin life on a cattle ranch and then appear in Stockholm to receive the Nobel prize in physics. But it is much less improbable to me when I reflect on the good fortune I have had in the ambiance provided by my parents, my family, my teachers, colleagues and students. I have two sons from my marriage to Elise Cunningham who died in 1972. In 1976 I married Daisy Harper who brought with her three stepchildren into my life.

Honors and Distinctions:
I am a fellow of the American Physical Society and the American Association for the Advancement of Science, a member of the American Academy of Arts and Sciences and the National Academy of Sciences. I hold the Cyrus Fogg Brackett Professorship of Physics at Princeton University and since 1976 have served as chairman of the Physics Department. I received the E. 0. Lawrence award in 1968. In 1967 Jim Cronin and I received the Research Corporation award for our work on CP violation and in 1976 the John Price Witherill medal of the Franklin Institute.

THE DISCOVERY OF CHARGE-CONJUGATION PARITY ASYMMETRY

Nobel lecture, 8 December, 1980

by

VAL L. FITCH

Princeton University, Department of Physics,
Princeton, New Jersey 08540

Physics as a science has made incredible progress because of the delicate interplay between theory and experiment. Astonishing predictions based on theories devised to account for known phenomena have been confirmed by experiment. Experiments probing previously unexplored areas often reveal physical effects which are completely unanticipated by theoretical conjecture. The incorporation of the new effects into a theoretical framework then follows.

This year Prof. Cronin and I are being honored for a purely experimental discovery, a discovery for which there were no precursive indications, either theoretical or experimental. It is a discovery for which after more than 16 years there is no satisfactory accounting. But showing as it does a lack of charge-conjugation parity symmetry and, correspondingly, a violation of time-reversal invariance, it touches on our understanding of nature at its deepest level.

The discovery of failure of CP symmetry was made in the system of K mesons. This observation is especially interesting because it was the study of these same particles that led to the overthrow of parity conservation, the notion that interactions and their mirror-reflected counterparts must be equal.

My own interest in K particles started in 1952-53 while I was at Columbia working with Jim Rainwater on μ--mesonic atoms. At that time the strange behavior of the particles newly discovered in cosmic rays[1] was a major topic of conversation in the corridors and over coffee. By strange behavior I am referring to the copious production but slow decay. Protons bombarded by pions would result in the production of Λ^0's at 10^{13} times the rate of their decay back to pions and protons. Pais came to Columbia and talked of his ideas on associated production to explain this anomaly.[2] Gell Mann visited and discussed the scheme which he and independently, Nakano and Nishijima, had devised to account for associated production.[3,4]

Their idea was implausible and daring in the face of available data. The scheme assigned the K mesons to two doublets, K^+K^0, and the antiparticles K^- and $\overline{K}^0$. The natural assignment would have been the same as for pions, a triplet of particles K^+, K^0, K^-. Nishijimaalso assigned quantum numbers, subsequently called stangeness, which were conserved in the strong interaction but not in the weak. The K^+K^0 were assigned + 1, the $\overline{K}^0K^-$ as well as the Λ^0, - 1

Standing alone among the particles with positive strangeness were the K^+ and K^0 mesons, and I idly thought that if the situation was ever to be understood these objects might be the key. Most often experiments in physics are long and difficult. It takes some special tweaking of interest to make the commitment to a new area of research. The original motivation is, in the end, apt to appear naive. However, I did in fact join the Princeton Cosmic Ray Group headed by George Reynolds, and spent the summer of 1954 on a mountain in Colorado learning about the ongoing experiments. During the same period the energy of the cosmotron at Brookhaven was being raised to 3 GeV. Associated production was clearly seen by Shutt and his group at Brookhaven(and K mesons produced in the cosmotron were identified in photographic emulsion.(“) By the end of the summer I reluctantly decided the future was not in studying cosmic rays in the mountains I loved, but with the accelerators.

The following fall, with Bob Motley, a graduate student, we began to design an apparatus to detect K mesons using purely counter techniques at the cosmotron. As this work progressed the cascading interest in the tau-theta puzzle [7] led us naturally to explore the lifetime of the K particles as a function of their decay mode. We were successful with our detectors and Motley and I published our results simultaneously with those from the Alvarez group at Berkeley which was using the bevatron as a source. [8,9] These results showed the degeneracy in the lifetime of the tau and theta mesons. Independently the masses of tau and theta had been shown to be the same to within 1 %. [10] The situation then set the stage for the famous work of Lee and Yang[11] followed by the experiments with the striking results showing maximal parity violation in the weak interactrons.[12] This remarkable story was told by Lee and Yang on this occasion in 1957.

At about this time there appeared a paper by Landau written before the results of the beta decay experiments were known.(“) Addressing the tau-theta problem, he observed that a simple rejection of parity conservation would create difficult problems in physics. However, with what he called “combined inversion”, that is, space inversion and the simultaneous transformation of particle into antiparticle, the difficulties would be avoided. Indeed, this is a path that nature appeared to take. Subsequent experiments showed parity violation was compensated by a failure of charge conjugation. The weak interactions were therefore invariant under the combined operations of particle-antiparticle interchange and mirror reflection, charge conjugation-parity, CP.

One symmetry had been shown to be invalid but had been replaced by a still deeper one. This new symmetry was especially appealing because of the CPT theorem. This theorem, which is based on little more than special relatively and locality and which is the foundation of all quantum field theory, says that all interactions must be invariant under C, P, and T, time reversal, all combined. If CP is good so also is T, in complete accord with all experimental data. The subject was left in a highly satisfactory state. “Who would have dreamed in 3953 that studies of the decay properties of the K particles would lead to a new revolution in our understanding of invariance

principles," wrote Sakurai in 1963.[14] But then in 1964 these same particles, in effect, dropped the other shoe.

It is difficult to give a better example of the mutually complementary roles of theory and experiment than in telling the story of the neutral K mesons. For a physicist the pleasures are special because there is scarcely a physical system which contains so many of the elements of modern physics. Two-state systems, of which this is an example, abound, but this one has special properties which give it a unique beauty. I hope that I can convey to you some of the reasons why this system has held such a fascination for us. The story begins with the isotopic spin, strangeness assignment of Gell Mann and Nishijima. The assignment of the K mesons to two doublets makes the K^0 and $\bar{K}^0$ distinct entities. But both particles decay to two π mesons. If the physicist sees π^+ and π^- mesons in his detector, which is the source, the K^0 or $\bar{K}^0$? The problem was solved through the remarkable insight of Gell Mann and Pais in their 1955 paper.[15] In the spirit of quantum mechanics it is necessary that the source of the $\pi^+\pi^-$ mesons be some linear combination of K^0 and $\bar{K}^0$ states. They observed that a $\pi^+\pi^-$ final state is even under charge conjugation. By even we mean that the wavefunction does not change its algebraic sign upon interchange of particle and antiparticle. This evenness condition is obviously met by the combination $K^0+\bar{K}^0$. This they called the K_1^0.[16] If this is the case, there must be another state equally probable, the $K^0-\bar{K}^0$, the K_2^0, which is odd under charge conjugation and, correspondingly, is forbidden to decay to $\pi^+\pi^-$. But it can decay to many other states, three-body states such as $\pi^+\pi^-\pi^0$ It was expected that the decay to the three-body states would be substantially inhibited compared to the two-body. The particle corresponding to the K_2^0 would have a longer lifetime than the K_1^0 by about 500. In addition, it was expected that the K_1^0 and K_2^0 would have somewhat different masses even though the masses of K" and $\bar{K}^0$ are strictly equal by the CPT theorem.

This long-lived neutral K meson, predicted by Gell-Mann and Pais, was then looked for and found by a Columbia group working at the Brookhaven cosmotron.[17] The theoretical model, based on the notion of charge conjugation invariance in the weak interactions, had been confirmed. Then suddenly parity was found to be violated in the weak interactions along with charge conjugation! This dark cloud was almost immediately removed with the observation that one had only to replace C with CP and the story of the neutral K mesons would remain the same.[13] With CP invariance the K_2^0 would continue to be absolutely forbidden to decay to two pions. The successful1 description of the neutral system of K mesons has been characterized by Feynman as "one of the greatest achievements of theoretical physics."[18]

Additional features of the $K^0\bar{K}^0$ system become evident if we write the wavefunction including the lifetime and energy terms for the case of production of a K^0 at t = 0.

$$\Psi(t) = \frac{1}{\sqrt{2}}\{|K_1^0> e^{-t/2\tau_1+i\omega_1 t} + |K_2^0> e^{-t/\tau_2+i\omega_2 t}\}$$

$$|K_1^0> = \frac{1}{\sqrt{2}}[|K^0> + |\bar{K}^0>]$$

$$|K_2^0> = \frac{1}{\sqrt{2}}[|K^0> - |K^0>]$$

It is seen that after a time, long compared to the K_1^0 lifetime and short compared to the K_2^0 lifetime, the state that was originally a pure K^0 will become a K_2^0 which in turn is an equal mixture of K^0 and $\bar{K}^0$. To give a measure of the magnitudes involved we should point out that the K_1^0 meson, in a typical experimental situation, travels an average of a few centimeters before it decays, whereas the K_2^0 travels tens of meters. At a distance greater than about one meter from the point ofproduction of a K^0a nearly pure K_2^0 beam will be present.

Another important characteristic of the system becomes apparent when we consider the interaction of K_2^0's with matter. The K^0's and $\bar{K}^0$'s, by virtue of their opposite strangeness, have quite different interaction cross sections. Passing a beam of K_2^0's through a block of material will result in a mixture of K^0 and $\bar{K}^0$'s which, because of differential absorption of the two components, is no longer 50-50, but instead a mixture equivalent to a new combination of K_1^0's and K_2^0's. The newly produced short-lived K_1^0's decaying to $\pi^+\pi^-$ will appear behind the material. This phenomenon is called regeneration.[19] In the case of the absorbing material being completely transparent to K^0's and opaque to $\bar{K}^0$'s the intensity of the K_1^0's after the absorber will be 1/4 the initial intensity of the K_2^0 incident on the absorber.

In the late 1950's M. L. Good[20] observed that with a very small mass difference between the K_1^0 and K_2^0 the regeneration phenomena just discussed would result in a coherent process. By coherent we mean that the scattering process of K_2^0 to K_1^0 would not be from individual nuclei but from the whole block of scattering material! That is, the block of material would remain in its initial quantum mechanical state during the scattering process. In this case, as with ordinary light passing through glass, the regeneration material could be treated as having an index of refraction. The K_1^0's regenerated coherently would have precisely the same energy as the incident K_2^0's and an angular distribution identical to the incident beam but broadened by diffraction effects determined by the size of the regenerating material perpendicular to the beam. A characteristic wavelength for the K_2^0 mesons in a typical experiment is about 10^{-13} cm. The transverse dimensions are typically 10 cm. The corresponding diffraction pattern has a width of the order of 10^{-14} radians! In addition, the coherent addition of K_1^0 waves has been observed over distances greater than 10^{14} wave lengths. The unique feature of this coherently regenerated K_1^0 beam is that it can be distinguished from the original beam since it decays with a short lifetime to $\pi^+\pi^-$. To my knowledge, it is the only instance where a forward coherently scattered beam can be distinguished from the original beam.

It has become evident to physics students in the audience that the $K_1^0K_2^0$ story has an analogy in polarized light. The K_1^0 and K_2^0 correspond to the left and right circularly polarized light, and the K^0 and $\bar{K}^0$ states are equivalent to the x and y components of linear polarization. The passage of a K_2^0 beam through a block of condensed material is equivalent to the passage or left circular polarized light through a doubly refractive medium like calcite which has a different index refraction for the x and y components of polarization. The general picture of regeneration, coherent and incoherent, was confirmed in a definitive bubble chamber experiment.[21]

There are many associated phenomena still to be explored. For example, experiments coherently regenerating K_1^0's from the planes in crystals have yet to be done. At the particle momenta commonly available the Bragg angles are exceedingly small, and the extinction factor, the Debye-Waller factor, comes into play at correspondingly small angles, but the experiments could be done.

Unexpectedly, the K" $\bar{K}^0$ system provides us with important and highly precise information about the gravitational interaction. It relates to the question of strong universality; that is, whether different objects, in this case particle and antiparticle, with the same inertial mass behave the same in a gravitational field. As observed by M. L. Good,[22] if the K" and its antiparticle, the $\bar{K}^0$, had an opposite gravitational potential energy, the K" $\bar{K}^0$ system would mix so quickly that the long-lived particle would never be seen. By analyzing the system in more detail one can show that if the gravitational interaction of particle and antiparticle differ by a fraction, κ, then κ must be less than 10^{-10} if we're dealing with the gravitational field of the earth, 10^{-11} for the solar system, and 10^{-13} for the galaxy.

Voyages of discovery can be made in new uncharted waters but also in the familiar bays close to port provided one has observing apparatus that can see familiar objects with detail greater than that previously possible. In 1963 we had the opportunity to investigate the neutral K meson phenomena with resolution greater than that permitted before. The introduction of spark chambers as charged particle detectors permitted precise track position determination, but also the chambers could be selectively triggered on appropriate classes of events.

Using such new devices with our colleagues, Jim Christenson and Rene Turlay, Jim Cronin and I initiated a systematic study of (1) the regeneration phenomena, (2) what we called CP invariance, and (3) neutral currents. We were interested in the regeneration phenomena in particular because of an anomaly that had just been reported by a group studying the passage of K_2^0's through a liquid hydrogen bubble chamber.[23] Not many of our colleagues would have given us much credit for studying CP invariance, but we did anyway, and neutral currents, of long interest, were discussed by Professor Glashow on this occasion one year ago.

A plan view of the apparatus we used for these studies is shown in Figure 1. It is a two-armed spectrometer, each arm with spark chambers before and after a magnet for track delineation. Cerenkov and scintillation counters in both arms operated in coincidence provided the signals to trigger the spark chambers, which were recorded photographically. The apparatus was placed in a beam of neutral particles at the Brookhaven A. G. S. at a distance such that K_1^0's would have decayed away leaving K_2^0's. The angle between the spectrometer arms was chosen to optimize the detection of K" mesons decaying to two π mesons. In the regeneration studies blocks of various solid materials were placed in the neutral beam. In the studies of the free decay of $K_2^0 \rightarrow 2$ pions, the decay volume was filled with helium gas to minimize the interactions.

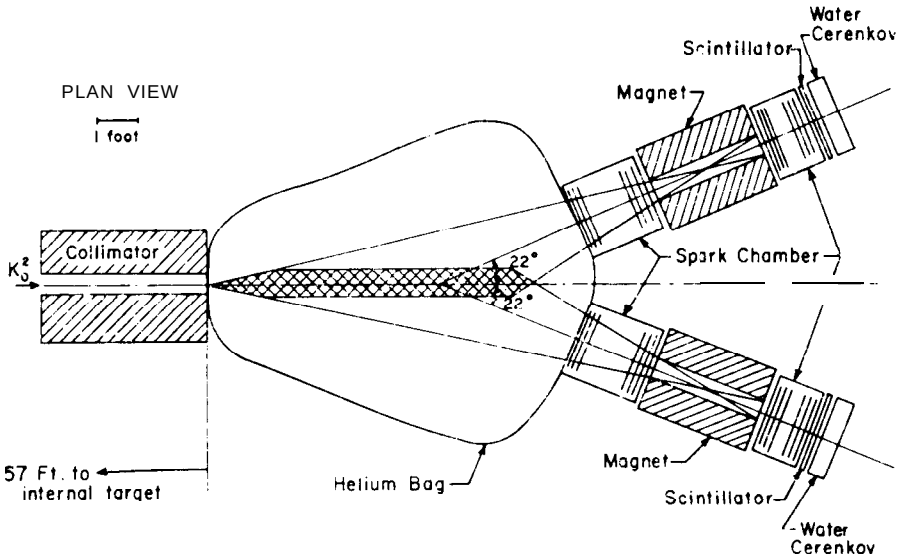


Fig. I. Plan view of the apparatus as located at the A. G. S.

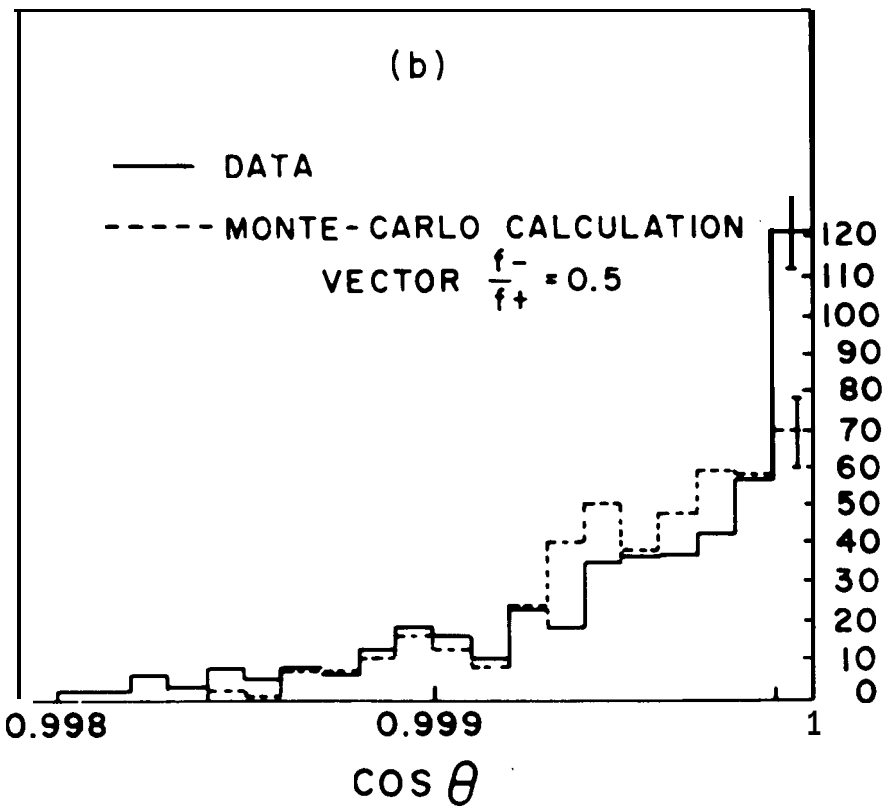


Fig. 2. Angular distributions of those events in the appropriate mass range as measured by a coarse measuring machine.

The decay to 2 pions is distinguished from the copious three-body decays in two ways. The sum of the momenta of the two detected particles must line up with the direction of the incident K_2^0's. In general this will not happen for three-body decay. In addition, the mass computed for the parent particle must match the mass of the K^0 meson. The original data are shown in Figure 2 and 3. Figure 2 shows the data after measurement of the photographic records on a relatively coarse measuring machine. The presence of the peaking of events along the beam line stimulated more precise measurements and these results are shown in Figure 3. Clearly there are about 56 events in the forward peak in the proper mass interval where the background is 11. From this data we established that the branching ratio of K_2^0 to 2 pions relative to all the charge modes decay was 2×10^{-3}. Here was the first evidence for the decay completely forbidden by CP conservation.[23] We were acutely sensitive to the importance of the result and, I must confess, did not initially believe it ourselves. We spent nearly half a year attempting to invent viable alternative explanations, but failed in every case.

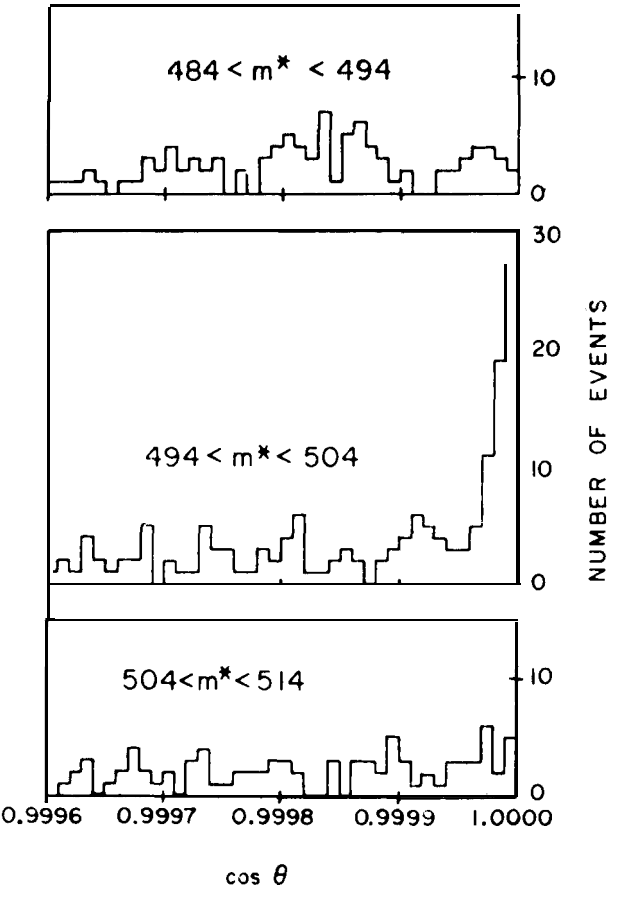


Fig. 3. Angular distribution of the events after measurement by a precise machine in three relevant mass regions.

The study of coherent regeneration was important for the CP measurement for several reasons. First, the results we found were entirely consistent with expectations; there were no anomalies. The measured coherent regeneration rates in tungsten, copper, carbon, and liquid hydrogen enabled us to show that coherent regeneration in the gaseous helium which filled the decay volume would produce a totally negligible contribution to the signal we observed. Second, the coherent regeneration of the K_1^0's, which subsequently decrayed to $\pi^+\pi^-$ mesons, provided an invaluable calibration of the apparatus.

It is appropriate now to look at the neutral K system in a somewhat more quantitative way.[24] Because of the mixing of the K^0 and $\bar{K}^0$ through the weak interaction, the time rate of change of a K^0 wave will not only depend on the K^0 amplitude, but also on the $\bar{K}^0$ amplitude, viz.,

$$-\frac{d\,K^0}{dt} = A\,K^0 + p^2\bar{K}^0$$

and

$$-\frac{d\,\bar{K}^0}{dt} = B\,\bar{K}^0 + q^2\,K^0.$$

We have let the particle symbol stand for the amplitude of the corresponding wave. With invariance under CPT, particle and antiparticle masses and lifetimes must be precisely identical. In terms of the above equations, A must be equal to B. Now, CP violation can, in fact, occur in two ways, either through terms in the set of equations above, or in the amplitudes for the decay. Subsequent experiments show that most, if not all, of the violation is in the equations above, involving the so-called mass-decay matrix. Professor Cronin will discuss the ramifications of the effect being present also in the decay terms. Suffice is to say here that any departure of p^2 from q^2 will result in the decay of the $K_2^0 \rightarrow$ 2 pions. With CP nonconservation the short and long-lived particles are no longer the K_1^0 and K_2^0 previously defined but rather

$$K_S^0 = \frac{1}{\sqrt{p^2+q^2}} \quad \{p|K^0> + q|\bar{K}^0>\}$$

and

$$K_L^0 = \frac{1}{\sqrt{p^2+q^2}} \quad \{p|K^0> - q|\bar{K}^0>\}$$

The fact that K_L^0 decays to 2 pions shows that the amplitude for particle to antiparticle transitions, in this case $K^0 \rightarrow \bar{K}^0$, does not quite equal the reverse, $\bar{K}^0 \rightarrow K^0$, and indeed we now know rather precisely that not only are the magnitudes somewhat different but that there is a small phase angle between the two amplitudes. See Figure 4.

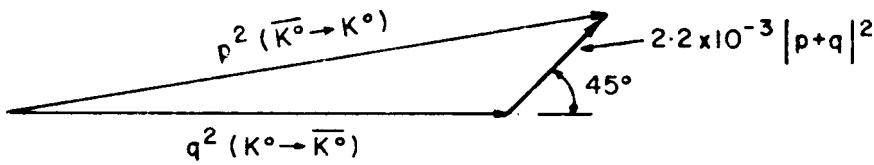


Fig. 4. Vector diagram showing schematically the difference in the amplitudes for $K^0 \rightarrow \bar{K}^0$ and $\bar{K}^0 \rightarrow K^0$.

We indicated earlier that, through the CPT theorem, a violation of CP is equivalent to a violation of time reversal invariance. As Professor Cronin will show, the CPT theorem has been shown to hold in the neutral K system independently, so in a self-contained way a violation of time reversal invariance is demonstrated.

We are all familiar with the time asymmetry associated with entropy. Entropy in a closed system increases with time. This kind of time asymmetry results from the boundary conditions. But for the first time we have in the neutral K mesons a physical system that behaves asymmetrically in time as a result of an interaction, not a boundary condition.

Since the microscopic physical laws had always been thought to be invariant under time reversal, this discovery opens up a very wide range of profound questions. Professor Cronin will go into some of these questions in greater detail. I will mention two. Can this effect be used to decrease the entropy of an isolated system? We look out from the earth and see a highly ordered universe. With entropy always increasing how can this be? Is CP violation an effect that can be used, in effect, to wind up the universe? The answers to these questions appear to be no.[25]

At the same time we look out from the earth and see the remains of an earlier much hotter universe. In that earlier time one expects that matter and antimatter would condense out in equal amounts and eventually annihilate to gamma radiation. However no evidence of antimatter is seen. The gauge theories described on this occasion one year ago allow for the possibility of proton (and antiproton) decay. This process, coupled with CP violation, drives the universe towards a preponderance of matter over antimatter and can account for the observed ratio of the amount of matter to radiation.[26]

Lewis Thomas, whose essays on science grace our literature, has written, “You measure the quality of the work by the intensity of the astonishment.” After 16 years, the world of physics is still astonished by CP and T non-invariance. I suspect that the Nobel Committee was motivated by considerations similar to those of Thomas in awarding to Professor Cronin and myself this highest of honors.

REFERENCES

I. For a review ca 19.53 see Rochester, G. D. and Butler, C. C., Reports Progress in Phys. 16, 364 (1953).

2. Pais, A. Phys. Rev. 86,663 (1952).

3. Gell-Mann, M. Phys. Rev. 92, 833 (1953).

4. Nakano, T. and Nishijima, K. Progr. Theoret. Phys. 10, 581 (1953).

5. Fowler, W. B., Shutt, R. P., Thorndikc, A. M. and Whittemore, W. L. Phys. Rev. 93,861 (1954).

6. Rochester Conference Proc. (1954).

7. Dalitz, R. H. Phil. Mag. 44, 1068 (1953); Fahri, E. Nuovo Cimento 11, 479 (1954).

Among the strange particles some were seen to decay to two and some to three pions. By using the analysis of Dalitz and Fahri, it was shown, with very few examples in hand, that the parity of the three pion system was opposite to that of the two pion system. If parity is conserved in the decay interaction then there must he distinguishable parents of opposite parity, the theta that decays to two and the tau that decays to three pions. The puzzle was in the question, if the particles are distinct entities why should they have the same mass and lifetime? Now with parity violation both are recognized as K mesons, $K_{\pi 2}$ and $K_{\pi 3}$.

8. Alvarez, L. W., Crawford, F. S., Good, M. L. and Stevenson, M. L. Phys. Rev. 101, 503 (1956); Harris, G., Orear, J. and Taylor, S. Phys. Rev. 100,932 (1955).

9. Fitch, V. and Motley, R. Phys. Rev. ***101, 496*** (1956); Phys. Rev. 10.5, 265 (1957).

IO. Birge, R. W., Perkins, D. H., Peterson, J. R., Stork, D. H. and Whitehead, M. N. Nuovo Pimento 4, 834 (1956).

11. Lee, T. D. and Yang, C. N. Phys. Rev. 104, 254 (1956).

12. Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D. and Hudson, R. P. Phys. Rev. 10.5, 1413 (1957); Garwin, R., Lederman, L. and Weinrich, M. Phys. Rev. 105, 1415 (1957); Friedman, J. I. and Telegdi, V. I.. Phys. Rev. 105, 1681 (1957).

13. Landau, L. Nucl. Phys. 3, 254 (1957).

14. Sakurai, J. J. Invariance Principles and Elementary Particles, Princeton University Press (1964), Princeton, N. J., p. 296.

15. Gell-Mann, M. and Pais, A. Phys. Rev. 97, 1387.

16. We have changed the notation to correspond to recent custom. Gell-Mann and Pais called them Θ_1 and Θ_2.

17. Lande, K., Booth, E. T., Impeduglia, J., Lederman, L. .M. and Chinowsky, W. Phys. Rev. ***103, 1901 (1956).***

18. Feynman, R. P. Theory of Fundamental Processes, Benjamin, W. A. Inc. New York, p. 50.

19. Pais, A. and Piccioni, 0. Phys. Rev. ***100,*** 1487 (1955).

20. Good, M. L. Phys. Rev. 106, 591 (1957).

21. Good, R. H., Matsen, R. P., Muller, F., Piccioni, O., Powell, W. M., White, H. S., Fowler, W. B. and Birge, R. W. Phys. Rev. ***124,*** 1223 (1961).

22. Good, M. L. Phys.Rev. 121, 311 (1961).

23. Christenson, J., Cronin, J. W., Fitch, V. I.. and Turlay, R. Phys. Rev. Letters 13, 138 (1964).

24. Lee, T. D., Cehme, K. and Yang, C. N. Phys. Rev. 106, 340 (1957).

25. Ne'eman, Y. Erice Summer School Lectures, June 16-July 6, 1972.

26. Sakharov, A. D. JETP Letters 5, 24 (1967). For a non-technical discussion see Wilczek, F. W. Scientific American 243, 82 (Dec. 1980).