TEKST NR 394 2001

Bent Sørensen

PHYSICS REVEALED

THE METHODS AND SUBJECT MATTER OF PHYSICS

an introduction to pedestrians (but not excluding cyclists)

PART II: PHYSICS PROPER

MARCH 2001

PRIS: 77.00
PHYSICS REVEALED:
9 789673 011599
14.03.2001
STUDIERABAT-10%

TEKSTER IMFUFA

fra Institut 2

ROSKILDE UNIVERSITETSCENTER
INSTITUT FOR STUDIET AF MATEMATIK OG FYSIK SAMT DERES FUNKTIONER I UNDERVISNING, FORSKNING OG ANVENDELSER

ROSKILDE UNIVERSITY, P O BOX 260, DK-4000 ROSKILDE, DENMARK
INSTITUTE OF STUDIES IN MATHEMATICS AND PHYSICS, AND THEIR FUNCTIONS IN EDUCATION, RESEARCH AND APPLICATIONS
TEL: +45 4674 2000, FAX: +45 4674 3020, WEBSITE: http://mmf.ruc.dk/energy

MARCH 2001

PHYSICS REVEALED: The methods and subject matter of physics, an introduction to pedestrians (but not excluding cyclists). PART II: PHYSICS PROPER

this text may be downloaded from http://mmf.ruc.dk/~boson/fysikE
printing history:
-as lecture notes for FYSIK E, 1983
-this version: part 2 of the material for the course FYSIK E, spring 2001

IMFUFA text 394, 140 pages ISSN 0106 6242

Abstract:

This is part 2 of 3, serving as reading material for the a course called Physics E. The course addresses students not aiming at becoming physicists (although some do after all!), as part of the two-year curriculum for the Natural Science base years of the Bachelor and Master's studies at Roskilde University. Part 2 covers the subject matter and methods of physics. Each Chapter has an introductory survey of a given field, followed by a short mathematical description of the methods used in that field.

Adjoining materials:
Part 1: Physics in society, deals with the place physics occupies in society
Part 3: Physics in philosophical context, deals with the recent history of physics, with emphasis on the period 1960-1990, and placed in a context of philosophy of science as well as sociology of the physics community.

Contents of Part II:

Bent Sørensen

PHYSICS REVEALED

THE METHODS AND SUBJECT MATTER OF PHYSICS

an introduction to pedestrians (but not excluding cyclists)

PART II: PHYSICS PROPER

MARCH 2001

PREFACE

This material grew out of a series of lectures held at Roskilde University since 1982. It has been somewhat updated for the 2001 lectures, but many of the issues used to illustrate physics are unchanged from the 1983 version, reflecting a belief that the most recent illustrations are not always the most instructive ones.

The course using this material is primarily aimed at first (two) year's natural science students not aiming at specialising in physics (although some do after all). It is, however, very different from the available textbooks for such courses, which generally try to make these people "small" ordinary physicists by presenting them a simplified version of "real" physics textbooks.

My treatment does not present a simplified version of physics. All the complexity is there, but it is not what the student is supposed to master. The student has achieved the purpose of the course if he or she at the end knows the breathtaking depth of the subject matter of physics and has a feeling for the kind of methods employed by physicists. In addition to this, the position of science in current society is presented for debate, and is placed in a historical and philosophical context.

Dyson once said: The difference between a text without problems and a text with problems is like the difference between learning to read a language and learning to speak it (Dyson, 1981). I shall invite the reader to join in two kinds of exercises. One I call "problems" and the other "discussion issues". Both are found at the end of each chapter. The problems may be solved using physical theory at some level. Some may be answered from everyday experience, but in such cases a physical reasoning should be sought. In many of the problems, the sharpening of the formulation into a well-defined set of questions is 90% of the solution. Such detailed formulations will be given for selected problems, but remember that it is only in textbooks that these precise problem formulations exist. In real life, problems are mostly diffuse and open-ended. Here is one to start you off: Read the short love story from my book *Superstrings* (Sørensen, 1987), which you can find on my web page

http://home9.inet.tele.dk/novator/Bent/SSTkap45.htm

and discuss what it has to do with relativity theory [at the moment, the story is only in Danish].

The other participation part consists of bringing up society-related issues connected to the text in each chapter. They may be discussed in groups or with yourself as the discussion partner. They may lead you anywhere, and there are no correct or false conclusions. I hope the reader will think of further discussion topics. People insisting that they see no problem in entrusting quantitative evaluations of scientific issues to politicians or some other decision-makers may of course skip these problems!

Gilleleje 2001, Bent Sørensen

PHYSICS REVEALED: The methods and subject matter of physics

CONTENTS

Chapter 7

What is physics really about?

The subject matter of physics can be categorised in more than one way. Conventionally, a division is made on subject areas such as solid state physics, atomic physics, nuclear physics, high energy or elementary particle physics and astrophysics. One may say that this is a division according to characteristic amounts of energy needed to study the phenomena in the different disciplines. The division used in the following chapters is more geared to the circumstances, under which we use physical concepts and formulate physical laws: the structure of matter comprise all the energy intervals from those characterising the chemical binding of atoms in molecules up to the high energies needed to probe into say the internal structure of a proton.

A completely different approach would be to divide physics according to the kinds of models used. A lot of models are used in physics. On one side are the physical models aimed at isolating a part of the physical universe, in order to better understand this subuniverse while at the same time disregarding other and perhaps prominent features of the known physical reality. On the other side are the mathematical models used to express nearly all physical laws. A mathematical model may be a formal expression (say a formula) of a physical model (say a pictorial model showing what should be related to what, and indicating the direction of causal implications). Often the mathematical model becomes synonymous with the physical law, and the physicist may formulate new ideas directly in terms of the mathematical models, sometimes guided by what is mathematically most simple or "nice", and then only later starting to interpret the mathematical models in terms of physical models,

Of course there are fundamental differences between mathematical and physical models. A mathematical model in itself can be correct or incorrect, while a physical model can be true or false or something in-between. The correctness of a mathematical model is a question of internal consistency, whereas the physical model is the closer to being "true", the more consistent it is with what goes on in the real world. Well, if the physical model is expressed in a mathematical form, one could say that the same set of equations at the same time constitute a physical and a mathematical model. Still, one can ask the separate questions of whether or not it is mathematically correct, and whether or not it is physically "true". "True" is perhaps not the best word, because we can only know that a model or a theory describes our experience well "so far" - there could come an experience, say the result of an experiment, tomorrow, which would disagree with the currently used model or the theory.

There is also a difference between the basic approach of a mathematician and a physicist, which one should note. The mathematician would like to express his model in the most general form, while the physicist contemplating to use a mathematical model tries to keep it as close to what is needed for the physics problem, as possible. This means looking at special cases of more general mathematical theories, for example applying a model valid for any dimension to a three-dimensional situation. Furthermore, the physicists may not always be interested in the full mathematical rigor. Useful physical approximations, such as sometimes considering a small object as a point particle having

mass but no extension, may lead to mathematical problems (singularities, divergences), which do not necessarily bother the physicist, because their cause is the approximate treatment of certain features of the physical objects, deemed useful for other reasons. Laws derived from one set of models may turn out to have much wider validity than the models from which they were derived. This is the case, as Feynman (1965) notes, for Kepler's laws and the law of angular momentum conservation. One of Kepler's laws states that the radius vector of planetary motion (a line from the Sun to the planet considered) sweeps equal areas in equal times. From this statement, the conservation of angular momentum (a quantity numerically equal to the product of that radius vector and the velocity of the planet, times the sine function of the angle between the directions of the radius vector and the velocity) can be deduced. However, while Kepler's laws are valid only in connection with gravitational forces, the law of angular momentum conservation has a much wider area of validity.

The physicist thus uses mathematical models in quite a different way from the one, in which a mathematician would use them. To the physicist, the transformation of a physical model into rigorous mathematical terms may only be a stepping stone for the derivation of new physical models having wider applicability.

7.1 Some general requirements for physical laws

Can we say something general about the form of mathematical models suited for the description of the physical nature? The answer seems to be "yes". There are a number of requirements that would seem very reasonable and in some cases absolutely necessary, at least for a certain class of models aimed at understanding nature.

One such proposed requirement is "time displacement invariance". It states that physical laws should not depend on an overall shift in the time co-ordinate. In other words, the course of a physical process will be the same independently of whether the process is started at one or at another time, as long as initial conditions other than the starting time are kept the same,

A requirement of similar form is "translational invariance", stating that physical laws should not depend on an overall displacement of position co-ordinates (space co-ordinates). This means that the course of a physical process will be the same at different space locations, as long as all other initial conditions are maintained.

One can go on and require the physical laws to be invariant under rotational displacement, so that a physical process will not change by rotating everything belonging to the system considered, including initial conditions. This means that there would be no absolute ("preferred") directions in space (that is in the universe). Similarly, the invariance under translations and time displacements can be said to exclude that there are preferred points in space (a "centre of the universe") or time (a "zero of time").

The displacement invariances considered above may be described in terms of a space-time co-ordinate system used as a reference frame for describing the physical processes. The requirements are, that the physical laws do not change, if we move or rotate the space part of the co-ordinate system, or if we move the origin of the time-axis. The spa-

tial co-ordinate system may consist of three co-ordinates x, y and z, representing directions perpendicular to each other. However, as long as we don't specify what these co-ordinates are, we don't have to define things like "perpendicular" (which could be complicated if the geometry of space is not Euclidian). The space co-ordinates are just a set of variables describing a position in space, just as the time co-ordinate describes a point in time.

So far I have considered co-ordinate systems displaced relative to each other, but fixed. What happens if the co-ordinate systems are moving relative to each other? Should one not go on to require, that the laws of physics are the same in any two co-ordinate systems moving relative to each other? This is called the "principle of general relativity". If it is valid, the same physical law should describe a particle moving relative to the Earth in a co-ordinate system following the rotation of the Earth, and in another co-ordinate system fixed relative to distant stars. Classical physics clearly does not have this property: in the rotating co-ordinate system, additional forces (sometimes denoted "fictitious forces") such as the centrifugal force and the Coriolis force have to be added, whereas in the fixed co-ordinate system they are absent. The centrifugal force is in the direction away from the Earth's axis of rotation, while the Coriolis force tries to deflect the particle away from the equator, whenever the particle is moving relative to the Earth.

Einstein proposed to accept the general principle of relativity and then simply to consider the forces called "fictitious" by Newton as real forces. They might derive from a kind of gravitational interaction between a rotating object and all the distant masses in the universe (stars, galaxies). Just as moving charges experience magnetic field interactions and static ones not, Einstein suggested that the interaction with celestial masses was of importance only in accelerated (such as rotated) frames of reference, and not in fixed or uniformly moving "inertial systems of reference" (Einstein, 1913). This led him to formulate his general theory of relativity.

A weaker form of the principle of relativity is to claim it only for frames of reference moving at constant velocity relative to each other. That is co-ordinate systems, which are not accelerated relative to each other. The transformation from one such co-ordinate system to another is called a "Lorentz transformation". Its form was derived 1905 by Einstein in his "special theory of relativity". If a right-angle co-ordinate system with the x-axis along the direction of the relative velocity, v, between the co-ordinate systems is selected, then the Lorentz transformation between space-time points (x, y, z, t) in one of the co-ordinate systems and the corresponding ones, (x', y', z', t'), in the other, may be written

$$x' = y\,(x - vt)$$

$$y' = y$$

$$z' = z$$

$$t' = \gamma\,(t - vx/c^2)$$

where

$$\gamma = (1-v^2/c^2)^{-1/2}$$

Here c is the velocity of light in vacuum. The requirement that physical laws are the same in any reference frames connected by Lorentz transformations is called "Lorentz invariance". Classical electromagnetic theory, formulated long before Einstein, is already Lorentz invariant (it also contains the velocity of light as a fundamental constant), whereas Newton's theory of mechanics is not.

It is seen that the Lorentz transformation contains both space and time co-ordinates. The transformation of the space co-ordinates depends on time, and the transformation of the time co-ordinate depends on space co-ordinates. It is thus impossible to consider invariance with respect to relative motion in space without also considering the change in the time co-ordinate. If one did that one would - instead of the Lorentz transformation - get

$$x' = x - vt$$

$$y' = y$$

$$z' = z$$

$$t' = t$$

This is called the Galileo transformation. Newton's equations of motion are invariant under Galileo transformations, so the assumption of separate relativity principles for space and time can be said to be underlying classical mechanics (there is a t-dependence in the transformation $x \rightarrow x'$, but as $t=t'$ is unchanged, there is no coupling between space and time co-ordinates). Galilean invariance may be considered as an approximation to Lorentz invariance. An approximation, which is indeed very good for the macroscopic objects of everyday mechanics, such as cannon balls, dinner tables, bricks and so on. It is not a very good approximation for studying electrons from large particle accelerators - but who could blame Newton or Galileo for that! In fact, it is easily seen, that the Lorentz transformation reduces to the Galileo transformation, if the velocity v in the above expressions is small compared to the velocity of light in vacuum, c. If terms containing the ratio v/c are neglected, then $\gamma = 1$ and furthermore the second term in the Lorentz transformation of the time co-ordinate vanishes, which gives precisely the Galileo transformation. Thus the more precise special theory of relativity has not rendered classical mechanics useless. Newton's equations still give a very good approximation to the description of a large class of physical phenomena: those not involving velocities close to that of light in vacuum (3×10^8 m/s). Similarly, the special theory of relativity is a good approximation for space-time transformations between non-accelerated systems of reference, but must be extended to a more general relationship for situations, where the general relativity principle is important, and the treatment of gravitational interaction must in this case fulfil the invariance principle of general relativity (Einstein, 1921).

We presently know that reality is much closer to Lorentz invariance than to Galileo invariance. But we don't know for sure if physical laws must be exactly Lorentz invariant, and even less if they should obey the general relativity invariance. An additional requirement in the latter case would be invariance under rotations. Rotation is one special

kind of accelerated motion, so rotational invariance would imply that the laws of physics should look the same in fixed and rotating co-ordinate systems. To each invariance principle corresponds a conservation law. Translational motion is driven by momentum, and translational invariance corresponds to linear momentum conservation. Rotational motion is driven by angular momentum, and rotational invariance corresponds to angular momentum conservation.

There are many more candidates for invariance principles, one class of such principles is connected with various types of reflections, invariance under inversion in a point in space (for example (x,y,z)=(0,0,0)) corresponds to parity conservation, while invariance under reversal of the direction of time has consequences for the possibility of distinguishing states of elementary particles (Sakurai, 1964). The usefulness of these invariance principles on a microscopic scale is not necessarily reflected in the macroscopic behaviour of objects. In everyday life, the reversal of the direction of time would not seem to involve any simple symmetry. Yet even Newton's equations indeed have the property of not changing, when the sign of the time variable is changed. With proper interchange of initial and final boundary conditions, any dynamical behaviour of a mechanical system could be time-reversed. No one should be able to identify one of the behaviours (forward or backward motion in time) as more correct than the other. The sense of a direction of time in everyday life is then caused by the difference in probability of different states. You drop a glass on the floor and it splinters, but the probability of all the pieces having the exactly reversed velocities - so that they could jump together to form the glass again - is nearly zero!

What about scale transformations? Are the physical laws invariant under transformations of the size of systems considered? Classical mechanics is scale invariant, but quantum mechanics not, precisely because of the quantization of properties such as energy and angular momentum: the system cannot be scaled up "a little bit", it must be scaled up at least to the next higher quantum state.

There are requirements to think of beyond those associated with invariance. One would be the condition, that physical laws always give physical solutions: that there are no singularities (space-time points where a physical property becomes infinite), and that solutions are reasonably bounded (if say a density stretches to arbitrarily large distances, it must decrease in magnitude fast enough, so that the integral of it over space becomes finite). Infinities do exist in classical mechanics and electromagnetism, if point particles and point charges are considered - but that is of course only a convenient approximation to finite, physical systems, or is it? More serious is the energy concept used in electromagnetic theory, because it leads to infinite "self energies" of charged particles, and modifications of the theory to allow a "smearing" of the charge over some finite volume have not succeeded in making the description free of contradictions (Feynman *et al*, 1964). In quantum theory, non-bounded solutions are frequently used, for instance plane waves stretching to infinity. They describe particles with definite momentum. But this is considered just a convenient approximation, because due to Heisenberg's uncertainty relations, there can be no such thing as a particle with no uncertainty in momentum.

Finally going into the contents of physical laws, there would have to be requirements

regarding the nature of forces and other quantities entering into the models used. Should one demand that forces are of limited range, or could they act between regions in space infinitely separated from each other? The Coulomb interaction is not limited in range but nuclear forces conceivably are. If the interaction depends on the values of variables in the vicinity of the point considered, the interaction is said to be local. There could be questions of locality in time as well as in space. Do interactions at a given time depend on the value of state variables at previous or later times? This question cannot be separated from that of space dependence, due to the coupled relationship between space and time co-ordinates (for example in the Lorentz transformation).

Two objects at the same point in space - if that is possible - are likely to interact by forces depending only on the current time (which is also the common time for the objects). But if they are separated, the interaction may take the form of a signal from one object reaching the other object at a later time, and only then causing a reaction. Electromagnetic fields are of this nature. Interaction is not instantaneous, but the field created at one location has to be retarded by the time needed to propagate it to the other object, which is then affected in a manner determined by the retarded field.

Are the forces in nature simple or complicated? Are they very specific or are they fairly arbitrary? This will be discussed further below. The question is how few assumptions we need for formulating a valid model of the physical reality. Maybe only the invariance principles are needed, and not any detailed description of particular forces.

7.2 Physical models and physical laws

The laws of physics are usually expressed in mathematical terms, typically by equations of the form

$A=0;$

$B=0;$

etc.

Here A, B, ... are (often complicated) functions, perhaps involving differential or other operators. Until we know how they should be written, the above equations do not give us any information.

The first problem is how we want to describe our picture of reality (or of some corner of reality). Assume that we want to use a set of functions of space and time co-ordinates to describe the system,

$$(\psi_i(\mathbf{x},t) \mid i = 1,2,\ldots,n),$$

where $\mathbf{x}=(x,y,z)$ is a three-dimensional vector containing the space co-ordinates of a certain location and t is the time. The n functions ψ_i in some way describe the system (say an elementary particle, a windmill, a boiling fluid or the entire universe). It may do so by describing the locations in space and time of each component of the system (each particle - each atom - as in classical physics), or by furnishing the probability amplitude for finding the system in a given state. In quantum mechanics, ψ_i may represent the

probability of finding the i'th particle in the system at location **x** at the time t, or *i* may label the different possible states of the system, and ψ_i would then represent the probability of finding the system in the i'th state. Still another formulation of quantum mechanics (called "double quantization") would have *i* label the possible states, in which the particles making up the system can be, and the probability amplitude ψ_i would then represent the average number of particles being in the i'th state at the time t (and location **x** unless the space variables have been excluded, in which case we get the number of particles in each state, no matter what their locations).

We are interested in formulating a model for the dynamic behaviour of the system. The mathematical version of this mode - one "law of physics" - will be a set of equations allowing the time development of the system to be calculated, once boundary conditions are specified. These boundary conditions may be a complete system description at a given time ("initial conditions"), from which the equations should allow a determination of the state of the system at a later time. For some systems, the initial condition approach is not convenient - either because there is fundamental uncertainty associated with some of the parameters, or because the particular system is not conveniently described in that way. In such cases, other types of boundary conditions can be imposed, such as prescribing the values of some variables at some locations for all times considered (for instance requiring the horizontal wind velocity to be zero at the ground, or asking that the wave pattern inside a box holding electromagnetic radiation be stationary).

The physical model for the system consists in part of general assumptions, e.g. deciding which invariance principles should hold, and in part more specific modelling of the particular system considered: determining its degrees of freedom, how many variables that are needed to describe the system, and which kinds of interaction one should include. The change in each variable ψ_i with time is found by differentiating ψ_i with respect to time†. Let the function resulting from this procedure be G_i:

$$d\psi_i(\mathbf{x},t) / dt = G_i(\psi_1, \psi_2, ..., \psi_n)$$

This equation is of course of the general form given in the beginning of this section, since is can be written

$$d\psi_i / dt - G_i = 0.$$

Here the arguments **x** and t have been left out in ψ_i (as on the right hand side of the equation above), and the arguments ($\psi_i \mid i=1,2,...,n$) have been left out in G_i. However, physical assumptions are already made in writing the time derivative in the above form. While G_i depends on the state variables, it has been assumed not to depend on space and time variables directly. It could depend directly on space variables, if they were chosen as some of the ψ_i's, but time is not a state attribute and hence cannot be a state variable. The lack of appearance of t among the arguments of G_i then means, that the time development of the system will be the same at two different times t_1 and t_2, if eve-

† Some of the mathematical methods used in this book were planned to be introduced in an appendix. This is deferred to later revision work, so for now each time there is a reference to "appendix" refer instead to standard books on mathematics.

rything else including the state variables ψ_i remains unchanged. This, however, is precisely the assumption of time displacement invariance, as discussed in the previous section.

By writing the time derivative of ψ_i, one mathematical assumption has been made: that the state variables are functions that can be differentiated. A certain smoothness is necessary to allow differential calculus to be used. Physically, the condition pertains to the requirement noted in the previous section, that there be no singularities in the physical quantities. In a way, this is a fairly strong assumption, implying that there can be no point particles or infinities in the real world, but only in some of our convenience approximations. One may find this a plausible assumption, but it has severe implications, for example for the interpretation of the early development of the universe (stating that the "big bang theory" can at most have approximate validity - see more on this in chapter 9). If we know the functions G_i, we have in our hand a fundamental law of physics. If we don't know them, we could piece by piece try to construct them from experience and experiments. This is more or less the approach taken in the history of physics. Here I shall use a different approach, suggested by S. Chadha, C. Litwin and H. Nielsen (Nielsen, 1978). They suggest, that as we don't really know G_i, we might as well use arbitrary guesses for these functions. More precisely, we may consider the entire class of smooth functions and then pick at random the n functions (G_i | i=1,2,...,n). Once they are selected, the differential equations for ψ_i are put into a computer and solutions are grinded out. This may be repeated many times with new sets of random G_i functions, and hopefully some regularity will appear in the results, which we can then interpret as common features of all possible physical laws.

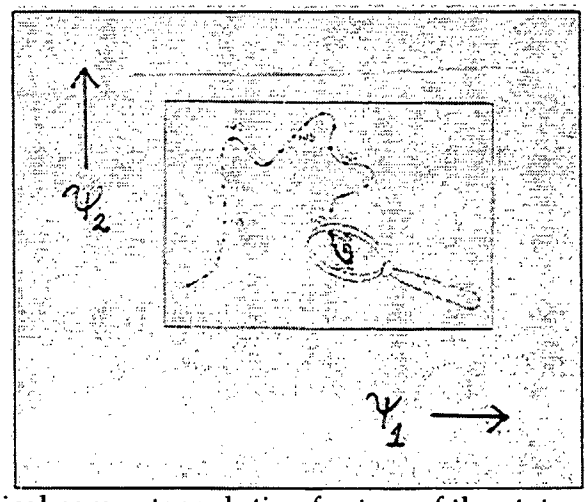


Fig. 7.1. Typical computer solution for two of the state variables ψ_1 and ψ_2, given as the path traversed in time, t, in the ψ_1, ψ_2-plane.

Now it is not so easy to construct functions containing operators and similar complications in an arbitrary way. What Chadha, Litwin and Nielsen did, was to construct G_i from their Fourier expansions (see appendix for definition), and assume random coefficients in these expansions. a typical solution is shown in Fig. 7.1 for two dimensions of state variables, say ψ_1 and ψ_2. Each point on the time path corresponds to connected values of the variables (ψ_1, ψ_2), which can be read off the coordinate axes by drawing

two lines from the point and perpendicularly onto the axes (that is, you can't read the value because there are no scale or units indicated!).

The interesting part of the time path is shown in the looking glass. The solutions start to spiral around a point in the ψ_1, ψ_2-plane, Let us call that point (ψ_1^0, ψ_2^0). It turned out that most of the computer solutions for arbitrary G_i's had this property: waiting sufficiently long time, the solution got closer and closer to a single value of the set of state variable (not just the i=1 and 2 components, but all n variables (ψ_i). The solutions did not move head on into the convergence point, but always spiralled around it†.

Close to the convergence point, the solutions can be written in a simpler form, and due to the spiralling behaviour, this simpler form is valid from the time, when the solutions get sufficiently close to the convergence point, and it stays valid forever after. This implies, that if we can consider our universe as old enough, we need only consider the simpler version of the physical laws deduced from the model considered here. But since we nave no way as yet of connecting the time scale of the model calculation with the physical time, it is really one additional assumption, we are making in taking the present age of the universe as being in the spiralling phase of the random function $d\psi_i/dt$ model.

In order to determine the behaviour of the ψ_i's in the spiralling phase, we may use a Taylor expansion around the convergence point (ψ_i^0 | i=1,2,...,n) (see appendix for definition of Taylor series),

$$G_i(\psi_k / k=1,2,...,n) = G_i(\psi_k^0 / k=1,2,...,n)$$
$$+ \sum_{j=1,2,...,n} \frac{\partial}{\partial \psi_j} G_i(\psi_k / k=1,2,...,n) \Big|_{(\psi_k=\psi_k^0 | k=1,2...,n)} (\psi_k - \psi_k^0)$$
$$+ \text{terms with } (\psi_k - \psi_k^0)^2 \text{ etc.}$$

The Taylor expansion gives the value of the function G_i for its argument equal to the set of variables ψ as a power series in $\psi - \psi^0$, with coefficients given in terms of the derivatives of G_i taken in the convergence point ψ^0 (this is what it means, when after the derivative $\partial G_i/\partial\psi$I write a vertical line and at its lower part note that the ψs should be set equal to the corresponding ψ^0's, after the derivative has been evaluated). I have used the summation sign, Σ, to avoid writing everything n times. This works as follows:

$$\sum_{i=1,2,...,n} X_i = X_1 + X_2 + \cdots + X_n$$

The particular kind of derivative $\partial/\partial x$ used here is called a partial derivative. It only works on explicitly occurring variables (here x) in the function it operates on, in contrast to the full differential quotient d/dx, which also works on any other variable (say y) in the function it acts on, provided that the variable indirectly depends on x. That is, $\partial/\partial x$ only works on the first argument in a function f(x,y(x)), while d/dx works on both arguments of f (see appendix).

---

† There has been some doubts, whether this behaviour was perhaps an artefact due to the use of sine functions truncated at some high but finite Fourier order, but for the sake of argument we shall assume

Now back to the Taylor expansion. The assumption that the universe is old enough for the solutions to be close to the convergence point can now be stated more precisely: we assume that the spiralling has brought us so close to the convergence point, that all terms containing $(\psi_i - \psi_i^0)^2$ (any "i") and higher than second power of $(\psi_i - \psi_i^0)$ can be neglected. In other words, we include only the first order terms - proportional to $(\psi_i - \psi_i^0)$ - in the Taylor expansion, an approximation that becomes better and better, the older the universe gets. The coefficient multiplying each $(\psi_i - \psi_i^0)$ is a partial derivative of G taken in the convergence point. Once the ψ_i^0's have been inserted, this is just a number, since it only depends on the two indices i and j we shall call it A with suffices i,j

$$A_{ij} = \frac{\partial}{\partial \psi_j} G_i (\psi_k \mid k = 1,2,\ldots,n)\Big|_{(\psi_k = \psi_k^0 \mid k=1,2,\ldots,n)}$$

The value of G_i in the convergence point I shall denote G_i^0, Then the original equation for the time development of the state variables may be expressed in the following way:

$$d\psi_i / dt = G_i^0 + \sum_{j=1,2,\ldots,n} A_{ij} (\psi_j - \psi_j^0) = B_i + \sum_{j=1,2,\ldots,n} A_{ij} \psi_j$$

In the second rewriting, the sum over ψ_i^0's has been lumped together with G_i^0 (both are just numbers) in B_i,

$$B_i = G_i^0 + \sum_{j=1,2,\ldots,n} A_{ij} \psi_j^0$$

It is no problem to redefine the state variables ψ_i in such a way, that the constant term B_i disappears from the equation for $d\psi_i / dt$ (say add $-B_i/A_{ii}$ to one ψ_i). Then we get what is called a homogeneous differential equation for our system:

(*) $$d\psi_i (\mathbf{x}, t) / dt = \sum_{j=1,2,\ldots,n} A_{ij} \psi_j (\mathbf{x}, t)$$

or rather a set of n such coupled differential equations, for i=1,2,...,n. They are called "coupled", because the time development of the i'th variable depends on an the other ψ-variables, not just on ψ_i.

The equation (*) is now our law of physics. But we still need to find out more about the quantities A_{ij}. Let me look at the simple case of only one variable $\psi_i = \psi$,

$$d\psi / dt = A\psi.$$

The solutions to this equation are exponential functions of time,

$$\psi(t) = a \exp(At),$$

where a is a constant or a function of the space co-ordinates. Now let me look at the possible values of A. A may be a real number, positive or negative, but it may also be a complex number (see appendix), if A is positive and real, $\psi(t)$ will grow with time, in contradiction to our assumption that the solution for late times should stay close to a constant value ψ. If A is negative and real, $\psi(t)$ goes to zero with time, and in a late uni-

---

that the behaviour is not in any fundamental way due to numerical uncertainty.

verse it will have practically died out. Thus the only possibility of obtaining finite solutions is if A is purely imaginary. We can then write A as $i=\sqrt{-1}$ times a purely real quantity, which I choose to write as $-H/\hbar$. This makes no difference, as $\hbar$ is just a constant compensated for by the magnitude of H. I select the value $\hbar = 6.63\times10^{-34}/(2\pi)$, which is Planck's constant divided by 2π. My differential equation for ψ now reads

$$i\hbar\, d\psi/dt = H\,\psi$$

or in the general case

$$(**) \qquad i\hbar\, d\psi_j/dt = \sum_k H_{jk}\,\psi_k$$

Here the indices j and k have been used rather than i, to avoid confusion with the imaginary unit $i=\sqrt{-1}$, and the possible values of k in the summation has not been specified (there could be a finite number n, or even an infinite number of state variables). The equation (**) is a Schrödinger equation, and we have arrived at this mathematical form of the fundamental law of physics with very few assumptions. However, it only gives the structure of the equation, The physical content is in the function H, and a physical model is required for determining H. For example, in quantum mechanics H is an energy function (called the Hamiltonian - more on this in Chapter 8). I clearly had this in mind when choosing the function name, but also other physical theories, including classical mechanics, can be written in one of the forms (*) or (**), as I will show later.

PROBLEMS AND DISCUSSION ISSUES

PROBLEM 7.1. Looking at yourself in a mirror, right and left are interchanged. Why not up and down?

PROBLEM 7.2. A train is approaching a pedestrian. The train sounds a whistle emitting sound at the frequency 440 Hz (cycles per second). The pedestrian hears the sound as a C (frequency 512 Hz). At what speed is the train going?

DISCUSSION ISSUE 7.3. Do you think the true laws of nature are simple?

DISCUSSION ISSUE 7.4. Do you think we will ever find the true laws of physics, or will they continue to become revised as more knowledge is accumulated?

DISCUSSION ISSUE 7.5. Does God play dice?

PROBLEM 7.6. Must one use relativity theory to calculate the motion of electrons in atoms? Give reasons! (textbook version of this problem is given below).

Textbook version of problem 7.6:

Consider a hydrogen atom, consisting of a proton of mass M and charge +e, and an electron of mass m and charge -e, located at the distance r from the centre of the proton.

A. How large is the electric force between the electron and the proton?

B. How large is the gravitational force between the electron and the proton?

C. Use tables of relevant constants to assure yourself, that the gravitational force is insignificant compared with the electric force.

D. Assume that the electron is moving in a uniform circular orbit with speed v. How large is its acceleration?

E. Find v expressed in terms of e, m, r and the constant $k=9x10^9$ (SI units) in Coulomb's law (force = $-k(e/r)^2$)

F. Determine the ratio between v and the speed of light, c (use tables of physical constants)

G. Is it necessary to use relativistic mechanics for calculating the motion of the electron in the hydrogen atom?

Chapter 8

The structure of matter

A picture of matter, which has very profound roots in the history of natural philosophy, is that of atomistic particles serving as "building stones" for all kinds of matter. Matter is made of atoms, and atoms should ideally be elementary, indivisible particles - either all identical, or at least of a few, fundamental kinds.

When the objects, that we today call atoms, were thought to be the smallest building bricks of matter, the variety of chemical and materials properties made it an untenable position to claim, that atoms were identical or simple. To understand chemistry and materials, the inner structure of atoms had to be investigated. Once realising, that atoms consisted of a nucleus containing most of the mass, plus a number of electrons, all interacting through the Coulomb force (because of their electric charges), there was no longer any mystery. The number of electrons in an atom determined the chemical properties and the relative positions of atoms determined the material structure. Other physical properties such as conductivity of heat or electric current became simple consequences of the electron and lattice structure. The theory needed to describe these phenomena accurately was quantum mechanics.

Quantum mechanics is a theory for non-relativistic particles (that is particles moving at speeds well below that of light), with energies in the range of electron volts to mega-electron volts (1 MeV = one million eV = 1.6×10^{-13} J) .

Each particle (electron in the atomic case) is described by a probability amplitude, that is a quantity which has to be squared to get the probability for finding the particle fit a given location for a given time. The smearing of the particle density over an extended region in space is necessary in order to obtain a quantum behaviour: the particles can only be in discrete energy states. All the energies between the allowed ones do not correspond to stable atomic configurations. This is the quantization of energy identified by Niels Bohr as being needed for explaining the observed emission of light during rearrangement of electrons in atoms. To reproduce this quantization of energy, the notion of describing a particle by one position in space as well as by its precise state (energy, momentum, and so on) had to be given up. Instead, the probabilistic theory of quantum mechanics was offered by Erwin Schrödinger and Werner Heisenberg. If a particle is forced to be at a given location, information on its state of motion will be lost, and vice versa. The product of the uncertainties in one momentum component (say along the x-axis) and the corresponding position (x) must always exceed the fundamental constant of Planck ($h = 6.63\times10^{-34}$ Js). This is called "Heisenberg's uncertainty relation".

The atomic nucleus was found to consist of protons and neutrons, which together with electrons became considered as elementary particles. However, just as the chemical forces were incomprehensible as long as the atoms were regarded as fundamental particles, the nuclear forces (forces between nucleons, nucleons being the common name for protons and neutrons) could not be understood until the internal structure of nucleons came under investigation.

A picture of the inner structure of protons and similar particles has been developed during the period 1960-80. It uses an analogy to the model for the interaction between electrons and light, which was formulated by Dirac around 1930 and perfected by Schwinger, Tomonaga and Feynman in the late 1940ies (using a renormatisation procedure to avoid infinite electron self-energies). The electrons are described by a quantum mechanical wave function, and the light is described by an electromagnetic field given by the classical theory of electromagnetism (Maxwell). The combination of the two is called "quantum electrodynamics", and it views the possible processes in terms of an elementary coupling, which is illustrated in Fig. 8.1. The elementary coupling is a vertex involving three particles. In Fig. 8.1, the electromagnetic field is represented by a light quantum (also called a "photon"), which is illustrated as a particle, b, which can be created or annihilated by interaction with the electron or its antiparticle (the positron). If the electron (positron) existed before the interaction, it is changed (scattered) by the process, because it gains energy from or loses energy to the photon. This is illustrated in Fig. 8.1 (a) and (b), where the electron (positron) is denoted "f" before the interaction, and "f'" after. In Fig. 8.1 (c) and (d), a particle-antiparticle (electron-positron) pair denoted "f" and "¬f'" is created or destroyed by the interaction with the light quantum. The transformation of the photon into a particle and an antiparticle is called "pair production". Neither this or the opposite annihilation process is possible in classical electrodynamics. These processes satisfy Einstein's equivalence principle for mass and energy - mass times the square of the velocity of light being equal to energy - in that the photon energy is always equal to or larger than the rest mass of the electron-positron pair. If there is a surplus, the particles can have a non-zero velocity.

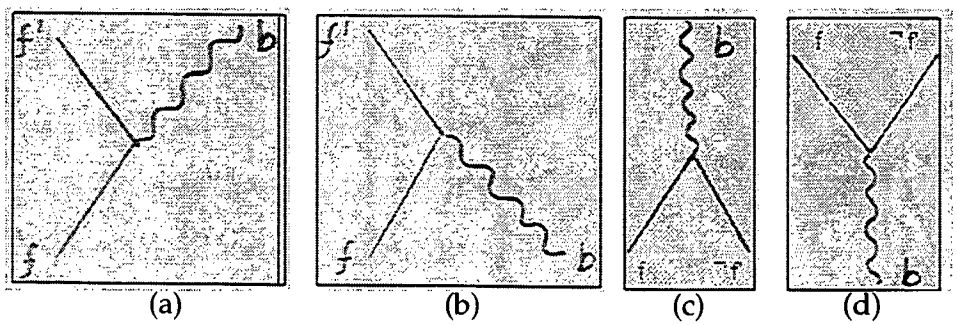


Fig. 8.1. Feynman diagram for elementary interactions between fermions (f) and bosons (b). The f-particles may be electrons or quarks, while the b-particles, which correspondingly would be photons and gluons, are usually called "fields". There is an implicit time axis going upwards in all the pictures.

The interaction between two electrons is now given by a diagram of the kind shown in Fig. 8.2, or a more complex diagram formed from the fundamental interactions of Fig. 8.1 (some examples of such "higher order" diagrams ore shown in Fig. 8.3). The electrons are seen to interact with each other by "exchanging" (emitting and receiving) photons.

It is in analogy to this, thpt the interactions between the protons and neutrons in nuclei were in the 1930ies suggested (by H. Yukawa) to involve the exchange of a pi-meson

(pion). There are three pi-mesons, of which the neutral one would account for interactions between like particles (proton-proton or neutron-neutron), while the pions with plus or minus one electron charge (the opposite of the proton charge) would be responsible for proton-neutron interactions.

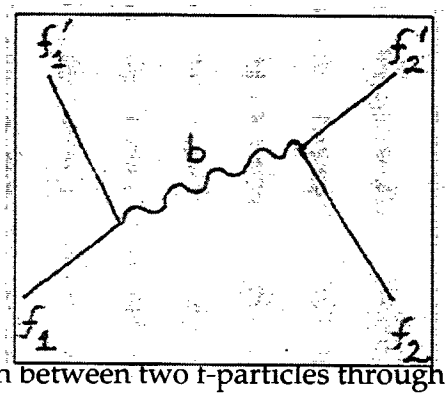


Fig. 8.2. Interaction between two f-particles through the exchange of a virtual b-particle.

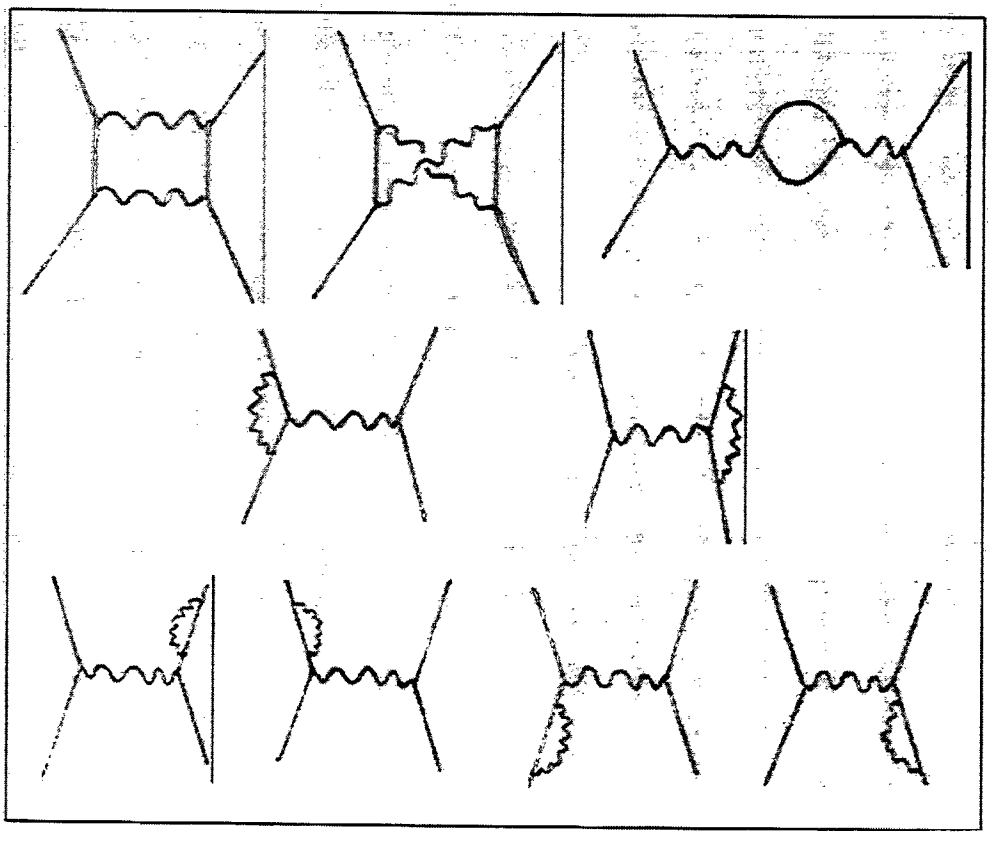


Fig. 8.3. Higher order contributions to the b-field interaction between two f-particles (Feynman, 1949).

This theory ran into trouble in about 1970, when bombardment of protons with energetic electrons revealed something hard inside the proton (Jacob and Landshoff, 1980). Proton-proton scattering had earlier shown similar "hard-core" effects, but this was no problem,

as it could be interpreted as part of the pion exchange interaction. Not so with the electron-proton scattering experiments, because electrons were not supposed to interact with protons by pion exchange, that is by so-called "strong interactions", but only by weak and electromagnetic interactions ("weak" interactions being the name for the forces active in processes such as radioactive beta-decay in nuclei, where a neutron transforms into a proton plus an expelled electron). Murray Gell-Mann proposed that protons were composed of three other particles, which he called "quarks", and so were neutrons and other similar particles, while pions and other "mesons" were composed of just two quarks. The assumption that electrons in the scattering experiment hit three separate, charged particles rather than one, could explain the observations.

*Table 8.1. Elementary particles according to vintage 1983 models**

Particle	symbol	mass (MeV**)	Charge (e)	spin ($\hbar$)
electron	e^-	0.51100	−1	½
e-neutrino	ν	0	0	½
muon***	μ	105.6595	−1	½
μ-neutrino	ν	0	0	½
tau***	τ^-	~1800	−1	½
τ-neutrino	ν	0	0	½
u-quark	u_1, u_2, u_3 (#)	~400	2/3	½
d-quark	d_1, d_2, d_3	~400	−1/3	½
s-quark	s_1, s_2, s_3	~500	−1/3	½
c-quark	c_1, c_2, c_3	~1550	2/3	½
b-quark	b_1, b_2, b_3	~4700	−1/3	½
t-quark	t_1, t_2, t_3	?	2/3	½
photon	γ	0	0	1
intermediate vector bosons	W^-/W^+	>50000	−1/1	1
intermediate vector boson (neutral)	Z^0	>50000	0	1
gluon	$g_1,\ldots,g_8$	0 ?	0	1
Higgs particle	H^0	large	0	0
graviton	G	0	0	2
gravitino##	?	large ?	0	3/2

* to each spin-½ particle p, there is an anti-particle $\bar{p}$ with the same mass and spin, but opposite charge.

**1 MeV (million electron volts) equals 1.6×10^{-13} J.

*** finite lifetime,.

color index.

if this exists, some of the ones above don't!

The interaction between two protons are therefore believed to be the combined picture of more complex processes, the elementary steps of which are quark-quark interactions.

These are described by introducing a new field in analogy to the electromagnetic field. This field is one that may be described by the pictures in Figs. 8.1 to 8.3, with quarks as the f-particles and the new field particle, called the gluon, as the b-particle. Some properties of these new, "elementary" particles are listed in Table 8. 1.

The exchanged particle in the picture shown, for instance, in Fig. 8.2, is called a "virtual" particle. What this means is that it can have more energy, a different charge, and different values for other properties usually obeying conservation laws, as compared with the particles from which it originates and the particles formed by the annihilation of the virtual particle. This "quantum fluctuation" property of virtual particles is analogous to quantum tunnelling, the fact that in quantum mechanics, a particle has a certain finite chance of penetrating a barrier higher than its total energy (for example in spontaneous alpha-decay or fission of heavy nuclei). The theory excludes that the particle should be observed in the middle of the wall-penetrating process. Similarly, virtual particles, such as the photons in the electron-electron interaction or the gluons participating in the quark-quark interactions, cannot be seen in any experiment. This does not exclude that real photons may be seen (in solar radiation, for example), or that real gluons might some day be seen (they almost have!), but in that case they must obey standard rules such as the conservation of energy, charge, angular momentum etc.

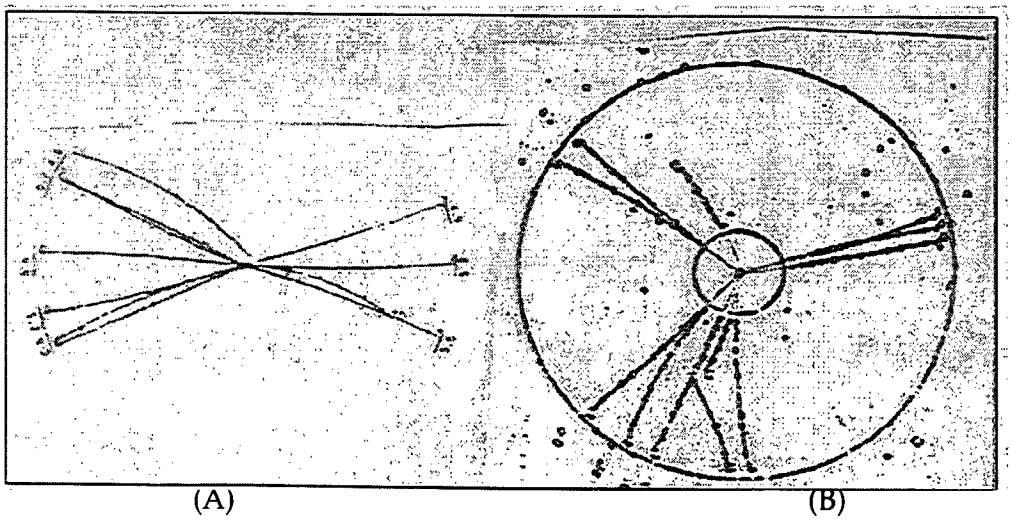
Table 8.2. Selected composite hadrons (particles made from quarks), according to vintage 1983 models

Particle	structure	mass×c^2 (MeV)	Charge (e)	spin ($\hbar$)
proton, p	uud	938.280	1	½
neutron, n	udd	939.573	0	½
ornega, Ω^-	sss	1672	–1	3/2
pion, π^+	u$\bar{}$d	139.567	1	0
phi, φ	s$\bar{}$s	1020	0	1
J or psi, J/ψ	c$\bar{}$c	3100	0	1
upsilon, Υ	b$\bar{}$b	9400	0	1

Free quarks have never been observed. They would be very conspicuous, due to their charge being one or two thirds of what is otherwise regarded as the fundamental quantum of electric charge (Table 8.1). The evidence for the existence of quarks is that the quark model predicted a number of new particles, which subsequently were found experimentally. For example, the 3-quark states ("baryons") that can be formed from the quarks of lowest energy, the u, d and s, could all be associated with known particles, except one – the sss-state with energy about 1500 MeV (3 × 500 MeV). Soon after, a particle with the predicted properties was indeed found, the omega-minus of rest mass 1672 MeV (see Table 8.2 - the mass is here given in energy units, implying that it is really the mass times the square of the velocity of light). Two-quark states (mesons) involving the u, d and s quarks could also be interpreted in terms of known particles, and among the simple quark-antiquark states based on one of the 3 heaviest quarks, two candidates have been observed for the first time during the 1970ies: the J- or phi-particle (with anticipated

quark structure $c\bar{c}$) and the upsilon-particle ($b\bar{b}$).

Evidence for the gluons is still more indirect. Since quarks are charged, it should be possible to produce them from sufficiently energetic photons by the process shown in Fig. 8.1(d), and possibly they would subsequently emit gluons by the process of Fig. 8.1(a). This would happen, if the quarks are accelerated, say by moving along a spiralling path in an electromagnetic field. Such experiments have been performed, obtaining the energetic photons by electron-positron annihilation (Fig. 8.1, process (c)) and looking at particles emerging from the magnet-surrounded collision area by a range of detectors suited for the detection of all charged, real particles emerging from the area. Two different types of results are shown in Fig. 8.4. They are called "two-jet" events and "three-jet" events. Their interpretation in the quark-gluon model is shown in Fig. 8.5. To the extent that no other interpretations could provide similar consistency, it may be said, that the observation of 3-jet events is indirect evidence for the existence of the gluon (Sutton, 1980). The pictures in Fig. 8.4 should be interpreted in the following way: The electron and the positron enter from opposite sides and collide in the centre of the picture (as entering particles, they are not triggering the detector and hence are not seen in the picture). The two quarks are not seen, because the forces holding them together grow rapidly with distance (this is a property of the gluon exchange force, Fig. 8.2), thus confining them to a very small volume - say about the size of the particles they can form. Therefore, what is seen in the picture (Fig. 8.4(a)) is only the baryons and - mesons formed from the initially created quarks by subsequent processes. Each quark gives rise to a bunch of particles, the two quarks thus produce two bundles ("jets"). In Fig. 8.5(b), one gluon escapes the primary interaction region and gives rise to an independent bunch of baryons and mesons (each bunch in Fig. 8.5 forms just one particle, for simplicity, but any number of quark-antiquark pairs consistent with the energy available can be formed within each interaction region). Thus the observation of 3-jet events may be taken as evidence that at least one gluon was formed.



(A) (B)

Fig. 8.4. Computer reconstruction (based on detector data) of two events recorded at 15000 + 15000 MeV positron-electron collisions at the TASSO experiment at the German electron-synchrotron laboratory (DESY) in Hamburg. (A) two-jet event. (B) three-jet event.

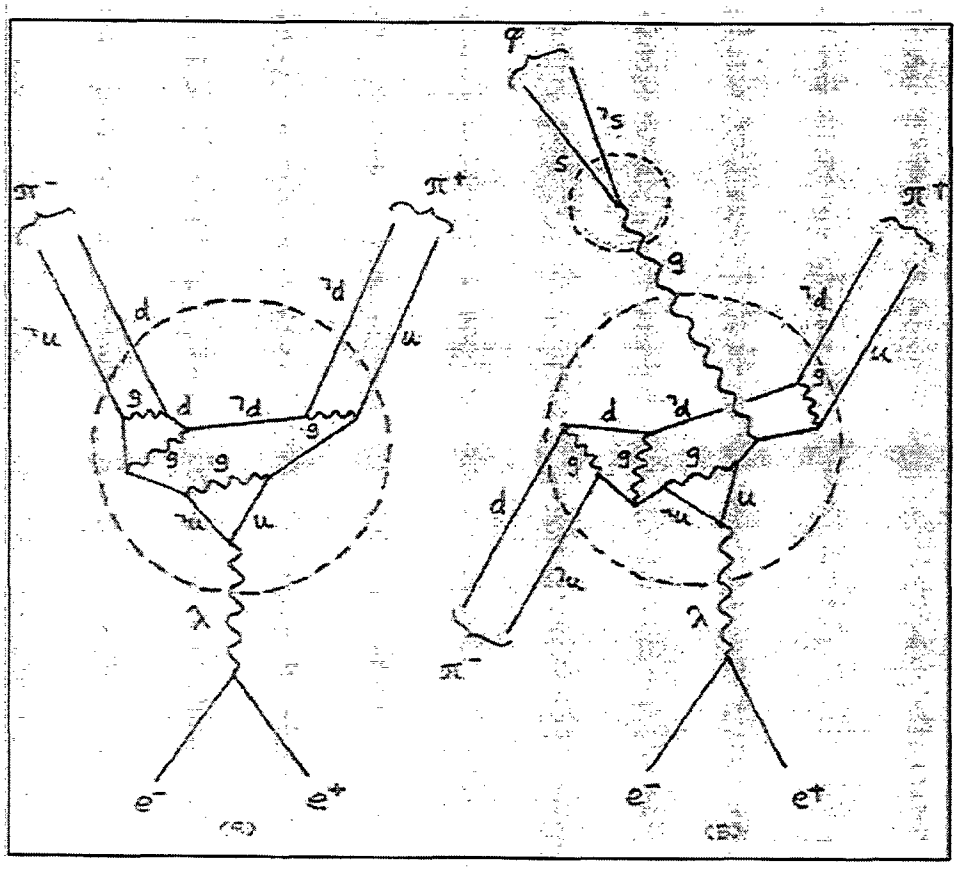


Fig. 8.5. Theoretical interpretation of the 2- and 3-jet events (A) and (B) of Fig. 8.4. The electron-positron pair is annihilated into a photon λ, and it subsequently forms a quark-antiquark pair. However, due to gluon interactions, the quarks cannot escape the region in which they were formed, until they have picked up other quarks to form integer charge entities. The gluons may escape (in B) and then form a third jet of particles.

So far I have discussed two forces of nature: the electromagnetic force and the "strong interactions" (those involving gluon fields). The strong interactions have a short range - not an infinite range such as the Coulomb force - but at short distances (say up to 10^{-15} metres), the strong forces are typically about 100 times larger than the Coulomb force. There are two additional forces we know of in nature. One is called "the weak interaction". Is is of even shorter range (10^{-18} metres) and five orders of magnitude (10^5) weaker than the strong interaction. Finally, there is the gravitational force, supposed to act between any two particles. It is of long range, proportional to the product of the masses of the two particles and the inverse square of their relative distance, and in contrast to all the other ones always attractive when acting between two like particles. It is 39 orders of magnitudes weaker than the strong interactions.

The weak interaction may be described by a field interaction of the type given in Fig. 8.2. There are three field particles, because the charge transfer can be +1, 0 or –1 e-units. They are called "intermediate vector bosons" and are denoted W or Z with the appropriate charge index. None of them had been observed, when the theoretical model was formu-

lated - in fact it was formulated before the quark-gluon model (work by many scientists contributed, including S. Glashow, C. Yang, R. Mills, J. Goldstone, P. Higgs, S. Weinberg and A. Salam). Recently, the vector bosons seem to have been found by proton-antiproton collision at 270000 + 270000 MeV at the European accelerator centre CERN at Genève. The processes involved are illustrated in Fig. 8.6, in terms of the quark and the antiquark inside the colliding proton and antiproton, which take active part in the weak interaction.

The theoretical development achieved by the people mentioned above made it possible to combine the theory of electromagnetic and that of weak interactions in a consistent way. The similarity between this theory and that of the strong interactions makes one wonder, if they could not all be combined. In fact, the fourth interaction, the gravitational one, can also be formulated in terms of an exchanged particle (called the graviton, see Table 8.1). So maybe all physical theories can be unified into one. In order to do this, symmetry relations between particles with different spin must be established, so that they may be transformed among each other.

Spin is an internal property of particles, which has the dimension of angular momentum (length times velocity). As seen in Tables 8.1 and 8.2, the spins of known particles are either integer or half-integer multiples of a fundamental unit. The half-integer spin particles have properties widely different from those of the integer spin particles, because the former obey the Pauli-principle and the latter do not.

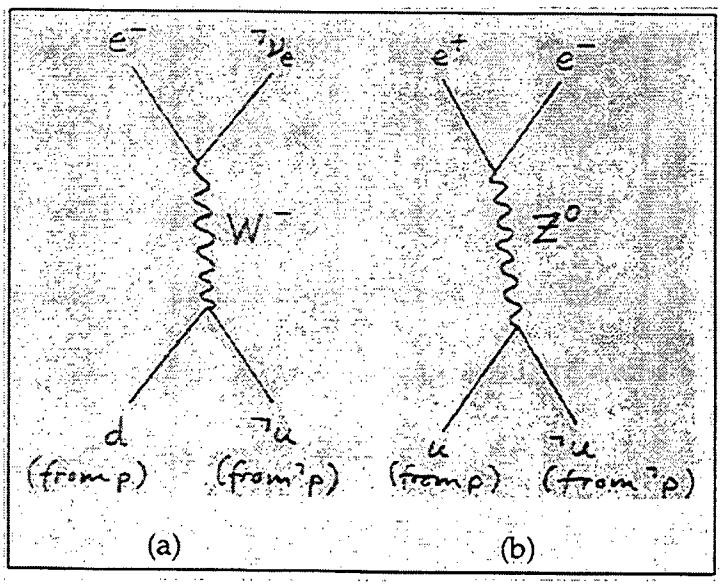


Fig 8.6. Diagrammatic model of intermediate vector boson formation and decay, following (p,p̄)-collisions.

The Pauli-principle pertains to identical particles. It states that in each quantum state, there can be at most one of the identical particles, if their spin is half-integer. The half-integer spin particles are called "fermions". The integer spin particles (called "bosons" - these notations were already introduced in Figs. 8.1 and 8.2) are not restricted. There can be as many identical bosons as one likes in each state. One example offered by light (photons, having spin one) is laser technology - in a laser, light is intensified by superposition of many light quanta in exactly coherent wave-forms (without phase displacement), that is in identical states.

At this stage it is evident, that proposing a super-symmetric theory attempting to put fermions and bosons in the same category is a very daring proposal. A super-symmetry transformation is one that transforms a boson into an adjacent-spin fermion or vice versa, for example, it transforms a spin-one boson into either a spin-1/2 or a spin-3/2 fermion. The remarkable feature of a transformation of this kind is, that if one applies two such transformations, say to go from boson to a fermion and back to the original boson, then the boson will reappear in a different point in space and not at its original position. Thus super-symmetry transformations can be used to move objects, and knowing that it is perhaps not so surprising that they include the gravitational forces in a natural way. The unified theory requires the existence not only of the (as yet unobserved) spin-2 graviton, but also of a spin-3/2 elementary particle, which may be called the "gravitino". On the other hand, other "elementary" particles appearing in the disjoint theories of the different forces are absent (or composite entities). This includes the muon and the intermediate vector bosons*.

It is rather clear, that there are still many problems with most of the above theories. For example, most field theories lead to infinities of one kind or another. The most accepted theories used to be those, for which it was been possible to circumvent the infinities, by rearranging sums so that the contributions that would otherwise sum up to infinity were cancelled by other terms of opposite sign. Furthermore, it turned out useful to attributing a structure to vacuum, by a method called "renormalisation". It consists in assuming a vacuum (that is the state with no particles) with an infinite number of particle-antiparticle pairs, which alter the motion of any real particle as if it had a different mass (the "effective mass"). In this way infinite contributions to the self-energy (= bare mass) can be cancelled out, so that the effective mass - which is interpreted as the particle mass physically observed - becomes finite. Most field theories initially give a zero mass to all the photon-like quanta responsible for the exchange forces. This was believed to be a major problem until the Higgs mechanism was developed. It allows some of the particles to be given a mass and is notably used to give large masses (beyond the capability of current accelerators) to those particles not yet found. Theories such as the super-symmetric ones seem to be having the opposite problem: to predict fewer particles than those already believed to exist.

A new family of unified theories are the superstring theories, which in 1984 were demonstrated to be able to avoid the problem of infinities marring earlier attempts. An account of the first years of development of such theories may be found in my broader audience book on the subject (Sørensen, 1986).

The fact that the proton is not an elementary particle according the combined theory of weak and strong interactions makes it important to look for its possible decay (processes such as splitting the proton into a positron plus a neutral pion). No such decays have been observed, and the experiments presently put a lower limit to the proton lifetime of

---

* If the reader wants to know more about the theories discussed, and how they derive from imposing symmetry requirements locally, there are excellent reviews by Freedman and van Nieuwenhuizen (1978) and by 't Hooft (1980).

close to 10^{30} years, this is much more than the estimated age of the universe (believed to be about 10^{10} years), so don't worry. More accurate measurements of the proton lifetime are required for dismissing the above theories, because they can be stretched to predict proton lifetime up to around 10^{33} years.

There is no way to tell, if the future of elementary particle theories will be a repetition of the past. Will the next generation of physicists be concerned with the inner structure of the quarks? Or what? There would seem to be a valid point in looking at models making different assumptions on what is important and what is less important in our present stock of knowledge. The present theories consider local symmetry as a requirement of overriding importance, even if it requires postulating the existence of particles with fractional charges, that no one has seen. What if the fundamental indivisibility of the electron charge was taken as the fundamental postulate. Could other types of models, other theories and laws, be developed, which would be equally or more consistent (that is consistent with our interpretation of the observations we make)?

Present theories of matter are all field theories, in that the interactions are field quantities, which become quantized into exchange particles such as the photon, the gluon, the intermediate vector bosons and the graviton. However, the particles constituting matter itself (quarks, electrons and so on) are treated strictly as particles. It is true that they are given by a quantum mechanical field quantity, the wave function. This does not, however, imply a deviation from a particle picture (such as the one I have implied in the figures 8.1 and 8.2), because each wave function corresponds to a precise particle number. The use of a probability amplitude as function of (x, y, z) does not mean, that one could expect to find one seventh of a quark in some part of space, where the probability density was then a seventh. If one rapidly placed a screen around that region in space, and then measured how many particles one had captured, one would find zero, or one, or two, etc. - never a seventh. This is the way quantum mechanics works according to the interpretation of Niels Bohr (his "complementarity principle" and the so-called "Copenhagen interpretation" of quantum mechanics, recently reinforced by the outcome of the Aspect experiments (Aspect, 1982)). So there are two kinds of fields in the current theories: the field quantities describing indivisible particles, and then those describing true fields, of behaviour similar to Maxwell's fields. The massless fields associated with electromagnetic and gravitational interactions are of this kind. If a region of space containing an electromagnetic field is suitably divided, one could have one region containing one seventh of the original field, and another containing the rest. Electromagnetic radiation may be described in terms of quantized photons, that is by a particle type of model, but this does not lend light all the properties expected for say an electron. At the heart of this difference is the distinction between bosons and fermions, bosons being inherently "field-like". The further distinction between massless and massive particles may lead to speculations that all truly fundamental particles are without rest mass, and that mass is only quantum fluctuations of something not described explicitly in the model (see for example Sørensen, 1986).

8.1. Quantum mechanics

Quantum mechanics is a model developed during the 1920ies for the description of the motion of electrons, notably in atoms. It deals with the then established duality of electron behaviour, sometimes looking like that of a point particle and sometimes producing interference patterns like waves, by assigning a probability density to each electron. This probability function, ρ, is a function of space and time co-ordinates, giving the probability of finding the particle (here the electron) at a given place at a given time. For fixed time, t, the integral of ρ over all space must be unity, since it is the probability of finding the particle at any place,

$$1 = \int \rho d\mathbf{x} = \int \rho(\mathbf{x}, t)\, dx_1 dx_2 dx_3 = \int \rho d^3x$$

Here the integral over space co-ordinates $\mathbf{x} = (x_1, x_2, x_3) = (x, y, z)$ is a three-dimensional integral (a few different notations are given, see also appendix).

It would have been simple, if the development of ρ with time could be written as a differential equation in ρ itself. However, the wave-like interference patterns observed in electron experiments require, that the basic function describing the particle is a kind of amplitude, that can be added to or subtracted from other amplitudes. In other words, the function we are looking for must be like the amplitude of a sine wave, which can combine with other amplitudes in constructive or destructive interference. In the wave case, the intensity is proportional to the square of the amplitude, so it is proposed in quantum mechanics to describe particles by a probability amplitude (called a "wave function"), the square of which gives the probability density written above. We use the symbol ψ for the wave function and demand that

$$|\psi(\mathbf{x}, t)|^2 = \rho(\mathbf{x}, t).$$

The absolute value should be taken before squaring, because then the expression is valid no matter whether ψ is a real or a complex function. If it is complex, the absolute value squared may be written

$$\psi^*\psi = |\psi|^2,$$

where ψ can be expressed through the real and imaginary parts,

$$\psi = \mathrm{Re}(\psi) + i\,\mathrm{Im}(\psi) \quad \text{and} \quad \psi^* = \mathrm{Re}(\psi) - i\,\mathrm{Im}(\psi)$$

As the functions Re and Im extract the real and imaginary parts of a complex number, respectively, it follows from that the probability density relation, that the wave function for a particle must satisfy the following normalisation condition:

$$\int |\psi(\mathbf{x}, t)|^2 \, d^3x = 1$$

If we now look at the expected value of some physical quantity A, such as the momentum or energy of the system considered (so far just a particle), then the usual rules for probability would suggest that the average value of such a quantity could be obtained by multiplying it with the probability density and integrating over space. The average value of A (called the "expectation value") would then be $\int \rho A\, d^3x$, which may also be written

$$\langle A \rangle = \int \psi^* A \psi \, d^3x$$

In fact, this is how it should be written in quantum mechanics, because A may be an operator, and then it is not the same, if it stands before or after the two ψ-functions, or in between. Let me consider below the simple example, where A is the momentum p, a quantity which in classical physics is mass times velocity, $p_i = m \, dx_i / dt$, for the three co-ordinate directions, i = 1, 2, 3.

Experiments show, that electrons scattered from crystals behave as sine waves with the wave number being $k = \hbar p$ (or $k_i = \hbar p_i$ in an arbitrarily oriented co-ordinate system). This means that p could be introduced in the sine functions $\sin(kx - \omega t)$ (or $\sin(\mathbf{k} \cdot \mathbf{x} - \omega t)$, where $\mathbf{k} \cdot \mathbf{x} = \sum k_i x_i$, summed from i = 1 to 3). Wave functions of this type will give the correct average value in the formula above, only if p is expressed by the following operator:

$$p_i = -i\hbar \, \partial / \partial x_i, \quad \text{for the index } i = 1, 2, 3.$$

The parameter ω multiplying t in the sine wave is related to energy by Planck's relation (which was originally derived for heat radiation and here postulated to be valid for the wave functions appearing in electron scattering experiments),

$$E = h\nu = \hbar\omega, \quad \text{where } \nu = \omega / 2\pi$$

In analogy to the argument for the momentum p, the formula for the expectation value now lets us identify the energy with a certain operator:

$$E = i \hbar \, \partial / \partial t$$

This more or less completes the physical part of the model of quantum mechanics according to Schrödinger. Forgetting about the sine wave arguments, the expressions above may be postulated to be valid in any system, which need to be treated in quantum mechanics. It is clear, that the postulates made above have something to do with quanta, because they introduce Planck's constant $\hbar$ as a fundamental unit in measuring energy and momentum. This is just as the appearance of the velocity of light in electrodynamics and in relativity theory tells us that this constant plays a special role.

Recalling that quantum mechanics is a non-relativistic theory, we can use the classical expression for the energy,

$$E = T + V = p^2 / (2m) + V = \sum p_i^2 / (2m_i) + V$$

Assuming that we are talking about a single particle, the summation in the last expression runs over the three components of momentum, but the expression is also valid for several particles, if the summation is extended to go over particle number as well (particle masses are denoted m_i). T is the kinetic energy and V the potential energy. Introducing the operator expressions for E and p from above, one gets

$$i \hbar \, \partial / \partial t = -(\hbar^2 / (2m)) \Delta + V = -(\hbar^2 / (2m)) \sum (\partial / \partial x_i) (\partial / \partial x_i) + V(\mathbf{x}).$$

Now let the operators on each side of the equality sign act on the wave function ψ and get

$$i \hbar \, \partial \psi / \partial t = -(\hbar^2 / (2m)) \Delta \psi + V \psi = H \psi$$

This is the time-dependent Schrödinger equation for a single particle of mass m. The operator Δ is called the Laplace operator. It follows from the derivation above, that it is defined by the scalar product of two gradient operators:

$$\Delta = \nabla \cdot \nabla = \sum_{i=1,2,3} (\partial / \partial x_i) (\partial / \partial x_i)$$

(where the dot product, has also been introduced - it is defined generally for two three dimensional vectors **a** and **b** as $\mathbf{a} \cdot \mathbf{b} = \sum a_i b_i = |a| \, |b| \cos\theta$ where θ is the angle between the directions of the two vectors). Back in the Schrödinger equation, the sum of the operators obtained for kinetic and potential energy has been denoted H. This operator, which is the quantum mechanical analogy to the classical energy expression, is called "the Hamiltonian". The potential energy part, V, is usually not an operator but a conventional function of position co-ordinates (and perhaps of time).

The Schrödinger equation as written above bears strong similarity with the general "law of nature" derived towards the end of chapter 7 (marked (**)). However, there is still a difference in form, because the above equation has a partial derivative with respect to time on the left hand side, and an operator H on the right hand side, where (**) has the full time derivative on the left hand side and a linear combination of ψs on the right hand side, with functions H(i,j) rather than operators multiplying them. However, the Schrödinger equation can be rewritten in precisely that form by mathematical transformation.

The mathematical concept needed here is that of complete sets of functions, which are (finite or infinite but numerable) sets of functions of given variables, with the property that any (reasonably smooth) function of these variables can be written as a linear combination of functions belonging to the set of basis functions,

$$\varphi(x) = \sum_i a_i \chi_i (x).$$

We can now use such a complete set of functions, depending on the space co-ordinates, to write the wave function of a given system ("expand" it on the basis functions):

$$\psi(\mathbf{x}, t) = \sum_i a_i(t) \chi_i (\mathbf{x}).$$

Here the expansion coefficients a_i are t-dependent, because the basis function were taken to depend only on **x**. In the following, the potential functions V in the Hamiltonian are assumed not to depend directly on time. Consider now a wave function, which has only one of the coefficients a_i in the sum above different from zero. The state described by such a wave function is called a "stationary state". Most bounded quantum mechanical systems possess one or (usually) several stationary states. Inserting $\psi(\mathbf{x}, t) = a_i(t) \chi_i (\mathbf{x})$ in the Schrödinger equation, it can be written

$$(\chi_i)^{-1} H \chi_i = i \hbar (a_i)^{-1} d a_i / d t$$

(note that i is used both for the imaginary unit and for labelling the state; sorry but such are "conventions"). Here the partial derivative with respect to time has turned into a full derivative, since $a_i(t)$ only depends on t. The form of the equation now obtained has the property, that the left hand side only depends on **x** and the right hand side only on t. This is possible only if both are equal to a common constant, which I shall denote E_i. The time-

dependent equation can be readily solved,

$$i\hbar\, da_i/dt = E_i a_i \quad \Rightarrow \quad a_i(t) = a\exp(-iE_i t/\hbar)$$

The space part of the Schrödinger equation now reads

$$H\chi_i(\mathbf{x}) = E_i\chi_i(\mathbf{x})$$

and the expectation value of the energy operator H is

$$\langle H\rangle = \int\chi^* H\chi\, d^3x = E_i\int\chi^*\chi d^3x = E_i$$

where the last equality assumes that the constant a in $a_i(t)$ has been chosen as one. Then by using exp(–ir) exp(ir) = 1 it follows that the space part of the wave function is separately normalised to unity, i.e. $\int|\chi|^2 d^3x = 1$.

Thus E_i is the energy of the stationary state i. From the definition of H, it and hence E_i are real quantities. This can be expressed in the following way:

$$E^* = \int(H\psi)^*\psi d^3x = \int\psi^* H^+\psi d^3x = \int\psi^* H\psi d^3x = E$$

The second equality is really the defining equation for the quantity H^+, which is called the hermitian adjoint operator to H. The reality of E (really E_i) can then be expressed as $H^+=H$. To prove the relation above mathematically would involve what is called integration by parts to move the derivatives contained in H around. Now we can use a similar technique to calculate the integral of H between two different stationary state wave functions χ_1 and χ_2,

$$0 = \int\chi_1^* H\chi_2\, d^3x - \int\chi_1^* H^+\chi_2\, d^3x = E_2\int\chi_1^*\chi_2\, d^3x - \int(H\chi_1)^*\chi_2\, d^3x = (E_2-E_1)\int\chi_1^*\chi_2 d^3x$$

This evaluation tells us, that for any two stationary states with different energy, the spatial wave functions are orthogonal, that is by definition of the short-hand notation on the left side,

$$\langle 1|2\rangle = \int\chi_1^*\chi_2 d^3x = 0$$

If there are more than one stationary state with the same energy, one talks of a degeneracy, in this case the wave functions may not in the first place be orthogonal, but there is a mathematical procedure (the Gram-Schmidt orthogonalisation procedure), which allows the basis functions for the degenerate energy value to be redefined, so that also in this case they are orthogonal. As stated above, the choice of the time-part of the wave function for each stationary state as a pure exponential function (a=l in front) also makes the basis functions normalised. We can then assume to have a complete set of basis functions with the properties

$$\langle j|n\rangle = \int\chi_j^*\chi_n d^3x = 0 \text{ for all } j \neq n$$

$$\langle j|j\rangle = \int\chi_j^*\chi_j d^3x = 0 \text{ for all } j$$

Now let me go back to the Schrödinger equation and insert the expansion $\psi = \sum a_n\chi_n$ on both sides:

$$i\hbar\,\partial\psi/\partial t = \sum_n i\hbar\chi_n\, da_n(t)/dt = H\psi = H\sum_n a_n(t)\chi_n$$

The second and fourth expressions are then multiplied from the left by $(\chi_i)^*$, the complex conjugate of χ_i, and integrated over space co-ordinates,

$$i\hbar \sum_n \langle i \mid n \rangle \, da_n(t)/dt = \sum_n \langle i \mid H \mid n \rangle \, a_n(t),$$

using $\sum_n$ for the sum over n, the definition of $\langle i \mid n \rangle$ given above, and

$$\langle i \mid H \mid n \rangle = \int \chi_i^* H \chi_n \, d^3x$$

Making use of the ortho-normality conditions, this reduces to

(§) $$i\hbar \, da_i(t)/dt = \sum_n \langle i \mid H \mid n \rangle \, a_n(t)$$

This is what I wanted to arrive at. The equation (§) is the Schrödinger equation in a form exactly equivalent to the expression (**) in Chapter 7. They become identical with the definitions

$$H_{in} = \langle i \mid H \mid n \rangle,$$

$$\psi_n = \langle n \mid \psi \rangle = \sum_i \langle n \mid i \rangle \, a_i(t) = a_n(t)$$

So let me write the basic law of quantum mechanics once more, with this new notation:

$$i\hbar \, d\psi_i / dt = \sum_n H_{in} \psi_n$$

The Hamiltonian can often be found by taking the classical expression for kinetic and potential energy, and then "quantize" it by substituting $-i\hbar \, \partial / \partial x_i$ for p_i. The potential V may for an electron rnoving around a nucleus of charge q be taken as the Coulomb potential $-(4\pi\varepsilon_0)^{-1} e q / r$. For a nucleon in a nucleus, V is the "strong interaction", and it would have to be derived from the gluon interaction between quarks in some way.

Quantum mechanics has led to a very accurate description of electron motion in light atoms. For heavier atoms, the number of electrons get too high for carrying through the full calculation. One then takes advantage of the shell structure of electron orbits, by considering the inner electrons as only influencing the wave functions of the outer ones through modifying the effective charge of the nucleus. Such calculations nave also been quite successful, as have calculations of electron wave functions for molecules. In the nucleon case, there is no "force centre", and the interaction is stronger and less well known. Therefore, calculations using methods similar to those used in electron calculations have only given a gross resemblance with measured quantities (position of energy levels, electric charge distributions (moments), reaction cross sections, and so on). If the nucleon number is large (say 200), collective behaviour such as rotations or vibrations can be well described, but still the motion of individual nucleons show up in the energy spectra, in contrast to the situation for collective phenomena in the electron systems of condensed matter. This should not be surprising, since in solids the number of electrons is of the order of 10^{24}, which certainly allows for statistical and collective treatment. In atoms and nuclei beyond the lightest ones, particle numbers of 20-250 are right in the middle - too big for individual particle description and too small for accurate collective description. The same is true for materials studied in recent nano-physics: single layer atoms or small conglomerates in the form of tubes or balls formed by a small number of atoms.

8.2 Second quantization

Consider again the complete set of quantum states $\chi_i = |i\rangle$ for a given system, with i = 1,2,... Let me define a vacuum, that is an empty state, as the state with no particles in any of the states i. I call this state $|0\rangle$. I then define an operator c_i^+ with the property of adding a particle in state i, when acting on a wave function. For example, acting on the vacuum, a one particle state is created with the particle in state i:

$$c_i^+ |0\rangle = |i\rangle$$

This also implies that if both the states and the c-operators are properly normalised, then

$$\langle i | c_i^+ |0\rangle = 1,$$

which by complex conjugation (write out the integral and use rules given earlier) gives

$$\langle 0 | c_i |i\rangle = 1.$$

In other words, the hermitian conjugate of c_i^+, that is c_i, takes a particle out of the state i:

$$c_i |i\rangle = |0\rangle$$

If there are no particles in state i, none can be taken out, so we must nave

$$c_i |0\rangle = 0 \quad \text{and} \quad c_i |j\rangle = 0 \text{ for all } j \neq i$$

A state with n particles can be written

$$c_{i_1}^+ c_{i_2}^+ \ldots c_{i_n}^+ |0\rangle = |i_1 = 1, i_2 = 1, \ldots, i_n = 1\rangle c_{i2}^+$$

In order to count the number of particles in a state, one just has to apply the operator $N = \sum_i c_i^+ c_i$, which according to the rules above gives

$$N \,|\text{state with n particies}\rangle \,\text{»}\, n \,|\text{same state}\rangle$$

or

$$\langle \text{n-particle state} \,|\, N \,|\, \text{same n-particle state}\rangle = n.$$

The name for this procedure, second quantization, should now be clear. The first quantization is to introduce the quantum states $|i\rangle$, using the Schrödinger equation with its fundamental constant $\hbar$. The second quantization is to introduce the particle number quanta, forcing the system to have an integer number of particles. The fundamental constant associated with this can be said to be the particle number "1". The use of the particle number formalism depends on whether the particles are fermions or bosons (half-integer or integer spins). Consider first the fermions. Here the Pauli exclusion principle should be imposed, stating that there can be at most one particle in each distinct state i. This is easily expressed in terms of the creation operators:

$$f_i^+ f_i^+ = 0,$$

where the letter f has been used instead of c, in order to remind you that these operators create fermions. A fermion multi-particle state must have all state indices i different. A generalisation of the exclusion principle is

$$f_i^+ f_n^+ + f_n^+ f_i^+ = 0,$$

$$f_i^+ f_n + f_n f_i^+ = \delta_{in}, \text{ where } \delta_{in} = 1 \text{ if } i=n, \text{ else } 0$$

These are called the "commutation relations" of the fermion operators. For $i \neq n$ they show that interchanging two fermions produces a sign shift in the wave function. This is the same as saying that the wave functions should be antisymmetric. This in fact is the defining property of fermions, a property confirmed by all experiments to date. I shall not at this point discuss why this is so. The sign change in the wave function has no effect on the probability density or any expectation value of physical quantities, because they involve the square of the wave function. But any experiment depending on the interference of more than one wave function will clearly exhibit an effect of the relative sign change. For i=n the first of the commutation rules is just the exclusion principle, while the second gives a very useful computational rule for the creation and destruction operators:

$$f_i f_i^+ = 1 - f_i^+ f_i$$

The role of the terms can be clarified by letting the left and right hand side of this identity operate on states with 0, 1 and 2 particles. The relation further allows any expression involving creation and destruction operators to be rewritten, so that all the creation operators are to the left of all the destruction operators (this is called "normal ordering").

For bosons (for which I shall call the second quantization operators b and b^+), the defining commutation relations read

$$b_i^+ b_n^+ - b_n^+ b_i^+ = 0 \quad \text{(trivial for i=n)}$$

$$b_i b_n^+ - b_n^+ b_i = \delta_{in}$$

These properties ensure that any boson state is totally symmetric, and there can be any number of bosons in each state i.

In order to show a little of the power of second quantization, I shall look at the form of the Hamiltonian for fermion and for boson systems. The fermion Hamiltonian must have the form

$$H(\text{fermion}) = \sum h^{(2)}_{12} f_1^+ f_2 + \sum h^{(4)}_{1234} f_1^+ f_2^+ f_3 f_4 + \ldots$$

where the terms with different numbers of creation and destruction operators nave been left out to indicate the conservation of the fermion number. There should be two or more fermions in the system, before the 4-fermion operator or higher terms come into play. By a linear transformation among the basis states,

$$|i'\rangle = \text{new } |i\rangle\text{'s} = \sum_n a_{in} |n\rangle,$$

the matrix $h^{(2)}$ may be brought to diagonal form, so that if higher order terms are absent or can be neglected, the fermion Hamiltonian reduces to (dropping the prime on i again)

$$H(\text{fermion}) = \sum_i F_i f_i^+ f_i$$

This Hamiltonian would describe a set of elementary fermion states with F_i equal to the mass times c^2 for each of them. For bosons, the Hamiltontan could be

$$H(\text{boson}) = \sum_i B_i b_i^+ b_i + \sum A_{123} (b_1^+ b_2^+ b_3 + b_3^+ b_2 b_1) + \ldots$$

where interactions can increase or decrease the number of particles. The leading order term has already been assumed to be diagonalised.

Hamiitonians representing interactions between fermions and bosons can equally easily be constructed. For example, the basic fermion interaction is believed to be an interaction involving a boson field, as we nave seen. Thus it should not just involve fermions, as the expression above does (although this could in some cases be a useful approximation). We can write the Hamiltonian for the four basic interactions shown in Fig. 8.1 (electron-photon or quark-gluon) in the following way:

$$H(\text{coupling}) = \sum C_{123} (b_1^+ f_2^+ f_3 + f_3^+ f_2 b_1) + \sum D_{123} (b_1^+ f_2 f_3 + f_3^+ f_2 b_1)$$

Note how the hermitian property of the Hamiltonian, $H^+ = H$, makes the coupling terms pairwise of equal strength. This means that the interaction strengths of diagram (a) and (b) in Fig. 8.1 are identical, as are those of diagrams (c) and (d).

For the gluons, interaction terms, such as the A_{123} terms in the expression for h(boson) above, are believed to be present. The corresponding diagrams are illustrated in Fig. 8.7.

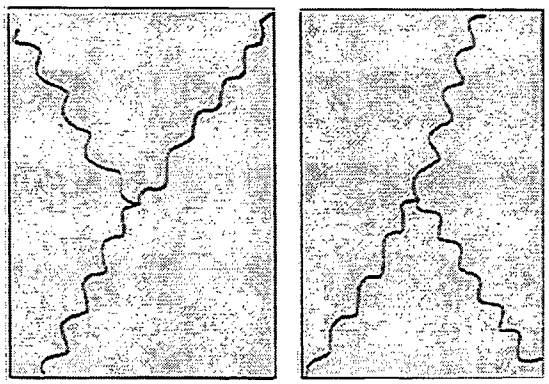


Fig. 8.7. Feynman diagrams for gluon interactions involving b^+b^+b and b^+bb.

8.3. Quantum electrodynamics

So far I have discussed non-relativistic quantum mechanics, or I have talked sufficiently general not to have to enter the question of relativistic corrections, for example in the precise form taken by the Hamiltonian operator. Quantum electrodynamics is a theory of relativistic electrons, satisfying requirements set by the special relativity theory. In many experiments involving electrons, the velocity is indeed close to that of light, and furthermore, the creation and destruction of light (photons) by accelerated electron motion is a prominent feature. The description of these phenomena requires a way of combining quantum mechanics, special relativity theory and electromagnetic theory.

I shall approach the problem in small steps. The electron is a particle with a non-zero mass and with spin ½ $\hbar$. Before dealing with the electron itself, I consider a relativistic particle without spin, and one with spin but without mass.

The Schrödinger equation was obtained by inserting $E = i\hbar\,\partial/\partial t$ and $p = -i\hbar\,\nabla$ in the classical energy equation, E = T+V (sum of kinetic and potential energy). For a relativistic particle, it then seems worth trying to make the same substitutions in the relativistic rela-

tion between energy and momentum:

$$E^2 = p^2 c^2 + m^2 c^4,$$

where m is the mass of the particle at rest and c is the speed of light in vacuum. This relation follows from assuming Lorentz invariance, and for a particle at rest (p=0), one recognises Einstein's formula $E = mc^2$. For particles without rest mass, such as photons, the relation reads $E = pc$, which is a relation well known in classical electrodynamic theory - one that relates the radiation pressure to the energy of a light beam. For a particle without spin, we may thus try a wave equation of the form

$$(i\hbar\, \partial / \partial t)^2\ \psi(\mathbf{x}, t) = (c^2 (-i\hbar\nabla)^2 + m^2c^4)\ \psi(\mathbf{x}, t)$$

This is called the Klein-Gordon equation, and it indeed is a valid description of free, spin-0 particles of mass m ("free" means not feeling any forces - there is no potential V in the equation as yet).

Now consider particles with spin. Spin is like an angular momentum (which in classical physics is $\mathbf{x}\times\mathbf{p}$, or xp times the sine function of the angle between the two directions). However, it does not depend on the space co-ordinates, but on some internal co-ordinates of the particle. It is not necessary to know what these intrinsic degrees of freedom are for writing an equation of motion. This is because the spin behaves exactly like other angular momenta in quantum mechanics. The central theorem is that for each angular momentum j of a system, there is a finite, complete set of 2j+1 basis states $|i\rangle$ in the sense of the previous section. The 2j+1 states corresponds to projections of the angular momentum on a fixed axis in space having one of the values $-2j-1, -2j, -2j+1, \ldots, 2j, 2j+1$. I shall not prove this theorem, which involves group theory applied to the possible rotations of the system around a fixed axis. So each state allowed for the system by other consideration corresponds to 2j+1 angular momentum states. For electrons of spin ½, this means that there are two spin states for each spatial wave function. Let me label them i=1 and i=2 and treat the spin degree of freedom by expanding wave functions on this two-dimensional basis, while still retaining the **x**-dependence of the wave function:

$$\Psi = (\Psi_1(\mathbf{x}, t), \Psi_2(\mathbf{x}, t))$$

The reason for having to cope with this complication is that the momentum operator $\mathbf{p} = -i\hbar\,\nabla$ is capable of transforming one spin state into the other. One says that there is a spin-orbit coupling. Now consider first a massless particle with spin ½. The classical energy equation is

$$E^2 = c^2 p^2 \quad \text{or} \quad E = \pm cp$$

Wolfgang Pauli (1933) showed, that the quantum form of the momentum, which would make the interaction Lorentz invariant, was not just $\mathbf{p} = -i\hbar\,\nabla$ for each of the spin states i=1 and 2, but rather a set of operators P_{in}, making the energy equation above become

$$i\hbar\, \partial \Psi_n / \partial t = \pm c \sum_i P_{ni}\, \Psi_n \quad \text{with n=1,2 and where}$$

$$\begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} p_z & p_x - ip_y \\ p_x + ip_y & -p_z \end{pmatrix}$$

are the scheme of quantities (called a "matrix") to be used for P_{ni} in the equation above. The momentum components p_i have to be inserted as $-i\hbar\,\partial / \partial x_i$ in the equation, known as the Weyl equation. Weyl suggested that the neutrino could be described by this equation if the minus sign were selected where the ± appears (Weyl, 1929).

Now the electron has a mass and spin ½. It is therefore tempting to describe a free electron by an equation formally of the Klein-Gordon kind (with $\partial / \partial t$ squared in contrast to the Schrödinger equations), but with the momentum terms written in the Pauli way. That is called the Dirac equation (Dirac, 1931):

$$(i\hbar\,\partial / \partial t)^2\,\Psi_n = c^2 \sum_{j=1,2} P_{nj} \sum_{k=1,2} P_{jk}\,\Psi_k + m^2c^4\,\Psi_n$$

Rewriting the equation as

$$\sum_{j=1,2} (i\hbar\,\delta_{nj}\,\partial / \partial t - cP_{nj}) \sum_{k=1,2} (i\hbar\,\delta_{jk}\,\partial / \partial t + c\,P_{jk})\,\Psi_k = m^2c^4\,\Psi_n$$

one may get rid of the messy second order term in $\partial / \partial t$ at the expense of doubling the number of equations. The trick is to introduce a new wave function χ, the j'th component of which (times mc^2) equals the result of the second summation acting on Ψ_k above. The two sets of (two) equations then become

(•) $$i\hbar\,(\partial / \partial t)\,\Psi_n + c \sum_i P_{ni}\,\Psi_i = mc^2\chi_n,$$

$$i\hbar\,(\partial / \partial t)\,\chi_n - c \sum_i P_{ni}\,\chi_i = mc^2\,\Psi_n$$

The two equations can even be combined into one, if one defines a 4-component wave function

$$\psi = (\psi_1, \psi_2, \psi_3, \psi_4) = (\chi_1 + \Psi_1, \chi_2 + \Psi_2, \chi_1 - \Psi_1, \chi_2 - \Psi_2)$$

It would have been as good to simply take $\psi = (\Psi_1, \Psi_2, \chi_1, \chi_2)$, but the above choice makes the form of the equations identical to that originally suggested by Dirac and still the preferred one. Introducing these definitions in the equation (•), the Dirac equation becomes a first order differential equation for a four-component wave function,

(••) $$mc^2\,\psi_n = i\hbar\,(\partial / \partial t) \sum_i \beta_{ni}\,\psi_i - c \sum_i \Gamma_{ni}\,\psi_i$$

where

$$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \\ \beta_{31} & \beta_{32} & \beta_{33} & \beta_{34} \\ \beta_{41} & \beta_{42} & \beta_{43} & \beta_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

and

$$\Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} & \Gamma_{14} \\ \Gamma_{21} & \Gamma_{22} & \Gamma_{23} & \Gamma_{24} \\ \Gamma_{31} & \Gamma_{32} & \Gamma_{33} & \Gamma_{34} \\ \Gamma_{41} & \Gamma_{42} & \Gamma_{43} & \Gamma_{44} \end{pmatrix} = \begin{pmatrix} 0 & 0 & p_3 & p_- \\ 0 & 0 & p_+ & -p_3 \\ p_3 & p_- & 0 & 0 \\ p_+ & -p_3 & 0 & 0 \end{pmatrix}$$

with

$$p_+ = p_1 + ip_2, \quad p_- = p_1 - ip_2 \quad \text{and} \quad (1,2,3) = (x,y,z)$$

Note that the m and the $\partial/\partial t$ terms have in (••) been moved to the other sides of the equality sign, as compared with the Hamiltonian type equation (•). The reason is that in relativity theory, it is customary to treat "ict" as the fourth component of a space-time coordinate $(x_1, x_2, x_3, x_4) = (x, y, z, ict)$. Then $-\hbar/c\,\partial/\partial t = -i\hbar\,\partial/\partial x_4 = iE/c$ becomes the fourth component of the momentum vector, introducing $p_4 = -\hbar/c\,\partial/\partial t$, the energy and momentum terms now collected on the right hand side may be rewritten so that the Dirac equation gets the compact form

$$(\otimes) \qquad -mc\,\psi_n = \Sigma_i M_{ni}\,\psi_i, \quad n=1,2,3,4$$

with

$$M = \begin{pmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{pmatrix} = \begin{pmatrix} ip_4 & 0 & p_3 & p_- \\ 0 & ip_4 & p_+ & -p_3 \\ p_3 & p_- & -ip_4 & 0 \\ p_+ & -p_3 & 0 & -ip_4 \end{pmatrix}$$

So much for the relativistic treatment of a free electron. If the electron is moving in an electromagnetic field, the classical energy function is modified in the following way (a discussion of classical electrodynamics will be taken up in one of the following chapters):

$$E = (2m)^{-1}\,(\mathbf{p} - q\mathbf{A})^2 + q\,\varphi,$$

where q is the electric charge of the particle, φ (**x**, t) the familiar electric potential (for example the voltage change across a battery or through an electric circuit), and **A**(**x**, t) the corresponding vector potential describing magnetic fields.

One interpretation of the equation above is, that the electromagnetic fields can be included by making the substitutions

$$E \rightarrow E - q\,\varphi$$

$$p_i \rightarrow p_i - qA_i, \quad \text{for } i=1,2,3$$

In order to use the relativistic 4-component description introduced above, the fourth component of the vector potential is defined as $A_4 = i\varphi/c$, whereby the substitution prescription can be written

$$p_i \rightarrow p_i - qA_i, \quad \text{for } i=1,2,3,4$$

Making this replacement of the p_i's in the matrix M above makes the Dirac equation valid for electrons interacting with an electromagnetic field.

Let me separate the physical (momentum and potential parts) and the geometrical components of the M-matrix,

$$M_{nm} = \Sigma_{i=1,2,3,4}\,(p_i - qA_i)\,\gamma_{nm}(i).$$

The new γ-matrices contain simple numbers:

$$\gamma(1) = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \qquad \gamma(2) = \begin{pmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & i & 0 & 0 \\ -i & 0 & 0 & 0 \end{pmatrix}$$

$$\gamma(3) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad \gamma(4) = \begin{pmatrix} i & 0 & 0 & 0 \\ 0 & i & 0 & 0 \\ 0 & 0 & -i & 0 \\ 0 & 0 & 0 & -i \end{pmatrix}$$

The next step is to use second quantization on both the electron wave functions and on the photon fields A_i. I expand each of the four components of ψ_i on a complete set of basis states for all but the spin degrees of freedom, $|\nu\rangle$,

$$\psi_i = \sum_\nu r_i(\nu)\ |i\nu\rangle = \sum_\nu r_i(\nu)\ f_i^+(\nu)\ |0\rangle, \quad i=1,2,3,4 \text{ and } \nu = \text{spatial electron states}$$

Since the photon fields are hermitian ($A^+ = A$), the simplest expansion on creation and destruction operators is

$$A_i = \sum_\mu s_i(\mu)\ (b^+(\mu) + b(\mu)), \quad i=i1,2,3,4 \text{ and } \mu = \text{spatial photon states}$$

The summations in the above expressions are over ν or μ. There is no i-index on the b-operators, since the photons have no spin and the μ-states form a complete basis for the spatial states.

I am now ready to write the interaction in second quantized form. The m- and p-dependent terms in the Dirac equation describe the free electron. Hence the A_i-dependent terms constitute the interaction Hamiltonian,

$$H(\text{int}) = \sum_{\substack{i,i_1,i_2=1,2,3,4 \\ \mu\nu_1\nu_2}} \langle i_1\nu_1 | (-q) s_i(\mu) \gamma_{i_1\ i_2}(i) | i_2\nu_2 \rangle f_{i_1}^+(\nu_1)\,(b^+(\mu)+b(\mu))\, f_{i_2}(\nu_2)$$

The brackets $\langle\ |\ \rangle$ are just numbers, so this is precisely the kind of electron-photon interaction anticipated in the "Second quantization" section in connection with Fig. 8.7 (the terms with coefficients C).

Since the Dirac equation originated in a form with $E^2 = (i\hbar\,\partial/\partial t)^2$, it is not surprising, that the solutions ψcome in pairs, where one has a positive energy and the other a negative one of same numerical magnitude. Dirac assumed, that normally all the negative states were filled with exactly one electron in each (exclusion principle). Occasionally, a negative energy electron would become excited up into the positive region (receiving energy say from an impinging photon). This would appear as the creation of a new electron, as well as of a hole in the negative energy "sea of states". Dirac interpreted the hole of energy $-E$ as equivalent to an anti-particle of energy E. This anti-particle, the positron, was later observed and gave credibility to the theory of Dirac.

This model suggests, that the creation operator for a negative energy state may be interpreted as a destruction operator for a positron, $f(\bar{\nu}) = f^+(\nu)$, and similarly that $f^+(\bar{\nu}) = f(\nu)$. In this way, the interaction Hamiltonian gets the pair creation and destruction terms an-

ticipated in the discussion of Fig. 8.7. Omitting the i=1,2,3,4 indices for clarity, this gives the form

$$H(\text{int}) = -q \sum_{\substack{\mu,\nu_1,\nu_2 \\ \text{positive energy} \\ \text{states only}}} \langle \nu_1 | S(\mu)\gamma | \nu_2 \rangle \{ f^+(\bar{\nu}_1)(b^+(\mu)+b(\mu))f(\nu_2) + f^+(\bar{\nu}_1)(b^+(\mu)+b(\mu))f(\bar{\nu}_2) + f(\bar{\nu}_1)b^+(\mu)f(\nu_2) + f^+(\nu_1)b(\mu)f^+(\bar{\nu}_2) \}$$

The description so far has included the free electrons and positrons and their interaction with photons. For completeness, the Hamiltonian should also contain a term describing the propagation of photons and their mutual interaction, if any.

In order to quantize the photon field we must know its form in classical electrodynamics. This will just be stated here. The classical energy density u in an electric field **E** and a magnetic field **B** is

$$u = \tfrac{1}{2} \varepsilon_0 (E^2 + c^2 B^2),$$

where $\varepsilon_0 = 10^7/(4\pi c^2)$ is the vacuum dielectric constant and

$$\mathbf{E} = -\nabla\varphi - \partial \mathbf{A}/\partial t, \quad \mathbf{B} = \nabla \times \mathbf{A}$$

(the cross product × is defined in the appendix). Through these expressions, the energy density of the photon field has been expressed in terms of the potentials already used in the quantum description of the photon-electron interaction. These potentials went unchanged from the classical to the relativistic quantum description, so we expect a Hamiltonian H = u to work. It can be rewritten in the four-component form of the Dirac equation:

(##) $H = \varepsilon_0 (\sum_i (F_{4i})^2 - (1/4)\sum_{in} (F_{in})^2)$, with sums from 1 to 4,

provided that I define an "electromagnetic field matrix" F by

$$\begin{pmatrix} 0 & cB_3 & -cB_2 & -iE_1 \\ -cB_3 & 0 & cB_1 & -iE_2 \\ cB_2 & -cB_1 & 0 & -iE_3 \\ iE_1 & iE_3 & iE_3 & 0 \end{pmatrix} = (F_{in}) = F$$

Inserting the expressions for the electric and magnetic fields, F can be expressed directly in terms of the potentials

(§§) $F_{in} = c\,(\partial A_n/\partial x_i - \partial A_i/\partial x_n)$, i,n = 1,2,3,4

One may note that the Hamiltonian for the photons themselves involves derivatives of the potentials A_i, in contrast to the electron-photon interaction, which involved just the fields themselves. Introducing second quantization, the terms with products of F-matrices will transform into terms with two boson operators. By suitable choice of the basis states, the ($bb + b^+b^+$) terms may be removed, so that the photon Hamiltonian gets the form

$$H(\text{photon}) = \sum (\hbar\omega_i)\, b_+(i)b(i),$$

where I have called the coefficients $\hbar\omega_i$, because they must be just the energies of the photons ($\hbar\omega = h\nu$). If there is a continuous interval of possible photon energies $\hbar\omega$, the sum should be replaced by an integral. However, throughout this chapter I have avoided this complication by always assuming a numerable set of disjoint basis states in my expansions. This would seem a fair approximation also in cases where it is not strictly valid, and it avoids problems of defining and using concepts such as orthogonality of continuum basis states and their normalisation.

The form of the photon hamiltonian shows us, that photons do not interact with themselves or with other photons. In other words, the diagrams of Fig. 8.7 are not valid for photons. Photons can only interact through charged particles with an interaction of the form H(int) given above and corresponding to diagrams of the type shown in Fig. 8.1.

I have second quantized the spin-½ electrons as fermions and the spin-1 photons as bosons. Could I not have done otherwise? No, according to a general theorem, first proved for free particles by Pauli (1940), choosing the "wrong" type of quantization - quantizing integer spin particles as fermions or half-integer spin particles as bosons - lead to inconsistency. Among the assumptions for proving this theorem is the requirement of Lorentz invariance.

8.4 Quantum chromodynamics

It is tempting to describe quark interactions in analogy to electron interactions, using a theory similar to quantum electrodynamics. The quarks are charged, spin-½ fermions, so they would correspond to electrons and positrons, while the gluons, which are massless spin-1 bosons, would correspond to photons. However, there are differences as well: since for example protons and omega-minus particles are supposed to have two or three quarks in the same state (Table 8.2), the exclusion principle for fermions cannot be satisfied. One way out of this would be to postulate an additional property for the quarks, which may be different for the otherwise identical states. This property is called "colour", and the three possible colour states may be denoted c=1,2,3. The different quarks (see Table 8.1) may be described by a property (denoted "flavour") f, which can take on 6 different values, corresponding to the u, d, s, c, b and t-quarks. A complete set of quark basis states must then be described by c- and f-indices as well as by the i=1,2,3,4 index of the relativistic spin-½ wave functions of the Dirac theory,

$$\psi(\text{quark}) = \psi_i(c, f; \mathbf{x}, t)$$

The gluons are described by potentials similar to the A_i's of the electromagnetic field. However, the gluon interaction is capable of changing the colour of quarks, say from c_1 to c_2, no matter whether the quark flavours are the same or different. The gluon potentials are then

$$A(\text{gluon}) = A_i(c_1, c_2; \mathbf{x}, t)$$

By inserting these quantities into the Dirac equation (⊗) in the previous section on quantum electrodynamics, with the M_{ri}'s given there, the quark equation of motion in quantum chromodynamics emerges,

$$-m(f)\, c\, \psi_n(c_i, f) = \sum_{i,j,c_2} (p_j\, \delta(c_1,c_2) - g\, A_j(c_1, c_2))\, \gamma_{ni}(j)\, \psi_i(c_2,f)$$

Here m(f) is the mass of the quark with flavour f and g the strength parameter of the quark-gluon interaction. δ was defined in the section on double quantization. The parameter g corresponds to the charge q, which measures the strength of the electromagnetic interaction. There could have been a color-dependence of g, as the gluon potentials and g might also nave been chosen to depend on the flavour f. However, there is no need to complicate things as long as there is no experimental evidence to distinguish these complications.

Since the quarks are charged, part of the interaction written above is electromagnetic and independent of colour numbers. This can be the sum over the diagonal elements,

$$g \sum_{c_1} A_i(c_1,c_1) = q\, A_i(\text{electromagnetic})$$

By the way this part of the interaction must depend on the flavour f, because the charges q do (being $\pm 1/3$ or $\pm 2/3$). Of the original 3·3 = 9 elements in the matrix $A(c_1,c_2)$, this leaves 8 to describe the colour-shift interactions caused by the gluon fields.

In second quantization, a complete set of quark operators f^+, f, and a complete set of gluon operators, b^+, b, are introduced (as in the previous section for the electromagnetic case). The quark-gluon interaction can then be evaluated in analogy to the electron/positron-photon interaction,

$$H(int) = \sum C(i_1,c_1,\nu_1,i_2,c_2,\nu_2,f,\mu)\{f^+(i_1,c_1,f,\nu_1)\,(b^+(c_1,c_2,\mu) + b(c_1,c_2,\mu))\, f(i_2,c_2,f,\nu_2)$$
$$+ f(i_1,c_1,f,\bar{\nu}_1)\, b^+(c_1,c_2,\mu)\, f(i_2,c_2,f,\nu_2) + f^+(i_1,c_1,f,\nu_1)\, b(c_1,c_2,\mu)\, f^+(i_2,c_2,f,\bar{\nu}_2)\}$$

This Hamiltonian describes the four interaction types of Fig. 8.1.

Each of the 8 gluons have a specific role. The $(c_1,c_2) = (1,3)$ gluon may be emitted by a colour=3 quark, if this quark changes colour to c=1 by the interaction. Later, this gluon will be absorbed by a colour=1 quark, thereby transforming it into a colour=3 quark.

The quark-gluon interaction is attractive and rapidly increasing with distance. This is because virtual gluons and quark-antiquark pairs are constantly formed and destroyed around any real quark. The gluon cloud around a quark tends to enhance the colour of the quark, so that around a c=1 quark, the effective c=1 colour may by many units exceed 1. This is felt as a similarly multiplied attraction by a distant quark. Only when it gets close to, that is inside part of the gluon cloud of the first one, the effective colour and hence the attraction diminishes. The quark-quark interaction via gluons (Fig. 8.2) is thus weak at short distances, but rapidly increasing for larger distances. This may serve as an explanation for the apparent lack of ability of the quarks escape the proton or pion or whatever hadron they are part of. Free quarks have never been observed! Fig. 8.8 illustrates the cloud of virtual parades surrounding a quark.

A reason for the large enhancement of colour around a quark is the ability of the gluons to multiply. Gluons can self-interact, in contrast to photons. This is because gluons carry definite colour indices (two each) and therefore are subject to colour-interactions similar

to the ones involving one gluon and one quark. Actually it if the other way round: gluons are assumed to have a colour-interaction, because this will explain their behaviour and hopefully the quark confinement in hadrons (for which only very idealised calculations have yet been carried through).

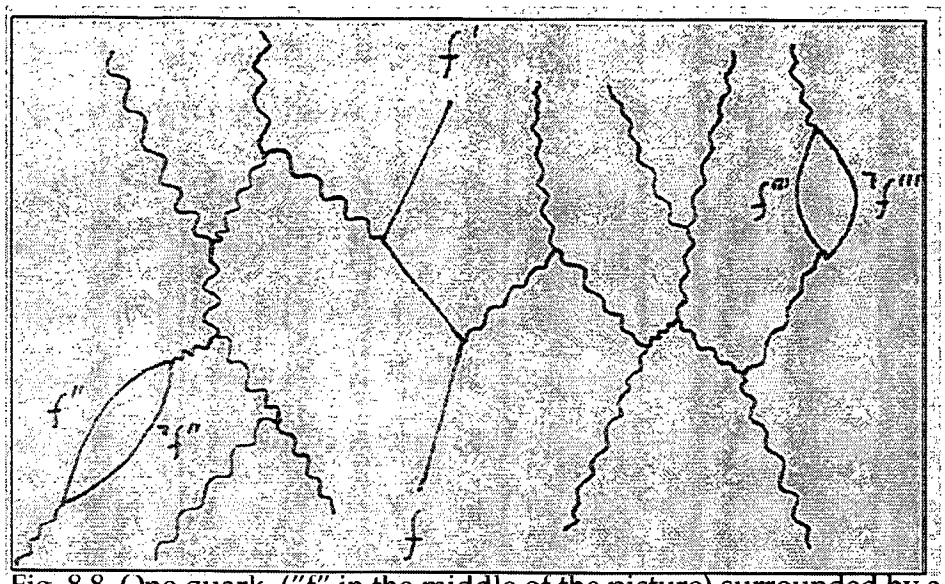


Fig. 8.8. One quark, ("f" in the middle of the picture) surrounded by a cloud of virtual gluons and quark-antiquark pairs.

As we have seen above, photons do not interact directly with themselves, and they also do not have any charge to "do it with". We begin to see a pattern, where a quantity like charge or colour is required for particles to interact in the way described by the field theories considered here. What these theories have in common can be stated more precisely: the theories are invariant under local symmetry operations of the Lorentz type (co-ordinate systems displaced or rotated, but moving with constant velocity relative to each other). That the symmetry should be local means that it may only pertain to a finite region in the space-time domain, not to all of it (see Fig 8.9). Invariance under local symmetry is a stronger requirement than invariance under the corresponding global symmetry operation. Local Lorentz type symmetry is also called "gauge symmetry". The transformation involved can be expressed in a very simple way for the potentials,

$$A_i \rightarrow A_i + \partial\vartheta/\partial x_i, \quad i=1,2,3,4$$

where ϑ is a function of space and time co-ordinates, common for the four components of A.

Returning to the gluon self-interaction, we may proceed partly as in the photon case, by choosing a field matrix depending on the potentials, and then construct the boson Hamiltonian in analogy to (##) in the previous section, but with the additional quantum numbers c_1 and c_2 attached to each F_{in}. In order to obtain a direct interaction between two gluons, a glance at the boson Hamiltonian in the section on "second quantization" tells us, that each F must contain both linear and quadratic terms in the potentials, that is

$$F_{in}(c_1,c_2) = \partial A_n(c_1,c_2)/\partial x_i - \partial A_i(c_1,c_2)/\partial x_n + \sum D(c_1,c_2,c_3,c_4,c_5,c_6)\, A_i(c_3,c_4)\, A_n(c_5,c_6)$$

where the D's are just coefficients and the sum is over c_3, c_4, c_5 and c_6. The two first terms

correspond to the Maxwell field in equation (§§) above. In second quantization, one boson creation or destruction operator should be inserted for each $A_i(c_1,c_2)$, leading to a Hamiltonian of the form

$$H = \sum m_1\, b^+_1 b_1 + \sum A_{123}\,(b^+_1 b^+_2 b_3 + b^+_3 b_2 b_1)$$
$$+ \sum B_{1234}\,(b^+_1 b^+_2 b_3 b_4 + b^+_1 b^+_2 b^+_3 b_4 + b^+_4 b_3 b_2 b_1)$$

In addition to the terms illustrated in Fig. 8.7, there are here terms of fourth order (but I have left out terms that would be inconsistent with energy conservation).

The order, in which the individual $A_i(c_1,c_2)$'s occur (say in H) is of importance. Interchanging two terms may give a different result. Such quantities are said to be "non-commuting" or "non-Abelian".

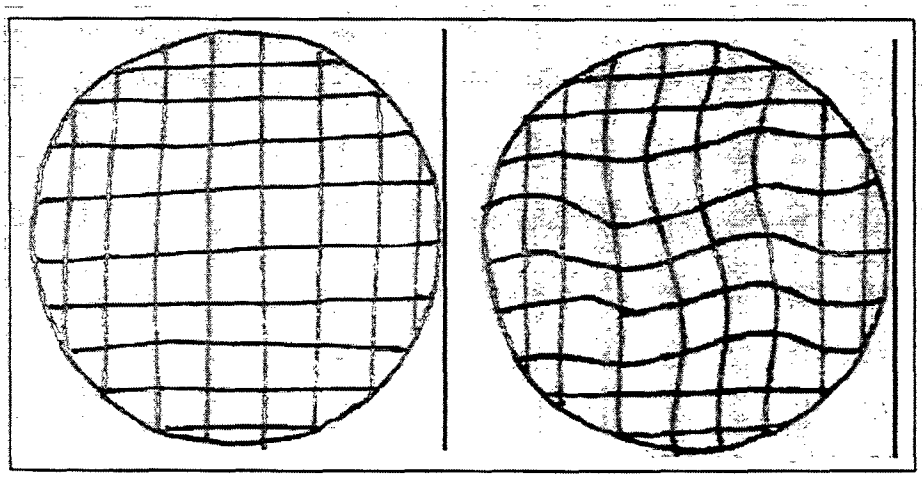


Fig. 8.9. In kindergartens, children sometimes make oil drawings on metal trays or cake forms, and then put the tray on a hot range and shake it a bit, to get nicely distorted pictures. This illustrates local symmetry: the pattern to the right in the picture is unchanged near the rims (compared with that to the left), but is twisted locally (near the centre). It is a stronger request to demand that natural laws are invariant with respect to such transformations than with respect to rotation or stretching of the whole pattern (idea based on Freedman and Nieuwenhuizen, 1978).

Weak interactions can be described by a theory of the same kind as quantum chromodynamics, that is a theory based on potentials A_i (with suitable variables) satisfying the principle of invariance under gauge symmetry transformations. One difference is that weak interactions are capable of changing the properties associated with the 6 flavour indices. In other words, the gauge potentials must contain such indices in a way similar to the colour indices of the gluon potentials.

PROBLEMS AND DISCUSSION ISSUES

PROBLEM 8.1. Investigations into the stability of the proton takes place in an abandoned goldmine in India, 2300 metres below the surface. In this way, interference from cosmic radiation can be minimised. The experiment deals with 140 metric tons of iron tubes,

inside which detectors of possible decay products from proton decay events are located.

A: estimate the approximate number of protons in the 140 tons of iron.

During 1981, three events indicating possible proton decay were observed. Answer the following questions under the assumption that they were really proton decays.

B. What is the decay constant k? (definition may be found below)

C. Estimate the lifetime of the proton.

For time intervals small compared with the lifetime, the number of decays is proportional to the number N of protons. The proportionality constant is k. After the time dt has gone, the number of protons has been reduced by kN dt. Defining the lifetime $T_{1/2}$ as the time it takes to reduce N to half, one may use $T_{1/2} = \log(2)/k$ (log or ln is the natural logarithm function).

DISCUSSION ISSUE 8.2. Did this chapter convince you, that matter consists of small particles, which move around and interact with each other only at discrete moments in time?

PROBLEM 8.3. In Table 8.2 it appears, that three identical quarks may be in the same state (forming an omega-minus particle). Why is this impossible, and why has the colour index been introduced on the quarks in Table 8.1?

PROBLEM 8.4. Show that the relativistic energy, $E = (p^2c^2 + m^2c^4)^{1/2}$, is approximately equal to $mc^2 + mv/2$ when v is much smaller than c.

PROBLEM 8.5. Consider stationary solutions to the one particle Schrödinger equation, which are zero outside a certain "box". In fact, look only at the one-dimensional case, say of x being the only space co-ordinate. In this case, the box is an interval. Call its length L. Show that the solution to the Schrödinger equation is of the form $a \cdot \sin(kx - \omega t)$. Determine the possible values of k and ω, and the corresponding energies of the system for allowed k and ω.

PROBLEM 8.6. (difficult!) Evaluate the commutation relations (as in section 8.2) for pairs of fermion operators. That is, instead of using b as in the text, use $f_1 f_2$. Neglect terms, which would give zero as expectation value for the state with no particles, and demonstrate that in this case, the fermion pair behaves as a boson. This may be useful in situations with many degenerate or close-lying fermion states, if the actual particle number is small compared with the number of states. Then the probability of an individual state being occupied is small, and the approximation of replacing fermion operator pairs by a boson operator may be valid.

DISCUSSION ISSUE 8.7. The new elementary particles predicted by early quantum theories (neutrinos, positrons, and so on) have all been found. So have many of the particles predicted by more recent theories of matter, for example those with s=strange, c=charm and b=bottom flavours. Three quark colour states have been introduced, along with eight gluons to be able to go from one colour state to any other. Again the purpose of suggesting these new fundamental particles was only to make the theory consistent (this time with the exclusion principle). If they are found, there would seem to be one more argument in favour of the direction, in which physicists are working.

Why are these people doing so well? Could it be, that there are many alternative ways of describing nature, and that experience will (nearly) always confirm an otherwise consistent theory with a sufficiently large number of parameters?

Did pre-Galileo astronomy not confirm the epicycle phenomena associated with the Earth-centred model (with occasional addition of a few extra parameters)? If this line of thinking is correct, the experimental support does not prove that theoretical thinking is heading in the right direction, there may be a much simpler model, which in a more convincing way explains the same experiments.

Maybe the way scientists are working today should be modified. Should they reject any theory, which is not simple? Do you think that a workable definition of "simplicity" could be found and agreed upon?

PROBLEM 8.8. Prove that the first sum in the photon Hamiltonian (§§) (section 8.3) is equal to $\varepsilon_0 E^2$ and the second sum equal to $\varepsilon_0 ((c\mathbf{B})^2 - \mathbf{E}^2)/2$.

Chapter 9

The universe

Throughout history, the universe has served as playing grounds for physics. Early Earth-centred models had to deal with the observation, that some celestial bodies (the "fixed stars") moved uniformly across the night sky, while other ones (the planets) sometimes changed direction and seemed to move "backwards" for a while, and then resumed the "standard" direction of motion. A fairly consistent model describing the observations would have the planets move on several separate circular orbits (epicycles) in addition to the main orbiting motion around the earth, the many epicycles introduced in order to "save" the earth-centred model implied a large number of parameters (radii, orbiting period and inclination for each epicycle). The solar centred models of Copernicus and Kepler could explain the same data with much fewer parameters. Kepler's model, which was based on accurate measurements by Tycho Brahe, showed that planetary orbits were ellipses and not circles. Studies of falling bodies at the Earth's surface led Galileo and Newton to the models now known as "classical mechanics" and comprising Kepler's laws. The first important test of the new models was again in astronomy and concerned the moons of Jupiter. Classical mechanics contains a law of motion (the position of a body changes with time in a given way) and a description of an important force, the gravitational one.

Astronomical observations after the development of good telescopes revealed new celestial objects, such as clusters of stars and galaxies. The brightness of objects in the sky could be estimated, and distances to objects measured. For the closer ones, this was done by measuring the direction to the object from two positions of the Earth in its orbit around the Sun. For more distant objects, the directional change was too small to measure, even for the largest difference in Earth positions (corresponding to semiannual spacing). In the case of distant galaxies, a study of the average brightness of their brightest stars can be used to deduce their average distance, provided that the absolute brightness (that is the amount of energy radiated from the star per unit of time) can be estimated. Such estimates are made on the basis of spectral analysis of the radiation. Stars are selected, which are of the same type (have same spectra) as the most common stars in our own galaxy, "the milky way". Since the average brightness of the latter stars is known, the assumption that it is the same for stars in the foreign galaxy allows its average distance to be calculated from its apparent brightness. For a few of the closer galaxies, this method can be checked against another one, based on a relationship between period of pulsation and brightness of certain variable stars.

Another variable that can be deduced from spectral data is the velocity of distant stars or galaxies, relative to us. This is because of the Doppler shift in the frequency of radiation received from a moving object. Just like the change in sound frequency say from the sirene of an ambulance moving towards or away from you. Early in this century, the Dutch physicist de Sitter suggested that galaxy velocity and distance should simply be proportional, and in the 1920ies E. Hubble's measurements showed that this indeed was the most consistent interpretation of his data (Hubble, 1929). The proportionality constant Hbetween velocity v and distance d, $v=Hd$, is called Hubble's constant. Here v is positive

for galaxies moving away from us. The value of Hhas turned out to be very difficult to pin down. Even today, it is uncertain within a factor of two: $1/(20\times10^9$ years) < H < $1/(10\times10^9$ years) (van den Bergh, 1981).

The Hubble law is valid for any homogeneous universe, that is a universe which is isotrophic (looks the same in any direction) for observers on any galaxy, the closer environment may not always be isotrophic, and there may be a few distant galaxies looking peculiar. But if these are exceptions and the bulk of galaxies are homogeneously distributed, then the Hubble law should be valid. Some of the difficulties in determining the Hubble constant are associated with the presence of a peculiar cluster of galaxies (the Virgo cluster) close to us. The Hubble constant needs not be constant in time.

The motion of galaxies would be expected to be determined by the gravitational forces, however, some of the galaxy velocities are likely to be close to the velocity of light. Therefore, a relativistic theory of gravitation has to be used, such as Einstein's general theory of relativity. The gravitational field equations in Einstein's theory contain ten potentials (as compared with one in the classical law of gravitation) but still one fundamental constant (Møller, 1952). The Soviet mathematician A. Friedmann found two classes of solutions to Einstein's equation, depending on whether the average density of matter in the universe is smaller than or larger than the value of $3H^2/<8G>$, where G is Newton's constant of gravitation, $G=6.67\times10^{-11}$) $m^3/kh/s$. If the density is smaller than the critical value, the universe will expand indefinitely. If the density is above the critical value, the expansion will eventually halt and contraction will start. Just as a ball thrown upwards from the surface of the Earth may escape (if its velocity is sufficient) or fall back (Weinberg, 1977). Due to the uncertainty in our knowledge of the Hubble constant, and its appearing in the second power in the critical density, we don't know for sure if our universe is forever expanding or not. Before looking at the possible fates of the universe in a distant future (on a human time scale), I shall extrapolate backwards based on the known present expansion.

If the Hubble constant has remained constant, its present value tells us, that 15 billion years ago (give or take 5), all matter was concentrated in a single point. Whatever happened then - I personally do not believe in point singularities - the matter must have been close together, and it is worthwhile to look into the implications of that. If the Hubble constant has changed in time, the time needed for the expansion would change, but we would still have to explain an original state of extreme density. The only simple law that would make the extrapolation invalid would be if new matter is created or destroyed at times. In other words, the line of thinking behind the backwards extrapolation rests on the assumption that matter, or let me say energy, is conserved. Matter in the form of fermion particles may transfer their energy to light (photons) or gluons or whatever, but the total amount of energy remains constant throughout the history of the universe.

The time for which simple backwards extrapolation would imply all matter at one point (and infinite density) is called the "big bang" singularity, and I shall denote the corresponding time t=0. Gamow suggested that from t=0 to $t=10^{-4}$ sec, most energy would be in the form of light ("Gamow's fireball"). There could be Fermion particles around, such as electrons and neutrinos and quarks. Processes like

$\gamma + \gamma \leftrightarrow q + \bar{q}$ (q = quark)

$\gamma + \gamma \leftrightarrow e + e$ (e = electron)

$e^+ + e^- \leftrightarrow \nu + \bar{\nu}$ (ν = neutrino)

would be in equilibrium, that is they would happen so often, that the proportions of energy carried by each of the particle types would be fixed. It would also be possible to define a common temperature for the system, because the equilibrium assumption will make temperatures defined from the energy distributions of photons, electrons and neutrinos identical. This temperature would exceed 10^{12} K (degrees Kelvin).

At this temperature, it is possible for the quarks to form protons (p) and neutrons (n), which would interact with the leptons by processes such as

$e^+ + n \leftrightarrow p + \bar{\nu}_e$

$p + e^- \leftrightarrow n + \nu_e$

$n \leftrightarrow p + e^- + \bar{\nu}_e$

If all the fermion particles were formed from photons, there should be an equal number of particles and anti-particles, and if this was so at the time of the big bang, there should still today be an equal number of particles and antiparticles. This is not the case: the present universe has a clear preponderance of particles. I shall return to this problem and give a possible explanation below, in connection with the discussion of black holes.

When time has reached about $t = 5\times10^{-3}$ sec, the neutrinos formed by the above processes are able to escape without experiencing any further interaction. Then the neutrino temperature no longer equals that of the other particles ($T \sim 10^{11}$ K). At t = 1 sec, the neutrons no longer experience enough collisions to stay in equilibrium with protons, so the ratio of neutron and proton abundances freezes at the value it has at that moment. This equilibrium value is determined by the proton-neutron mass difference (Table 8.2) and the last temperature ($T = 10^{10}$ K), for which there was equilibrium. The resulting neutron-proton ratio is about 0.2 (Peebles, 1971; Yang *et al.*, 1979).

For the next 180 seconds, the dominant reactions are nuclear fusion processes,

$n = {}^1_1H \leftrightarrow {}^2_1H + \gamma$ (*where* ${}^1_1H = p$; *upper index is atomic number, lower is proton number*)

${}^2_1H + {}^2_1H \rightarrow {}^3_1H + {}^1_1H$ (${}^2_1H$ *is deuterium,* ${}^2_1H$ *is tritium*)

${}^2_1H + {}^2_1H \rightarrow {}^3_2He + n$

${}^3_1H + {}^1_1H \rightarrow {}^4_2He + \gamma$ (${}^4_2He$ *is stable helium*)

${}^3_2He + n \rightarrow {}^4_2He + \gamma$

If all these reactions were completed at once, the relative abundance of helium-4 would be 0.33 (two mass units of helium-4 for each mass unit of neutrons, thus X(He-4) = 2X(n)/(X(p)+X(n)) with X(n)/X(p) = 0.2, if the abundances are denoted X). In reality, this does hot hold because the neutron is unstable, so that around 30% of the neutrons have decayed during the time interval t = l sec to t = 3 minutes. Taking this into account, the

prediction for the helium-4 abundance at t = 3 minutes becomes X(He-4) = 0.25.

Little of this helium is lost by subsequent reactions, because it so happens, that there are no stable nuclei with 5 or 8 nucleons. Therefore, those nuclei formed by fusion of He-4, protons and neutrons will soon again decay to lighter species. This means that the production of new nuclei essentially ceases three minutes after the big bang. Not until new fusion processes start in stars will there be produced more helium. The experimental evidence now available points to helium abundances of 25% or more, which lends strong support to the main aspects of the big bang theory.

What happens next in our expanding universe? For the following 300000 years, the universe consists of light and the particles already produced. All take part in frequent interactions, and a common temperature can be defined. This common temperature decreases as the universe expands, and at t = 300000 years it has reached about 6000 K. At this temperature, particles and photons no longer collide frequently enough to maintain a common temperature. Thus two different temperature curves follow. The temperature of light is decreased only due to the uniformly decreasing density of photons. Today, at t = 1.5×10^{10} years, this temperature is 2.7 K. The observation of this "cosmic background radiation" lends further support to the theory (Penzias and Wilson, 1965).

The matter part, consisting of light nuclei and electrons, cooled more rapidly after t = 300000 y, because nuclei and electrons now could form atoms without immediate disruption by thermal or photon-induced ionisation. Matter took the form of clouds of dust and gases, sometimes partly ionised. Such clouds can still be observed. Due to local fluctuations in the density of matter constituents, the universe no longer had a completely uniform distribution of matter (although on a large scale it was still uniform). Gravitational forces tried to pull the mass particles together, but local contraction increased the local temperature of the gases, causing their pressure to rise and to counteract the contraction. However, if the initial density anomali was sufficient for reaching the temperature at which molecules would break up and atoms become ionised again (expelling electrons), then the pressure forces would disappear and the gravitational contraction ("collapse") take place very rapidly. In the matter of 20 years, the density can increase by a factor of about a million. This is identified with the "birth" of a main sequence star (such as our Sun). Inside the star, energy is exchanged between the surface and the interior by convection of hot gases. The surface temperature stays roughly constant, while the centre temperature continues to rise as the star further contracts, until it reaches about 10^7 K. At this temperature, nuclear fusion processes involving hydrogen are capable of starting. The hydrogen burning leads to outward transport of radiation (the star starts to shine!), and after some time - about 50 million years for our Sun - an equilibrium situation is reached with no further contraction (Figure 9.1). Temperature and pressure gradients have formed, which exactly balance the gravitational force. The several nuclear reactions participating in maintaining this equilibrium are listed in section 9.1 later in this chapter. As far as the energy production is concerned, the reactions may be summed up to give

$$4p \rightarrow {}^4_2He + 2e^+ + 2\nu_e + 26.2\ \text{MeV} \quad (1\ \text{MeV} = 1.6\times10^{-13}\ \text{J}).$$

The stability will continue until most of the hydrogen has been transformed into helium. When this situation is reached, there is again nothing to prevent a contraction. Hydrogen

will first be depleted in the centre of the star, where contraction and temperature rise will start. Hydrogen will still burn in a shell around the centre. Because of the higher centre temperature, more heat must be transported to the surface. This causes the exterior of the star to find a new equilibrium state with lower density. The star thus expands, its radius may increase fifty times and the surface temperature drops some 1500 K. The star has become a "red giant". The increased energy transport typically makes the luminosity 1000 times greater. A star with mass equal to the sun will reach the red giant stage about 10^{10} years after its birth (our Sun is halfway there!) (Iben, 1970).

As the centre temperature rises, it eventually become high enough to sustain enough collisions of more than two helium nuclei. Then the helium to carbon fusion reaction becomes possible (at $T = 10^8$ K),

$$3\,{}^{4}_{2}\mathrm{He} \rightarrow {}^{12}_{6}\mathrm{C} + \gamma + 7.28\ \mathrm{MeV}$$

Oxygen would also be formed by the reaction

$${}^{12}_{6}\mathrm{C} + {}^{4}_{2}\mathrm{He} \rightarrow {}^{16}_{8}\mathrm{O} + \gamma + 3.2\ \mathrm{MeV}$$

After some instability, a stable situation of helium fusion in the centre of the star and still hydrogen 'burning' in a shell around the core is maintained for a period of about 100 million years. The helium has then become depleted from the core, and new contraction plus temperature rise follows. At $T = 10^9$ K, carbon and oxygen fusion, by which Mg-24 and Si-28 are formed, will take place in the centre, while helium starts to fuse in an inner shell, and hydrogen now in a second shell.

When the temperature is above 3×10^9 K, photons hitting Mg or Si nuclei may knock out helium nuclei, which may then be captured by other Mg and Si nuclei, thereby leading to the formation of a range of nuclei with mass numbers between 30 and 60 (that is proton plus neutron numbers equal to atomic number ranging from 30 to 60).

If temperatures op 4×10^9 K and densities around 10^{11} kg/m^3 persisted for extended periods of time, processes such as

$$2\,{}^{28}_{14}\mathrm{Si} \rightarrow {}^{56}_{26}\mathrm{Fe} + 2e^{+} + 2\,\nu_e + 17.6\ \mathrm{MeV}$$

would eventually transform all the lighter elements to iron (and no higher atomic numbers). This is because iron-56 is the most stable nucleus there is. Lighter ones can gain energy (per nucleon) by fusion processes, and heavier ones by fission processes.

It is possible that some stars end their life as cold iron stars. There would be little radiation, because the iron nuclei would become packed in a dense but regular pattern and would cool slowly.

More likely, the shell structure and the diversity of nuclear reactions will lead to instabilities and temperature variations across the star. If the temperature in any region should reach 7×10^9 K or more, iron would "melt" by the reaction

$$\mathrm{Fe} \rightarrow 13\mathrm{He} + 4\mathrm{n} - 124.3\ \mathrm{MeV},$$

This would lead to a rapid collapse of the star, and any element lighter than iron will start to burn explosively. This is believed to be what happens in a supernova outburst (Pines, 1980). Fractions of the stellar matter become ejected into interstellar space, proba-

bly as the result of large numbers of neutrinos formed in the central region.

What happens then depends on the mass of the material left behind after the supernova outburst. If it is small (below 1.5 solar masses), the matter will contract to a faint "white dwarf" star, like the pure iron star but not necessarily made of iron.

If the mass of the remainder is between 1.5 and 3 solar masses, the contraction will be able to make electrons and protons coalesce to form neutrons,

$$p + e^- \rightarrow n + \nu_e$$

The star has become a neutron star. The density of the entire neutron star is similar to that inside a nucleus. The violent formation of the neutron star often entails that a rotation and a magnetic field are associated with the matter. After the contraction, these properties cause the neutron star to have a very rapid rotation (because angular momentum is conserved) and a very large magnetic field (1-1000 radians per second and 10^{14} to 10^{17} tesla). Such a neutron star will form a co-rotating magnetosphere of charged particles ripped off from its surface, and these accelerated, charged particles will emit electromagnetic radiation with a pulse period equal to that of rotation. Such stars are called pulsars (Pines, 1980). Other neutron stars emit intense but unpulsed X-rays. These are electromagnetic waves of frequencies similar to those emitted by electrons accelerated in cyclotrons. The neutrons in many neutron stars are believed to be in a superfluid state (a phase transition to superfluidity could have resulted from forces favouring pairs of neutrons, forces known from studies of atomic nuclei (Bohr *et al.*, 1958)). Evidence that some parts of neutron stars (the interior) may be superfluid and other parts (the crust) not, comes from the sudden speed-ups of rotation observed in neutron stars in the Crab and Vela nebulae. Such response would be expected from slow equilibration approach in a two-phase system.

If the mass of the supernova remnants exceeds some 3 solar masses, the gravitational collapse would continue for the neutrons, and the density will become so high that the gravitational escape velocity exceeds the velocity of light. According to the theory of relativity, this means that nothing can carry energy out through the surface of the star: it has become a "black hole".

The escape velocity is a concept well known from satellite launching. Anything below this velocity will fall back towards the surface. For situation leading to the formation of a black hole, it is of course necessary to use the general theory of relativity to calculate an escape velocity, in order to ask to which size the star should collapse for the escape velocity to become equal to the velocity of light in vacuum.

The result is that there is a definite radius for a black hole, defined as the maximum distance from the cent that a photon will be able to reach, it is called the Schwarzschild radius and is given by

$$R = 2GM/c^2$$

where G is Newton's constant of gravitation, M the mass of the black hole and c the velocity of light in vacuum. For M equal to ten solar masses, the Schwartzschild radius becomes about 30 kilometres.

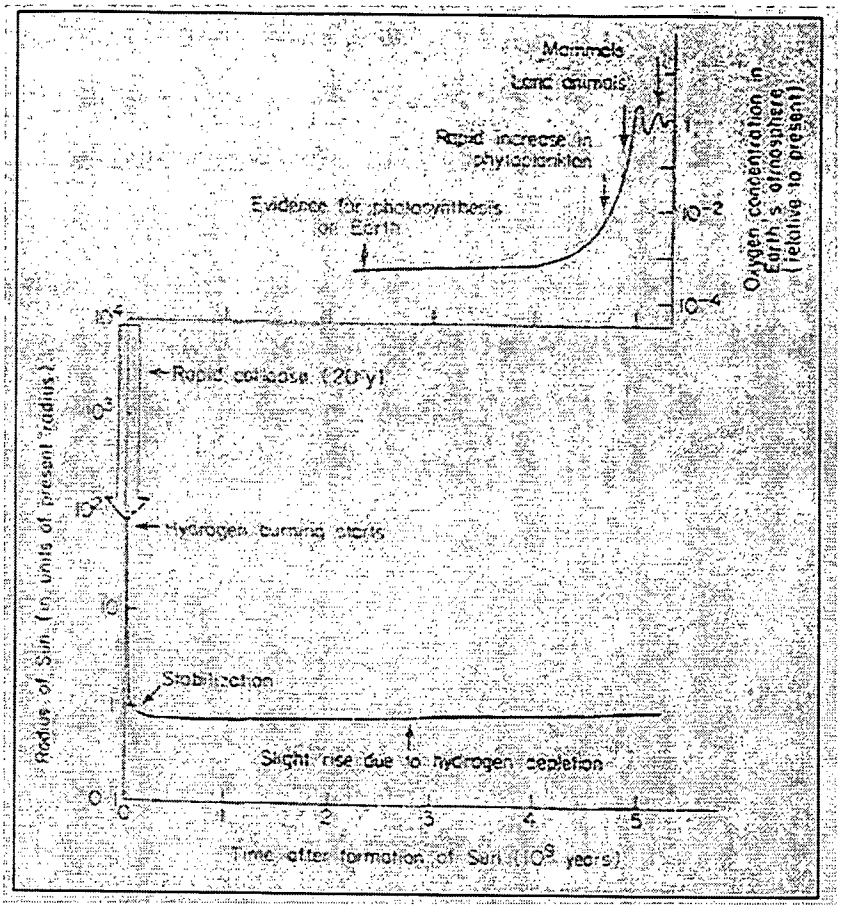


Figure 9.1. Variations in solar radius (lower part), and the building up of oxygen in the Earth's atmosphere (upper part). Between oxygen concentrations of 1/100 and 1/10, the ozone shield is building up and the ultraviolet radiation is gradually reduced, whereby first phytoplankton development and later life on land are being made possible (Sørensen, 1979).

Formation of black holes by gravitational collapse leads to a considerable loss of information, because many of the properties of the matter forming the black hole will be lost. Only three quantities remain known some time after the collapse: the total mass, the total angular momentum and the total electric charge (Hawking, 1977). Other previous knowledge, such as the total number of baryons and leptons (proton- or electron-like fermions) will be lost. This means that a black hole would not know if a particle or an anti-particle fell into it. This open up for the only presently accepted way of changing the ratio of particles and anti-particles.

What may happen is that a particle or a field present just outside the Schwarzschild radius may form a virtual particle-antiparticle pair, and one member may fall into the black hole before having time to recombine with the other one. As we have seen in Chapter 8 (for example in Fig. 8.8), virtual pairs of particles and antiparticles are formed in great numbers around any quark or electron-like particle. This being a random process, one might think that the chance of losing particles and antiparticles into the black hole were identical, and that for this reason the number of particles and antiparticles lost into the hole would be identical. This may indeed be true for neutral particles, but not for pairs with two electrically charged members (plus and minus charge), provided that the black hole itself is charged. In this case, it will preferably attract charges opposite to its own, and their chance of being captured is higher than that of their antiparticles with charges of the same sign as the black hole. The particles capable of escaping from the region near

the black hole surface are thus predominantly of the same charge sign as the black hole.

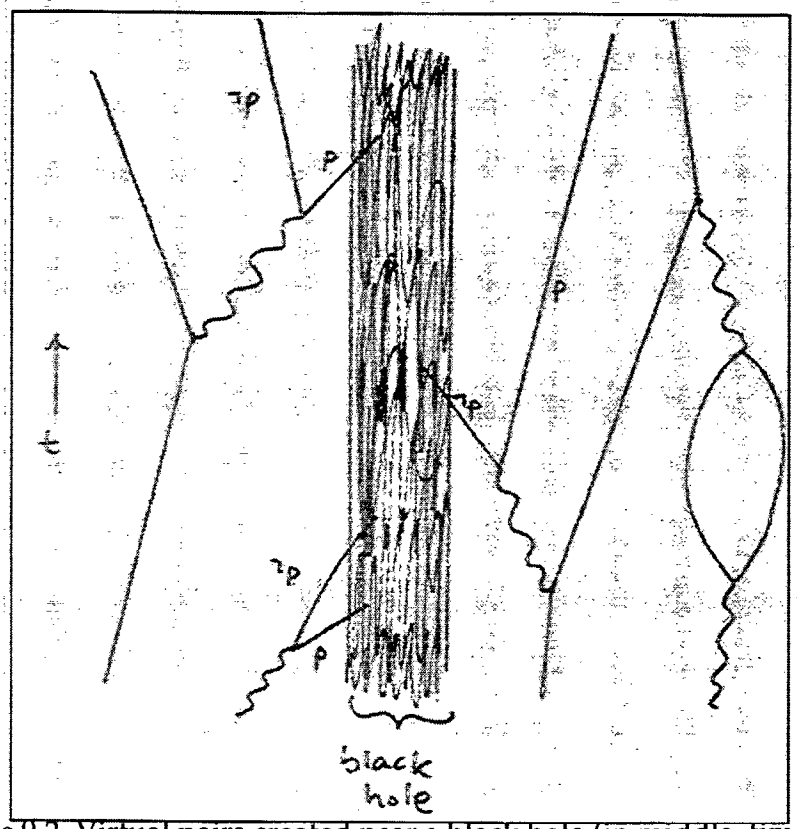


Figure 9.2. Virtual pairs created near a black hole (in middle, time increases upwards) may lose one member to the black hole, which then appears to be emitting the other member of the pair from its interior.

The black hole seems to prefer to "emit" particles with the same sign of charge as itself. For a negative black hole, this means that electrons are emitted more often than positrons, $\bar{u}$-quarks ($\bar{u}$ is the shorthand for anti-u) more often than u-quarks, and so on. This may then serve as a mechanism for changing the ratio of particles and antiparticles in the part of the universe near a black hole.

The most conclusive evidence for the existence of black holes is the radiation received from a double star in the X-ray source Cygnus X-1. The two stars are rotating around each other, but only one is visible. From the period of the radiation received, it can be calculated that the mass of the invisible star is eight solar masses. This excludes white dwarfs and neutron stars, leaving only one known possibility: that the invisible star is a black hole. That would also explain the X-ray emission, because volatile matter from the visible star would emit X-rays, as it is accelerated and sucked into the black hole.

One may now ask if a black hole will exist forever, or if it can transform into something else. The question has been answered by Stephen Hawking (1975) and is deeply related to the virtual particle-antiparticle phenomena discussed above.

However, the easiest way of illustrating the emission from black holes may be in analogy to well-known quantum tunnelling effects. A quantum system, say a particle, confined

behind a potential barrier of finite height and finite width, may nevertheless penetrate the barrier and appear on the outside. This is what happens in fission of uranium nuclei. Although energy can be gained by pulling the fragments apart, there is an initial barrier created by the nuclear forces, which must be passed in order to get from the ground state of U-235 to any fragmented state. The possibility of tunnelling through a potential barrier is a straightforward consequence of the Schrödinger equation of quantum mechanics. The black hole is incapable of emitting anything according the general relativistic mechanics. But when quantum theory is combined with relativity, the inability to escape from the interior of black holes gets modified, and Hawking was able to show, that all kinds of elementary particles could be emitted, and that the distribution of their energies was as for thermal radiation from a black body of a temperature T given by

$$T = hc^3/(16\pi^2 kGM),$$

Where h is Planck's constant, k Boltzmann's constant (1.38×10^{-23} J/K), c the velocity of light in vacuum (3×10^8 m/s), G Newton's constant of gravitation (6.67×10^{11} $m^3/kg/s^2$) and M the mass of the black hole.

This answers the question that naturally arises in connection with the virtual pair interpretation. Since virtual pairs need not fulfil the principle of energy conservation (this being the definition of "virtual"), it would seem that if the black hole has swallowed one member, the other one could have any energy. Yet, to become a real particle, energy conservation has to become restored. Should the "emitted" particle suddenly change its energy, or what? Hawking's model gives a plausible answer. Recall that particle-antiparticle pairs are really a positive and a negative energy solution to the quantum theory equations. The negative energy solutions were interpreted as holes in a sea of filled quantum states, and the holes were interpreted as positive energy antiparticles. If we go back to the interpretation in terms of negative energy particles, then the virtual pairs in Fig. 9.2 consist of one positive energy particle and one negative energy particle. Assume that the negative energy particle gets sucked into the black hole. It then adds a negative amount of energy to the black hole, or in other words, the black hole loses energy. This energy loss is then the energy gain of the free particles outside the black hole. If it is the positive energy virtual particle that gets sucked into the black hole, we just use an interpretation entirely in terms of antiparticles. The particle is equivalent to a negative-energy antiparticle and it removes energy from the black hole when it is absorbed. The positive energy antiparticle becomes a free antiparticle and it gets an energy determined by insisting on energy conservation.

The Hawking temperature for a black hole like that in Cygnus X-1 is only about 10^{-7} K (which is even below the background temperature 2.7 K of the interstellar photons). Such a black hole will lose energy very slowly, and the time required for emission of all its energy is

$$\Delta t = 2\pi G^2 M^3/(hc^4),$$

which for the Cygnus X-l case is of the order of 10^{68} years. The emission is not uniform over time, because as the mass decreases due to matter already emitted, the Hawking temperature increases, and so does the emission rate. The final part of the process will be an enormous explosion, with most of the emitted particles being photons. Small black

holes would reach this stage very soon. It has been speculated that black holes the size of a proton (mass 10^{12} kg) might have been formed in the big bang. If so, their radiation, which corresponds to a Hawking temperature around 10^{11} K should be detectable at the Earth. The photons would create showers of electron-positron pairs at the top of our atmosphere, and we should be able to detect electromagnetic shock waves ("Cerenkow radiation") from the deceleration of such particles. Attempts to find such evidence of black holes from the big bang have so far been unsuccessful.

It would therefore seem, that attempts to associate the big bang origin of the present universe with some kind of black hole explosion are not warranted. Large-mass black holes have tiny temperatures, and if many small black holes of sufficient temperature had been present, we ought to be able to see traces of them now.

If the density of the present universe is such, that it will continue to expand, there will be a succession of supernova outbursts, as the stars reach the end of the nuclear reaction stages described above. The resulting dwarfs and neutron stars will presumably transform into black holes by some kind of quantum tunnelling effect, but with an extremely low probability (so that it will take very long time - maybe 10 raised to 10^{26} years (Dyson, 1979) - before half of them have become black holes). The black holes radiate matter, which may be forming new stars elsewhere, and finally should explode into photon radiation. According to Hawking's theory a minimum black hole mass exists. It is $(hc/(2\pi G))^{-1/2} = 2\times10^{-8}$ kg. When this minimum mass M(min) is reached, the hole explodes. Eventually, all objects above M(min) will become photons, wh ile dust grains below this mass may exist forever.

This lead Dyson to ask, if life in the universe could also last forever (Dyson, 1979). He defines life as a certain organisational pattern, that is a certain structure and not a certain form of matter. If life were matter-linked, it could not exist in the late states of the universe as described above. It would require a certain temperature and definite materials such as liquid water perhaps. This is excluded because the universe gets colder as it expands. So argues Dyson, but this is basically an average argument, because local conditions for life could be maintained in analogy to greenhouses with temperatures above that of the environment. However, a continuing sophistication of the "greenhouses" would be required, as the average temperature of the universe continues to decline.

Dyson favours the hypothesis of life as a certain structural organisation, because then it is not bound to present forms. When only dust particles are left in the universe, then life could organise itself as a pattern of positive and negative charges attached to such dust grains and communicating with each other by means of electromagnetic signals (Hoyle, 1957).

Dyson assumes that life may adjust to declining temperatures by correspondingly reducing the speed of life processes. He argues that the Hamiltonian operator is scale invariant, so that if it is multiplied by a factor T'/T then it describes a subjectively identical form of life, but with life processes slowed down by the factor T'/T. This hypothesis is probably wrong, since at least quantum field theories are not scale invariant (Nielsen *et al.*, 1979). Still, it may be possible to define a new suitably varying set of Hamiltonians as function of time or temperature, which can describe if not identical then at least equally

acceptable forms of life.

To be living and to maintain a never-declining level of intelligence, Dyson assumes that a creature must process at least as much information per unit of subjective time as average present humans, and estimates this as 10^{23} bits per second (one bit being defined as the information contained in a "1" or "0", or a "yes" and a "no"). By slowing down subjective time as described above, and using the minimum of energy required (by thermodynamics) to achieve the assumed rate of information processing, Dyson is able to estimate the maximum duration of intelligent life in the universe. He initially finds that it is finite, although the "life" of the universe itself is infinite. He then gets the idea to make intelligent information processing intermittent, but waste heat removal continuous. By introducing such hibernation periods of increasing lengths, Dyson succeeds in making life continue indefinitely (in subjective time units) and yet use only a finite amount of energy. There is even room for (electromagnetic) communication between such intelligent societies (on their respective dust clouds), and for maintaining a record of the history and intellectual achievements of the societies (using analog types of memory because digital ones would require at least one material particle per bit of information, and only a limited number of particles are available in the late universe).

The considerations made above are just natural extensions of the questions about life in the universe outside our planet, which have been asked throughout our history. Most questions seem to be static: is there life on other planets in our solar system, in other solar systems, on other galaxies? If yes, is that life similar to ours, or could it be entirely different? The "static" questions soon become time-dependent, when we realise that the light reaching us from the most distant galaxies has been travelling for more than ten billion years between there and here. Present knowledge of the universe makes it highly probable, that life as well as intelligent life exist numerous places in the universe. Seen in this perspective, it really isn't much of a catastrophe, if our planet should become uninhabited as a result of the nuclear holocausts that we are in the process of staging. I do not know if that is - or if it should be - of any consolation to you. One suggested explanation for the evidence that such extraterrestrial, intelligent life has not in recent years visited us on Earth is, that it exists only for such a short period (before destroying itself by nuclear wars), that it never has had time to make space travel, except possibly to nearby space objects. I am not convinced by this argument, as there would always be a tail distribution of civilisations that succeed for a long time without destroying themselves. So maybe the lack of visitors indicate that we are really a precious one-of-a-kind miracle in the entire universe.

The question of whether or not life elsewhere in the universe would be similar to that on Earth is a more difficult one. Evolution history has shown to us, that species adapt to new conditions very rapidly (in the range of at most millions of years), that grossly different kinds of creatures may evolve in matters of a few hundred million years, and that the development from scratch to the human race took less than 10 billion years. It would then seem very probable, that any life forms in other inhabited parts of the universe may look substantially different from ours, due to likely differences in the external conditions. Yet, we also expect basic similarities between life forms, if they are based on material properties (organic molecules, water, and so on). If Dyson is right in assuming that life is struc-

ture not matter, then there is no reason at all to expect life forms in different parts of the universe to show any similarity. On Earth we find organisms depending on oxygen, other ones depending on light, and so on. In other parts of the universe, structures organised in a way deserving to be called life according to Dyson's definition may be based on entirely different molecular structures and sources of energy, or not on atoms and molecules at all. Dyson and Hoyle's dust cloud inhabitants are charge patterns (they could be formed by electrons and positrons), and their only source of energy in the late universe would be radiation from black holes. So maybe our lack of observation of visitors is due to our lack of knowing what to look for.

Let me get down to Earth for a moment. The Sun being a hydrogen burning star, there is no source of elements heavier than helium. All the elements on Earth with atomic number higher than helium must therefore have come from a different source. The presence on Earth of elements up to iron clearly tells us, that the material must derive from a supernova explosion. One thus has to assume, that the material for the planets was either delivered by close encounters between the Sun and a star in its supernova phase, or that material from supernova stars has slowly been accreted from interstellar matter by our Sun.

The formation of elements above iron is possible by neutron capture. Excess neutrons are not available from nuclear reactions in stars until the multi-shell burning stages. If instabilities sometimes mix hydrogen into the helium and carbon containing shells of the star, then reactions like

$$^{12}_{6}C + {}^{1}_{1}H \rightarrow {}^{13}_{6}C + e^{+} + \nu_e$$

$$^{13}_{6}C + {}^{4}_{2}He \rightarrow {}^{16}_{8}O + n$$

will produce large numbers of neutrons. By various sequences of neutron capture and beta-decay (electron or positron emission and p to n or n to p transformation) processes, all the known heavier elements can be formed.

9.1 Stellar atmosphere theory

Attempts of understanding the structure of stellar atmospheres and their development are to a large extent based on methods of classical physics (such as fluid mechanics), but with quantum effects put in where they are important. It is also possible to understand most features of normal stars (not very dense ones such as neutron stars and black holes), by using non-relativistic approximations. For stable stars, one can assume fixed distributions of matter and quantities such as pressure and temperature. The goal is then to calculate these distributions, and thereby be able to calculate the luminosity of the star - a quantity that can be compared with measurements. Even stable stars may have regions of instability (say density variations), for instance convective zones, and in general, the evolution of stars involve many dynamic changes, as we have already seen.

A model of an evolving star must contain a description of the mechanism for nuclear energy release, for gravitational collapse, for transport of matter and radiation within the

star. The matter in stars is in most cases in the form of gases. This is the reason for using the term "stellar atmospheres" for the entire star. I shall formulate the theory of such atmospheres on the basis of a general set of transport equations, first formulated by the mathematician Euler. For any quantity A=A(**x**,t), the time variations may be expressed by the Euler equation,

(•) $\partial(\rho A)/\partial t + O(A) = S(A)$,

where A pertains to a unit volume and ρ is the density (mass per unit of volume) at the location considered. O(A) is the net outflow of the quantity A from the volume element considered, that is the outflow minus the inflow. Similarly, S(A) is the net creation of the quantity A inside the volume element considered. If there are no sources or sinks for the quantity A in the volume considered, then S(A) = 0.

Let me now look at some definite quantities A. The simplest one is to put A = 1, in which case the Euler equation deals with the time change of the density ρ itself. The outflow term O(1) in this case has to describe matter removed from the volume under consideration, and it can be non-zero only when there is a motion of matter. Let (v_i, i=1,2,3) be the velocity components of matter at the unit volume U considered. Then ρv_i is the amount of matter flowing away from U in the i'th direction (or into U if the quantity has the opposite sign), per unit of time. The three components ρv_i are the mass flows through unit areas in the directions of the co-ordinate axes. If the mass flows are the same in neighbouring volume elements, the situation is steady and there is no net outflow from U. However, if $\partial(\rho v_i)/\partial x_i$ is non-zero, there is an outflow in the i'th direction. The total outflow is the sum of outflows in the three co-ordinate directions,

$$O(1) = \sum_i \partial(\rho v_i)/\partial x_i$$

The right hand side of Euler's equation, S(1), represents matter created or destroyed in the volume element U, for example by chemical or nuclear reactions. If light is absorbed within U and gives rise to formation of an electron-positron pair contributing to the density ρ, then this is also part of S(1). The entire Euler equation for density is then

(••) A = 1 and $\partial\rho/\partial t + \sum_i \partial(\rho v_i)/\partial x_i = S(1)$ = net new density formation.

This equation is called the "continuity equation". It states that the change in density with time, for a volume element, is equal to the net inflow (minus the net outflow) from neighbouring volume elements, plus any net creation of new mass within the unit volume considered. Note that the left-hand side is *not* equal to the total time derivative

$$d\rho/dt = \partial\rho/\partial t + \sum_i (\partial\rho/\partial x_i)(dx_i/dt) = \partial\rho/\partial t + \sum_i v_i \partial\rho/\partial x_i$$

Next, I choose A as one component of the local velocity, $v_i = dx_i/dt$. The left hand side will contain two terms completely analogous to those in (••). What about the right hand side - what are the sources and sinks of velocity? Forces, of course! The forces make things accelerate or slow down. We must then expect an Euler equation of the form

$$\partial(\rho v_i)/\partial t + \sum_n \partial(\rho v_i v_n)/\partial x_n = F_i$$

The force F_i would be the sum of all forces acting on the material in the volume element U. However, since the level of description, on which it is meaningful to talk about density

and velocity of volume elements, is largely restricted to molecular or macroscopic models of the atmosphere (considering it a gas or a plasma:), then the most important forces will be the gravitational force and the molecular pressure- and viscosity-related forces. Nuclear and sub-nuclear forces may give rise to motion of individual particles, but they are not directly related to the gross motion except for special cases such as neutron stars (which I have excluded from the present discussion). So let me be explicit and write F_i as the sum of the gravitational and molecular forces.

The gravitational force can be written

$$F_i(\text{gravity}) = -G \int \rho(\mathbf{x})\rho(\mathbf{x}') (x_i' - x_i) \, |\mathbf{x}' - \mathbf{x}|^{-3} \, d\mathbf{x}'$$

and the molecular force is often expressed in terms of the stress tensor (τ_{in}; i,n = 1,2,3), which has 3×3 components, because the force on each side of the volume element can have any of the three co-ordinate directions,

$$F_i(\text{molecular}) = \Sigma_n \partial\tau_{in}/\partial x_n$$

The stress tensor partly contains a scalar term (a term independent of the co-ordinate indices i and n), which by definition is the pressure of the atmosphere in the volume element considered, and partly terms which depend on derivatives of the velocity components and on two viscosity parameters. These latter terms are zero if the atmosphere is non-viscous, that is if molecular friction forces can be neglected (see e.g. Sørensen, 1979),

$$\tau_{in} = -P \, \delta_{in} + \eta_{in}$$

Since F_i(molecular) is given as a sum over space derivative terms, just like $O(v_i)$, it is often considered as an additional outflow term and placed on the left-hand side of the Euler equation. One can then say that it represents molecular scale flow out of U, or "diffusion" from U, which is then added to the macroscopic flow out of U. On the right-hand side are then only "external" forces,

(¤) $\quad A = v_i$ and $\partial(\rho v_i)/\partial t + \Sigma_n \partial(\rho v_i v_n)/\partial x_n + \partial P/\partial x_i - \Sigma_n \partial\eta_{in}/\partial x_n = F_i(\text{gravity})$

This is the equation of motion for a continuous medium, corresponding to Newton's equation of motion for particles. If the medium is a gas, the study of solutions to this equation is often called "aerodynamics", if it is a fluid it may be called "hydrodynamics" or "fluid mechanics".

Solutions to (¤) may exhibit all the complexity we expect for a gas or a fluid: there may be regions of turbulence, other regions of ordered (laminar) flow, and large scale eddies (circular flow patterns). It is therefore of interest to determine the gross behaviour and possibly avoid detailed calculation of small-scale deviations. This is done by some kind of averaging.

One may consider time averages,

$$(A)_0 = \frac{1}{\Delta t}\int_{t_1}^{t_1+\Delta t} A \, dt$$

and density weighted averages,

$$\langle A \rangle = (\rho A)_0 / (\rho)_0$$

whereby a quantity a may be written as one of the averages plus a remainder,

$$A = \langle A \rangle + A' = (A)_0 + A''$$

Introducing $v_i = \langle v_i \rangle + v_i'$ in (¤) and then taking time averages of each term, $(...)_0$, one gets

(¤¤) $$\partial ((\rho)_0 \langle v_i \rangle)/\partial t + \Sigma_n \partial((\rho)_0 \langle v_i \rangle \langle v_n \rangle)+(\rho v_i' v_n')_0)/\partial x_n = (F_i(\text{molecular})+F_i(\text{gravity}))_0$$

where several terms have been eliminated by using the relation $(\rho A')_0 = 0$. In fact, except for one term, (¤¤) is an equation only dealing with the averaged quantities. It therefore describes the gross motion of the atmosphere, and it tells us, that the gross motion cannot be determined entirely without knowing the microscopic motion, which appears in the coupling term

$$-F_i(\text{turbulence}) = \Sigma_n \partial(\rho v_i' v_n')_0)/\partial x_n$$

The turbulent terms play an important role in some regions of stellar atmospheres, where convective transport, that is long range motion of matter due to turbulent processes, is the dominant mode of motion (usually near the centre and near the surface of the star). A detailed calculation of these effects is rarely possible, and instead the terms F_i(turbulence) are parametrised. By that is meant, that they are replaced by a simplified expression with only a few unknown parameters, and these parameters are then left to be determined empirically. The standard approximation is to give F_i(turbulence) the same form as the molecular friction terms F_i(molecular), but to replace the viscosity parameters by so-called diffusion parameters. The viscosity parameters may be interpreted as telling how far a piece of matter on average moves before being held back by molecular friction forces, and the diffusion parameters similarly tells how far a piece of matter on average travels before being affected by the turbulent mechanisms described by F_i(turbulence). The diffusion parameters are typically several orders of magnitude larger than the corresponding viscosity parameters.

I now return to the general Euler equation, where I insert the temperature T for A. In order to find the sources and sinks for temperature, one should recall that flow of mass with a temperature is equivalent to heat flow. In other words, we should be looking for sources and sinks of heat. They are chemical and nuclear reactions releasing or consuming heat (the S(T) terms), and furthermore terms describing the emission and absorption of thermal radiation. The latter terms may be considered as outflows arid inflows of temperature (that is of heat), and may be included as O(T) terms on the left hand side:

(§) $$\partial ((\rho)_0 \langle T \rangle)/\partial t + \Sigma_n \partial((\rho)_0 \langle T \rangle \langle v_n \rangle)+(\rho T' v_n')_0)/\partial x_n + (O(T))_0 = (S(T))_0$$

where

$$O(T) = C^{-1} (E + V + R - A)$$

contains terms connected with expansion of the volume element considered (the E-term depending on pressure and density; it may be derived by thermodynamic considerations of the pressure and volume change caused by heating), terms connected with viscosity (the V-term representing heat formed by molecular friction), and terms associated with radiation from (R) or absorption by (A) the volume element considered. E may be derived from thermodynamic theory, V from statistical mechanics and the Bolzmann-Planck laws for thermal radiation and absorption from quantum statistical theory. These theories

will be briefly discussed in Chapter 12.

The common factor C^1 is one over a heat capacity, introduced in order to reconcile units. In choosing the numerical value of C, it should be made clear whether the equation (§) is to be evaluated per unit volume or per unit mass. The heat capacity takes on different values if either pressure or volume is forced to remain constant, according to thermodynamic theory. Therefore a properly varying C has to be used in the general situation where neither pressure or volumes can be considered constant.

The three Euler equations considered so far often constitute a reasonably closed model, that is, one that contains relatively few "external" parameters to be derived from somewhere else. The main variables would be the velocities v_i, The density ρ, the pressure P and the temperature T. Diffusion and viscous terms may be parametrised, and the heat added by nuclear fusion processes could be estimated externally. There is still one more variable than equations, but the quantities P, T and ρ must be related by thermodynamics. Often it is possible simply to consider the stellar atmosphere as an ideal gas, in which case

$$P = (R/\mu)\,\rho T,$$

where the constants are R = 8.32 J/K/mole (the "gas constant") and μ the mean molecular weight, that is the mass per mole of free particles in the gas constituting the atmosphere in question (for example electrons and protons in case of an ionised hydrogen gas).

If the constituents of the atmosphere are changing, for instance if the long-term conversion of hydrogen to helium is to be modelled, then Euler equations for the abundances should be added. This means putting A=X(H), A=X(He) and so on, for all the important abundances X. Each gives an equation similar to the continuity equation, but with a source term derived from the nuclear reaction schemes leading to the disappearance of some elements and formation of others. The energy release (or absorption) by the reactions are still to be kept track of in (§).

In the Sun, the nuclear processes involved in hydrogen to helium fusion are

(1a, 99.75% of the cases)	$^1_1H + {}^1_1H \rightarrow {}^2_1H + e^+ + \nu$
(1b, 0.25% of the cases)	$^1_1H + e^- + {}^1_1H \rightarrow {}^2_1H + \nu$
(2)	$^2_1H + {}^1_1H \rightarrow {}^3_2He + \gamma$
(3a, endpoint in 91% of cases)	$^3_2He + {}^3_2He \rightarrow {}^4_2He + 2\,{}^1_1H$
(3b, 9% of cases)	$^3_2He + {}^4_2He \rightarrow {}^7_4Be + \gamma$
(4a, 8.9% of cases)	$^7_4Be + e^- \rightarrow {}^7_3Li + \nu$
(5a, endpoint)	$^7_3Li + {}^1_1H \rightarrow 2\,{}^4_2He$
(4b, 0.1% of cases)	$^7_4Be + {}^1_1H \rightarrow {}^8_5B + \gamma$
(5b)	$^8_5B \rightarrow {}^8_4Be + e^+ + \nu$
(6, endpoint after 5b)	$^8_4Be \rightarrow 2\,{}^4_2He$

The calculation of stellar atmospheres from the Euler equations will be much simplified,

if one can assume the atmospheres to have spherical symmetry. In that case, the three space co-ordinates get replaced by a radial co-ordinate r, and quantities such as ρ, P, T become functions of r only (and t unless a stationary situation is considered). The gravitational force gets a much simpler form, because a mass element located at the distance r from the centre feels a gravitational attraction as if it only interacted with a single mass, located in the centre and equal to the integral over all mass inside the shell with radius r, $M(r) = \int 4\pi (r')^2 \rho dr'$, integrated from 0 to r,

$$F(\text{gravity}) = -G\, M(r)\, \rho(r) / r^2$$

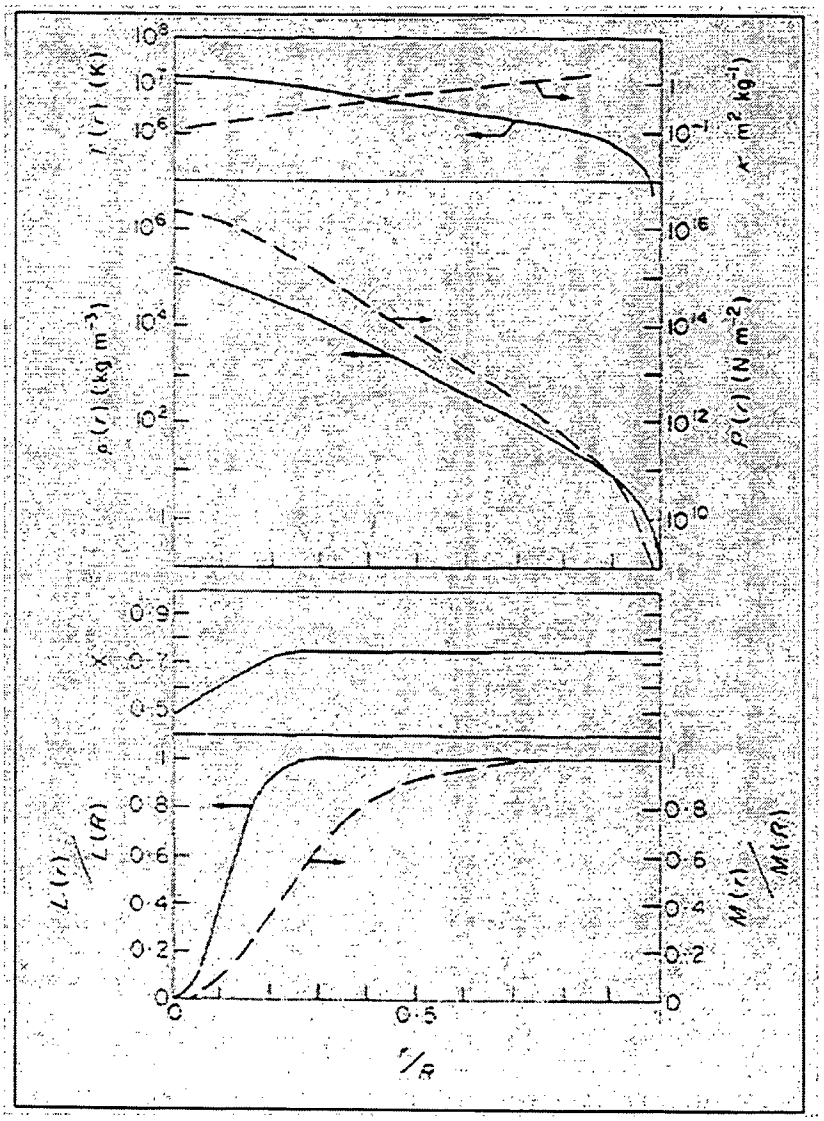


Fig. 9.3. Calculated radial variation in various properties for the Sun (Schwarzschild, 1958, cf. Sørensen, 1979): temperature (T), density (ρ), hydrogen fraction (X), luminosity (L), mass within shell of radius r (M), pressure (P) and opacity (κ).

For a star in equilibrium, radial velocities v(r) should be zero (excepting regions of turbulence, but even in such cases, averaging over sufficiently large regions will produce zero mean radial velocity). This means that the terms containing <v> in (¤¤) may be neglected, if there is no turbulence, and viscous terms are neglected, the only remaining terms are

$$dP(r) / dr = F(\text{gravity})$$

This equation tells that the pressure gradient must balance the force of gravity in every point. Together with the temperature equation and auxiliary equations explaining how to calculate the luminosity L(r) as function of r from the light emission theory (Planck's law, depending on T) and light absorption theory (often parametrised in terms of an opacity function $\kappa(r)$), a closed set of equations is obtained. A nuclear energy production function S(T) has to be generated from known or estimated reaction rates (some of the reactions listed above can be investigated in laboratories on Earth). One example of an early equilibrium calculation for the Sun is given in Fig. 9.3.

9.2 General theory of relativity and gravitation

According to the general principle of relativity, a field of gravity is entirely equivalent to an accelerated motion (see Fig. 9.4). The gravitational forces may therefore be treated by postulating suitable accelerations of the co-ordinate systems following different objects, relative to each other. One may say that by going to suitable co-ordinate systems, the gravitational forces may be "transformed away". The equivalence between gravity and acceleration is in fact already underlying classical physics, by its assumption that inertial mass (the one that multiplies acceleration in the equation of motion) is equal to gravitational mass (the one appearing in the expression for gravitational attraction (Einstein, 1921).

If there are many masses distributed in the universe, it is unlikely that all the gravitational forces could become "transformed away" simultaneously. In other words, we could only expect this possibility to work locally: for each point in the universe, there is a co-ordinate transformation which makes the gravitational forces disappear locally (around the point), this means that the special theory of relativity, which disregards acceleration (or gravitational forces, which is the same according to the principle of relativity), is valid locally, but not globally.

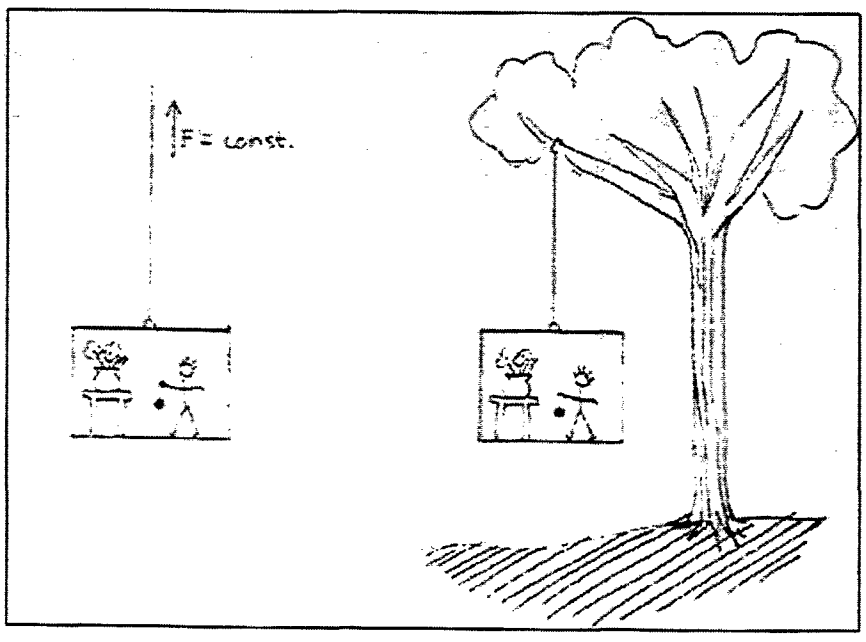


Fig. 9.4. Left: the box is in empty space, pulled in one direction with a constant acceleration. Right: the box is in the gravitational field of a large mass (say a planet). If the acceleration on the left times the mass of the box equals the gravitational force on the right, the man in the box will not be able to distinguish the two cases.

I now consider the motion of a particle within such a local region (Fig. 9.5). Since the velocity must be smaller than c, the time-space track of the particle (called its "world line") must be inside the cone depicted in Fig. 9.6, and only particles of zero mass (such as photons) may move along the sides of the cone, which is therefore called the "light cone".

Small changes in the time measured with a clock following the particle may be calculated using the Lorentz transformation (section 7.1) for the time co-ordinate. This is possible only for small changes in the intrinsic time, τ, (called the "proper time" for the particle), because the special theory of relativity is valid only locally, as stated above,

$$d\tau = (1-v^2/c^2)^{1/2}\, dt$$

The velocity of the special Lorentz transformation is $v = dx/dt$ and for a general case $\mathbf{v} = (dx/dt, dy/dt, dz/dt)$, so that

$$(d\tau)^2 c^2 = (dt)^2 c^2 - (dx)^2 - (dy)^2 - (dz)^2$$

As in Chapter 8, this may be written in four-dimensional notation, we here define $x_4 = ct$ rather than ict, in order to avoid complex numbers. We then have to accept, that the terms on the right hand side of the above expression have different signs. The equation may be rewritten as

$$c^2 (d\tau)^2 = -\sum_{in} g_{in}\, dx_i\, dx_n, \quad i,n = 1,2,3,4$$

where

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

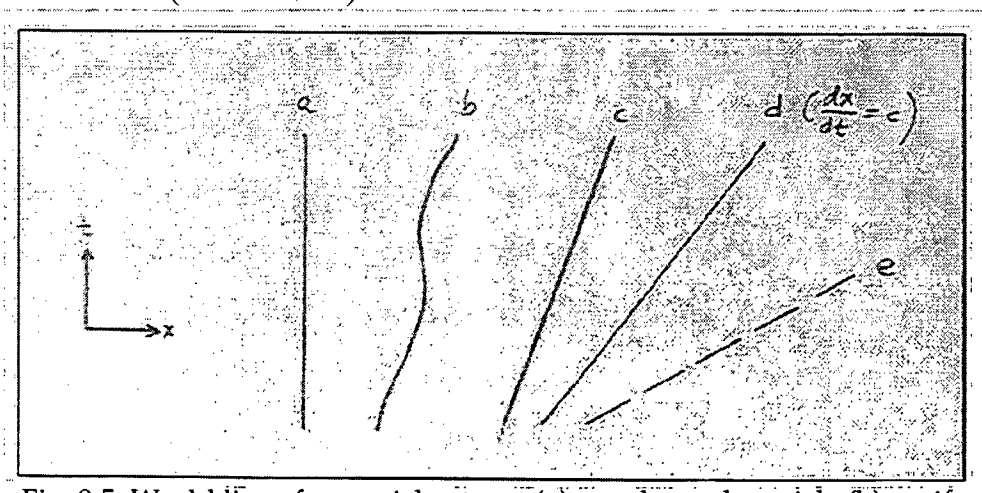


Fig. 9.5. World lines for particle at rest (a), accelerated particle (b), uniformly moving particle (c) and photon (d). (e) is a "forbidden" world line, corresponding to a velocity above c.

Defining a four-dimensional distance between points, s, as being equal to ic times the proper time change for small distances,

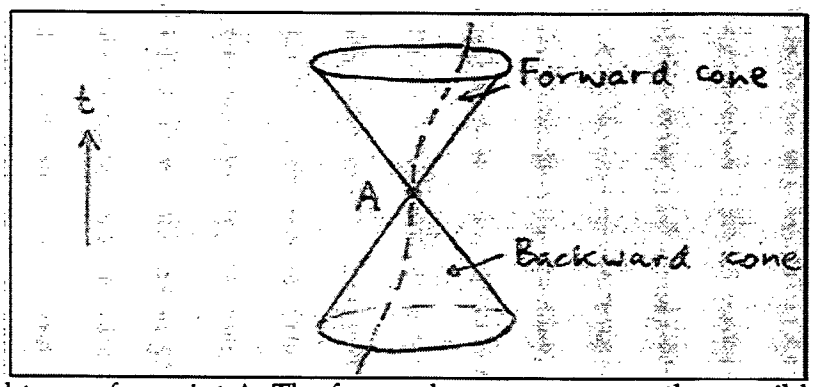


Fig. 9.6. Light cone for point A. The forward cone represents the possible futures of a particle at A, the backward cone its possible pasts. The path of a light signal emitted from a will lie on the surface of the forward cone. The inclination of the cone is the same for all light cones in the special theory of relativity.

$$(ds)^2 = -(c\,d\tau)^2$$

it is seen that $[g_{in}]$ may be interpreted as representing the metric of the space-time structure. A *metric* is a set of rules for measuring distances. In one dimension it could be a measuring tape with which to compare lengths. The metric of the special relativity theory is the 4×4 tensor (or matrix) g, which tells us which coefficients to multiply the different increments $dx_i dx_n$ by,

$$(ds)^2 = \Sigma_{in}\, g_{in}\, dx_i\, dx_n$$

The above form of the matrix $[g_{in}]$ is valid for the cartesian co-ordinate system following the point of the particle considered. If a relativistic particle is to be described in a different co-ordinate system, for example one that is fixed and independent of how the particles moves around, then the g_{in}'s change. However, the requirement of invariance under any co-ordinate transformation tells us that the general form of the expression for $(ds)^2$ is unchanged, only the values of the g_{in}'s change. In a general co-ordinate system, in which particles (or matter) perform accelerated motion, $[g_{in}]$ may have elements outside the diagonal, and the elements may no longer be simple numbers such as 1 or -1.

Consider now a particle not affected by any force in one co-ordinate system, but view the particle in a second co-ordinate system moving with a small, constant acceleration relative to the first one. One can then derive (Weinberg, 1972) the leading order expression for the acceleration of the particle as seen in the second co-ordinate system,

$$d^2x_i/dt^2 = \tfrac{1}{2}\,\partial g_{44} / \partial x_i, \quad i=1,2,3$$

This equation describes the left hand side of Fig. 9.4 for small velocity and small acceleration. According to the equivalence principle, adding a gravitational force in the first co-ordinate- system must give an identical description. Due to the assumption of slow velocities and accelerations, the Newtonian expression for the gravitational force can be used,

$$d^2x_i/dt^2 = -GM/r^2 = -\partial\varphi/\partial x_i, \quad i=1,2,3$$

where a potential function φ has been defined in terms of the gravitational constant and

the mass of and distance to the object responsible for the gravitational pull,

$$\varphi = -GM/r, \text{ with } r = (x^2 + y^2 + z^2)^{1/2}$$

If the two descriptions should be identical, then

$$-2\varphi = g_{44} + \text{constant}$$

Since g_{44} should be -1 (see above) if there is no gravitational potential ($\varphi = 0$), then the constant must be 1 and thus

$$g_{44} = -1 - 2\varphi$$

If the gravitation field is not due to a point mass M, but to a uniform mass density ρ extending (at least) out to the distance r, then according to the discussion in section 9.1

$$\partial\varphi / \partial x_i = G\, M(r)\, x_i / r^3$$

and

$$\Sigma_i\, \partial^2\varphi / \partial x_i^2 = G \Sigma_i\, (\partial M(r) / \partial x_i)\, x_i / r^3 = G \Sigma_i\, 4\pi\rho r^2\, (x_i/r)\, (x_i/r^3) = 4\pi G \rho$$

This implies that for the accelerated system, g_{44} can be determined from the equation

$$\Sigma_i\, \partial^2 g_{44} / \partial x_i^2 = -2 \Sigma_i\, \partial^2\varphi / \partial x_i^2 = -8\pi G \rho$$

This equation may be interpreted as saying, that the (i,n) = (4,4) component of a tensor G_{in}, with components generally equal to second order derivatives of g_{in}, is equal to the density times some constant. If we can find a 4×4 tensor having ρ as its (4,4) component, the invariance of physical laws under general relativity transformations - which may change g_{44} into any other g_{in} - tells us that the relation given above is valid for all 16 components (i,n).

Now ρ is indeed member of a 4×4 tensor, called the energy-momentum tensor T_{in}. For a particle, T is simply a generalisation of the classical kinetic energy expression to include a fourth index,

$$T_{in} = \rho U_i\, U_n$$

with

$$(U_1, U_2, U_3, U_4) = \gamma\, (v_1, v_2, v_3, c)$$

This form of the four-velocity U_i is dictated by the Lorentz transformation (see section 7.1, where γ is also defined). For a continuous fluid (but without viscous (friction) forces), T_{in} also contains terms depending on the proper pressure p (defined in analogy to proper mass and proper density) (Møller, 1952),

$$T_{in} = (\rho + p/c^2)\, U_i U_n + p\, \delta_{in}$$

These expressions for T_{in} are valid in standard co-ordinate systems of the kind used in the special theory of relativity (x, y, z, ct). In general co-ordinate systems, the transformations determined by g_{in} should be applied. However, in the form given above, we readily recognise T_{44} as ρc^2, that is mass (density) times c squared, which is the rest energy, that is the energy of the particle at rest. The i,n = i,2,3 components are (2 times) the usual kinetic energy expressions $\rho v_i v_n$, and for i or n equal to 4, the classical momentum compo-

nents are found. All are multiplied by γ, because $\gamma\rho$ is the mass density of a particle not at rest, as one can see by using the requirement of invariance under Lorentz transformation on the classical equations of motion. These are valid locally but the conclusions based on invariance requirements are valid generally as argued before.

The fundamental equation expressing the relationship between accelerations and gravitational forces can now be written (Einstein, 1915)

$(\blacklozenge)$ $\quad G_{in} = -8\pi G\, T_{in}$

This is the field equation of the general theory of relativity and gravitation. The Einstein tensor G_{in} depends on up to second order derivatives of the metric tensor g_{in}. That is, it depends only on the geometry of the space. In the simple example considered above, G_{in} was simply c^2 times the second derivatives of g_{in}. On the right hand side of the field equation is the energy-momentum tensor, which depends on the matter variables ρ, p and U_i, and in general also on g_{in}. In the simple case above, there was no dependence on g_{in}. The only parameters in the theory are c and G, Newton's constant of gravitation. The field equations in their general form have not just one but many possible solutions. Different classes of solutions are obtained by adding assumptions to the physical model of the universe. Such an assumption could be isotrophy (that the universe looks the same in any direction). This would seem a proper assumption for spherically symmetric systems, and may serve as a good approximation for stars. In this case it is simplest to use spherical co-ordinates (r, θ, φ) instead of (x, y, z). The distance from the centre of symmetry is r, while θ and φ are the polar and azimuth angles well known from geography on the Earth sphere (φ is longitude, θ is $\pi/2$ minus Northern latitude), the metric may in this case be written (Schwarzschild, 1916)

$$(ds)^2 = f(r)^{-1}\,(dr)^2 + (r\,d\theta)^2 + (r\sin\theta\, d\varphi)^2 - f(r)\,(cdt)^2$$

where

$$f(r) = 1 - 2MG/(rc^2)$$

Since f(r) appears in the denominator, proper solutions would have to be sought either inside or outside the Schwarzschild radius $r=2MG/c^2$.

For models of the entire universe, not only isotropy but also homogeneity is often assumed. This may be stated by saying, that the universe should look the same for any observer, no matter what her or his location and the direction being looked in. This model (Robertson, 1935) leads to the metric

$$(ds)^2 = g(t)\left\{(dr)^2 / (1-kr^2) + (r\,d\theta)^2 + (r\sin\theta\, d\varphi)^2\right\} - (cdt)^2,$$

where g(t) is a function of time taking only positive values, and k is a constant which may be positive, zero or negative. The set of solutions to Einstein's field equations satisfying the principles leading to this metric are called Friedmann models (Friedmann, 1922). They exhibit Hubble type expansion, and if $k \leq 0$ they are forever expanding, while for $k > 0$ they will eventually again contract.

In general relativity, the light cones of different points do not have the same opening angles ("inclinations"), in contrast to the situation locally, where the special theory of

relativity made all light cones similar. This is illustrated in Fig. 9.7, which may be compared with Figs. 9.6 and 9.5.

The distortion of light cones, or equivalently, that the world lines of photons are no longer straight lines, imply that light rays will bend near strong gravitational fields, an effect that has been observed for photons reaching us from directions close to the periphery of the Sun.

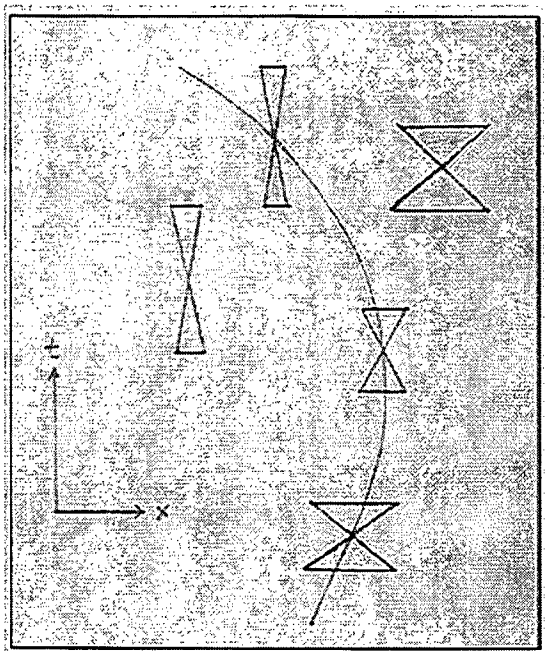


Fig. 9.7. Photon path and light cones in general relativity, that is in the presence of gravitational fields.

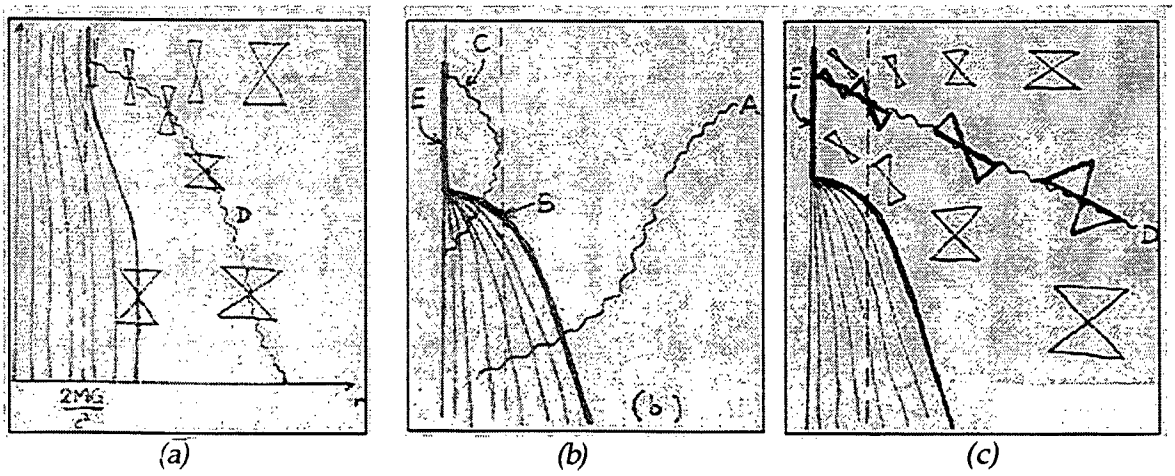


Fig. 9.8. Black hole formation and properties. (a): Schwarzschild diagram (cartesian coordinates; full lines represent particles in the star, wavy lines photons), (b) and (c): Eddington-Finkelstein diagrams; A: escaping photon. B: star becomes black hole, C: photon unable to escape, D: impinging photon becoming absorbed by black hole. E: singularity at centre.

Looking now at a black hole, that is a star, which has collapsed to a radius below the

Schwarzschild radius, one finds that the light cones have smaller and smaller openings, as that radius is approached from the outside. If the "maps" are distorted, so that incoming photon world lines become straight lines, then the light cones will become inclined towards the black hole, and at the Schwarzschild radius, only inward motion is possible (such "maps" are called Eddington-Finkelstein diagrams (Hawking and Ellis, 1973). Once inside the critical radius, the photon, as well as any piece of matter, will continue until it reaches the centre (Fig. 9.8). Inside the black hole there is a point singularity. It has been demonstrated by Penrose (1965), that this is not due to particular assumptions made in the calculations. It is a general result that follows from the Einstein field equations as they are given above.

The presence of a fundamental singularity tells that Einstein's general theory breaks down for the interior of black holes. In some vague way, is it easy to identify the cause of the break down: Since quantum effects are not considered, the effects of uncertainty relations or "quantum fluctuations" are not included, and they certainly will modify the solutions close to the singularity, which then no longer will be a singularity, if relativity and quantum theories are properly combined. So far, this has not been done (except possibly in superstring theory), but the importance of quantum effects, and the possibility of calculating some of them, has been demonstrated by Hawking (1975, 1977) in his evaluation of the tunnelling effect (cf. problem 8.10).

PROBLEMS AND DISCUSSION ISSUES

PROBLEM 9.1. If the Sun suddenly became a black hole, how soon would we find out?

DISCUSSION ISSUE 9.2. Give a number of reasons for taking energy conservation as a very fundamental assumption.

PROBLEM 9.3.

(A) Calculate the escape velocity for the Earth's gravitational field. What happens when a rocket is fired vertically with exactly this velocity (disregard friction)?

(B). How much energy must a spaceship expend to re-enter the Earth's atmosphere and descend to the ground?

(C) Recall problem 2.4, where an object was orbiting the Earth near ground level. At what height will the centrifugal force due to the speed of the object in its orbit balance the gravitational force and the speed furthermore be such that the object follows the rotation of the Earth? (this is the geosynchronous orbit, where the satellite is over the same point on Earth all the time).

PROBLEM 9.4. What would be the energy released by a collision between 1 kg electrons and 1 kg positrons?

DISCUSSION ISSUE 9.5. Do you consider life as bound to matter, or is it structure/organisation?

PROBLEM 9.6. Can you formulate a simple model, which would explain the rise in

temperature, when stellar matter contracts (due to gravitation), and the decrease in temperature during expansion?

PROBLEM 9.7. Use the special theory of relativity to derive the Doppler-shift formula

$$\frac{\nu_0}{\nu} = \frac{\sqrt{1-v/c}}{\sqrt{1+v/c}}$$

Compare with the one derived by classical mechanics.

DISCUSSION ISSUE 9.8. Suppose we got into contact with living organisms from elsewhere in the universe. Do you think they would be friendly?

If the Earth was visited by organisms from space with a form entirely different from ours, or perhaps by structure- and not form-like organisms, then do you think that we could communicate with them?

How would we know that they are here at all? What size do you think they have? How could we detect their non-Earth origin and determine whether or not they are intelligent beings?

PROBLEM 9.9. Consider section 9.1 and prove that $(\rho A')_0 = 0$, and show how to get from equation (□) to equation (□□).

PROBLEM 9.10. Consider the Cygnus X-1 double star and assume that the visible star has mass M and is rotating around a point at a distance R with period T. Can you derive a formula for the mass of the invisible partner?

Chapter10

Our surroundings

The Earth is a planet with a solid crust, surrounded by a thin gaseous layer called the atmosphere (Fig. 10.1). The inside of the Earth is molten or partly molten, but the precise composition is poorly known. This is due to imprecise knowledge of the processes of accreting the materials forming our planet. As mentioned in Chapter 9, the material used to form the planets must derive from supernovae outbursts. It is likely, that this material has been accreted over an extended period of time, say 10^8 years. The accreted matter would form a gas cloud with temperatures declining with distance from the centre. The cloud would slowly cool as it emits heat radiation, and would at characteristic temperatures form definite molecules such as calcium and aluminium oxides (1600 K), nickel-iron and magnesium-silicium alloys (1200-1300 K), feldspar (1000 K), troilite (FeS, 680 K) and water ice (273 K). The condensed material would in most cases be collected into planets, due to inhomogenuities and the gravitational interaction (with exceptions such as the asteroids). However, it is not certain if the planets were formed quickly, as compared with the overall cooling of the cloud, or more slowly. In the first case, a planet such as the Earth would have a composition characteristic of the condensation state at a definite temperature (which for the Earth would be around 660 K). In the other case, the planet would have layers with compositions characteristic of different condensation temperatures.

The planet would have a hot interior for two reasons. One is that backward extrapolation of radioactivity based on present abundance of radioisotopes (the decay schemes of which are known) tells us that the early Earth must have been highly radioactive. The second source of heat would be the gravitational contraction forming the dense core (which we can detect by seismology). The centre temperature is believed to be of the order of 4000 K. The solid crust of the Earth would have formed very early in the formation process, due to heat radiation from the surface and the associated cooling. The temperature gradient between the centre and the surface has been rather stable, although convective transport does occur (as evidenced by volcanoes) and probably was more pronounced in earlier stages. The present total heat flow between core and mantle is 25×10^{12} W, while that between mantle and crust is 36×10^{12} W (higher due to radioactive elements in the crust) (Chapman and Pollack, 1975).

The formation of the planets must have taken place along with the formation of the Sun or earlier. The age of the Earth is thus at least 5×10^9 years.

The Earth is surrounded by an atmosphere of nitrogen, oxygen and a number of minor constituents (Fig. 10.2). Its density decreases rapidly, and so does its pressure, while the temperature exhibits a more complex variation, as the (annual) average values shown in Fig. 10.3 indicate.

The properties of the atmosphere can be understood from its constituents and two main features: the rotation of the Earth and the solar radiation received (1353 W/m^2 at the top of the atmosphere and perpendicular to the direction to the Sun, with ±3% excursions

due to the seasonal change in the Earth-Sun distance). In the absence of solar radiation, the atmosphere would follow the rotation of the Earth. In other words, there would be no winds. If there were no atmosphere, and the surface of the Earth absorbed all incoming solar radiation, then the average surface temperature would be 277k (4°C). However, the real Earth-atmosphere system is not a perfect absorber. Some 30% of the solar radiation is reflected back into space. The fractional reflection is called the "albedo". Such an albedo would change the average temperature of the atmosphere-less system to 254 K. If water is present, such a planet would freeze, and as the albedo of ice is even higher (some 90%), this glaciated Earth would have an average temperature well under 200 K.

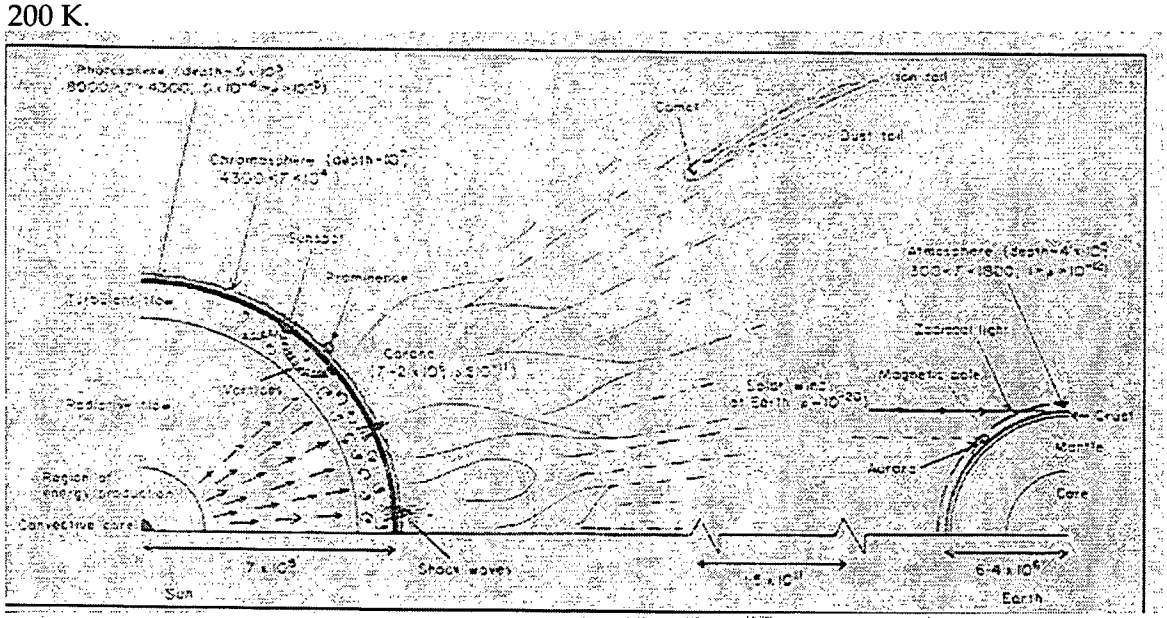


Fig. 10.1. Schematic picture of the structure of the Sun and the Earth (Sørensen, 1979). The Earth mantle is believed to consist of an outer part (silicates of Mg and Fe) and an inner part (oxides of Mg and Fe). Also the core has an outer part (presumably liquid FeS) and an inner part (a solid iron-nickel alloy).

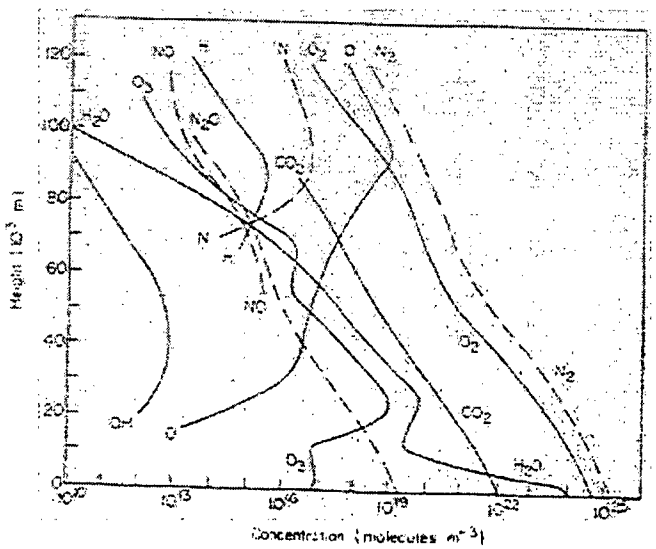


Fig. 10.2. Height variation of mean concentrations of some gaseous constituents of the atmosphere, based on estimates by Almquist (1974).

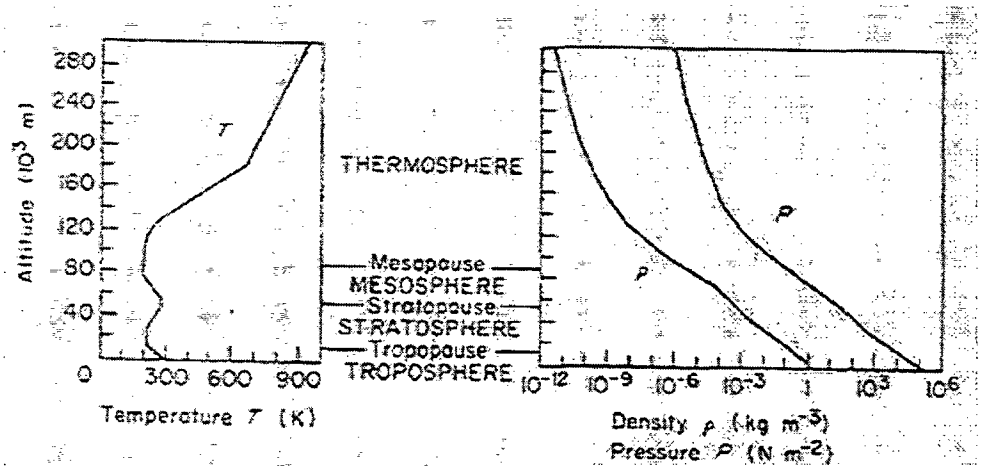


Fig. 10.3. Altitude dependence of temperature, pressure and density for typical state of the atmosphere. Names used for the different temperature intervals (declining or increasing temperatures) are indicated in the middle.

However, the actual average temperature at the surface of the Earth is 287 K, because of the greenhouse effect of the atmosphere. This can be understood by looking at the fate of the solar radiation intercepted, as illustrated in Fig. 10.4. The radiation absorbed by the Earth's surface is partly re-radiated as thermal radiation corresponding to the temperature of the surface, and partly transferred to the atmosphere as other forms of energy (heat capacity, evaporation energy, etc.). The thermal radiation is to a large extent absorbed in the atmosphere, and then re-radiated in part back towards the surface and in part to space. This makes it possible for the surface to have an average temperature above the naked planet, while the atmosphere has an average temperature decreasing upwards. The atmosphere thus acts as the glass panes of a greenhouse.

The absorption of solar radiation is not the same everywhere oh the surface of the Earth. At latitudes from -25 to +25 degs., the average value is about 225 W/m^2. It then drops towards the poles, until about 60 degs., and remains at about 100 W/m^2 between 60 and 90 degs., but of course with larger and larger seasonal variations relative to this average value. If the surface temperature followed the absorption pattern, it would be much warmer hear the Equator and much colder near the poles than it actually is. It follows that there must be some mechanism that transports energy from low to higher latitudes.

There are two important mechanisms of such transport. One is wind and the other is ocean currents. The ocean currents carry heat directly, while the winds carry evaporated water from latitudes around 20 degs. N or S, and release the energy again by condensation (leading to precipitation) at latitudes around 50 degs. N or S (Sørensen, 1979).

These features of the circulation in the atmosphere and oceans can be modelled by the same methods as those discussed in Chapter 9 for stellar atmospheres. Since the atmosphere of the Earth is only a thin shell around the surface, the vertical transport tend to be considerably smaller in magnitude than the horizontal transport. This and the absence of nuclear reactions make it possible to simplify the equations of motion. On the other hand, a number of equations describing the abundance of minor constituents and

their distribution in the atmosphere must be considered, because of the importance of constituents such a water vapour, dust, carbon dioxide, ozone and several others in determining the reflection and absorption of radiation as function of position. Also the oceanic motion can be determined from the atmospheric models, which describe fluids as well as gases (Sørensen, 1979).

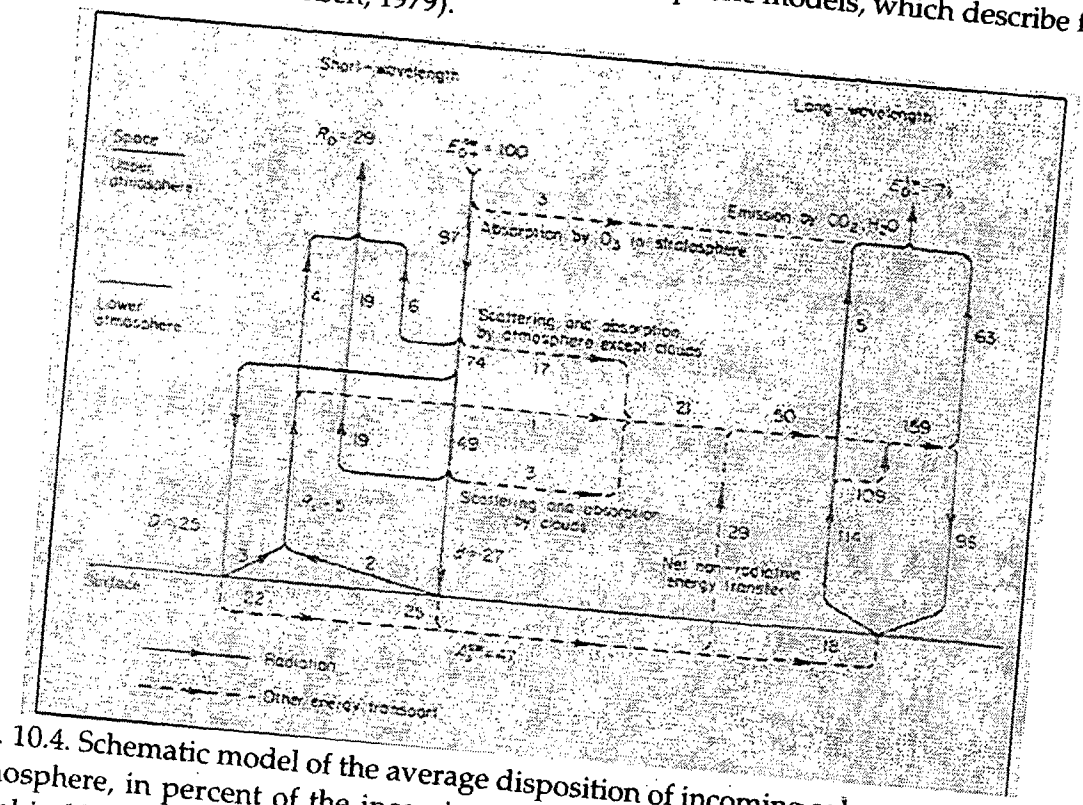


Fig. 10.4. Schematic model of the average disposition of incoming solar radiation in the atmosphere, in percent of the incoming radiation. Averaging is over time and geographical location. (D and d stand for direct and scattered radiation (Sørensen, 1979).

The energy transformations in both atmosphere (Fig. 10.4), hydrosphere and in the solid crust of the Earth have been summarised in Fig. 10.5.

The balance between energy absorption and transfers can be altered in several ways. When the continents were combined (their slow motion is due to convection in the mantle), the ocean currents were unable to transport heat to some of the polar regions reached at present. This accounts for the extended glaciation 300 to 240 million years ago. After the combined continents Laurasia and Gondwana broke up and let warm ocean currents penetrate towards the poles, there has been a periodic oscillation between glaciation and no glaciation at both poles. The oscillation period is 0.5 million years, and we are presently in the warmest part of the average temperature curve. The other ways of changing the climate would be to change the constituents of the atmosphere or the albedo of the atmosphere or the Earth surface. Higher absorption of long wavelength radiation (heat absorption) results from adding carbon dioxide to the atmosphere, while the addition of dust particles would change both short and long wavelength absorption. The reflection of solar radiation by dust with particle size typical of combustion products (power plants, cars, etc.) is likely to lead to a cooling of the Earth,

which exceeds the warming that is due to containment of heat radiation (within the Earth-atmosphere system) due to the same dust. The larger grain size of dust from desert storms may on the other hand predominantly lead to warming. Finally, changing the albedo of the surface to a less absorbing (less black) colour will definitely lead to a cooling. This actually happens when the Earth is covered by ice or snow. Such white surfaces reflect 80-95% of the solar radiation. Glaciation is a self-reinforcing process: ice cover lowers the albedo, it then becomes colder and the ice cover expands, leading to further lowering of the albedo, new cooling, and so on. Budyko (1974) estimates that if the glaciation reaches 50 deg. latitude, it will continue until the entire Earth is covered. The temperature would then be about 156 K (for an albedo of 0.9). It should be stressed, that this "white Earth state" is an energetically more stable state than the present one.

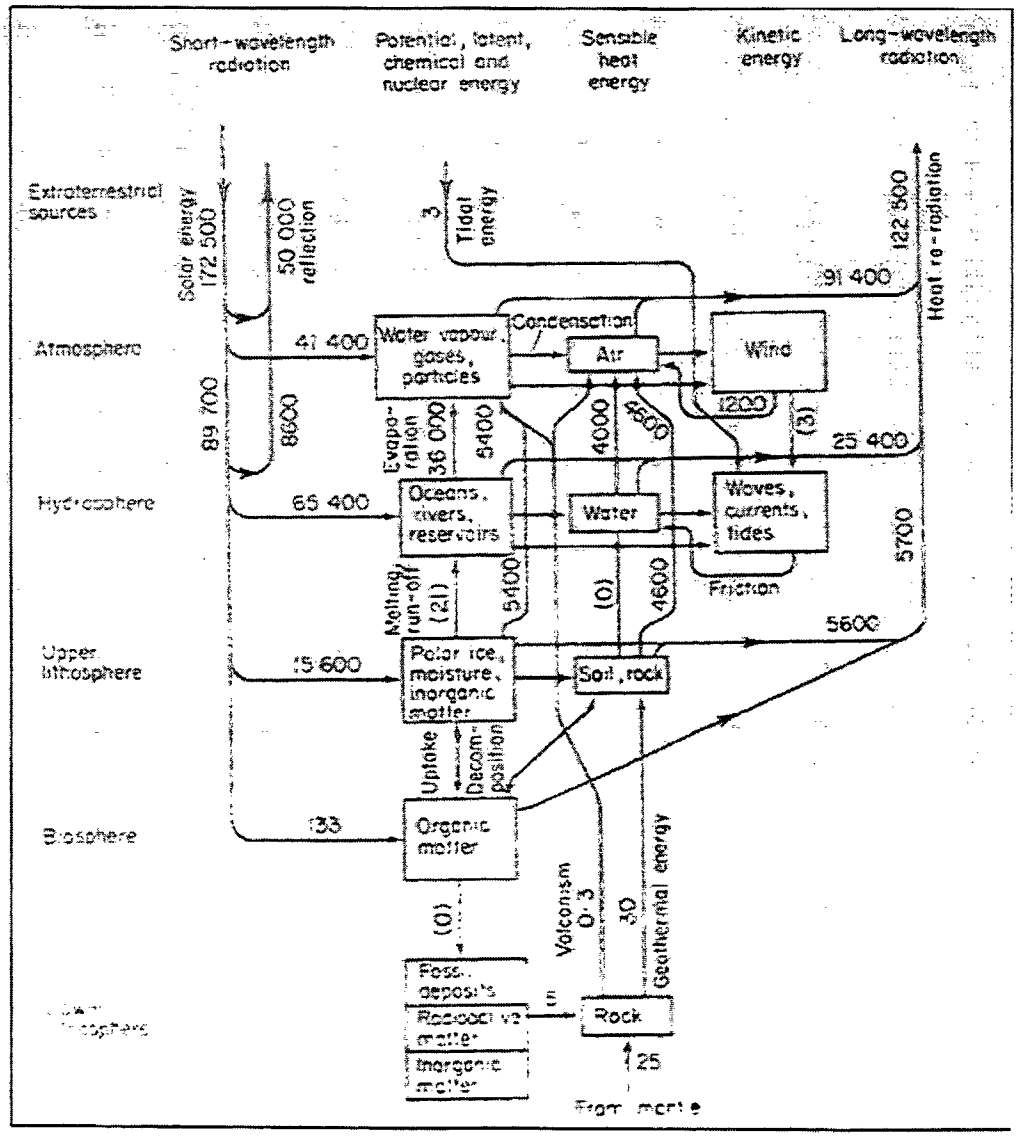


Fig. 10.3. Average energy flows near the surface of the Earth (in TW = 10^{12} W) (Sørensen, 1979).

The cause of the periodic glaciation is not fully known, but it is believed to be connected with oscillations in the eccentricity and inclination of the Earth's orbit (Milankovich, 1941).

Life on Earth has been developed on the basis of organic molecules, notably depending on the availability of water and oxygen. Fig. 9.1 showed the change in oxygen abundance. It can be explained by assuming that all oxygen in the atmosphere has been formed by photosynthesis, that is by absorption of solar radiation in specific organic molecular structures capable of forming an energy-rich protein molecule, which subsequently can assimilate carbon dioxide and thereby synthesise glucose or starch. These sugar molecules serve as energy for processes in the organism, which can then be carried out in the absence of light. The photosynthetic reaction may be written

$$2\,H_2O + \text{light} \rightarrow 4\,e^- + 4\,H^+ + O_2$$

where a membrane system in the plant is capable of keeping the electrons and hydrogen ions apart (else they would recombine). The subsequent carbon-dioxide assimilating reactions may be summarised in the form

$$6\,H_2O + 6\,CO_2 + 4.7\times10^{-18}\,J \rightarrow C_6H_{12}O_6 + 6O_2$$

Water in its fluid phase is believed to have been present on Earth during practically all its existence. The other prerequisite for green plants is carbon dioxide. It is formed by weathering of limestone but could never have exceeded 1% of the atmosphere, because of the rate at which it would recombine with calcium to form limestone again. The reason for believing that green plant assimilation of carbon dioxide and release of oxygen is responsible for all oxygen in the atmosphere is, that the only alternative way of forming oxygen would be by photodissociation, and a calculation of the oxygen concentration in air resulting from this process gives a result 1000 times lower than the actual one (Berkner and Marchall, 1970).

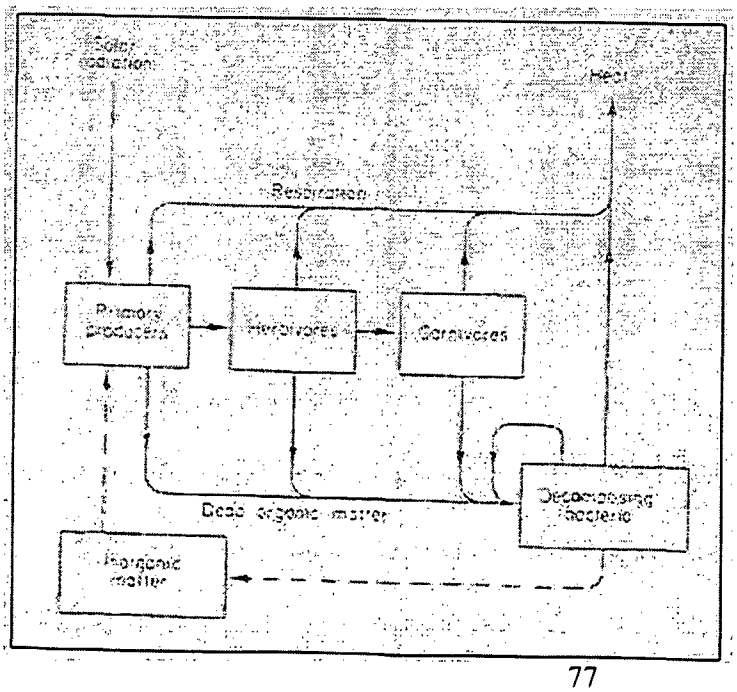


Fig. 10.6. Model of ecological system on Earth. Solid lines indicate flows of energy or organic matter, dashed lines flows of inorganic matter. There could be photosynthetic bacteria, but they have not been included (Sørensen, 1979).

Based on the photosynthetic plants in water or on land, other living organisms have developed. They form predator chains of the general form shown in Fig. 10.6. These organisms have a finite life span, and the organic material (proteins and so on) left over from the predator chain is decomposed by bacteria, where the elements are brought back to inorganic molecular form. This makes the relevant nutrients soluble, so that they become available for renewed uptake by (for instance the root systems of) primary photosynthesizing producers, in a closed, ecological cycle. Energy-wise, the biosphere does not use more than a tiny fraction of the incoming solar radiation (see Fig. 10.5).

From the first formation of life to the epoch of man as an influential inhabitant of the Earth, almost 3×10^9 years elapsed. Man has changed several of the ecological cycles dramatically over a period of time very small compared with that required for the evolution of such ecosystems. Does this mean that the rules of the game - the physical laws of evolution - have changed? Or do the interventions of man constitute short-term fluctuations around a smooth line of long-term development?

The development of life can be understood in a probabilistic model. There is a finite chance that molecules of carbon, hydrogen, oxygen, nitrogen, etc. will find each other and form the molecules needed for the formation of organic materials, once the elements are available and there are processes which might bring them together (ocean currents, wind), and once there is energy for the chemical reactions to go in a given direction (hot lava from volcanoes, solar radiation). It is then a matter of probability, when organic molecules such as proteins will become formed. Once they are there, we know that certain combinations of such molecules may form self-reproductive systems. Even if the probability that this will happen is small, it will indeed happen if we wait long enough. The self-reproducing and self-organising systems appearing in this way are living organisms. In the later stages, where their survival is determined by competition for food and the rate of reproduction, Darwin's theory of evolution can explain the selective principle leading to long-term survival of certain types of organisms in periods of certain external conditions (notably climatic ones). In the early period of precursor molecules and simple self-reproductive organisms, a more fundamental principle has to take the place of Darwin's theory. Prigogine has proposed, that such a principle may derive from thermodynamic considerations (Prigogine *et al.*, 1972). However, in contrast to the near-reversible processes near an equilibrium situation considered in classical thermodynamics, the evolutionary systems are far from equilibrium and they must be assumed to take part in irreversible processes (or "non-linear" processes, as they may be called because they cannot be treated by linear differential equations (an example of a linear differential equation is (*) in section 7.2.

Near equilibrium, a system is seeking to minimise its dissipation of energy, but far from equilibrium, the model of Prigogine expects it to attempt to maximise its dissipation of energy (analogy: a newborn baby has a very high emission of heat, relative to its weight, and a growing child still dissipates heat at a higher rate than an adult person, which may be said to have reached the equilibrium size).

Returning to the present human society, the question I posed above was, if any such

general rules apply. The fluctuations, that is the range of statistically possible outcomes, were necessary features for the evolution to proceed. But are the present behavioural patterns of mankind fluctuations of this kind, or has the complexity of the human brain, that is what we would call our conscious mind, changed the situation. In other words, can we consciously change our fate in a way not in agreement with the rules of evolution, or with a more precise rephrasing: may short-term excursions take us so far away from the evolutionary track, that evolution theory fails even in the long run? These discussions of man's "free will" have been undertaken throughout the history of science, philosophy and theology, in various forms.

Our man-made environment includes a number of tools and devices, that we have produced on the basis of natural resources. It includes managed ecosystems (agriculture, silviculture, animal husbandry), construction of shelters (buildings) and roads of communications, and manufacture of tools, vehicles, equipment, weapons and all kinds of toys. To sustain this production, a special industry of minerals and energy resource extraction has been developed. The basis for this development has been an understanding of the properties of materials and the laws determining their motion and interactions.

The physical sciences involved are thus materials science, including chemistry, mechanics and electrodynamics. These theories have had to be developed on several levels of aggregation. It is not practical always to have to go back to quantum mechanics in order to describe phenomena such as heat capacity, phase change energy, elasticity, tensile strength or conductivity. Macroscopic theories have been developed, in order to give the answers to everyday questions. They are based on individual sets of assumptions, which in many cases have been derived from basic theory (say quantum mechanics), but often much later than their macroscopic formulation. Such theories include classical chemistry, thermodynamics, gas theory (a statistical molecular theory), and a number of partial theories describing selected properties (for example electromagnetic ones) of materials (for example solids) or their dynamical behaviour (fluid mechanics, aerodynamics and so on).

Macroscopic theories can be said to aim at a description of collective behaviour of a large number of electrons or atoms. A mole of a gas or a metal piece you can hold and feel in your hand will consist of something like 10^{24} atomic particles. Writing the quantum mechanical equations for every single one of them and solving the equations, including mutual couplings in terms of fundamental interactions, is impossible. There would be problems even if it was possible, because one would have to decide which of the many possible solutions would be interesting to look closer at. Therefore some kind of "next higher level" theory must be invoked in order to sort the interesting from the uninteresting types of behaviour. Among the interesting ones are situations where a large number of the particles move in a coherent way. By this is meant that they maintain a particular kind of correlation among them. Such behaviour is denoted "collective behaviour". Most of the interesting phenomena in macroscopic systems are collective phenomena, such as conduction of heat or electrons (forming an electric current), wave motion or vibration (oscillation about an equilibrium situation). Also macroscopic mo-

tion is of course a macroscopic phenomenon, in which an atomic structure (say a three-dimensional lattice) is being displaced or rotated collectively, without losing the internal structure.

One question that arises out of this is, whether the microscopic behaviour uniquely determines the macroscopic behaviour. Could there be macroscopic ("holistic") laws of nature, which could not be derived from the microscopic ones? We know that the probabilistic interpretation of the quantum mechanical laws are not inconsistent with macroscopic causality. This is a result of the largeness of the number of particles. The probability for non-causal behaviour becomes so small that it cannot be detected. Another well established fact is that the symmetries on the microscopic level (for example rotation or mirror symmetry in atoms) need not be found microscopically. The not very symmetric objects in our surroundings may be formed in a very natural way from components, which are all symmetric. Quantum fluctuations on the microscopic level correspond to peculiar structures on the macroscopic level, because only those fluctuations which are found in a coherent way among the individual particles are able to survive and become macroscopic features.

In order to understand these forms of behaviour, it is necessary to supplement the microscopic laws of physics with macroscopic ones. These are derived from system theories, that is theories of collective behaviour. Some of the system theories are derived from microscopic models, other ones are not. This does not mean that such system theories are not consistent with the microscopic theories (quantum mechanics and quantum field theories, for example). The existing incomplete theories of irreversible thermodynamics, and of structure and stability of thermodynamic system (Glansdorff and Prigogine, 1971), may be based on probabilistic models of atomic constituents of such systems, and may make different assumptions on the behaviour of the systems, which might all be consistent with current microscopic theories for the constituents.

10.1 Classical mechanics

The equation of motion for a particle of mass m is in classical mechanics given by Newton's 2nd law,

$$m\, d^2x_i/dt^2 = F_i \qquad i = 1,2,3$$

Where F_i is the i'th component of the force. An example is the gravitational force

$$F_i = -G\, m\, m'\, (x_i - x_i') \,/\, |\mathbf{x} - \mathbf{x}'|^3$$

between m and another mass m'.

The Newtonian equation of motion can be rewritten in the form of two linear equations, by introducing the velocity v_i or the momentum p_i as a new variable,

$$dp_i/dt = F_i,$$

$$dx_i/dt = p_i/m \; (= v_i)$$

Two important functions are the Lagrange function L and the Hamilton function H, de-

fined through the kinetic energy T and the potential energy V of the system (here restricted to consist of one or several particles),

$$L = T - V$$

$$H = T + V$$

$$T = \sum_i p_i^2/(2m)$$

V is normally a function of only space coordinates. If this is the case and T is a function of p_i only, then V is simply minus the gradient of the force (as for the general theory of relativity outlined in section 9.2).

The two, coupled first order equations of motion above have the form anticipated in Chapter 7 (section 7.1). The second order equation of motion may be written in terms of the Lagrange function,

$$d(\partial L / \partial v_i)/dt = \partial L / \partial x_i$$

with

$$F_i = \partial L / \partial x_i \text{ and } p_i = \partial L / \partial v_i$$

This does not at first sight look like a convenient form, but it is! The reason is that it is valid no matter which co-ordinates are used to describe the system, while the original Newton equation is valid only for Cartesian co-ordinates (x_i, i=1,2,3). Once you have tried to work out a problem in spherical or cylindrical co-ordinates you will appreciate this. Not only the forces and momenta can be expressed in terms of L, but also the energy function H,

$$H = \sum_i v_i (\partial L / \partial v_i) - L$$

The use of the Lagrange function makes classical mechanics correspond nicely with the field theories used in the description of matter (Chapter 8). In both cases, the fundamental equations can be derived from a variation principle. This is the reason for presenting classical mechanics in this form, rather than to take a historical approach. Let me explain what a variational principle is. Consider the integral

$$S = \int_{t_1}^{t_2} L \, dt$$

(called the "action integral"). The time development of the system would be determined, if the time development of all the variables describing the system (such as the x_i's for a particle described by Cartesian coordinates) were known. Trying all possible developments of these variables between the times t_1 and t_2, corresponding Lagrange functions could be evaluated and consequently the values of S for all the possible paths of the system. The variations principle now states, that the true development of the system is one that gives an extremum (a minimum or a maximum) for S. This means that the change in S for an infinitesimal change in the path (that is a change in the variables so small, that second order derivatives in the Taylor expansion of S, cf. section 7.2, can be neglected) is zero,

(♥) $\delta S = 0$

The equation of motion can be derived from (♥) alone. Since L does not change (is invariant) under Lorentz transformations, the variational principle is also used in relativistic theory.

The total energy, that is H in classical mechanics, is constant for a closed system (a system not interacting with its surroundings in any way). We have seen that the Hamilton function H is the link between classical and quantum theory. The classical theory can also be expressed in terms of H alone, but H is not invariant under Lorentz transformations and is therefore not used much in relativistic theories. From the definition of the action integral one gets the relation

$$H + \partial S / \partial t = 0$$

10.2 Classical electrodynamics

The Lagrange function in electrodynamics is given in terms of the electric and magnetic fields F_{in}. F_{in} is a 4×4 matrix given in section 8.3 (just before equation (§§)), both in terms of the x,y,z-components of the electric and magnetic field, and in equation (§§) in terms of derivatives of the potentials (A_i, i=1,2,3,4). L may be written

$$L = -(1/4) \sum_{in} F_{in} F_{in}, \quad i,n = 1,2,3,4$$

and the field equations derived from the variation principle applied to the action integral (as in mechanics) are

(♣) $\partial F_{jn}/\partial x_i + \partial F_{ni}/\partial x_j + \partial F_{ij}/\partial x_n = 0$ and $\sum_i \partial F_{ni}/\partial x_i = J_n/(c\varepsilon_0)$

These are called Maxwell's equations. The constant ε_0 is the dielectric constant in vacuum as in section 8.3, and the four-component current vector J is

$$J = (j_1, j_2, j_3, ic\rho) = \rho\,(v_1, v_2, v_3, ic)$$

where the j_i's are the components of the current vector (ρ times the velocity) and ρ is the electric charge density. The four co-ordinate components are (x, y, z, ict). The equations (♣) are invariant under gauge transformations of the potentials (see section 8.4).

10.3 Collective electron phenomena

Consider first one electron in one atom. The Schrödinger equation for stationary states (section 8.1) reads

$$-(\hbar^2/m)\,\Delta\chi_i - Ze^2/(4\pi\varepsilon_0 r)\,\chi_i = E_i\,\chi_i$$

where the Hamiltonian has been written in the same form as in section 8,1, with V equal to the Coulomb potential between the electron of charge -e and an atomic nucleus of charge +Ze. The mass of the electron is denoted m, and r is its distance from the centre of mass of the atom. The potential is spherically symmetric, as it depends on x, y and z only through the combination r. This makes the solutions especially simple when ex-

pressed in spherical co-ordinates, because they can be written as a product of a function depending only on r, and one depending only on the polar angle θ and the azimuth angle φ,

(♠) $\chi_i = u_{nl}(r)\, Y_{lm}(\theta,\varphi)$

The functions Y_{lm} are standard functions called "spherical harmonics". Each of them corresponds to an electron having a given total angular momentum (equal to $\hbar$), and a given projection of the angular momentum on the polar axis (equal to $m\hbar$), where m of course has to lie between $-l$ and l). The radial part u(r) does not depend on the direction of angular momentum, but may depend on l. It is similar to the wave function calculated in problem 8.5, but is not quite a sine function, because the Coulomb potential is not quite like the box potential.

Still the general picture of the stationary states is the same. Fig. 10.7 gives an idea of the energy levels allowed in the quantum treatment of the atom. The expression above shows, that not only energy, but also angular momentum is quantized, so that it can have only certain discrete values. The number i labelling the stationary states is now $i = (n, l, m)$, or more correctly $i = (n_i, l_i, m_i)$. It is a general feature that i is not just an index, but it can be given a physical interpretation in terms of angular momentum, etc. Also the index n appearing in the radial function u has a physical interpretation. It is the number of times, the radial wave function u passes through zero. The fewer times it does, the lower is the energy, according to Fig. 10.7. The energy values depend only on n and l, not on m

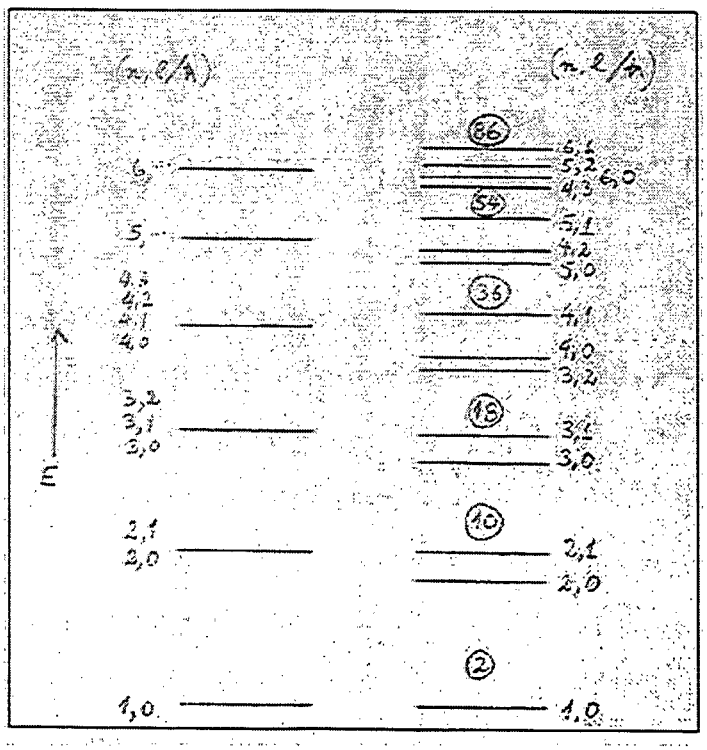


Fig. 10.7. Atomic shell model. To the left are one-electron energy levels, and to the right many-electron levels. No scale is given, because it depends on the atomic number Z, while the structure of the diagram is general. Numbers in circles are the numbers of electrons, which would fill all states up to the position of the circle. These "closed shell" atoms are particularly stable, because excitation requires provision of enough energy to lift an electron up through the "gap" to the next shell. They are the "noble gases".

The electron spin ($s = \hbar/2$) has been neglected in the treatment given above. This is a

good approximation as long as there is no magnetic fields around. If there are, the treatment sketched in the quantum electrodynamics part of Chapter 8 can be used to demonstrate, that each (n, l, m) level splits into two, which have the projection of the spin on the polar axis equal to $-\hbar/2$ and $+\hbar/2$, respectively. This means that the solutions obey the Pauli exclusion principle exactly, as they should because electrons are fermions (cf. section 8.2).

Now, consider the case of many electrons in one atom. The Schrödinger equation for one particular electron will be as above, except for additional terms in the potential V, expressing the repulsion of the electron in question by interaction with other electrons, $+e^2/|r-r'|/(4\pi\varepsilon_0)\chi_i$. The result of these additional terms is shown in Fig. 10.7. States with same *n* but different *l* have now been split. Filling up electrons from the bottom, with due respect to the Pauli principle, gives the electron structure. Groups of levels with similar energies are called "shells", and the model hence the "shell model".

The energy "gaps" between shells and the number of electrons in close-lying states form the basis for the group and valence numbers used in chemistry.

I shall now turn to the case of several atomic nuclei and several electrons, that is molecules. For a molecule of N atoms, I shall denote the centre of mass positions of the i'th atomic nucleus B_i, i = 1,...,N, and the position of a given electron **r**. As a first approximation, the atomic nuclei may be taken to be at rest with fixed distances to each other. This is a reasonable approximation due to the large mass of the nuclei as compared with that of the electrons. The motion of an electron is then given by the sum of the fixed Coulomb potentials of the nuclei, plus Coulomb interactions with the other electrons.

The solution of the Schrödinger equation is still prohibitively difficult for large molecules, so one additional approximation will be needed. It emerges from the assumption that when an electron gets close to one of the nuclei, its wave function ψ will look like one of the stationary solutions ψ_n for that particular atom,

$$\psi(\mathbf{r}) \rightarrow \psi_n(\,i;\,\mathbf{r}-\mathbf{R}_i) \quad \text{for } \mathbf{r} \rightarrow \mathbf{R}i$$

Here $\psi_n(\,i;\,\mathbf{x})$ is the n'th stationary electron state in the i'th atom. One may then try to write the electron wave function for an arbitrary position as a linear combination of the form

$$\psi(\mathbf{r}) = \sum_i c_i\ \psi_n(\,i;\,\mathbf{r}-\mathbf{R}_i)$$

When the electron looked at is close to a nucleus i_0, the argument of the stationary wave functions for other i's will be large and the numerical value of the wave function small. All the stationary wave functions go towards zero far from the nucleus, and out in the "tail region" they are small. In that case ψ is then approximately equal to c_{i0} times the value of the stationary wave function of the i_0'th atom. If the electron moves to the region around another nucleus, its wave function will approximately become that of a stationary state in that atom, and so on. In between nuclei, the electron wave function given above will be a linear combination of small contributions, with coefficients determined by inserting the linear combination wave function into the Schrödinger equation. In the region away from atoms, the wave function is probably a poor approximation to

the true one.

One may now relax the condition that the positions of the nuclei are fixed, in order to see if certain configurations are more stable than other ones. The next-simplest approximation (the simplest being to assume positions totally fixed) is to repeat the electron calculation for a variety of different relative position of the atomic nuclei. One would then be able to determine equilibrium states of the molecule, namely configurations for which the electron energy is lower than for neighbouring configurations. There may be more than one energy minimum, of different values but separated by potential barriers (which could be quite complex three-dimensional ones for molecules comprising several atoms). The method outlined here allows a calculation of the most probable molecular configurations. For a diatomic molecule there is only one configuration parameter, the distance $R = |\mathbf{R}_1 - \mathbf{R}_2|$. Fig. 10.8 shows the electron energy for the two lowest states in the ionised hydrogen molecule H_2^+ (that is two protons but just one electron). The lowest (n=1, l=0) stationary states of the hydrogen atoms (cf. Fig. 10.7) have been used to form an electron wave function,

$$\psi(\mathbf{r}) = c_1 \psi_{10}(\mathbf{r}-\mathbf{R}_1) + c_2 \psi_{10}(\mathbf{r}-\mathbf{R}_2)$$

Where the subscripts are n,l and where the index i on the wave functions nave been left out, since they are both hydrogen wave functions. More states (at higher energy) could be formed with $n > 1$ and/or $l > 0$, but there are only two states in the molecule, which can be formed from the lowest state in each atom. These two states have different energies, as seen in Fig. 10.8. The reason is that the component H_{12} (see section 8.1) is non-zero,

$$< 1 \mid H \mid 2 > \neq 0 \text{ for } \mid i > = \chi_{10}(\mathbf{r}-\mathbf{R}_i),\ i=1,2$$

The lowest energy level has a minimum value for a separation of 1.3×10^{-10} m, which then corresponds to an equilibrium state of the molecule. The other, excited state does not exhibit any energy minimum.

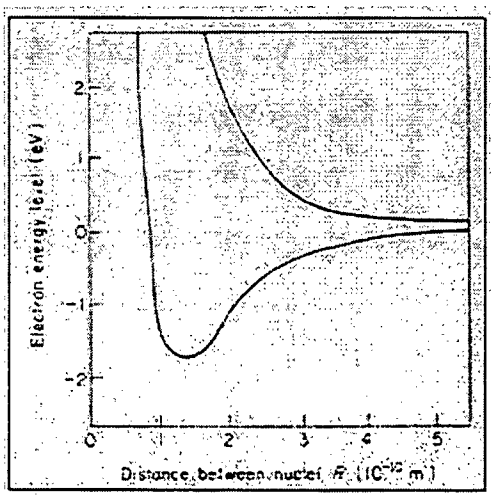


Fig. 10.8. Energy levels in ionised hydrogen molecule, as function of the separation distance between the hydrogen nuclei.

If there are several electrons in the molecule, some of them may approximately be in the lowest states of individual atoms, that is, localised around a definite nucleus. Only the electrons in the highest energy levels are moving around among the atoms, in linear

combinations of stationary states, of the form given above. If the atoms are different, it is likely that states of different *n,l* values are roughly at the same energy level, and since levels are getting closer at higher energies (see Fig. 10.7), it is possible that more than one stationary state in each atom may play a role. This means that the electron wave-function should be sought in terms of a linear combination more general than the one given above, with summation not just over atoms *i*, but also over individual levels n = n(i) in each atom. This situation is in chemistry called "hybridisation".

The states above the lowest one are called "excited electron states". However, there are two other types of excited states of a molecule: One is oscillations ("vibrations") of the distances between nuclei around their equilibrium value and the other is rotations of the whole molecule around some axis.

Let me illustrate vibrations by the example of the diatomic molecule of Fig. 10.8. The energy of the lowest state is minimum for some atomic distance R_0, and as seen, it is given by an unsymmetric function of the deviation of R from R_0. Near the minimum, this curve can be approximated by a second-power relationship (parabola). The Hamiltonian then can be written

$$H = p^2/(2M) + a(R-R_0)^2$$

where M is the vibrating mass and *a* is a constant defining the best parabolic approximation to the lower curve of Fig. 10.8. Such a Hamiltonian, which is second order in both momentum and position co-ordinates, can be written as a boson Hamiltonian (see section 8.2) of lowest order,

$$H = \hbar\omega\,(b^+b + \tfrac{1}{2})$$

by applying the transformation

$$p = t\,(b^+ + b)$$

$$R - R_0 = s\,(b^+ - b)$$

Here s and t are constants and (b^+,b) boson operators satisfying the commutation relations discussed in section 8.2. Direct insertion shows that the reformulation of H is emerging by chosing the following values for s and t;

$$s^2 = \hbar\omega\,/\,(4a)$$

$$t^2 = \hbar\omega\,M\,/\,2$$

This shows that the vibrational states are quantized, and that they can be viewed in terms of boson quanta quite analogous to the photons discussed in Chapter 8. The energies are given by

$$H\,|n\rangle = n\,\hbar\omega\,|n\rangle, \qquad E_n = n\,\hbar\omega$$

In analogy to the photon case, the vibrational states are called "phonon states". The molecular ground state is the zero-phonon state $|0\rangle$, the first excited vibrational state is the one-phonon state $|1\rangle$ with excitation energy $\hbar\omega$. The second excited state is the two-phonon state $|2\rangle$, with energy twice that of $|1\rangle$, and so on. The ground state energy is

not zero (or whatever the minimum in Fig. 10.8 is), but $\hbar\omega/2$. This is the effect of virtual phonon-excitations in the ground state, just as in quantum electrodynamics.

If the energy in Fig. 10.8 had not been approximated by a quadratic term, there would be higher order terms in the phonon-Hamiltonian (cf. section 8.2), and the energy levels of the vibrational states would no longer have equal spacing.

Systems with equally spaced, vibrational states are called "harmonic oscillators", while deviations from equal-distance energy levels are called "anharmonicities".

The other kind of collective molecular state is rotation (these states of motion are called "collective", because all the constituents of the molecule take part in them). Using the spherical co-ordinates (r,θ,φ) as in the beginning of this section, it is possible to show that the kinetic energy $p^2/2m$ in addition to the term involving two derivatives with respect to r also contains a term proportional to the square of the orbital angular momentum L (this is true for an electron, as well as for the entire molecule, with corresponding interpretation of m, p and L). The angular momentum operator has something to do with rotation. If an operator $D_i(\alpha)$ is defined, so that it rotates any state it operates upon by the angle α around the i'th axis,

$$| \text{ rotated state} > = D_i(\alpha) \; | \text{ fixed state} >$$

then D is given in terms of the angular momentum component along the i'th axis,

$$D_i(\alpha) = \exp(-i\alpha L_i/\hbar) \sim 1 - i\alpha L_i/\hbar$$

The expression to the right is valid for very small angles α, and the general expression is obtained by integration (Brink and Satchler, 1962). The exponential function of an operator (L_i) is defined by its series expansion only.

If a molecule is rotated, without touching inter-atom spacings or electron states, relative to an internal co-ordinate system (that is one which follows the rotation), then L^2 is the only part of the Hamiltonian being active. The wave functions must then be stationary with respect to action by L^2, that is L^2 acting on the wave function must give a number times the same wave function again. We have already seen such wave functions, namely in (♠) above. L acts only on the spherical harmonic functions Y, and from the properties of these standard functions it follows that

$$L^2 \, Y_{Lm}(\theta,\varphi) = L(L+1) \, \hbar^2 \, Y_{Lm}(\theta,\varphi)$$

Note that the length of the quantum mechanical angular momentum vector is not L but $\sqrt{L(L+1)}$. For a rotating molecule, the energy is then

$$H\,\chi(L;...) \; = \; wL(L+1)\,\chi(L;...) \; \Rightarrow \; E(L;\text{rot}) = wL(L+1)$$

The constant w contains $\hbar^2$ from L^2 and a factor from the ratio of $p^2/(2M)$ and L^2, where the effective mass M for the entire rotating system is related to the moment of inertia. The θ and φ depending parts of the stationary states χ of rotation are equal to the spherical harmonics only in certain cases. Generally they are more complex, because they depend on two angular momentum projections, m on an axis fixed in space and m' on an axis following the rotation but fixed relative to the molecular configuration. Only

when $m' = 0$ are the solutions Y-functions. However, the energies do not depend on angular momentum projections, and they increase quadratically with L, that is with the "speed of rotation". The energies wL(L+1) must be small compared with the energies needed to excite the internal molecular structure (such as the energy of the second electron state shown as the upper curve in Fig. 10.8), because otherwise the rotation could not be assumed to proceed with unaltered internal structure.

A system comprising several molecules can be in a number of states, such as solid crystal, solid amorphous, liquid or gaseous states. At high temperature, the molecules are usually forming gases ("high temperature" being still below the temperatures where molecules will disintegrate and atoms become ionised, transforming the gas into a "plasma", examples of which were described in the stellar atmosphere section of chapter 9). Gases can be described in terms of molecules moving freely among each other, occasionally hitting each other and becoming scattered, but without disintegrating. Such models are further described in Chapter 12.

At lower temperatures, the matter becomes a fluid or solid (depending on pressure). A model for the overall motion of a fluid was described in Chapter 9, including pressure and viscous forces, as for gases. Other properties of fluids, such as conductivity, may be described by models similar to those of amorphous solids. Electric conductivity depends on the ability of electrons to get from one molecule to another one, in order to bring about a macroscopic displacement of charge. It is clear that this possibility is small for non-ionised gases, but larger for fluids, because in the latter case the molecules are closer and interact more frequently.

Solids may be viewed as gigantic molecular structures. The number of molecules is of the order of 10^{24} (Avogadro's number) or higher. These molecules may form an ordered structure, such as a lattice. There may be irregularities in the construction, such as occasional "faults" or "edges" where the lattice regularity breaks down over an entire area. If not even local volumes of lattice structure can be identified, the solid is said to be "amorphous".

A lattice structure is characterised by periodicity: the atomic nuclei are located in a pattern, which repeats itself. It is invariant under translations in certain directions and of certain magnitudes. For example, a cubic lattice with atom centres separated a distance *a* (the "lattice constant") along the x, y as well as z-axis, is invariant under translations by the distance *a* along either the x, y or z-axis.

Consider for instance a lattice made up of identical atoms. If it were made up of just two atoms, it would be like the hydrogen molecule, and for each atomic energy level, there would be two molecular levels such as the ones shown in Fig. 10.8. If there were not two but five atoms, there would be 5 curves in Fig. 10.8, giving the positions of molecular energy levels as function of separation distance. For a lattice, the lattice constant *a* will play the role of the diatomic separation variable R. If the lattice periodicity is different in different directions, there will be more than one lattice parameter *a*. For N atoms in a lattice, there will be N curves in what corresponds to Fig. 10.8, say with the two ones shown being the highest and the lowest. For a solid with $N \sim 10^{24}$, there will be such a

large number of curves between the limiting ones, that it will appear as a continuum. In other words, the energy levels allowed by quantum mechanics will for given lattice constant(s) form a continuous interval between a lower and an upper value. The allowed energies will lie in "bands", and between the bands there will not be any allowed energy levels. The formation of bands is illustrated in Fig 10.9.

If a band is "filled" with electrons, that is if there is precisely one electron per state (the 10^{24}-odd states), then the least energy, which will be capable of exciting the system, is the energy difference between the next higher (empty) band and the filled one. On the other hand, a lattice structure with a partly filled band on top will allow electrons to be moved to empty levels at very small energy expenditure. Structures of this kind give electrons a chance to move appreciable distances, while structures of the first kind will very rapidly recapture any excited electron back into the band, which is filled in the lowest energy state. The structures allowing large-scale electron transport are "conductor", those which do not are "isolators".

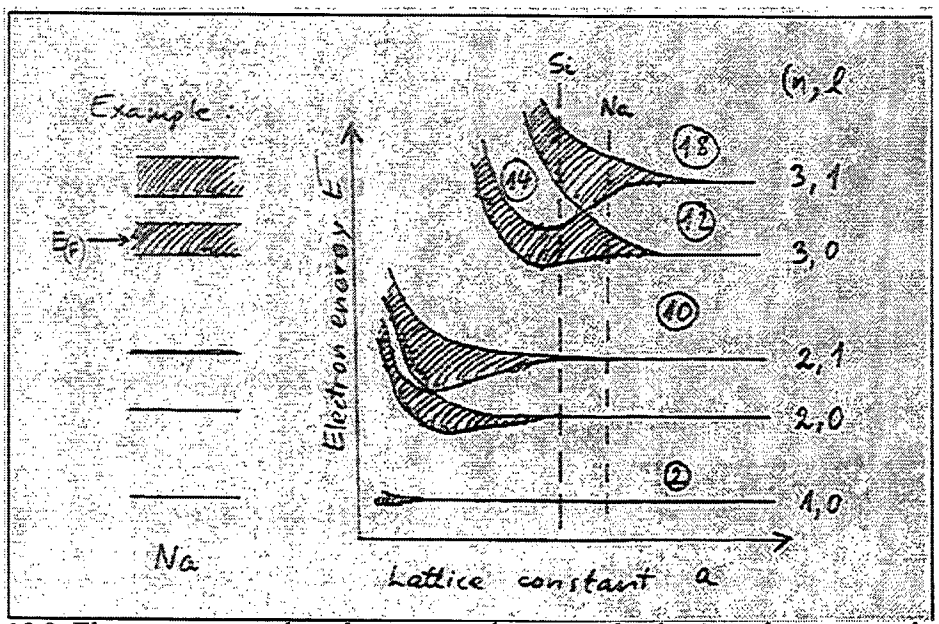


Fig. 10.9. Electron energy band structure for a regular lattice of atoms, as function of the lattice constant *a*. For large *a*, electrons in different atoms do not interact. The picture is highly schematic, and the energy scale and precise form of the band-limiting curves are different for different materials.

For electrons in a partly filled ("conduction") band, the variations in the Coulomb potential, as the electron moves across the lattice with its discrete sites of nuclear charges, may as a first approximation be neglected. In that case the conductor is like the box treated in problem 8.5 (or the similar one in three dimensions). The wave functions will be sine functions that are zero on the external boundaries, and the lowest energy state will have the electrons filled from the bottom up to an energy value E(F) determined by the number of electrons present. E(F) is called the Fermi level. Only at zero temperature does the filling rigorously stop at E(F). At higher temperatures, there is kinetic energy corresponding to a distribution of electron velocities (cf. Chapter 12). Those electrons

having kinetic energy above average will have a chance of being in states above E(F), and the number of electrons per unit of energy will be modified as shown in Fig. 10.10.

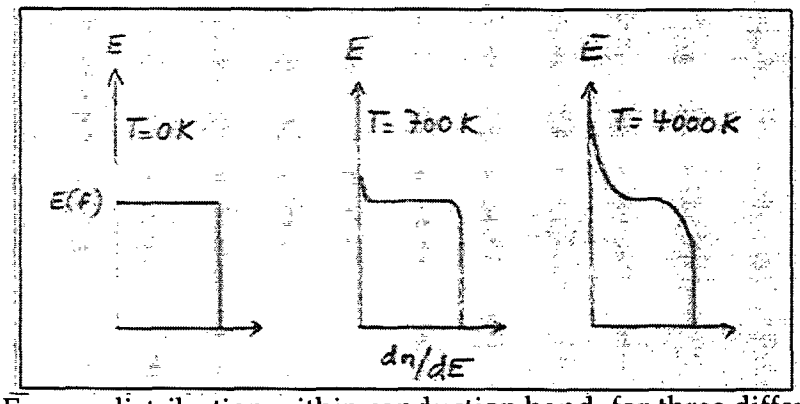


Fig. 10.10. Energy distribution within conduction band, for three different temperatures. The number of electrons per unit energy interval is denoted dn/dE.

Corresponding to the vibrations of molecules, there are excitations in solids characterised by periodic oscillations around some equilibrium value. These excitations are sometimes called "phonons". The quantity oscillating may be electron density, spin direction or inter-atom distance. The latter case corresponds to the oscillation in distance between atoms in molecules, discussed above. The spin vibrations arise from the part of the electron interaction involving spins. This interaction is appreciable only for electrons close to each other (say in adjacent atoms), but still the excitation it collective: spin directions are varying periodically throughout the lattice (Fig. 10.11). All the different phonon interactions may be described by the boson formalism discussed above.

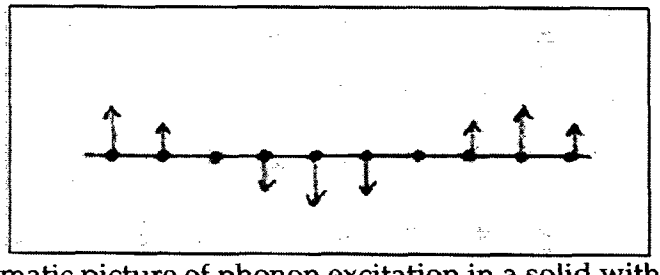


Fig. 10.11. Schematic picture of phonon excitation in a solid with lattice structure.

If external forces exerted on the surfaces of the lattice structure are capable of changing the lattice constants temporarily, the material is "elastic". If the change is permanent, the structure of the material has been changed, and the lattice possibly "broken". Deformation properties and material strengths play an important role in industrial and other manufacture. However, only in rare cases can a quantum mechanical lattice calculation be carried out with sufficient accuracy to predict material properties. In some materials of lattice structure, the lattice constants change strongly with temperature. These liquid crystals have several applications, notably due to the temperature dependence of light diffraction (display screens, etc.).

The properties of materials, such as heat capacity and strength, are of major importance in everyday applications. So are electromagnetic properties. Outstanding among these

is the phenomenon of permanent magnets. Individual atoms have magnetic moments, deriving in part from their orbital angular momentum and in part from their spin. The possibility of aligning the magnetic moments of solids is related to quantum mechanics, because classically, the moments would be randomly oriented for material in thermal equilibrium.

At low temperatures, many substances exhibit superconductivity. This phenomenon is caused by weak interactions, which are overshadowed by the thermal excitations at higher temperatures. They may be called "pairing interactions", as they favour states with pairs of electrons, having opposite momenta and spins, but otherwise being in the same quantum state. In the absence of magnetic fields, states of opposite spin directions have the same energy, and so have states of equal but oppositely directed momenta. The spatial wave function of the electrons overlap strongly, and a two-particle interaction of the following form has been suggested (Bardeen, Cooper and Schrieffer, 1957),

$$H(\text{int}) = -G \sum f^{+}(p,\uparrow)\, f^{+}(-p,\downarrow)\, f(-p',\rightarrow)\, f(p',\leftarrow),$$

written in second quantization. The electron creation operators are denoted f^{+} (as they create fermions), p stands for all three momentum components, and the spin directions have been indicated by pairs of opposite arrows. The summation is over some interval of momenta p and p', or alternatively the constant G is replaced by a state-dependent factor declining with increasing momentum difference $|p-p'|$ (and then placed inside the summation sign).

The interaction energy gained by lifting pairs of the kind considered here from below the Fermi level to above it may now exceed the energy spent in lifting the pair up. As a result, a new ground state-has been formed with energy lower than the original one. To excite an electron from the new ground state will involve breaking one pair and losing the pair correlation energy given above. A gap of a certain magnitude (say Δ per electron) has been created within the energy band (see Fig. 10.12). The usual loss mechanisms interpreted as electric resistance appear to have been quenched, because they involve a transfer of energy from a moving electron to the lattice vibrations (heat) by collisions shifting the electron between close-lying states. With a minimum excitation energy of 2Δ (breaking a pair), these collisions are no longer able to change the state of electrons. This means that there is no resistance to electric current: the material has become superconducting. At higher temperatures, the temperature excitations exceed 2Δ and the pairs no longer hold together: superconductivity is lost above a certain critical temperature.

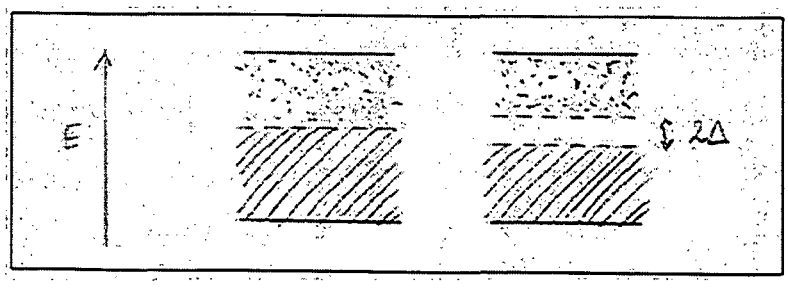


Fig. 10.14. Normal (left) and superconducting (right) structure of conduction band. Unoccupied band area is dotted, occupied one cross-hatched.

10.4 System theory

System theory is an approach, in which the time behaviour of aggregated quantities is described by differential equations. The change in such quantities is given by rates of input and output. Denoting these I and O, and the quantities Q, the equations to solve are of the form

$$dQ_i/dt = I_i - O_i$$

The rate functions I_i and O_i may be given externally, they may describe sources and sinks, and they may depend on Q_i as well as on the other Q_n's ($n \neq i$), taken at the actual time t or at some previous time t' (thus describing delays). Some inputs may equal other outputs (thus constituting a transfer between two compartments Q_i and Q_n), and some inputs may increase with increasing levels of certain of the quantities Q ("feedback").

An example of a compartment model amenable to system dynamical modelling is the one shown in Fig. 10.6. The general problem of compartment models is, that compartments have to be wisely chosen. If they are not, then models for the input and output rates cannot meaningfully be established, for instance because parts of what constitutes the compartment develop in ways that cannot be determined from the selected compartments alone.

PROBLEMS AND DISCUSSION ISSUES

DISCUSSION ISSUES 10.1. Consider the following two statements:

(A): All that is possible will eventually happen. Evolution theory tells us that anything with a finite probability will occur, if we wait long enough, and if other traits of evolution have not changed the probability to zero in the meantime. We have made nuclear extinction possible. Some day it will happen.

(B): A lot of things, which were theoretically possible, have not actually happened. Examples of this are numerous throughout history. This shows that we have a choice. We can choose not to do something which is technically possible. There are a range of possible futures. We should devote our efforts to selecting the best and work for its realisation. It does not come by itself.

Can you identify with any of these views?

PROBLEM 10.2. Describe the water cycle on Earth, that is the cyclic transformation and transport of water. You may use Fig. 10.13.

Can you give a gross estimate of the total hydropower potential of the Earth, based on the data given in the figure and the average elevation over sea level of the total land surface, which is about 840 meter.

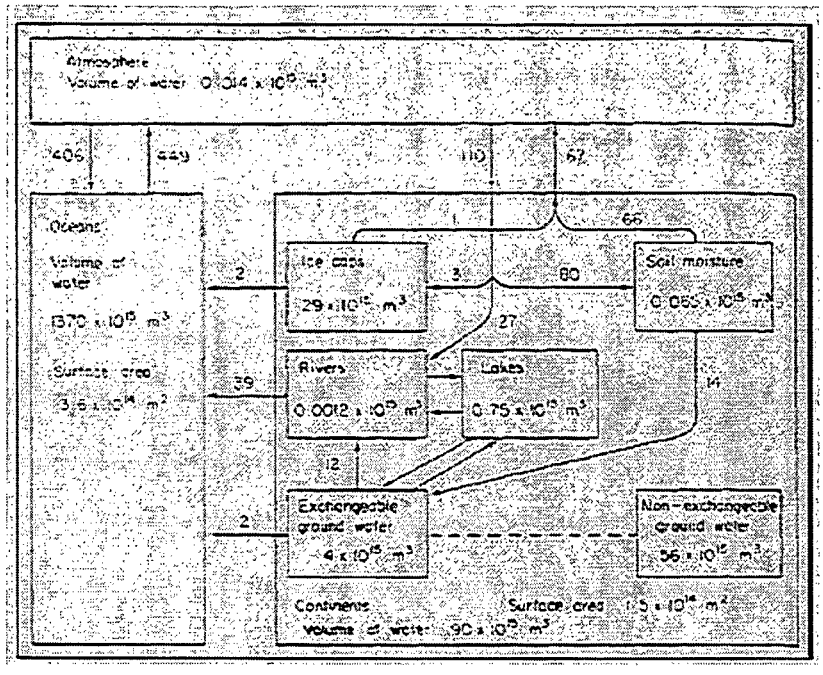


Fig. 10.13. Average water cycle on Earth, including free water to a depth of about 5 meters below the surface. The transfer rates at the arrows are in units of 10^{12} m^3/y (Sørensen, 1979).

PROBLEM 10.3. Can you find an expression for the energy of the electron in the ionised hydrogen molecule, when the wave function is assumed to be of the form given in section 10.3. Assume the integrals of H with wave functions for the same or different atomic centres to be known (the first is constant and the second a function of R). You will need also to assume the integral of the differently centred wave functions themselves to be a known function of R.

PROBLEM 10.4. An early suggestion for the cause of cancer, and perhaps also a cause of ageing, has been errors induced in DNA molecules by irradiation or by chemically induced changes. The DNA molecule is shown schematically in Fig. 10.14. If the model proposed above is correct, one might think that induced changes not destroying the general structure of DNA would be most likely to be copied to new cells and thus to have the stated effects. It has been proposed (Löwdin, 1961), that an important change of this kind would involve two proton (H) transfers through potential barriers (by quantum mechanical tunnelling). For example, in the Guanine-Cytosine bond (Fig. 18.14), the H in NH...H may move and form N...HN, at the same time as the H in O...HN moves and forms OH...N. It is favourable to move both protons at the same time, because else there would be a ionisation process, at a higher energy expense. The energy as function of the distance travelled by one of the protons (the other one moving similarly in opposite direction) is shown in Fig. 18.15.

The probability that thermal energy spread will cause the tunnelling to occur is 1000 times less than the observed mutation rate of 10^{-10} in 30 years. However, the probability of exciting a Guanine-Cytosine pair or an Adenine-Thymine pair by background radiation is around 10^{-24} in 30 years. Once excited, there is a chance of 10^{-5} that the protons will tunnel to the opposite positions and thus cause copy errors when each strand is transferring its code to a messenger RNA molecule. Estimate the total probability of background radiation induced mutation by this process and compare with the measured one (a human person may be assumed to have 10^{12} cells, each with 24 DNA

molecules, each of which again has 3×10^6 G-C or R-T pairs).

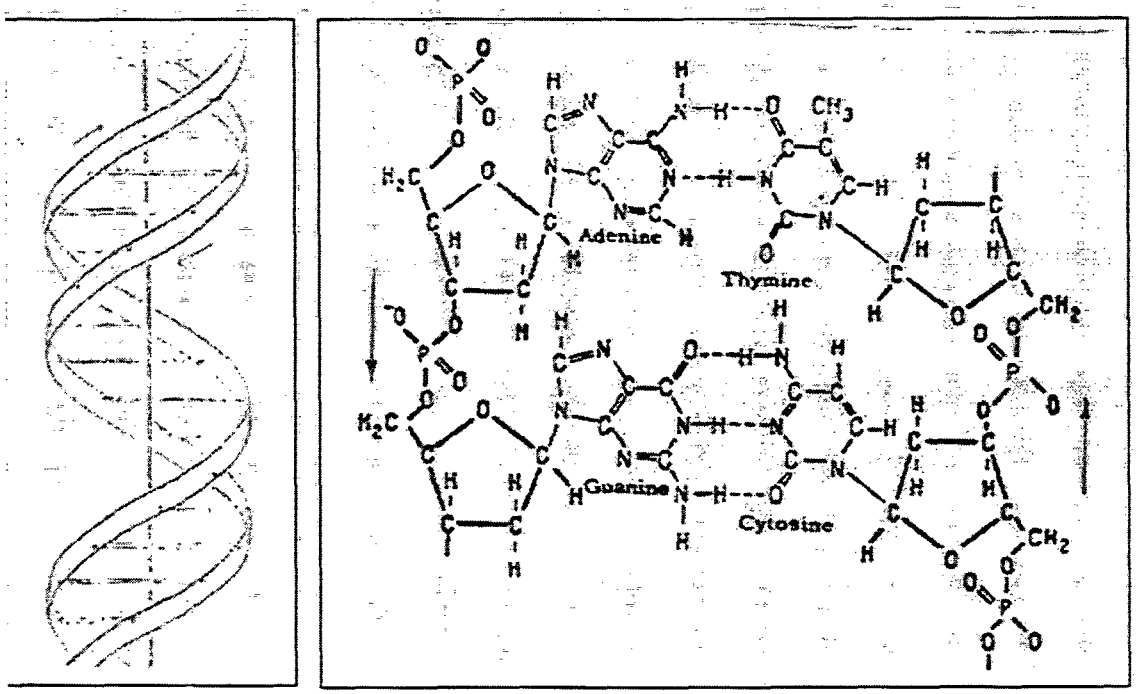


Fig. 10.14. Schematic structure DNA (left), with the binding of the two strands illustrated to the right.

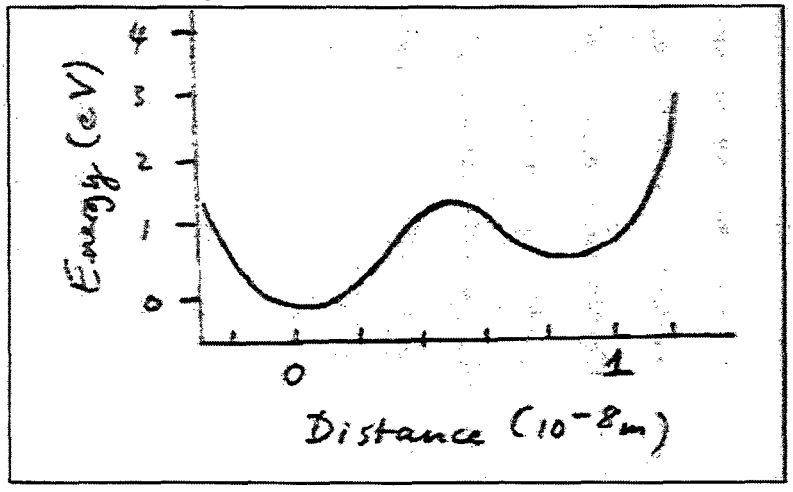


Fig. 18.15. Energy variation as function of proton position along line of binding (see text for details).

DISCUSSION ISSUE 10.5. Do you think the "microscopic" explanation of cancer and ageing offered in problem 10.4 is plausible, or do you rather believe in a macroscopic, system-type explanation model?

PROBLEM 10.6. Single crystals made from isolators are usually transparent, while metals are totally opaque. Explain these facts.

PROBLEM 10.7. Explain why the electric resistance in metals increases with temperature.

DISCUSSION ISSUE 10.8. Do you think it is possible to explain conductivity of solids without using quantum mechanics?

PROBLEM 10.9. Work out the system dynamical equations for a simplified version of the ecological system depicted in Fig. 10.6. Include for example only two compartments: the herbivores and the carnivores. The food for the herbivores may be taken as produced at a constant rate, and the birth and natural death rates of herbivores and carnivores may be taken as two different sets of constants. The birth surplus should be taken as higher for the herbivores than for the carnivores, in order to make room for predation. The predation rate can be modelled in different ways. For example, it could be proportional to the number of carnivores, but it could also depend on both the number of herbivores and carnivores, in a way reflecting the chance of finding a prey in a given area. The birth rate may also be modified to depend on whether food is adequate or not.

If you have a computer at hand, try to make a program for the description of the time development of this system, and run it with different sets of data. Try to get realistic results for cases such as a hare and fox population in Australia, initially being 99% hares and 1% foxes.

Chapter 11

Military technology

In order to predict the path traversed by an object thrown through the air, physics has to be used. A stone or an arrow approximately follows a parabolic orbit, when travelling near the surface of the Earth. If the gravitational force is the only one important, this is the exact solution. If air resistance is considered, the projectile is slowed down along its course, and turbulence and wind may change the course of the projectile, and its orientation (for example in case of the arrow). Physical theory of great complexity may thus be needed in order to predict the path, or alternatively, physical theory may be used to shape the projectile, in order to minimise friction and hence make a simplified trajectory calculation more reliable. These applications of physics have been numerous throughout history, and of course they are still used extensively for calculating trajectories of missiles, bombs and so on. When a missile of larger range (for example an intercontinental one) is considered, the approximation of a constant gravitational force is no longer sufficient, and Newton's full law of gravitation (Chapter 9) has to be employed, possibly even including the effects of uneven mass distribution near the Earth's surface.

Missiles in the current arsenal of the majority of belligerent nations are charged with one or more bombs. The destructive power of the bombs varies greatly according to the planned use. Stocks comprise smaller bombs for destroying enemy tanks or planes, intermediate size weapons for use against military targets, and bomb of devastating destructive power for wiping out entire cities. Other main characteristics of the weapon systems are range and precision, and whether they are to be launched from land, sea or air.

The destructive power of smaller weapons may derive from chemical reaction energy. From gun powder to TNT (tri-nitro-tuluol), but for the larger bombs, and for some of the smaller ones too, nuclear energy is used. I shall deal in a bit more detail with the nuclear weapons, which are strongly related to physics research, but in passing I would like to remind you, that there are other weapon types, such as chemical and biological weapons, which may have comparable damaging effects on human lives and the ecological environment.

The energy released by a nuclear bomb is often described in terms of the TNT equivalent. One tons of TNT corresponds to about 4.4×10^9 joule. A bomb which releases 1000 tons of TNT equivalent energy would then be described as a 1 kt (kiloton) bomb.

The physical process used in nuclear bombs is fission of heavy atomic nuclei, or fusion of light ones. As mentioned in Chapter 9, nuclei around iron are the most stable ones, that is they have the highest binding energy per nucleon. A heavy nucleus, say plutonium, could therefore gain energy by splitting into a number of smaller nuclei. However, before getting to the more bound, more stable iron region, the system must pass a potential barrier (the "fission barrier"). In other words, there must be supplied energy sufficient to climb the barrier, before the system can enter the next valley.

The very heavy nuclei may fission spontaneously, because a quantum mechanical sys-

tem always has a finite probability for tunnelling through a barrier of finite height and width. However, the probability of fission increases strongly, if the nucleus is excited to a level near to or above the "rim" of the potential barrier. This is achieved by bombardment with neutrons. An example of the reactions that may occur is

$$^{235}U + n \rightarrow {}^{142}Xe + {}^{92}Sr + 2n$$

Among the uranium isotopes (that is nuclei with different number of neutrons, but 92 protons), uranium-235 most easily absorbs a neutron and fissions. The plutonium isotope Pu-239 has the same property. The reaction above releases about 200 MeV, of which some 165 MeV is kinetic energy of the fission fragments (xenon and strontium in the example above). It is seen that two neutrons are formed but only one spent. In some other fission reactions the number of neutrons emitted per neutron absorbed is even greater than two.

The excess neutrons may produce additional fission reactions in a lump of a fissionable material, such as U-235 or Pu-239. These fissions also create excess neutrons, which again induce fission, and so on. This is called a chain reaction. Each step in the chain takes about 10^{-8} seconds, and after n steps, there will be $N = \exp(xn)$ neutrons for each initial one (Glasstone, 1962). The number of excess neutrons, x, is negative for small lumps of uranium or plutonium, because the chance that a neutron escapes from the lump is high. When the size of the lump increases, its surface to mass ratio diminishes, and for a certain "critical mass", the number of excess neutrons, x, becomes positive. The critical mass is about 10 kg for U-235 and about 2 kg for Pu-239. For larger bombs, x is very close to 1.0.

Figs. 11.1 and 11.2 show the principles used in the first two fission bombs, dropped by the United States over the cities of Hiroshima, August 6th, and over Nagasaki, August 9th, 1945.

The Hiroshima bomb used two sub-critical U-235 masses. One sits in a cannon tube and is fired into the other one by a conventional explosive. Around the second uranium lump is a mantel made of a material capable of reflecting many neutrons back into the uranium lump (e.g. beryllium). In the region where the two uranium lumps collide, there is also a neutron source, that is a material, which above a certain temperature (provided by friction heat during the collision) starts to emit neutrons, which can then initiate the fission chain reactions. This nuclear bomb type is not in use any more, because the slow assembly of the subcritical masses to an above-critical one makes the process difficult to control, so that an explosion of only part of the material may disrupt the bomb before all the fission processes have taken place.

The Nagasaki bomb shown in Fig. 11.2 uses plutonium-239, but in a form, which makes the overall density too small for criticality. The excess neutrons will initially not be able to escape, because they become re-absorbed in the non-plutonium atoms present. A series of directionally exploding conventional charges surround the plutonium core, and when fired they give rise to an implosion, that is a pressure wave moving inward. It presses the plutonium together, so that it reaches the critical density. Then the fission

chain reaction is started, again with assistance from a neutron source.

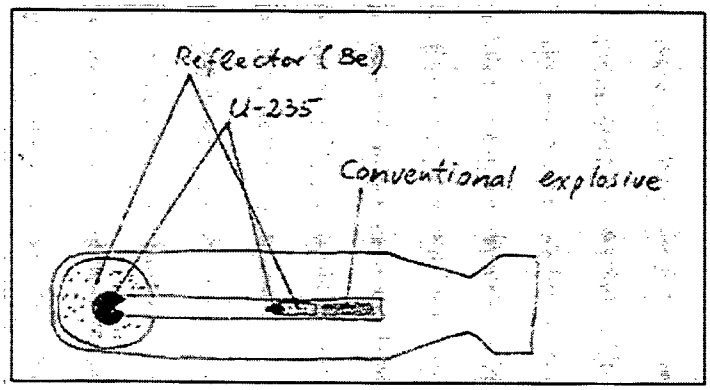


Fig. 11.1. Schematic picture of Hiroshima fission bomb.

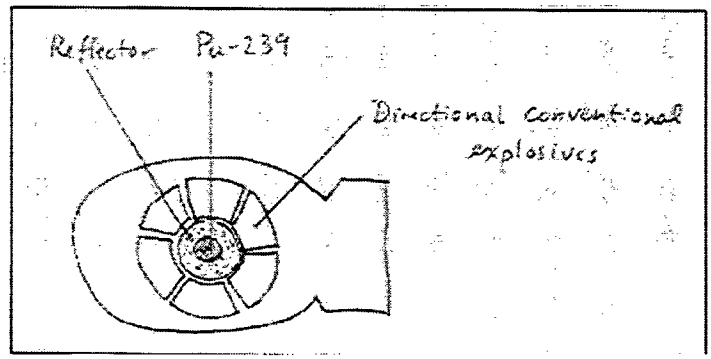


Fig. 11.2. Schematic picture of Nagasaki bomb (implosion bomb).

The plutonium "ignition" time is short for the implosion bomb. This reduces the risk of early ignition caused by the spontaneously fissioning plutonium isotope Pu-248. Plutonium is produced by placing uranium-238 in a nuclear reactor. In a reactor, the U-235 fission reaction is controlled, by arranging the density and presence of neutron absorbers in such a way that the number of excess neutrons can be controlled and put at x=0 once the desired process rate is obtained. The fuel is natural uranium (mostly U-238) with an enrichment of the isotope U-235, usually to 0.7%. Thus U-238 will absorb a good deal of the excess neutrons and form U-239, which can emit two subsequent electrons to form Pu-239. When Pu-239 is formed, it may also absorb neutrons and either fission or form Pu-240. It follows that unless a new enrichment process is performed, the Pu-239 will be "contaminated" by Pu-240. The too early disintegration of bombs by Pu-240 initiated fission, may as said be prevented by using a rapidly imploding device.

The controlled fission reactors are in use not only for the production of plutonium for bombs, but also for electric power production. In this case the fission products are slowed down in some absorbing material and the heat transferred to a fluid by means of a heat exchanger (for instance pipes going through the absorbing material). The hot fluid is used to drive a conventional thermodynamic cycle in a power plant (see Chapter 12).

The nuclear energy associated with fusion of light nuclei has been used for nuclear bombs since 1952. The processes are similar to those occurring in the interior of stars (see Chapter 9). In hydrogen bombs, the naturally occurring H-2 (deuterium - present as 3.4×10^{-3}% mass fraction in sea water) reacts with H-3 (tritium - the natural abundance

of which is zero),

$$^2H + ^3H \rightarrow ^4He + n + 17.6\ MeV$$

This process starts at temperatures above 4×10^7 K, that is considerably below the temperatures needed for the fusion processes occurring in stars (see Chapter 9). Still the establishment of this temperature inside bombs is difficult, and the tritium required is expensive to produce. These two problems are solved by using a lithium hydride, (Li-6)-(H-2), as the basic fuel, and to initiate the fusion process by a fission bomb. Lithium hydride is much easier to compress than a hydrogen gas, and the set-up may be as depicted in Fig. 11.3 (Morland, 1979). The primary fission explosion emits a wave of intense radiation and a pressure wave. The radiation wave reaches the fusion material some 10^{-6} sec before the pressure wave disrupts the whole thing. With suitable arrangement of geometry, this is enough to get the fusion process going. Neutrons formed in the internal jacket gets absorbed by Li-6 and form tritium for the fusion process written above,

$$^6Li + n \rightarrow ^3H + ^4He$$

In other words, there does not have to be any initial tritium present. Lithium-6, obtained from natural lithium (mostly Li-7) by isotope separation, furnishes the tritium, and the primary fission bomb furnishes the pressure and temperature for the H-2 + H-3 process. Furthermore, the whole bomb is surrounded by a uranium-238 blanket. Instead of escaping, many neutrons get absorbed here and produce fission. A typical hydrogen bomb may derive 14% of its explosive yield from the primary fission process, 43% from the fusion process and the remaining 43% from the second U-238 blanket. Furthermore, the explosive yield may be increased at a very low expense by just putting more and more U-238 around the bomb. Most present "hydrogen bombs" are therefore not pure hydrogen or fusion bombs, but often mostly fission bombs. However, a pure fission bomb is limited to around 500 kt (the Hiroshima bomb was 20 kt), while the fission-fusion-fission bombs can be built to much higher yields (60 Mt bombs exist in present arsenals).

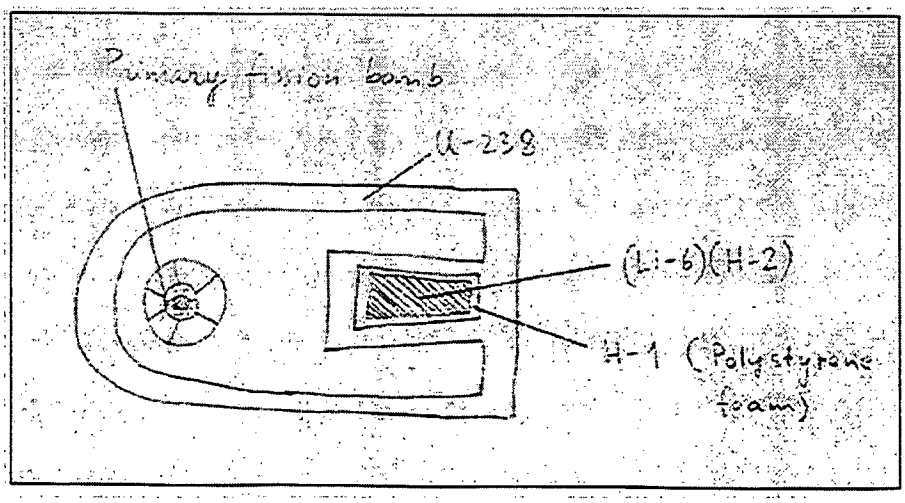


Fig. 11.3. Schematic picture of hydrogen bomb (or better "fission-fusion-fission" bomb)

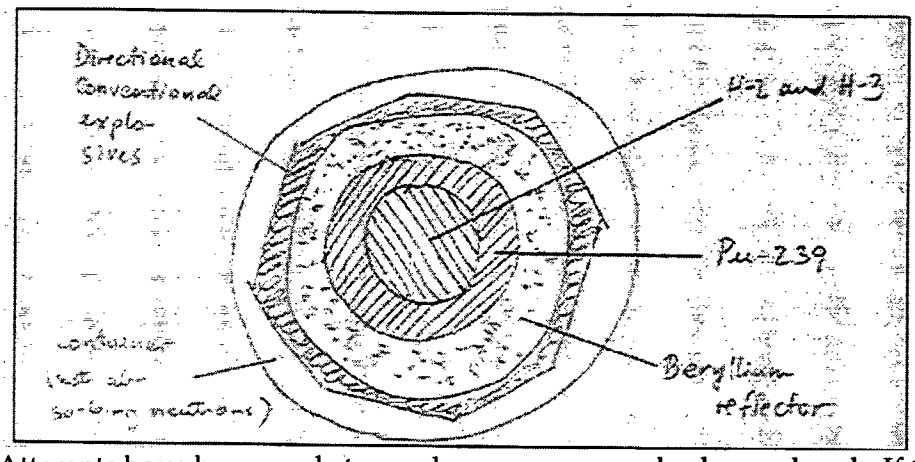


Fig. 11.4. Schematic picture of neutron bomb

Attempts have been made to produce a more pure hydrogen bomb. If the U-238 blanket is left out, more neutrons escape, but the explosive yield of course goes down. If the purpose is to enhance radiation at the expense of physical destruction (i.e. to kill people and destroy plants and food, but preserve buildings and vehicles), also the neutron-absorbing Li-6 process must be eliminated. This is done in "neutron bombs" by using a mixture of deuterium and tritium, which is brought to the required temperature by a fission bomb, as illustrated in Fig. 11.4 (Vitale, 1982).

Neutron bombs with 75% of the yield deriving from fusion have been resigned. The major problem has been to produce tritium at reasonable cost. High neutron-flux research reactors (which exist for instance in the United States and in France) may be used to produce tritium by the Li-6 + n process, just as in standard hydrogen bombs, but here the tritium is made in advance. Due to the finite lifetime of tritium (half-life 12 years), the fuel of stockpiled neutron bombs will have to be replaced at regular intervals.

The effects of nuclear weapons depend on their type and size, and on how they are used. The initial radiation is absorbed by the air, and an increasing "fireball" of hot air is formed. It re-re-radiates power at declining temperatures. This thermal radiation causes third-degree burns at distances of 5, 11 and 22 km for bomb yields of 100 kt, 1 Mt and 10 Mt (Rotblat, 1981). The pressure wave formed by the heated air often contains over 50% of the energy released. It destroys structures and kills people to distances of 2.5, 5 and 10 km (50% survival rate at those distances), for the three bomb yields quoted above. Ionisation of the air in the initial phase of the explosion causes an intense pulse of electromagnetic radiation to be emitted. For a high-altitudes burst this is likely to destroy electronic equipment (and hence limit communication), which has not been protected by for example Faraday (metal grid) cages.

Radioactive material from the bomb (notably fission products) and from material activated, for instance by neutrons from the bomb explosion, will fall back towards the Earth's surface, with a time and area distribution pattern difficult to predict accurately. If the burst takes place close to the ground, a lot of soil and other material evaporates and gets radioactive by activation processes. The subsequent radioactive fallout in this case increases dramatically.

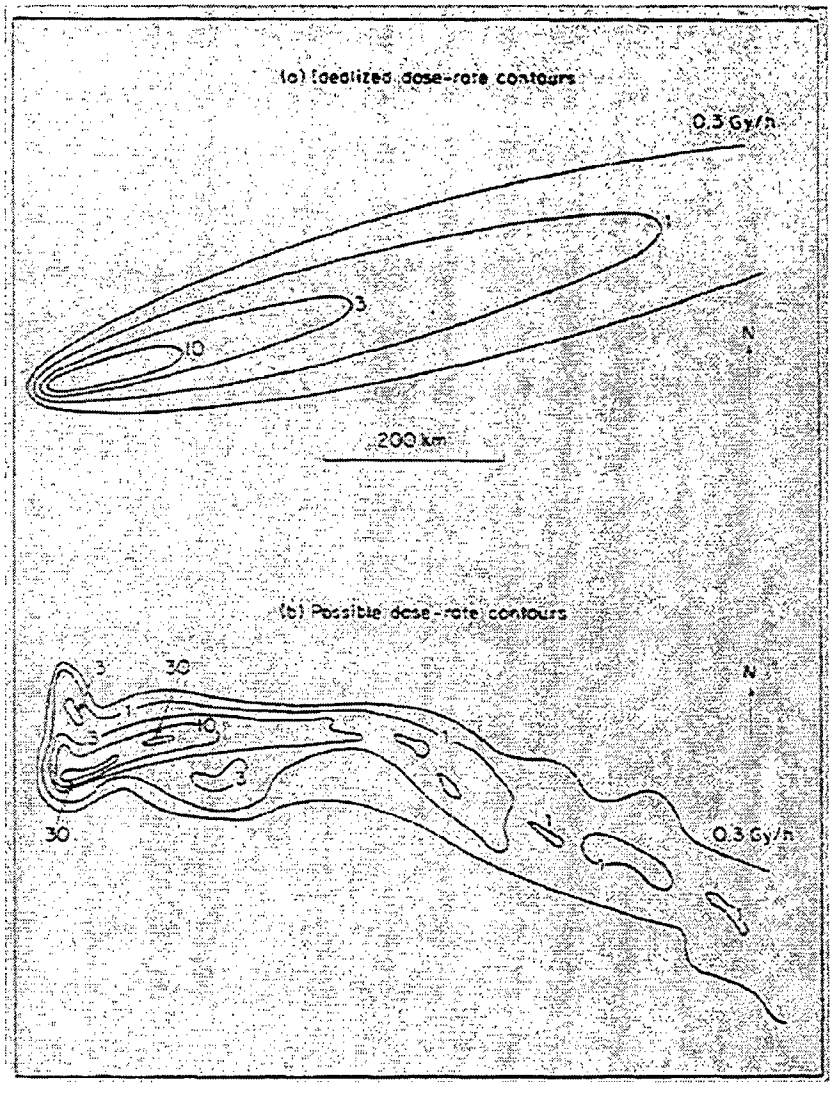


Fig. 11.5. Idealised (a) and realistic (b) dose-rate contours for a 10 Mt surface burst and wind velocity 50 km/h (from Rotblat, 1981)

The fallout composition and magnitude depends on whether the bomb is mainly fission or fusion, and on whether or not the bomb has been seeded with material enhancing the lethality of fallout. For large bombs, a considerable fraction of the radioactive debris may get into the stratosphere, from where it will slowly precipitate back into the lower atmosphere over periods of several years. The rest of the debris will reach the ground around the ground zero, in a pattern determined by wind, turbulence, gravitational settling (depending on mass of the dust particles), and on precipitation (rain, snow, etc.). Fig. 11.5 shows an example of a possible fallout pattern, along with the idealised one obtained from calculations assuming fixed turbulence parameters, fixed wind speed and -direction, and no precipitation.

The unit for radiation dose, used in Fig. 11.5, is gray (Gy). One gray is the energy in joules, deposited in one kg of a human body, due to the radiation received. Gy per second is thus a measure of the dose rate - the rate at which radiation energy is deposited in the body. Fig. 11.6 gives an idea of the harm induced by different dose levels.

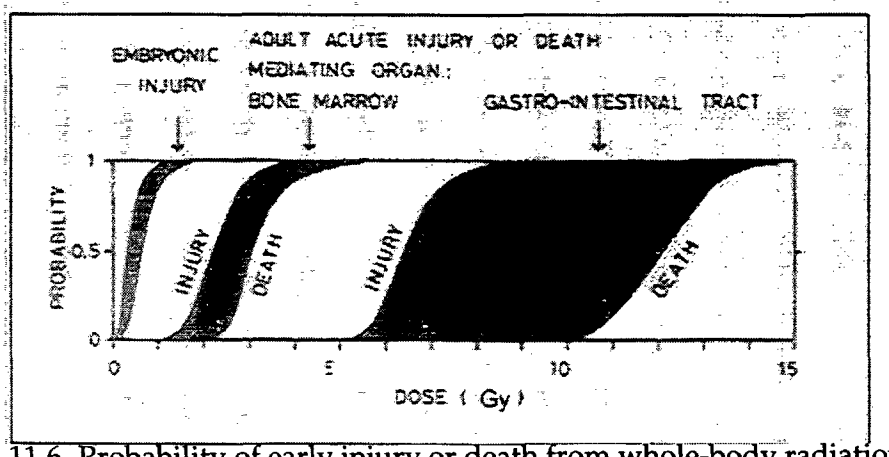


Fig. 11.6. Probability of early injury or death from whole-body radiation exposure, by various mechanisms. For children, the curves lie between those of embryos and those of adult persons (Sørensen, 1979b).

Knowing the energy absorbed by human tissue or organs is not sufficient for predicting the biological effects. The nature of the radiation is also important. A joule deposited by electromagnetic radiation (X-rays or gamma-rays) and by electrons (beta-radiation) has about the same biological effect, but a joule deposited by neutrons, protons or helium nuclei (alpha-radiation) causes more damage.

The range of doses given in Fig. 11.6 gives the probability, say in the case of an adult person, for various effects, including early death (death within hours, days or a few weeks). The data are meant to describe a "standard" situation. If exceptionally good medical treatment is available, the survival probability increases. This could be the case if a dozen workers at a nuclear power or fuel reprocessing plant got high doses and were rushes to well-equipped hospitals prepared for this kind of accident. An essential part of the medical treatment is repeated replacement of the entire blood volume. It is clear that this option will not be available to victims of a nuclear war, even if such victims were brought to a surviving hospital. The curve for "early death of adult person" in Fig. 11.6 shows a fifty percent probability of survival at just over 3 Gy. If the persons in question are weak, due to shock, lack of sanitation, infections, poor shelter, etc., ss one must expect in a nuclear war, then the probability of death will rise in the area below 3 Gy, that is for doses which could be survived under "normal" circumstances.

The fallout from one 1 Mt nuclear bomb will produce accumulated doses over 2 Gy in an area of around 5500 square kilometres, a 10 Mt bomb in an area of about 52000 km^2.

Any "hit" by radiation may affect a biological system in an adverse way. Thus also radiation levels below those causing early injury should be considered. They may lead to biological changes, which could cause cancers or genetic damage. Leukaemia cases

have been seen to occur at increased rates during the years following radiation exposure. Other cancers typically have a latency period of 20 years, and they follow radiation exposure with this delay. All together, the accumulated increase in the average risk of dying from delayed cancers is 1/100 per Gy. This may be taken as independent of the distribution of the radiation over time (only the accumulated value counts), the dose level and the number of people, over which the dose is spread. This is of course only approximately true, but available evidence does not support any more detailed model for the radiation effects. Only at very high doses, the value of one cancer death per 100 Gy cannot be used, because each person can die only once.

Nuclear bombs may be dropped from airplanes or fired from conventional artillery (in the latter case, small tactical weapons). However, missile delivery systems make it possible to shoot across continents from land or sea-based ramps, and possibly from space. Missiles are aerodynamically shaped rockets carrying one or several bombs. The rockets are accelerated by use of hydrogen or a fossil fuel. The fuel is burned with oxygen (from the air or brought along), and a stream of hot gases are expelled from the tail end of the rocket. The rocket will experience a force equal to, but of opposite direction, compared to that of the expelled material.

It is important to look at the way, in which nuclear weapons are planned to be used. In theory, the plan may be not to use them. They should only deter the enemy. This way of thinking was prominent in the 1950ies, where a sufficient number of nuclear armed missiles were directed at enemy cities to be capable of inflicting unacceptable damage. This was mutually seen as a guarantee that the superpowers would never dare to attack each other. The problem with this "balance" is the possibility of a surprise attack. In consequence, enormous efforts have been made on both sides to develop early warning systems, using land, sea and air-based radar systems, and from the sixties satellite surveillance. The ability to detect and interpret satellite pictures and other intelligence information has also been greatly enhanced from the late 1960ies. The philosophy seems at present to be to ensure, that an enemy attack can be answered before the actual destruction has happened. This "mutual destruction" tactics does not consider what is best for mankind as a whole. However, the political decision process may in any case be too slow to allow a counterattack decision to be taken after the first series of missiles have left the enemy ramps.

Several responses have been taken to this situation. The underground missile sites on land have been reinforced and missile launching from deeply immersed submarines has been made possible. As a countermeasure, new nuclear missile systems have been developed on both sides, which are believed to be able to destroy enemy missiles in their reinforced underground sites. Thus the emphasis has been moved onto considerations of how to actually fight a nuclear war. Special missiles are targeted on enemy missile sites. They have high precision and sufficient destructive power to render the reinforcement useless. On the other hand, this is not 100% certain, and the game, at least up to the 1990ies, has been to increase the number of nuclear weapons to a level so high, that the number surviving a first strike by the enemy will still be sufficient to cause unacceptable damage. The nuclear submarines are currently the least vulnerable launching

sites, and it is clear that great efforts are being made to develop anti-submarine intelligence and weaponry (Wit, 1981).

Other efforts are in the direction of decreasing the enemy's warning time (for example by deploying (placing) United States intermediate range missiles in Western Europe, a few minutes travel away from major targets in the former Soviet Union). Alternatives to perpetual increases in the number of missiles and bombs in this strategic game would be to keep a limited number of weapons, but to guard them better. This could be achieved by placing a large number of mini-rockets around each missile site, for intercepting and disabling any enemy missile (Feld and Tsipis, 1979). This would be on example of a non-offensive solution, because it would not increase the risk to the enemy. By contrast, the solutions of increasing numbers, precision and speed of missiles and warheads, that is the solutions that have been selected by both sides, have had the effect of decreasing the security of the other side by making a successful first-strike attack more probable. The same would be the case if a space-based missile defence system could be developed. In the long range, such policy of course diminishes the security of everyone.

I think that more thought should be given to the development in general of defence systems with as little offensive capacity as possible. This is true also for battlefield situations and what has been termed "limited nuclear war". Considering that, like the Soviet Union probably had larger armies than the West, so will China or other future adversaries, the idea of using smaller ("tactical") nuclear weapons against a conventional army has been brought up, particularly with reference to a war in Europe. Tactical nuclear weapons include a number of short range missiles, and may be expanded with neutron weapons (for use against tanks) and cruise missiles (that is unmanned airplanes with computerised navigation and bomb dropping). Again this is an offensive solution to a problem, which could also have been solved in a way not lowering the threshold for escalation into a full-scale nuclear war. A number of small, conventional weapons have been developed, which are estimated to be very efficient in stopping a conventional army attack (Morrison and Walker, 1978). Because of their small range, they are not offensive, and because they are non-nuclear and will be used by a dispersed defence, they will not invite a nuclear counterattack. One may say that the technological development in conventional weaponry presently favours the defending side, much in contrast to earlier situations.

The emphasis on satellite surveillance and on communication via satellites during war has made anti-satellite weapons interesting to the satellite-depending nations. Killer satellites, which can destroy other satellites by some kind of missile, are being developed. Laser beam weapons are considered, along with other radiation type weapons (charged particle beams), because they would perhaps be easier to use in space than rocket-borne weapons. First of all, hardening the electronics in satellites against the effects of an electromagnetic pulse from a high-altitude hydrogen bomb explosion would seem most urgent. In any case, there seem to be simple countermeasures against beam weapons (Hussain, 1978; Parmentola and Tsipis, 1979).

11.1 Collective nucleon phenomena

Atomic nuclei are structures containing from one to about 100 protons (the actual number being denoted Z), and a number of neutrons, N, similar to or somewhat higher than the number of protons. To a first approximation, a nucleon (that is a proton or a neutron) may be described as an independent particle moving in a potential V. This potential, which should be used in the Schrödinger equation, is the average interaction with all the other nucleons. There is no "centre of force" as in the atomic case. Still, the energy levels in the nuclear case are very similar to the electron levels in atoms, shown in Fig. 10.7. Only the positions of the energy gaps are changed (and of course the scale of energies is in the MeV not eV range). The closed shells (large gaps) are not at 2, 10, 18, 36, 54 and 86 particles, but at 2, 8, 20, 28, 50 and 82 protons or neutrons.

Nuclei exhibit collective phenomena in analogy to molecules: if they are not spherical, they may rotate, and in any case they can vibrate in a number of phonon modes. These phenomena are easiest to recognise and describe for heavy nuclei.

The general problem with atomic nuclei is, that there are too few nucleons to produce collective phenomena of the kind seen in the electron states in solids. In the latter case, there are perhaps 10^{24} particles, in nuclei at most a few hundred. Still, collective models are used, and naturally, they do best for the heaviest nuclei.

A heavy nucleus can be viewed as a drop of "nuclear matter", that is of smeared out protons and neutrons, for which only a density and a surface shape have to be known. This model (Bohr and Kalckar, 1937) gives a reasonable account of the trends in ground state energies of different nuclei (and hence binding energy per nucleon, cf. Fig. 11.7), of surface vibrations and of rotational spectra ($E = aL(L+1)$) found in the rare earth and actinide elements. The nuclei with rotational energy levels must be deformed, since rotation of a spherical object would not give a different wave function and therefore would not describe a new state.It is not so easy to calculate the energy levels of systems with few nuclei, as it is for molecules with few atoms. In the latter case, the interaction is perfectly known (the electromagnetic one), while in the former case it is poorly known. In the quark theory, the nucleon interactions should be described in terms of the quark-gluon interaction, but it is not known in detail. In the model of Yukawa (1935), nuclei interact by exchanging pions (as in Fig. 8.2). Approximate calculations using such forces have been made for few-nucleon systems, with moderate success. For somewhat larger but still light nuclei (Be to Si), such calculations can only be carried out with several simplifications, and the agreement with measured properties is modest. The reason this is so much more difficult than for molecules is, that the nuclear interaction is stronger and exhibit sign changes at small distances. Small errors in the model interaction may produce results with no resemblance to reality. As one goes to heavier nuclei ($N+Z \geq 30$), the use of simpler, phenomenological interactions has proven more successful. With such interactions, deformations, surface vibrations as well as the energy states associated with adding or subtracting one nucleon from the nucleus can be reasonably calculated. For nuclei with an even number of protons or neutrons, pairing-type interactions may create a kind of superconductivity (Bohr, Mottelson and Pines, 1958).

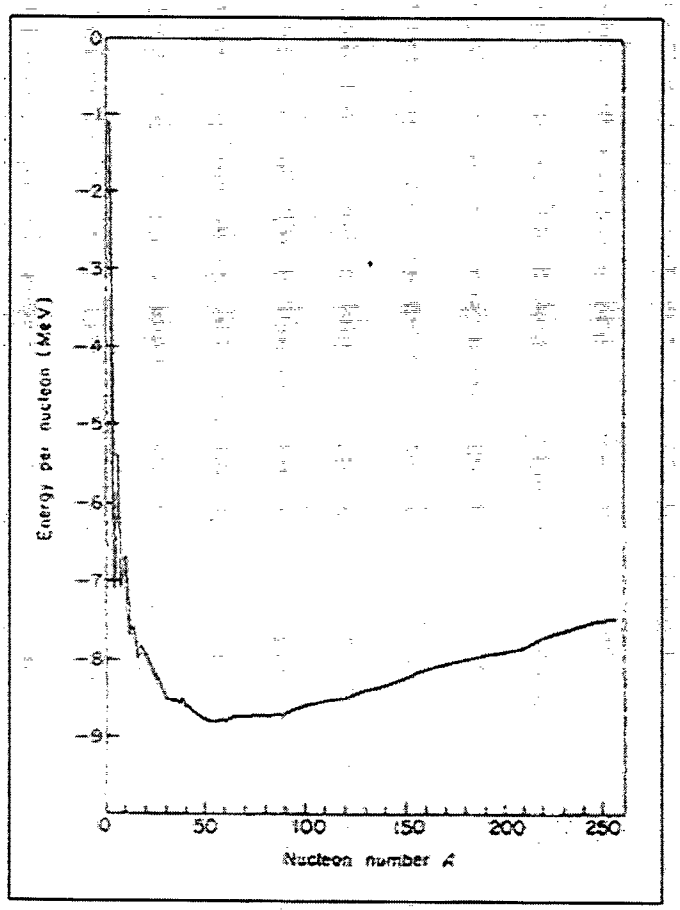


Fig. 11.7. Energy per nucleon, for lowest states of atomic nuclei. The zero on the energy scale corresponds to a situation with all nucleons separated from each other. The energy shown is thus minus the binding energy per nucleon.

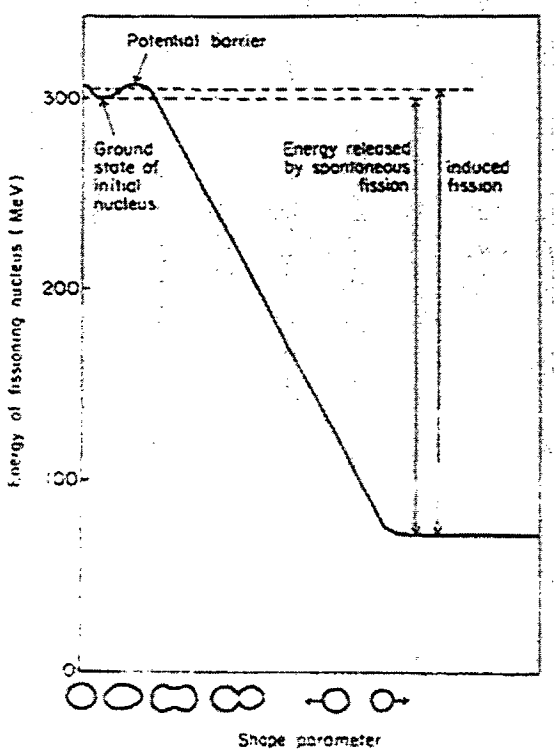


Fig. 11.8. Schematic picture of fission process, described in terms of a one-dimensional "shape-parameter". Zero on the energy scale corresponds to the energy of a (fractional) number of Fe-56 nuclei with same total mass as the fissioning nucleus (Sørensen, 1979).

The binding energy per nucleon variation, which is given in Fig. 11.7 for the most stable isotopes for each A = N + Z, shows the possibility of gaining energy by fusion and by fission.

When fission was discovered in 1939 (Meitner and Frisch, 1939), it could readily be interpreted in the drop model (Bohr and Wheeler, 1939). As schematically illustrated in Fig. 11.8, the ground state of a nucleus such as U-235 may be ellipsoidal, with energy a few MeV lower than that of a spherical shape. Once past the fission barrier, energy can be gained by forming an indention in the middle of the ellipsoid, by further deepening it and by finally cutting the nucleus in two.

PROBLEMS AND DISCUSSION ISSUES

PROBLEM 11.1. How many uranium-235 fissions are needed to release 100 kt?

How many steps in the chain reaction does it take to produce that many fissions? (assume X=1.0 excess neutrons in each fission process).

How many steps back from the last, in the chain of fission steps, should one go to have 99.90% of the energy yield in those last steps?

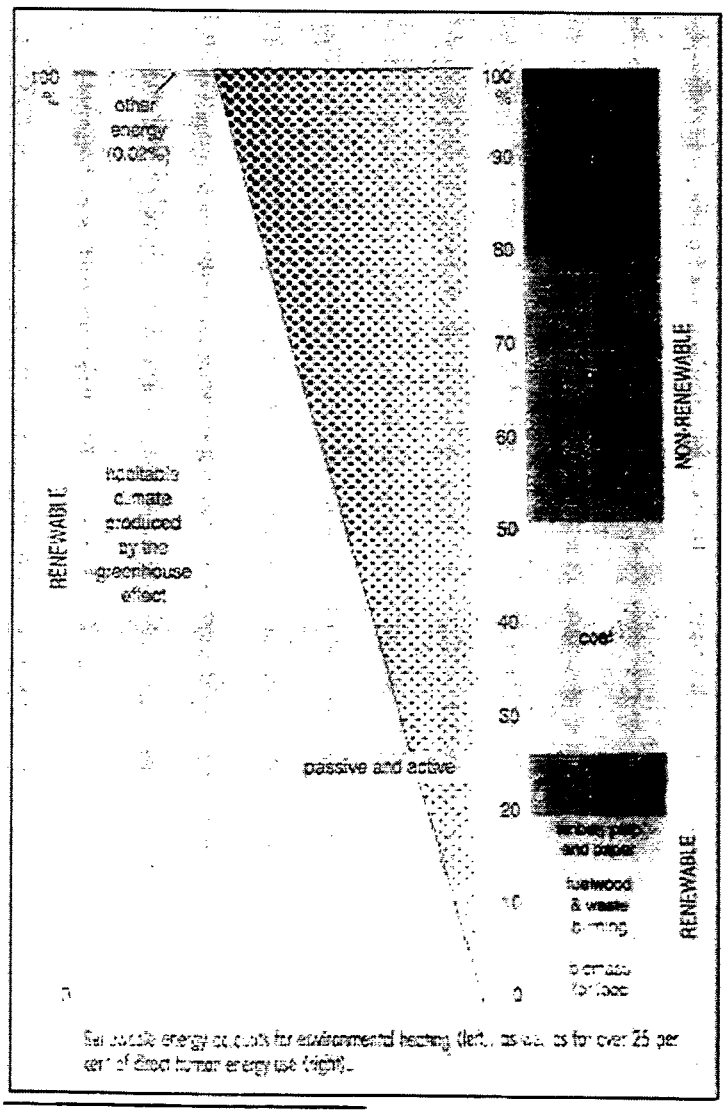
PROBLEM 11.2. What magnification is needed to get a satellite picture of an enemy missile silo to fill a page in this book? More precisely formulated: compare the angle spanned by lines from your eye to the edges of this page to the angle spanned by lines from a satellite to the rims of a missile field on the Earth (say they are 50m apart). Assume that the satellite is in a geo-synchronous orbit (see problem 9.3)

DISCUSSION ISSUE 11.3. Do you think that it is possible to convert weapon systems, so that they become non-offensive? Is it at all possible to work out ways of classifying weapon systems as being predominantly defensive or predominantly offensive? Inspiration may be found in Sørensen (1985).

Chapter 12

Energy technology

Energy is usually characterised as depletable or renewable, A practical definition of renewable energy is *a flow of energy, that is not exhausted by being used.* The primary renewable energy source on earth is thus solar radiation, because the earth-atmosphere system receives an amount of solar energy, which is essentially independent of the conversion processes that we may apply before allowing that energy to become re-radiated into space in the form of heat. On a more fundamental level, energy is of course a conserved quantity, but one that is degraded in quality by being converted into heat of lower temperature. Solar radiation is associated with the depletion of nuclear fuels in the sun, and it is basically the hugeness of this resource, that may allow us for practical purposes to consider solar radiation as a renewable energy source*.



Figur 12.1. Renewable energy's place in society.

* The exposition in this chapter draws on Sørensen, 1974; 1991 and 1992.

In general, renewable energy may be taken to include the use of any energy storage reservoir, which is being refilled at a rate comparable to that of extraction. In this way, geothermal energy use is renewable, as long as heat can flow into the sediment in question as fast as heat is being extracted. Fossil fuels would similarly be renewable energy sources, if they were only used at the average rate of fossilisation of biomass. Since, however, this rate is very small, compared to current use, fossil fuels do not at present qualify as renewable. Figure 12.1 shows the main renewable energy flows available on earth. The various forms are based on solar energy in its different converted forms, on heat stored in the interior of the earth or being created by radioactive processes, and on gravitational energy in the planetary system. An estimate of the amounts of renewable energy that may be recoverable in practice is presented in Table 12.1, with conventional fuel figures included for comparison.

The flow of solar and solar-derived energy forms is not independent from the activities of man. The radiation fluxes are modified by changing the reflectivity of the earth surface, e.g. by urbanisation and agricultural practices. Man's activities also change wind patterns and modify cloud coverage, thus again changing the radiation balance. Furthermore, the injection of pollution into the environment, and manmade structural changes, influence both radiation, heat and water flows.

The earth's atmosphere generates a greenhouse effect by retaining an amount of solar energy equal to five to six thousand times the sum of all current human energy use (i.e., conversion). This estimate is arrived at by calculating the difference between the amount of heat lost from the earth to outer space in the presence of the greenhouse effect and in its absence. Clearly, tampering with the greenhouse effect has a larger consequence than almost anything else humankind has thought of doing.

Aside from providing light and a climate suitable for human habitat, solar energy supports plant growth and animal subsistence, and it fuels general circulation patterns in the atmosphere and hydrosphere. Some of the solar energy directly benefits people, and should be counted in the human energy balance. Other portions of it may be said to constitute important external factors for the human endeavour.

Present use of renewable energy

In addition to the atmospheric trapping of solar energy involved in the greenhouse effect, renewable energy sources are currently utilised in a number of ways:

- Human society harvests biomass for food, providing energy and essential nutrients for our bodies.
- Human society burns fuelwood and waste to provide heat for comfort and for processing materials.
- Human society uses biomass to produce cut wood as well as pulp and paper. Other energy inputs would be required for mining and processing replacement materials.
- Human society uses hydropower for electricity production, plus modest amounts of wind power and a slight bit of power derived from solar cells.

• Human society uses solar thermal collectors for active heating and hot water supply. Much higher energy fluxes are associated with passive uses of solar energy in buildings, for example, by capture through windows in climates where heating is needed (in warmer climates, buildings are made to reject solar heat).

• Human society uses wind energy (air infiltration) in place of active air exchange and air conditioning systems in buildings.

• Human society uses biomass for biogas and biofuel production.

• Human society uses tidal energy for electricity production (in very small amounts).

Altogether, renewable energy sources provide over one quarter of the energy currently used around the world – aside from the energy involved in maintaining a habitable climate through the greenhouse effect.

Table 12.1. Estimate of global energy resources at the surface of the earth.

Resources	*Estimated as recoverable*	*Resource base*
Solar radiation	1 000 TW	90 000 TW
Wind	10 TW	1 200 TW
Wave	0.5 TW	3 TW
Tides	0.1 TW	30 TW
Geothermal flow		30 TW
Salinity gradients		3 TW
Biomass standing crop		450 TWyears
Geothermal heat stored	50 TWyears to ?	10^{11} TWyears
Kinetic energy stored in atmospheric and oceanic circulation		32 TWyears
Oil	300 to 2 500 TWyears	
Natural gas	180 to 2 500 TWyears	
Coal	930 to 7 000 TWyears	
Fission resources	90 to 9 000 TWyears	
Fusion resources	0 to 10^{11} TWyears	

Note: For renewable energy sources, flows are given in TW, while for non-renewable resources, an estimated resource range (from proven and possible to ultimately minable) is stated in TWyears (Jensen and Sørensen, 1984).

State of renewable energy technology

After a long period of neglect, the 1973 oil crisis spurred efforts to develop a broad range of renewable energy technologies. Today, several of these technologies are economically viable, while the ability of others to succeed in the marketplace is still uncertain.

Use of *flat-plate solar collector systems* for space and water heating has reached technical and economic viability in several parts of the world. Hot water systems are widely used in areas where there is little call for space heating. At higher latitudes, total heating sys-

tems containing solar components have been perfected. These use selective surface panels for high performance, and collectors able to withstand harsh, humid climates. Solar space heating systems still are more expensive than fuel-based alternatives (if the hard-to-measure cost of adverse environmental effects is excluded), and so their market penetration is low. Nevertheless, demand is large enough in several countries to ensure ongoing production and improvement.

Continuing technical improvement has allowed *photovoltaic collectors* to gain an increasing number of marketplace niches. However, solar cells are still too expensive to be considered for use in bulk power generation. Along with conventional crystalline silicon cells, research is under way on carious high-efficiency thin-film cells. A fair-sized market exists in the area of remote power supply.

Concentrating solar thermal devices are still in the experimental stage. Demonstration plants have been built, but no clear economic advantage has been demonstrated over photovoltaic devices – whose durability and maintenance-free operation give them a wider appeal.

The modern variety of *horizontal-axis wind turbines* has been technically perfected, gaining a fair marketplace position in areas such as Europe and the US. Under good wind conditions, these turbines are competitive or near-competitive relative to fuel-based alternatives for bulk production of electricity. Denmark is the lead manufacturing country supplying over 50% of the world market, including turbines mounted off-shore at water depths up to about 15m.

Harnessing waves and tides has been explored in a few prototype installations. Several problems have been encountered, and further development is needed in order to evaluate the potential, especially in the case of wave energy.

A number of installations have been built to use *geothermal energy* in a renewable or near-renewable fashion. These have been intended mostly for heating and hot water delivery, unlike the conventional, non-renewable geothermal steam turbine plants producing electric power.

Hydroelectric technology continues to be widely used. Hydropower plants of extremely large size have been built, but some unfortunately with negative environmental effects and forced displacement of populations from flooded areas, with little or no compensation. Considerable progress has in recent years been achieved in making small-scale plants cost-effective, and breaking up large installations into a number of environmentally acceptable, smaller units.

Biomass burning continues to be in widespread use, in developing countries as well as in forest-rich residential areas of industrialised nations, such as Sweden, Canada and the North-eastern United States. This practice has not diminished despite the adverse effects of over-exploitation of forest resources in developing countries, and the severe environmental problems associated with small-scale, intermittent-cycle burning of woodfuel.

Biogas production continues in widely varying circumstances. In Asia, a large number

of primitive, labour-intensive biogas plants were installed during the period 1960-80. Not all of them are in working order; for those that are, their utility depends on the extent to which various population groups have access to them. Prototypes of high-technology, fully automated biogas plants have also been built in countries such as Denmark. The results have been a bit mixed, but the best installations has worked to full satisfaction and producing energy at near the current prices.

Production *of liquid biofuels* (alcohols) is popular in a few countries, notably Brazil. To serve as a gasoline replacement, the alternative fuel must be introduced into the distribution network – for example, in the form of gasoline-ethanol blends – and automobile engines have to be adapted for running on this new fuel. These factors have slowed plans for biofuel production in countries not forced to by trade imbalances – as has the uncertainty about recovering the investment necessary for such a large-scale project.

If renewable sources are to account for a considerable portion of total energy use – say, more than 25 per cent of electricity production in a given region – energy backup (such as by import-export arrangement) or storage must be part of the system. It already is for reservoir-based hydroelectric power, and a combination of these or pumped hydro reservoirs with intermittent renewable energy sources (such as windturbines and photovoltaic panels) will allow greater utilisation of the variable sources. Hydro reservoirs are available only in a few regions of the world, but liquid or gaseous biofuels or hydrogen could play a similar energy storage role, and would ideally allow a 100% utilisation of renewable energy technologies, even in the electricity sector. A number of other energy storage technologies are available for short-and longer-term storage, but usually at a high cost.

Barriers

There are barriers to renewable energy technology. One is financing, which is rarely available at prime rates – typically because the investments are not made by the same groups as those involved in conventional energy supply (power utilities, oil and gas distribution companies, etc.). This is particularly true for highly decentralised systems, such as rooftop solar cell panels. Investments may be needed in times when interest rates are high, and ordinary people cannot afford financing. Another barrier is that the running costs for a new technology are high if maintenance and service networks are limited. However, if demand is great enough, these obstacles can surely be overcome.

A more formidable barrier is the widespread distorted pricing of energy options. Full costing should include not only economic but also environmental and social factors: land use, noise and visual impact, pollution, climate impact, impact on health, work environment and institutions, risks and effects of large accidents, effects of system failures, degree of supply security and safety against terrorist actions, adaptability to planning uncertainties and uncertainties in evaluating any of the above – to mention only a few. If all these factors were universally reflected in costing, renewable energy would become an overnight winner.

12.1. Principles of energy conversion and thermodynamics

For a number of energy forms, Table 12.2 lists some examples of energy conversion processes or devices currently in use or contemplated, organised according to the energy form emerging after the conversion. In several cases more than one energy form will emerge as a result of the action of the device, e.g. heat in addition to one of the other energy forms listed. Many devices also perform a number of energy conversion steps, rather than the single ones given in the figure. A power plant may, for example, perform the following conversion process chain between energy forms: chemical → heat → mechanical → electrical. Diagonal transformations are also possible, such as conversion of mechanical energy into mechanical energy (potential energy of elevated fluid → kinetic energy of flowing fluid → rotational energy of turbine), or of heat into heat at a lower temperature (convection, conduction). A process in which the only change is that heat is transferred from a lower to a higher temperature, is forbidden by the second law of thermodynamics. Such transfer can be established if at the same time some high-quality energy is degraded, e.g. by a heat pump (which is listed as a converter of electrical into heat energy in Table 12.2).

Initial energy form	Converted energy form				
	Chemical	*Radiant*	*Electrical*	*Mechanical*	*Heat*
Nuclear					reactor
Chemical			fuel cell, battery discharge		burner, boiler
Radiant	photolysis		photovoltaic cell		absorber
Electrical	electrolysis, battery charging	lamp, laser		electric motor	resistance, heat pump
Mechanical			electric generator, MHD	turbines	friction, churning
Heat			thermionic & thermoelectric generators	thermodynamic engines	convector, radiator, heat pipe

Table 12.2. Examples of energy conversion processes, listed according to initial energy form and one particular, converted energy form (the one primarily wanted).

The efficiency with which a given conversion process can be carried out, i.e. the ratio between the output of the desired energy form and the energy input, depends on the physical and chemical laws governing the process. For the heat engines, which convert heat into work or vice versa, the description of thermodynamic theory may be used in order to avoid the complication of a microscopic description on the molecular level (which is of course possible, e.g. on the basis of statistical assumptions). According to thermodynamic theory (again the "second law"), no heat engine can have an efficiency higher than that of a reversible Carnot process, which is depicted in Fig. 12.2, in terms

of different sets of thermodynamic state variables, (P, V) = (pressure, volume), (T, S) = (absolute temperature, entropy), and (H, S) = (enthalpy, entropy).

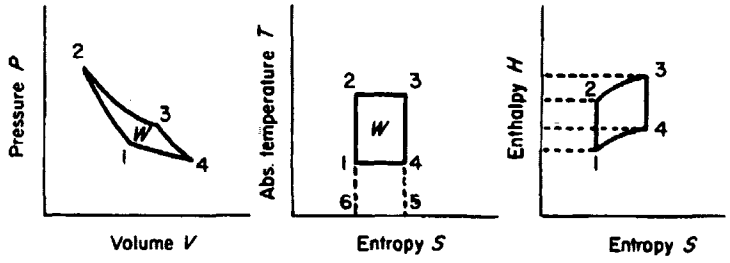


Figure 12.2. The cyclic Carnot process in different representations. Traversing the cycle in the direction 1→ 2→ 3→ 4 leads to the conversion of a certain amount of heat into work (see text for details).

The entropy S or rather its change is defined by the *second law of thermodynamics*, for a process that brings the system from state 1 to state 2 by a hypothetical succession of infinitesimal, reversible steps expressed by the integral

$$\Delta S = \int_{T_1}^{T_2} \mathrm{T}^{-1}\,\mathrm{d}Q$$

This definition is unique and does not depend on the actual path of the process, which may be irreversible or chaotic. The absolute value of S needs addition of an integration constant, which is fixed by the *third law of thermodynamics* (Nernst's law), which states that S can be taken as zero at zero absolute temperature ($T = 0$). The enthalpy H is defined by

$$H = U + PV,$$

in terms of P, V and the internal energy U of the system. According to energy conservation, called the *first law of thermodynamics*, U is a state variable given by

$$\Delta U = \int \mathrm{d}Q + \int \mathrm{d}W,$$

in terms of the amounts of heat and work added to the system (Q and W are not state variables, and the individual integrals depend on the paths of integration). The equation determines U up to an arbitrary constant, the zero point of the energy scale. Using the definition given above,

$$\mathrm{d}Q = \mathrm{T}\,\mathrm{dS} \text{ and } \mathrm{d}W = -P\,\mathrm{d}V,$$

both of which are valid only for reversible processes, the following relations are found among the differentials,

$$\mathrm{d}U = T\,\mathrm{d}S - P\,\mathrm{d}V,$$

$$\mathrm{d}H = T\,\mathrm{d}S + V\,\mathrm{d}P.$$

These relations are often assumed to have general validity, that is also for non-reversible processes.

If chemical reactions occur in the system, additional terms $\mu_i \, dn_i$ should be added on the right-hand side of both relations for dU and dH, in terms of the chemical potentials.

For a cyclic process such as the one shown in Fig. 12.2, $\int dU = 0$ upon returning to the initial locus in one of the diagrams, and thus according to the expression for dH one has $\int T\,dS = \int P\,dV$. This means that the area enclosed by the path of the cyclic process in either the (P, V) or the (T, S) diagram equals the work $-W$ performed by the system during one cycle (in the direction of increasing numbers on Fig. 12.2).

The amount of heat added to the system during the isothermal process 2-3 is $\Delta Q_{23} = T(S_3 - S_2)$, if the constant temperature is denoted T. The heat added in the other isothermal process, 4-1, at a temperature T_{ref}, is $\Delta Q_{41} = -T_{ref}(S_3 - S_2)$. It follows from the ($T$, S) diagram that $\Delta Q_{23} + \Delta Q_{41} = -W$. The efficiency by which the Carnot process converts heat available at temperature T into work, when a reference temperature of T_{ref} is available, is then

$$\eta = \frac{-W}{\Delta Q_{23}} = \frac{T - T_{ref}}{T}.$$

The Carnot cycle (Fig. 12.2) consists of four steps: 1-2, adiabatic compression (no heat exchange with the surroundings, i.e. $dQ = 0$ and $dS = 0$); 2-3, heat drawn reversibly from the surroundings at constant temperature (the amount of heat transfer ΔQ_{23} is given by the area enclosed by the path 2-3-5-6-2 in the (T, S)-diagram); 3-4, adiabatic expansion; and 4-1, heat given away to the surroundings by a reversible process at constant temperature ($|\Delta Q_{41}|$ equal to the area of the path 4-5-6-1-4 in the (T, S)-diagram).

The (H, S)-diagram is an example of a representation in which energy differences can be read directly on the ordinate, rather than being represented by an area.

It requires long periods of time to perform the steps involved in the Carnot cycle in a way that approaches reversibility. As time is important for man (the goal of the energy conversion process being power rather than just an amount of energy), irreversible processes are deliberately introduced into the thermodynamic cycles of actual conversion devices. The thermodynamics of irreversible processes is described below using a practical approximation, which has found some application for energy conversion devices. Readers without specific interest in the thermodynamic description may go lightly over the formulae.

Irreversible thermodynamics

The degree of irreversibility is measured in terms of the rate of energy dissipation,

$$D = T\,dS/dt,$$

where dS/dt is the entropy production of the system while held at the constant temperature T (i.e. T may be thought of as the temperature of a large heat reservoir, with which the system is in contact). In order to describe the nature of the dissipation process, the concept of "free energy" may be introduced.

The free energy, G, of a system is defined as the maximum work that can be drawn from the system under conditions where the exchange of work is the only interaction between the system and its surroundings. A system of this kind is said to be in thermodynamic equilibrium if its free energy is zero.

Consider now a system divided into two subsystems, a small one with extensive variables (i.e. variables proportional to the size of the system) U, S, V, etc., and a large one with intensive variables T_{ref}, P_{ref}, etc., which is initially in thermodynamic equilibrium. The terms "small system" and "large system" are meant to imply that the intensive variables of the large system (but not its extensive variables U_{ref}, S_{ref}, etc.) can be regarded as constant, regardless of the processes by which the entire system approaches equilibrium.

This implies that the intensive variables of the small system, which may not even be defined during the process, approach those of the large system, when the combined system approaches equilibrium. The free energy, or maximum work, is found by considering a reversible process between the initial state and the equilibrium. It equals the difference between the initial internal energy, $U_{init} = U + U_{ref}$, and the final internal energy, U_{eq}, or it may be written (all in terms of initial state variables) as

$$G = U - T_{ref} S + P_{ref} V,$$

plus terms of the form $\Sigma \mu_{i,ref} n_i$ if chemical reactions are involved, and similar generalisations in case of electromagnetic interactions, etc.

If the entire system is closed it develops spontaneously towards equilibrium through internal, irreversible processes, with a rate of free energy change

$$\frac{dG}{dt} = \frac{d}{dt}(U_{init} - U_{eq}(t)) = \left(\frac{\partial}{\partial S(t)} U_{eq}(t) \right) \frac{dS(t)}{dt},$$

assuming that the entropy is the only variable. $S(t)$ is the entropy at time t of the entire system, and $U_{eq}(t)$ is the internal energy that would be possessed by a hypothetical equilibrium state defined by the actual state variables at time t, i.e. $S(t)$ etc. For any of these equilibrium states, $\partial U_{eq}(t)/\partial S(t)$ equals T_{ref} according to the equation for dU, and by comparison with the definition of D above it is seen that the rate of dissipation can be identified with the loss of free energy, as well as with the increase in entropy,

$$D = -dG/dt = T_{ref}\, dS(t)/dt.$$

For systems met in practice, there will often be constraints preventing the system from reaching the absolute equilibrium state of zero free energy. For instance, the small system considered above may be separated from the large one by walls keeping the volume V constant. In such cases the available free energy (i.e. the maximum amount of useful work that can be extracted) becomes the absolute amount of free energy G, minus the free energy of the relative equilibrium which the combined system can be made to approach in the presence of the constraint. If the extensive variables in the constrained equilibrium state are denoted U^0, S^0, V^0, etc., then the available free energy becomes

$$\Delta G = (U - U^0) - T_{ref}(S - S^0) + P_{ref}(V - V^0),$$

eventually with the additions involving chemical potentials, etc. In this form, G is called the Gibbs potential. If the small system is constrained by walls, so that the volume cannot be changed, the free energy reduces to the Helmholtz potential $U - TS$, and if the small system is constrained so that it is incapable of exchanging heat, the free energy reduces to the enthalpy H. The corresponding form for ΔG give the maximum work that can be obtained from a thermodynamic system with the given constraints.

A description of the course of an actual process as a function of time requires knowledge of "equations of motion" for the extensive variables, i.e. equations that relate the currents such as

$$J_s = \mathrm{d}S/\mathrm{d}t \text{ (entropy flow rate) or } J_Q = \mathrm{d}Q/\mathrm{d}t \text{ (heat flow rate),}$$

(#) $$J_m = \mathrm{d}m/\mathrm{d}t \text{ (mass flow rate) or } J_\theta = \mathrm{d}\theta/\mathrm{d}t \text{ (angular velocity),}$$

$$J_q = \mathrm{d}q/\mathrm{d}t = I \text{ (charge fiow rate or electrical current), etc.}$$

to the (generalised) forces of the system. As a first approximation, the relation between the currents and the forces may be taken as linear (Onsager, 1931),

$$J_i = \sum_j L_{ij} F_j.$$

The direction of each flow component is J_i / J_i. The arbitrariness in choosing the generalised forces is reduced by requiring, as did Onsager, that the dissipation be given by

$$D = -\mathrm{d}G/\mathrm{d}t = \sum_i J_i \cdot F_i.$$

Examples of the linear relations for J_i are Ohm's law, stating that the electric current J_q is proportional to the gradient of the electric potential ($F_q \propto$ grad ϕ), and Fourier's law for heat conduction or diffusion, stating that the heat flow rate $E^{sens} = J_Q$ is proportional to the gradient of the temperature.

Considering the isothermal expansion process required in the Carnot cycle (Fig. 12.2), heat must be flowing to the system at a rate $J_Q = dQ/dt$, with $J_Q = LF_Q$ according to the expression above in its simplest form. Using the equation for D, the energy dissipation takes the form

$$D = T\,\mathrm{d}S/\mathrm{d}t = J_Q F_Q = L^{-1} J_Q^2.$$

For a finite time Δt, the entropy increase becomes

$$\Delta S = (\mathrm{d}S/\mathrm{d}t)\,\Delta t = (LT)^{-1} J_Q^2\,\Delta t = (LT\Delta t)^{-1} (\Delta Q)^2,$$

so that in order to transfer a finite amount of heat ΔQ, the product $\Delta S\,\Delta t$ must equal the quantity $(LT)^{-1}(\Delta Q)^2$. In order that the process approaches reversibility, as the ideal Carnot cycle should, ΔS must approach zero, which is seen to imply that Δt approaches infinity. This qualifies the statement made in the beginning of this subsection that, in order to go through a thermodynamic engine cycle in a finite time, one has to give up reversibility and accept a finite amount of energy dissipation and an efficiency which is smaller than the ideal Carnot efficiency.

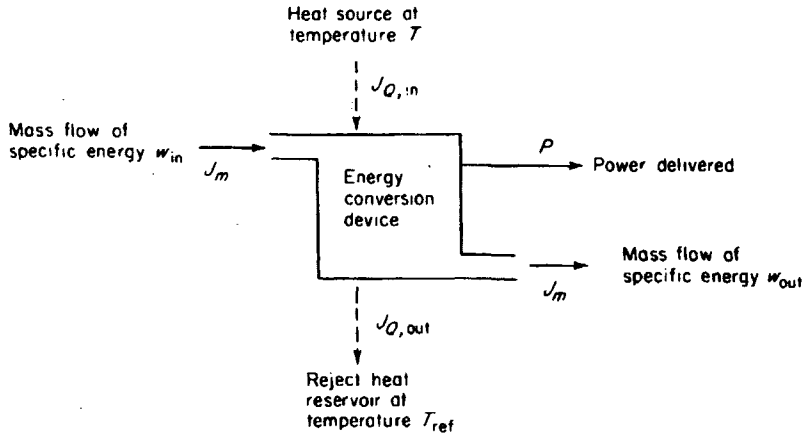


Figure 12.3. Schematic picture of an energy conversion device with a steady-state mass flow. The sign convention is different from the one used in (4.2), where all fluxes into the system were taken as positive.

Efficiency of an energy conversion device

A schematic picture of an energy conversion device is shown in Fig. 12.3, sufficiently general to cover most types of converters in practical use (Angrist, 1976). There is a mass flow into the device and another one out from it, as well as an incoming and outgoing heat flow. The work output may be in the form of electric or rotating shaft power.

It may be assumed that the converter is in a steady state, implying that the incoming and outgoing mass flows are identical, and that the entropy of the device itself is constant, i.e. that all entropy created is being carried away by the outgoing flows.

From the first law of thermodynamics, the power extracted, E, equals the net energy input,

$$E = J_{Q,in} - J_{Q,out} + J_m (w_{in} - w_{out})$$

The magnitude of the currents is given by (#), and their conventional signs may be inferred from Fig. 12.3. The specific energy content of the incoming mass flow, w_{in}, and of the outgoing mass flow, w_{out}, are the sums of potential energy, kinetic energy and enthalpy. The significance of the enthalpy to represent the thermodynamic energy of a stationary flow is established by Bernoulli's theorem. It states that for a stationary flow, if heat conduction can be neglected, the enthalpy is constant along a streamline. For the uniform mass flows assumed for the device in Fig. 12.3, the specific enthalpy, h, thus becomes a property of the flow, in analogy with the kinetic energy of motion and, for example, the geopotential energy,

$$w = w^{pot} + w^{kin} + h$$

The power output may be written

$$E = -\boldsymbol{J}_\theta \cdot \boldsymbol{F}_\theta - \boldsymbol{J}_q \cdot \boldsymbol{F}_q$$

with the magnitude of currents given by (#) and the generalised forces given by

$$F_\theta = \int \boldsymbol{r} \times \mathrm{d}\boldsymbol{F}_{mech}(\boldsymbol{r}) \qquad \text{(torque)},$$

(§) $$F_q = -\mathrm{grad}(\phi) \qquad \text{(electric field)}$$

corresponding to a mechanical torque and an electric potential gradient. The rate of entropy creation, i.e. the rate of entropy increase in the surroundings of the conversion device (as mentioned, the entropy inside the device is constant in the steady–state model), is

$$\mathrm{d}S/\mathrm{d}t = (T_{ref})^{-1} J_{Q,out} - T^{-1} J_{Q,in} + J_m (s_{m,out} - s_{m,in}),$$

where $s_{m,in}$ is the specific entropy of the mass (fluid, gas, etc.) flowing into the device, and $s_{m,out}$ the specific entropy of the outgoing mass flow. $J_{Q,out}$ may be eliminated by use of the equation for E, and the rate of dissipation obtained from the expression for D above,

$$D = T_{ref}\, \mathrm{d}S/\mathrm{d}t = J_{Q,in}(1 - T_{ref}/T) + J_m (w_{in} - w_{out} - T_{ref}(s_{m,in} - s_{m,out})) - E = \max(E) - E.$$

The maximum possible work (obtained for $\mathrm{d}S/\mathrm{d}t = 0$) is seen to consist of a Carnot term (closed cycle, i.e. no external flows) plus a term proportional to the mass flow. The dissipation found here may be brought to the Onsager form,

$$D = J_{Q,in} F_{Q,in} + J_m F_m + \boldsymbol{J}_\theta \cdot \boldsymbol{F}_\theta + J_q \cdot F_q$$

by defining generalised forces

$$F_{Q,in} = 1 - T_{ref}/T,$$

$$F_m = w_{in} - w_{out} - T_{ref}(s_{m,in} - s_{m,out})$$

in addition to those of (§).

The efficiency with which the heat- and mass-flow into the device is converted to power is, in analogy to the expression used previously,

$$\eta = \frac{E}{J_{Q,in} + J_m w_{in}},$$

where the expression for D may be used to eliminate E. This efficiency is sometimes referred to as the "first law efficiency", because it only deals with the amounts of energy input and output in the desired form, and not with the "quality" of the energy input related to that of the energy output.

In order to include reference to the energy quality, in the sense of the second law of thermodynamics, account must be taken of the changes in entropy taking place in connection with the heat and mass flows through the conversion device. This is accomplished by the "second law efficiency", which for power generating devices is defined by

$$\eta^{(2.law)} = \frac{E}{\max(E)} = -\frac{\mathrm{J}_\theta \cdot \mathrm{F}_\theta + \mathrm{J}_q \cdot \mathrm{F}_q}{J_{Q,in} F_{Q,in} + J_m F_m},$$

where the second expression is valid specifically for the device considered in Fig. 12.3, while the first expression is of general applicability, when max(E) is taken as the maximum rate of work extraction permitted by the second law of thermodynamics. It should be noted that max(E) depends not only on the system and the controlled energy inputs but also on the state of the surroundings.

Conversion devices for which the desired energy form is not work may be treated in a way analogous to the example in Fig. 12.3. In the final form of D, no distinction is made between input and output of the different energy forms. Taking, for example, electrical power as input (sign change), output may be obtained in the form of heat, or in the form of a mass stream. The efficiency expressions must be altered, placing the actual input terms in the denominator and the actual output terms in the numerator. If the desired output energy form is denoted W, the second law efficiency can be written in the general form

$$\eta^{(2.\,law)} = W/\max(W).$$

For conversion processes based on other principles than those considered in the thermodynamic description of phenomena, alternative efficiencies could be defined using this form, with max(W) calculated under consideration of the non-thermodynamic types of constraints. In such cases the name "second law efficiency" would have to be modified.

12.2 Conversion of wind flow

Conversion of wind energy into linear motion of a body has been utilised extensively, in particular for transportation across water surfaces. A large sail-ship of the type used in the nineteenth century may have converted wind energy at peak rates of a quarter of a megawatt or more.

The force on a sail or a wing (i.e. profiles of negligible or finite thickness) may be broken down into a component in the direction of the undisturbed wind (drag) and a component perpendicular to the undisturbed wind direction (lift). When referring to an undisturbed wind direction it is assumed that a uniform wind field is modified in a region around the sail or the wing, but that beyond a certain distance such modifications can be disregarded.

In order to determine the force components, Euler's equations (see Chapter 9) may be used. If viscous and external forces are neglected, and the flow is assumed to be irrotational (so that Bernoulli's equation is valid) and steady (so that the time-derivative of the velocity potential vanishes), then the force on a segment of the airfoil (sail or wing) may be written

$$\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}z} = \oint_C P\,\mathbf{n}\,\mathrm{d}s = -\tfrac{1}{2}\rho \oint_C (\mathbf{v}\cdot\mathbf{v})\,\mathbf{n}\,\mathrm{d}s.$$

Here $\mathrm{d}z$ is the segment length (cf. Fig. 12.4), C is a closed contour containing the airfoil profile, $\boldsymbol{n}$ is a unit vector normal to the contour (in the (x, y)-plane) and $\mathrm{d}s$ the pathlength

increment, directed along the tangent to the contour, still in the (x, y)-plane. Taking advantage of the fact that the wind velocity $\boldsymbol{v}$ approaches a homogeneous field $\boldsymbol{W}$ (assumed to be along the x-axis) far from the airfoil, the contour integral may be reduced and evaluated (e.g. along a circular path),

$$d\boldsymbol{F}/dz = \rho W \Gamma \boldsymbol{e}_y,$$

$$\Gamma = \oint_C \mathbf{v} \cdot d\mathbf{s} \approx \pi c W \sin \alpha$$

Here $\boldsymbol{e}_y$ is a unit vector along the y-axis, c the airfoil chord length and α the angle between the airfoil and $\boldsymbol{W}$. In the evaluation of the circulation Γ it has been assumed that the airfoil is thin and without curvature. In this case c and α are well-defined, but in general the circulation depends on the details of the profile, although an expression similar to the right-hand side of (4.114) is still valid as a first approximation, for some average chord length and angle of attack. This is known as the theorem of Kutta (1902) and Joukowski (1906).

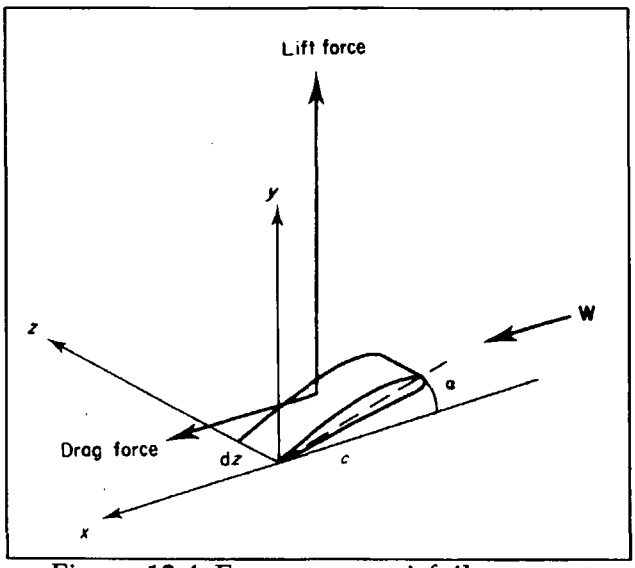


Figure 12.4. Forces on an airfoil segment.

The expressions above are valid in a co-ordinate system fixed relative to the airfoil (Fig. 12.4), and if the airfoil is moving with a velocity $\boldsymbol{U}$, the velocities $\boldsymbol{v}$ and $\boldsymbol{W}$ are to be interpreted as relative ones, so that

$$\boldsymbol{W} = \boldsymbol{u}_{in} - \boldsymbol{U},$$

if the undisturbed wind velocity is $\boldsymbol{u}_{in}$.

The assumption that viscous forces may be neglected is responsible for obtaining in the Kutta-Joukowski equation only a lift force, the drag force being zero. Primitive sailships, as well as primitive windmills, have been primarily aimed at utilising the drag force. It is possible, however, with suitably constructed airfoils, to make the lift force one or two orders of magnitude larger than the drag force, and thereby effectively approach the limit where the viscous forces and hence the drag can be neglected. This

usually requires careful "setting" of the airfoil, i.e. careful choice of the angle of attack, α, and in order to study operation at arbitrary conditions the drag component should be retained.

It is customary to describe the drag and lift forces on an airfoil of given shape, as a function of α, in terms of two dimensionless constants, $C_D(\alpha)$ and $C_L(\alpha)$, defined by

$$dF_x/dz = \tfrac{1}{2}\rho C_D W^2 c,$$

$$dF_y/dz = \tfrac{1}{2}\rho C_L W^2 c.$$

The constants C_D and C_L are not quite independent of the size of the system, which is not unexpected since the viscous forces (friction) in air contribute most to turbulent motion on smaller scales. Introducing the Reynolds number,

$$Re = Wc/\eta,$$

where η is the kinematic viscosity of air (a measure of the ratio between "inertial" and "viscous" forces acting between airfoil and air), the α-dependence of C_D and C_L for fixed Re, as well as the Re-dependence for the value of α which gives the highest lift-to-drag ratio, $L/D = C_L/C_D$, may look as shown in Figs. 12.5 and 12.6. The contours of these "high-lift" profiles are indicated in Fig. 12.5.

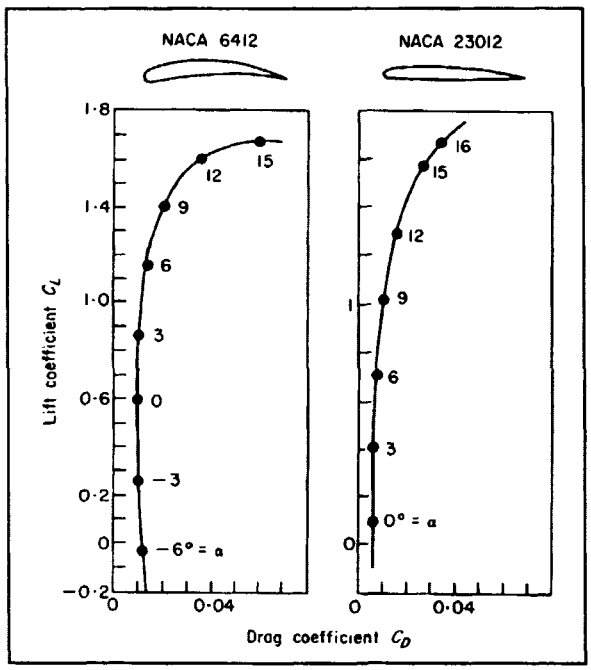


Figure 12.5. Lift and drag forces as function of the angle of attack, for two NACA airfoils (National Advisory Committee for Aeronautics; cf. e.g. Betz, 1959). The Reynolds number is Re = 8×10^6.

Assuming that C_D, C_L and W are constant over the area $A = \int c\,dz$ of the airfoil, the work done by a uniform (except in the vicinity of the airfoil) wind field $\mathbf{u}_{in}$ on a device (e.g. a ship), moving with a velocity $\mathbf{U}$, can be derived,

$$E = F \cdot U.$$

The angle β between u_{in} and U (see Fig. 12.7) may be maintained by a rudder. The power coefficient (4.50) becomes

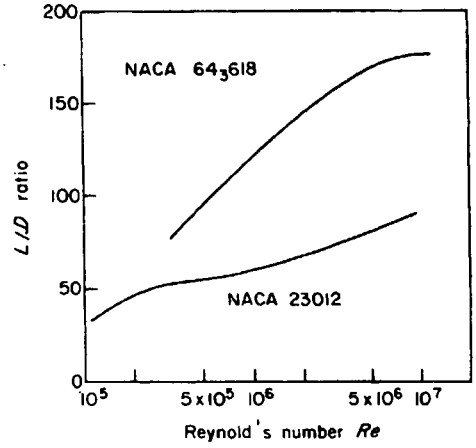


Figure 12.6. Reynolds number dependence of the lift-to-drag ratio, defined as the maximum value (as a function of the angle of attack) of the ratio between the lift and drag coefficients C_L and C_D (based on Hütter, 1977).

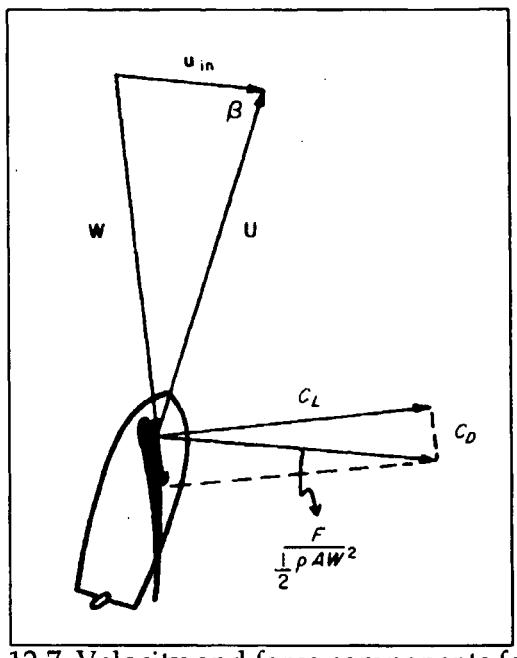


Figure 12.7. Velocity and force components for sail-ship.

$$C_p = f(C_L \sin\beta - C_D(1 - \sin^2\beta + f^2 - 2f\cos\beta)^{1/2})(1 + f^2 - 2f\cos\beta)^{1/2},$$

with $f = U/u_{in}$. For $C_L = 0$ the maximum C_p is $4C_D/27$, obtained for $f = 1/3$ and $\beta = 0$, whereas the maximum C_p for high lift-to-drag ratios L/D is obtained for β close to $\frac{1}{2}\pi$

and $f$around $2C_L/(3C_D)$. In this case, the maximum C_p may exceed C_L by one to two orders of magnitude (Wilson and Lissaman, 1974).

It is quite difficult to maintain the high speeds U required for optimum performance in a linear motion of the airfoil, and it is natural to focus the attention on rotating devices, in case the desired energy form is shaft or electric power and not propulsion. Wind-driven propulsion in the past (mostly at sea) has been restricted to U/u_{in}-values far below the optimum region for high L/Dairfoils (owing to friction against the water), and wind-driven propulsion on land or in the air has received little attention.

Propeller-type converters

Propellers have been extensively used in aircraft to propel the air in a direction parallel to that of the propeller axis, thereby providing the necessary lift force on the aeroplane wings. Propeller-type rotors are similarly used for windmills, but here the motion of the air (i.e. the wind) makes the propeller, which should be placed with its axis parallel to the wind direction, rotate, thus providing the possibility of power extraction. The propeller consists of a number of blades which are evenly distributed around the axis, each blade having a suitable aerodynamic profile usually designed to produce a high lift force, as discussed above. If there are two or more blades, the symmetrical mounting ensures a symmetrical mass distribution, but if only one blade is used it must be balanced by a counterweight.

A simple calculation allows the evaluation of the gross performance of such free stream flow turbines, i.e. turbines placed in open air as opposed to ventilation shafts etc. Consider a free stream flow passing horizontally through a converter. In this case the potential energy does not change and it is only necessary to consider kinetic energy. The pressure may vary near the converting device, but far behind and far ahead of the device the pressure is the same if the stream flow is free. Thus

$$w = w^{kin} = \tfrac{1}{2}\,(u_x^2 + u_y^2 + u_z^2) = \tfrac{1}{2}\,\mathbf{u}\cdot\mathbf{u},$$

and

$$w_{in} - w_{out} = \tfrac{1}{2}\,(\mathbf{u}_{in} - \mathbf{u}_{out})\cdot(\mathbf{u}_{in} + \mathbf{u}_{out}).$$

This expression and hence the efficiency would be maximum, if $\mathbf{u}_{out}$ could be made zero. However, the conservation of the mass flow J_m requires that u_{in} and u_{out} satisfy an additional relationship. For a pure, homogeneous streamline flow along the x-axis, the rate of mass flow is

$$J_m = \rho A_{in}\, u_{x,in} = \rho A_{out}\, u_{x,out},$$

in terms of areas A_{in} and A_{out} enclosing the same streamlines, before and after the passage through the conversion device. In a more general situation, assuming rotational symmetry around the x-axis, there may have been induced a radial as well as a circular flow component by the device. This situation is illustrated in Fig. 12.8. In the simple

case treated here, the radial and tangential components of the velocity field, u_r and u_t, which could be induced by the conversion device, are neglected.

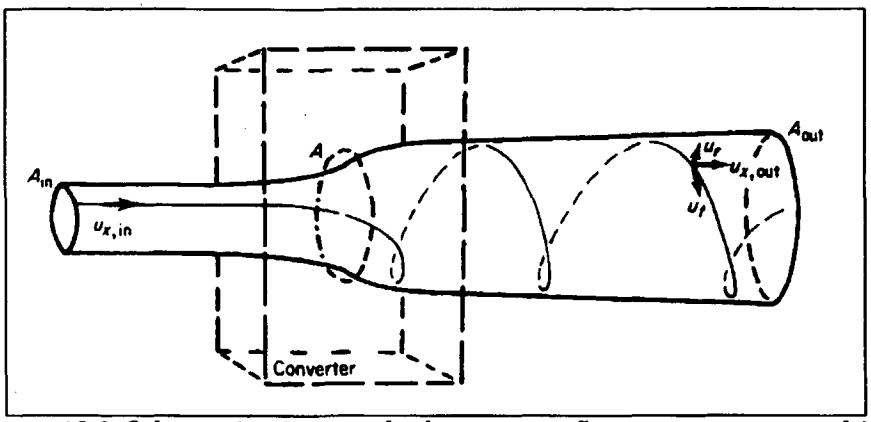


Figure 12.8. Schematic picture of a free stream flow converter or turbine. The incoming flow is a uniform streamline flow in the x-direction, while the outgoing flow is allowed to have a radial and a tangential component. The diagram indicates how a streamline may be transformed into an expanding helix by the device. The effective area of the converter, A, is defined in the text.

The axial force ("thrust") acting on the converter equals the momentum change,

$$F_x = J_m (u_{x,in} - u_{x,out}).$$

If the flow velocity in the converter is denoted u, an effective area of conversion, A, may be defined by

$$J_m = \rho A u_x.$$

according to the continuity equation for J_m. Dividing F_x by ρA, one obtains the specific energy transfer from the mass flow to the converter, within the conversion area A. This should equal the expression for the change in specific energy w_{in}, specialised to the case of homogeneous flows u_{in} and u_{out} along the x-axis,

$$u_x (u_{x,in} - u_{x,out}) = \tfrac{1}{2} (u_{x,in} + u_{x,out}) (u_{x,in} - u_{x,out})$$

or

$$u_x = \tfrac{1}{2} (u_{x,in} + u_{x,out}).$$

The physical principle behind this equality is simply energy conservation, and the assumptions so far have been the absence of heat exchange (so that the energy change becomes proportional to the kinetic energy difference) and the absence of induced rotation (so that only x-components of the velocity needs to be considered). On both sides of the converter, Bernoulli's equation is valid, stating that the specific energy is constant along a streamline. Far from the converter, the pressures are equal but the velocities are different, while the velocity just in front of or behind the converter may be taken as u_x, implying a pressure drop across the converter,

$$\Delta P = \tfrac{1}{2} \rho (u_{x,in} + u_{x,out}) (u_{x,in} - u_{x,out}).$$

The area enclosing a given streamline field increases in a continuous manner across the converter, at the same time as the fluid velocity continuously decreases. The pressure, on the other hand, rises above the ambient pressure in front of the converter, then discontinuously drops to a value below the ambient one, and finally increases towards the ambient pressure again, behind ("in the wake of") the converter.

It is customary (see e.g. Wilson and Lissaman, 1974) to define an "axial interference factor", a, by

$$u_x = u_{x,in}\,(1 - a),$$

in which case the above expression for u_x implies, that $u_{x,out} = u_{x,in}\,(1 - 2a)$. With this, the power output of the conversion device can be written

$$E = J_m\,(w_{in} - w_{out}) = \rho A\,(u_{x,in})^3\,2a\,(1 - a)^2,$$

and the efficiency

$$\eta = E\,/(J_m w_{in}) = 4a\,(1 - a).$$

It is seen that the maximum value of η is unity, obtained for $a = ½$, corresponding to $u_{x,out} = 0$. The continuity equation then implies an infinite area A_{out}, and it will clearly be difficult to defend the assumption of no induced radial motion.

In fact, for a free stream device of this type, the efficiency η is of little relevance since the input flux may not be independent of the details of the device. The input area A_{in}, from which streamlines would connect with a fixed converter area A, could conceivably be changed by altering the construction of the converter. It is therefore more appropriate to ask for the maximum power output for fixed A, as well as fixed input velocity $u_{x,in}$, this being equivalent to maximising the "power coefficient" defined by

$$C_p = E\,/(½\,\rho A\,(u_{x,in})^3) = 4a\,(1 - a)^2.$$

The maximum value is obtained for $a = 1/3$, yielding $C_p = 16/27$ and $u_{x,out} = u_{x,in}/3$. The areas are $A_{in} = (1 - a)A = 2/3\,A$ and $A_{out} = (1 - a)A/(1 - 2a) = 2A$, so in this case it is not unlikely that it may be a reasonable approximation to neglect the radial velocity component in the far wake.

The maximum found above for C_p is only a true upper limit with the assumptions made. By discarding the assumption of irrotational flow, it becomes possible for the converter to induce a velocity field, for which rot($\boldsymbol{u}$) is no longer zero. It has been shown that if the additional field is in the form of a vortex ring around the converter region, so that it does not contribute to the far wake, then it is possible to exceed the upper limit power coefficient 16/27 found above.

12.3 Photovoltaic conversion

Conversion of radiant energy (light quanta) into electrical energy can be achieved with the use of semiconductor materials, for which the electron excitation caused by impinging light quanta has a strongly enhancing effect on the conductivity.

It is not sufficient, however, that electrons are excited and are able to move more freely, if there is no force to make them move. Such a force would arise from the presence of a gradient of electrical potential, such as the one found in a *p-n* junction of doped semiconductor materials (a *p-n* junction is a junction of a *p*-type and an *n*-type semiconductor, as further described below). A *p-n* junction provides an electrical field which will cause the electrons excited by radiation (such as solar) to move in the direction from the *p*-type to the *n*-type material, and cause the vacancies (holes) left by the excited electrons to move in the opposite direction. If the electrons and holes reach the respective edges of the semiconductor material, the device is capable of delivering electrical power to an external circuit. The motion of electrons or holes receive competition from recombination processes (electrons being recaptured into vacancies), making such factors as overall dimensions and electron mobility in the material used of importance.

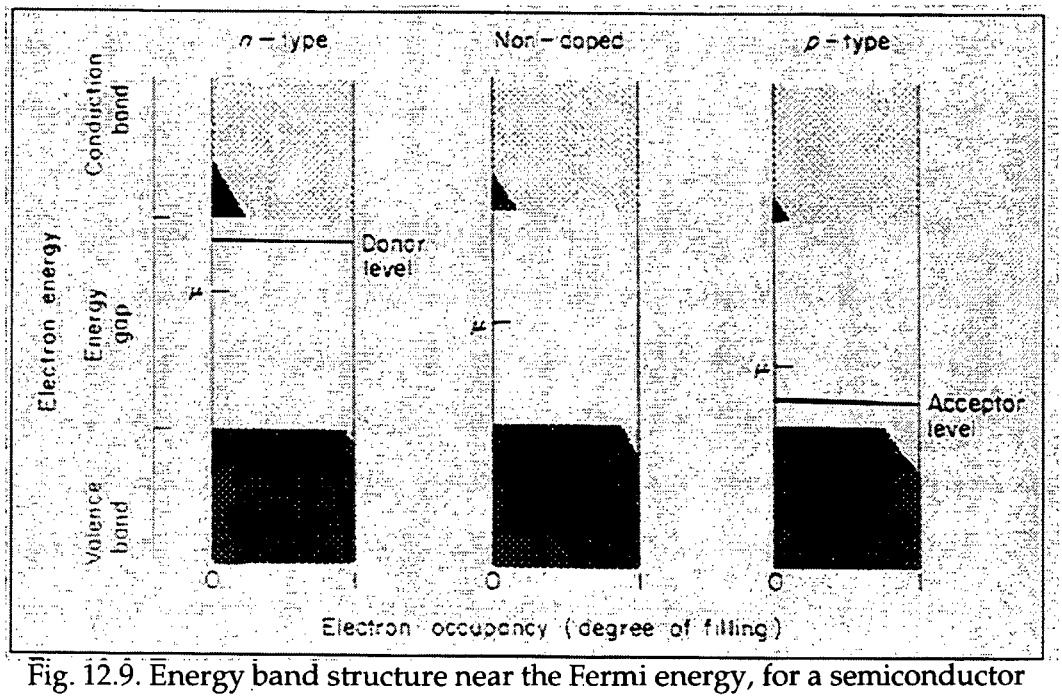


Fig. 12.9. Energy band structure near the Fermi energy, for a semiconductor material without impurities (middle), and with *n*- or *p*-type impurities. The dark shading indicates electron occupancy at a finite temperature.

The p-n junction

An essential constituent of photovoltaic cells is the *p-n* junction. *p*- and *n*-type doped materials are materials where minute amounts of a foreign atom is introduced into a solid state structure otherwise composed of just one type of atoms (a unique element with number Z in the periodic table). If the foreign atom has a higher charge (usually Z+1) then the doping is called *n*-type, and if it is lower, *p*-type. The effect of doping on the energy levels is illustrated in Fig. 12.9. It follows from the quantum models discussed in section 10.3.

When a *p*-type and an *n*-type semiconductor are joined to form a *p*-*n* junction, so that they acquire a common surface, then this will initially cause electrons to flow in the *n* to *p* direction because, as seen in Fig. 12.9, the electron density in the conduction band is higher in *n*-type than in *p*-type material, and because the hole density in the valence band is higher in the *p*-type than in the *n*-type material (the electron flow in the valence band can also be described as a flow of positive holes in the direction *p* to *n*).

This electron flow builds up a surplus of positive charge in the *n*-type material, and a surplus of negative charge in the *p*-type material, in the neighbourhood of the junction (mainly restricted to distances from the junction of the order of the mean travelling distance before recombination of an electron or a hole in the respective materials). These surplus charges form a dipole layer, associated with which is an electrostatic potential difference, which will tend to hinder any further unidirectional electron flow. Finally an equilibrium is reached in which the potential difference is such that no net transfer of electrons takes place.

Another way of stating the equilibrium condition is in terms of the Fermi energy (cf. Fig. 12.9). Originally the Fermi energies of the *p*- and *n*-type materials, μ_p and μ_n are different, but at equilibrium $\mu_p = \mu_n$. This is illustrated in Fig. 12.10, and it is seen that the change in the relative positions of the conduction (or valence) bands in the two types of material must equal the electron charge, $-e$, times the equilibrium electrostatic potential.

The number of electrons in the conduction band may be written

$$n_c = \int_{E_c}^{E_c'} n'(E) f(E)\, \mathrm{d}E,$$

where E_c and E_c' are the lower and upper energy limit of the conduction band, $n'(E)$ is the number of quantum states per unit energy interval (and, for example, per unit volume of material, if the electron number per unit volume is desired), and finally $f(E)$ is the Fermi–Dirac distribution $f(E) = (\exp(E-\mu_i)/kT)+1)^{-1}$. If the electrons in the conduction band are regarded as free, elementary quantum mechanics gives (see e.g. Shockley, 1950)

$$n'(E) = 4\pi h^{-3} (2m)^{3/2} E^{1/2},$$

where h is Planck's constant and m the electron mass. The corrections for electrons moving in condensed matter, rather than being free, may to a first approximation be included by replacing the electron mass by an "effective" value.

If the Fermi energy if not close to the conduction band,

$$E_c - \mu >> kT,$$

the Fermi–Dirac distribution may be replaced by the Boltzmann distribution,

$$f_B(E) = \exp(-(E - \mu)/kT).$$

Evaluating the integral then gives an expression of the form

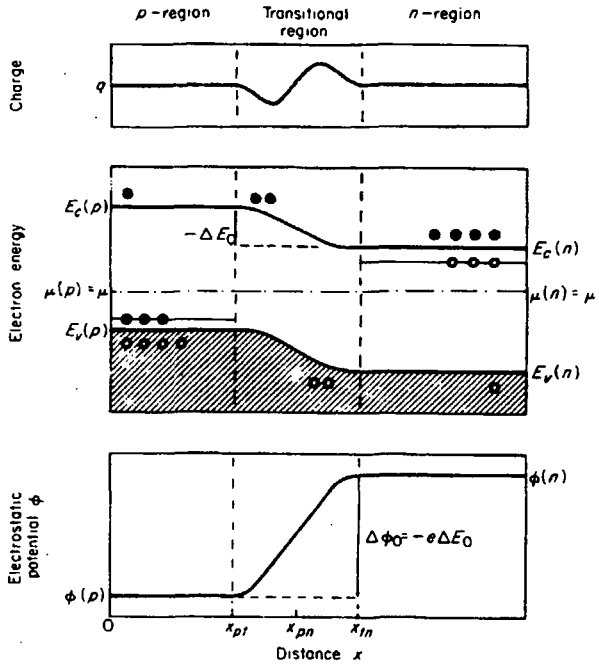


Figure 12.10. Schematic picture of the properties of a *p-n* junction in an equilibrium condition. The *x*-direction is perpendicular to the junction (all properties are assumed to be homogeneous in the *y*- and *z*-directions). The charge (top) is the sum of electron charges in the conduction band and hole (positive) charges in the valence band, and charge excess or defect associated with the acceptor and donor levels. In the electron energy diagram (middle), the abundance of minority charge carriers (closed circles for electrons, open circles for holes) is schematically illustrated. The properties are further discussed in the text.

$$n_c = N_c \exp(-(E_c - \mu)/kT).$$

The number of holes in the valence band is found in an analogous way,

$$n_v = N_v \exp(-(\mu - E_v)/kT).$$

where E_v is the upper limit energy of the valence band.

The equilibrium currents in a *p-n* junction such as the one illustrated in Fig. 12.10 can now be calculated. Considering first the electron currents in the conduction band, the electrons thermally excited into the conduction band in the *p*-region can freely flow into the *n*-type materials. The corresponding current, $I_0^-(p)$, may be considered proportional to the number of electrons in the conduction band in the *p*-region, $n_c(p)$, given above,

$$I_0^-(p) = \alpha N_c \exp(-(E_c(p) - \mu(p))/kT),$$

where the constant α depends on electron mobility in the material and on the electrostatic potential gradient, grad ϕ. The electrons excited into the conduction band in the *n*-type region will have to climb the potential barrier in order to move into the *p*-region. The fraction of electrons capable of doing this is given by a Boltzmann factor of the form

given above, but with the additional energy barrier $\Delta E_0 = -\Delta \phi_0 / e$ ($-e$ being the electron charge),

$$n_c(n) = N_c \exp(-(E_c(n) - \mu(n) - \Delta E_0)/kT).$$

Using $-\Delta E_0 = E_c(p) - E_c(n)$ (cf. Fig. 4.6) and considering the current $I_0^-(n)$ as being proportional to $n_c(n)$, the corresponding current may be written

$$I_0^-(n) = \alpha' N_c \exp(-(E_c(n) - \mu(n))/kT),$$

where α' depends on the diffusion parameter and on the relative change in electron density, $n_c^{-1}\text{grad}(n_c)$, considering the electron motion against the electrostatic potential as a diffusion process. The statistical mechanical condition for thermal equilibrium demands that $\alpha = -\alpha'$ (Einstein, 1905), so that the net electron current,

$$I_0^- = I_0^-(p) + I_0^-(n),$$

becomes zero precisely when

$$\mu(p) = \mu(n),$$

which is then the condition for thermal equilibrium. The same is true for the hole current,

$$I_0^+ = I_0^+(p) + I_0^+(n),$$

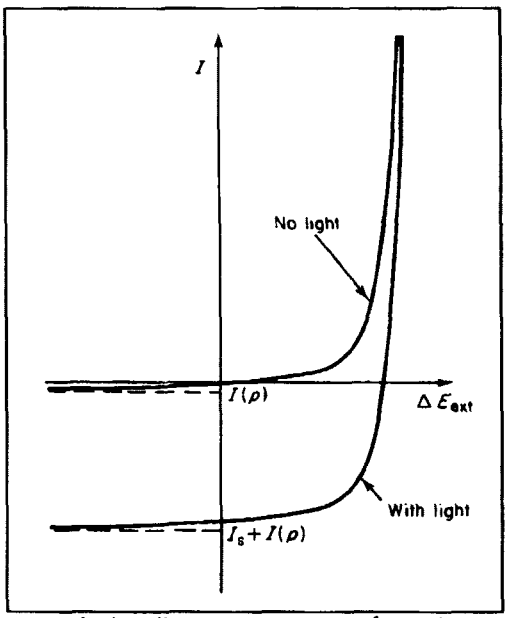


Figure 12.11. Characteristics (i.e. current as a function of external voltage) of a *p-n* junction, in the dark and with applied light. The magnitude of the short-circuit current, I_s, is a function of light intensity and spectral distribution.

If an external voltage source is applied to the *p-n* junction, in such a way that the *n*-type terminal receives an additional electrostatic potential $\Delta\phi_{ext}$ relative to the *p*-type termi-

nal, then the junction is no longer in thermal equilibrium, and the Fermi energy in the p-region is no longer equal to that of the n-region, but satisfies

$$\mu(p) - \mu(n) = e^{-1}\Delta\phi_{ext} = \Delta E_{ext}$$

if the Boltzmann distributions of electrons and of holes are to maintain their shapes in both p- and n-region. Similarly $E_c(p) - E_c(n) = -(\Delta E_0 + \Delta E_{ext})$, and assuming that the proportionality factors in the expressions for $I_0^+(p)$ *and* $I_0^+(n)$ still bear the relationship $\alpha = -\alpha'$ in the presence of the external potential, the currents are connected by the expression

$$I(n) = -I(p)\exp(\Delta E_{ext}/kT).$$

The net electron current in the conduction band then becomes

$$I = I(n) + I(p) = -I(p)(\exp(\Delta E_{ext}/kT) - 1).$$

For a positive $\Delta\phi_{ext}$, the potential barrier which electrons in the n-region conduction band (see Fig. 12.10) have to climb increases and the current $I(n)$ decreases exponentially (ΔE_{ext} negative, "reverse bias"). In this case, the net current I approaches a saturation value equal to $I(p)$.

For negative $\Delta\phi_{ext}$, (positive ΔE_{ext}, "forward bias"), the current $I(n)$ increases exponentially with the external potential. In both cases $I(p)$ is assumed to remain practically unchanged, when the external potential of one or the other sign is applied, considering that $I(p)$ is primarily limited by the number of electrons excited into the conduction band in the p-type material, a number which is assumed to be small in comparison with the conduction band electrons in the n-type material (cf. Figs. 12.9 and 12.10).

The contributions to the hole current, I^+, behave similarly to those of the electron current, and the total current I across a p-n junction with an external potential $\Delta\phi_{ext} = -e\Delta E_{ext}$, may be written

$$I = I^- + I^+ = -I(p)(\exp(\Delta E_{ext}/kT) - 1).$$

The relationship between current and potential is called the "characteristic" of the device, and the relation (4.62) for the p-n junction is illustrated in Fig. 12.11 by the curve labelled "no light". The constant saturation current $I(p)$ is sometimes referred to as the "dark current".

Solar cells

A p-n junction may be utilised to convert solar radiation energy into electric power. A solar cell is formed by shaping the junction in such a way that, for example, the p-type material can be reached by incident solar radiation, e.g. by placing a thin layer of p-type material on top of a piece of n-type semiconductor. In the dark and with no external voltage, the net current across the junction is zero, as was shown in the previous subsection, i.e. the intrinsic potential difference $\Delta\phi_0$ is unable to perform external work.

However, when irradiated with light quanta of an energy $E_{light} = h\nu = hc/\lambda$ (h is Planck's constant, c the velocity of light and ν and λ the frequency and wavelength of radiation),

which is larger than the energy difference between the conduction and valence band for the *p*-type material,

$$E_{light} \geq E_c(p) - E_v(p),$$

then electrons may be photo-excited from the valence band into the conduction band. The absorption of light quanta produces as many holes in the valence band of the *p*-type material as electrons in the conduction band. Since in the dark there are many fewer electrons in the *p*-type conduction band than holes in the valence band, a dramatic increase in the number of conduction band electrons can take place without significantly altering the number of holes in the valence band. If the excess electrons are sufficiently close to the junction to be able to reach it by diffusion before recombining with a hole, then the current in this direction exceeds $I_0(p)$ by an amount I_s, which is the net current through the junction in case of a short-circuited external connection from the *n*-type to the *p*-type material. The photo-induced current is not altered if there is a finite potential drop in the external circuit, since the relation between the current and the external potential drop $e\Delta E_{ext}$ was derived with reference only to the changes in the *n*-region.

An alternative *n*-*p* type of solar cell may consist of a thin *n*-type layer exposed to solar radiation, on top of a *p*-type base. In this case, the excess holes in the *n*-type valence band produce the photo-induced current I_s.

The total current in the case of light being absorbed in the *p*-type material and with an external potential drop is then

$$I = I_s - I(p)\,(\exp(-\Delta\,\phi_{ext}/kT) - 1).$$

The short-circuit current I_s depends on the amount of incident light with frequencies sufficient to excite electrons into the conduction band, on the fraction of this light actually being absorbed, and on the conditions for transporting the excess electrons created in the conduction band, in competition with electron–hole recombination processes. I_s may be written as the sum of a conduction and a diffusion type current, both related to the number of excess electrons in the conduction band, n_c^{ind}, induced by the absorption of light,

$$I_s = e(m_t E_e n_c^{ind} + k_t\, dn_c^{ind}/dx),$$

where e is the numerical value of the electron charge (1.6×10^{-19} C), m_t is the mobility of conduction band electrons (e.g. 0.12 $m^2 V^{-1} s^{-1}$ for silicon), E_e is the local electrical field, k_t the diffusion constant ($k_t = 10^{-3}$ $m^2 s^{-1}$ is a typical value estimated by Loferski, 1956), and x is the depth below the solar cell surface, assumed to be the only significant coordinate (as in Fig. 12.10).

The excess electron number density, $n_c^{ind}(x)$, at a depth x, changes when additional electrons are photo-excited, when electrons are carried away from x by the current I_s, and when electrons recombine with holes,

$$\frac{\partial\, n_c^{ind}(x)}{\partial\, t} = \int \sigma(\nu) n_{ph}(\nu) \exp(-\sigma(\nu)x)\, d\nu + \frac{I}{e}\frac{\partial I_s}{\partial x} - n_c^{ind}(x)\frac{1}{\tau_c}.$$

Here $\sigma(\nu)$ is the cross section for absorption of light quanta ("photons") in the *p*-type material, and $n_{ph}(\nu)$ is the number of photons at the cell surface ($x = 0$) per unit time and unit interval of frequency ν. The absorption cross section is zero for photon energies below the semiconductor energy gap, $h\nu < E_c(p) - E_v(p)$, i.e. the material is transparent to such light. The most energetic light quanta in visible light could theoretically excite more than one electron per photon (e.g. 2–3 in Si with an energy gap slightly over one electronvolt), but the probability for exciting just one electron to a higher energy is higher, and such a process is usually followed by a transfer of energy to other degrees of freedom (e.g. lattice vibrations and ultimately heat), as the excited electron approaches the lower part of the conduction band, or as the hole left by the electron de–excites from a deep level to the upper valence band. Thus in practice the quantum efficiency (number of electron–hole pairs per photon) hardly exceeds one.

The last parameter introduced in density equation, τ_c, is the average lifetime of an electron excited into the conduction band, before recombination (τ_c may lie in the interval 10^{-11} to 10^{-7}, 10^{-9} being a typical value (Wolf, 1963)). The lifetime τ_c is connected to the cross section for recombination, σ_c, and to the mean free path l_c of electrons in the conduction band by

$$l_c = \sigma_c^{-1} = v_c \tau_c N_a,$$

where v_c is the average thermal velocity of the electrons, $v_c = (2kT/m)^{1/2}$ (m being the electron mass, k Boltzmann's constant and T the absolute temperature) and N_a is the number of recombination centres ("acceptor impurities", cf. Fig. 12.9).

The boundary conditions for solving the density equation may be taken as the absence of excess minority carriers (electrons or holes) at the junction $x = x_{pn}$,

$$n_c^{ind}(x_{pn}) = 0,$$

and a prescribed (material dependent) excess electron gradient at the surface $x = 0$. This gradient, $(dn_c^{ind}/dx)|_{x=0}$, is often expressed in terms of a surface recombination velocity, s_c, through (4.64) by writing the left-hand side

$$I_s = s_c n_c^{ind}(0).$$

Typical values of s_c are of the order of 10^3 m s^{-1} (Wolf, 1963). For *n-p* type solar cells, expressions analogous to the above can be used. Once $n_c^{ind}(x)$ has been found, I_s can be calculated. The short-circuit current I_s increases linearly with intensity of light, if the spectral composition is kept constant, for the entire interval of intensities relevant for applications of solar cells at or near Earth.

For an open-ended solar cell (i.e. no external circuit), the difference in electrical potential between the terminals, $V_{oc} = \Delta\phi_{ext}(I=0)$, is obtained by putting I equal to zero in the expression above,

$$V_{oc} = kTe^{-1} (\log(I_s/I(p)) + 1).$$

The amount of electrical power, E, delivered by the irradiated cell to the external circuit is obtained by multiplication of I by the external voltage,

$$E = (\Delta\phi_{ext})I = \Delta\phi_{ext}\,(I_s - I(p)(\exp(-e\Delta\phi_{ext}/kT) - 1)).$$

From $\partial E/\partial(\Delta\phi_{ext}) = 0$ the external voltage V_{opt} may be found, which leads to the maximum value of power, E_{max}. In the situations of interest, V_{opt} is a slowly varying function of the amount of incident radiation. The corresponding current may be denoted I_{opt}.

The efficiency of solar cell radiant-to-electrical energy conversion is the ratio of the power E delivered and the incident energy, denoted E_+^{sw} (eventually for a tilted orientation of the solar cell), $\eta = E/E_+^{sw}$. In terms of the flux of photons of given frequency incident on the solar cell, the non-reflected energy flux at the surface may be written (a is the albedo of the cell surface)

$$E_+^{sw}\,(1-a) = \int_0^\infty h\nu n_{ph}(\nu)\,d\nu.$$

where h is Planck's constant. For a given semiconductor material, the maximum fraction of the energy, which can be absorbed, is

$$\int_{h\nu=E_c(p)-E_v(p)}^\infty h\nu n_{ph}(\nu)\,d\nu.$$

The part of the integral from zero up to the energy gap (i.e. the part above a certain wavelength of light) constitutes a fundamental loss. The same can be said of the energy of each light quantum in excess of the semiconductor energy gap $E_c(p) - E_v(p)$, assuming a quantum efficiency of at most one, i.e. that all such quanta are indeed absorbed (which may not be true if their energy is, say, between the upper limit of the conduction band and the lower limit of the following band), and that all excess energy is spent in exciting lattice degrees of freedom (vibrational phonons) that do not contribute to the photovoltaic process. In that case the energy flux available for photoconversion is only

$$(E_c(p) - E_v(p))\int_{h\nu=E_c(p)-E_v(p)}^\infty n_{ph}(\nu)\,d\nu = E^{avail}.$$

Further losses in addition to reflection and insufficient or excess photon energy may be associated with imperfections in the junction materials or in the current extraction system, causing heat formation or light re-emission rather than electrical power creation. Both heat creation (in the lattice) and re-radiation may take place in connection with the recombination of photo-excited electrons and holes. Since many of these processes are highly temperature dependent, the maximum efficiency that can be obtained in practice is also temperature dependent.

Rather than being p- and n-doped materials of the same elemental semiconductor, the solar cell junction may be based on different materials ("heterojunction") or on a metal and a semiconductor ("Schottky junction").

PROBLEMS AND DISCUSSION ISSUES

PROBLEM 12.1. Estimate the energy gain by a thermal solar collector, operating on a sunny day and using water of temperature 20°C or 60°C as its working fluid.

PROBLEM 12.2. Sketch the Carnot efficiency for a power plant converting heat to electricity, when the temperature of the heat energy varies from ambient to 6000°C.

PROBLEM 12.3. Discuss how you would steer a sail ship in a direction roughly against the wind (discuss angle of sail to wind, angle of rudder).

DISCUSSION ISSUE 12.4. Mention some of the "externalities" associated with different forms of energy and conversion technology, i.e. costs not paid directly by the consumers.

DISCUSSION ISSUE 12.5. Discuss the energy resources available in your region of the world. Do you estimate that they could cover all energy needs, or would import be required? Can the renewable energy sources available do the job? What are the problems that should be solved in order to accomplish an all-renewable energy supply?

Chapter 13

Communication, micro- and nanotechnology

An additional chapter is planned for future revisions, dealing with communication technology, microprocessors and nanotechnology. At present, these subjects are touched upon by showing and discussing the television documentary *Jo mindre des bedre* (Ernest Film for DTU, KU, OU, AU, DFH, RUC og KVL, 1999), med opgavesider og øvelser på http://www.sciencesite.dtu.dk

References

Almquist, E. (1974). Ambio, vol. 3, pp. 161-167

Angrist, S. (1976). "Direct Energy Conversion", 3rd edn. Allyn and Bacon, Boston.

Aspect, A. (1982). Physical Review Letters, vol. 49, p. 91 and 1804; see also review by F. Rohrlich (1983). Science, vol., 221, pp. 1251-1255

Bardeen, J., L. Cooper and J. Schrieffer (1957). Physical Review, vol. 106, p. 162; vol. 108, p. 1175

Berkner, L. and L. Marshall (1970). In "The encyclopedia of geochemistry and environmental sciences" (ed. R Fairbridge), vol. 4A, pp. 845-861. Van Nostrand, New York

Betz, A. (1959). "Stromunglehre", G. Brown Verlag, Karlsruhe, BRD.

Brink, D. and G. Satchler (1962). Angular Momentum. Oxford University Press, Oxford

Budyko, M. (1974). Climate and Life. Academic Press, New York

Bohr, A., B. Mottelson and D. Pines (1958). Physical Review, vol. 110, pp. 936-938

Bohr, N. and F. Kalckar (1937). Matematisk Fysiske Meddelelser, Danish Academy of Sciences, vol. 14, no. 10

Bohr, N. and J. Wheeler (1939). Physical Review, vol. 56, p. 426

Chapman, D, and H. Pollack (1975). Earth and Planetary Science Letters, vol. 28, pp. 23-32

Dirac, P. (1931). Proceedings of the Physical Society of London, vol. A126, p. 360

Dyson, F. (1979). Reviews of Modern Physics, vol. 51, pp. 447-460

Einstein, A. (1905). *Ann. der Physik* **17**, 549–560.

Einstein, A. (1913). Phys. Zeitschrift, vol., 14, p. 1249; C. Møller (1952): The theory of relativity. Clarendon Press, Oxford

Einstein, A. (1915). Berliner Berichte, p. 778, 799 and 844, Annalen der Physik (1916), vol. 49, p. 769

Einstein, A. (1921). On the special and the general theory of relativity. Fr. Vieweg & Sohn, Brauhschweig (in German)

Feld, B. and K. Tsipis (1979). Scientific American, November, pp. 44-55

Feynman, R. (1949). Physical Review, vol. 76, pp. 769-789

Feynman, R. (1965). The character of physical law. The MIT Press, Cambridge.

Feynman, R., R. Leighton and M. Sands (1964). The Feynman lectures on physics. Addison-Wesley Publ., Reading, MS.

Fowler, W. and F. Hoyle (1964). Astrophysical Journal, Supplement to vol. 9, # 91

Freedman, D. and P. van Nieuwenhuizen (1978). Scientific American, February, pp. 126-

Friedmann, A. (1922). Zeitschrift der Physik, vol. 10, p. 377; ibid. (1924), vol. 21, p. 326

Glansdorff, P. and I. Prigogine (1971). Thermodynamics of Structure, Stability and Fluctuations. Wiley-Interscience, New York; Prigogine, I. (1978). Science, vol 201, pp. 777-785

Glasstone, S. (ed.) (1962). The effects of nuclear weapons. United States Department of Defense/Atomic Energy Commission, Washington DC (other versions: 1957 and 1977)

Hawking, S. (1975). Communications in mathematical physics, vol. 43, p. 199

Hawking, S. (1977). Scientific American, February, pp. 34-40

Hawking, S. and G. Ellis (1973). The large-scale structure of space-time. Cambridge University Press, Cambridge; B. Felsager (1982), Gamma, No. 48, pp. 3-31 (in Danish)

Hoyle, F. (1957). The black cloud. Harper & Row, New York

Hubble, E. (1929). Proceedings of the National Academy of Sciences (USA), vol. 15, p. 168

Hussain, F. (1978). Nature, vol. 271, pp. 293-294

Iben, I. (1970). Scientific American, July, pp. 26-39

Jacob, M. and P. Landshoff (1980). Scientific American, vol 242, March, pp. 46-55

Joukowski, N. (1906). *Bull. de l'Inst. Aeronaut. Koutchino*, Fasc. I, St. Petersburg.

Kutta, W. (1902). "Auftriebkrafte in stromende Flüssigkeiten", *Ill. aeronaut. Mitteilungen*, July.

Lewis, J. (1974). Scientific American, March, pp. 51-65

Loferski, J. (1956). *J. Appl. Phys.* **27**, 777–784.

Löwdin, P. (1961). Advances in Quantum Chemistry, vol. 2, p. 213. Academic Press, New York

Meitner, L. and O. Frisch (1939). Nature, vol. 143, p. 239

Milankovich (1941). Royal Serbian Academy, Beograd. Special publication No. 132 (English translation 1969 by Israel Program for Scientific Translations, Jerusalem)

Morland, H. (1979). The Progressive. November issue

Morrison, P. and P. Walker (1978). Scientific American, October, pp. 36-49; P. Walker (1981). Scientific American. August, pp. 21-29

Møller, C. (1952). The theory of relativity. Clarendon Press, Oxford

Nielsen, H. (1978). Do we need fundamental laws of nature? Gamma no. 36, pp. 3-16, the Niels Bohr Institute, Copenhagen (in Danish).

Nielsen, H., J. Randrup and J. von Boehm (1979), Quantum chromodynamics, asymp-

totic freedom, and the bag model. Lecture motes 79/5, Nordita, Copenhagen

Onsager, L. (1931). *Phys. Rev.* **37**, 405–426.

Parmentola, J. and K. Tsipis (1979). Scientific American, April, pp. 38-49

Pauli, W. (1933). Handbuch der Physik, vol. 24. J. Springer, Berlin

Pauli, W. (1940). Physical Review, vol. 58, p. 716

Peebles, P. (1971). Physical cosmology. Princeton University Press, Princeton

Penrose, R. (1965). Physical Review Letters, vol. 14, pp. 57-59; Chapter 12 in General Relativity: An Einstein Centenary (S. Hawking and W. Israel, eds.). Cambridge University Press, Cambridge

Penzias, A. and K. Wilson (1965). Astrophysical Journal, vol. 142, p. 419

Pines. D. (1980). Science, vol. 207, pp. 597-606

Prigogine, I., G. Nicolis and A. Babloyantz (1972). Physics Today, November, pp. 23-28; December, pp. 38-44

Robertson, H. (1935). Applied Journal, vol. 82, p. 284; ibid. (1936), vol. 83, p. 187 and 257; A. Walker (1936). Proceedings of the London Mathematical Society, vol. 42(2)

Rotblat, J. (1981). Nuclear radiation in warfare. Stockholm International Peace Research Institute. Taylor & Francis, London

Sakurai, J. (1964). Invariance principles and elementary particles. Princeton University Press

Schwarzschild, K. (1916). Sitzungsberichte der Preussischer Akademie der Wissenschaften, p. 424

Schwarzschild, M. (1958) Structure and evolution of the stars. Princeton University Press, Princeton

Shockley, W. (1950). "Electrons and Holes in Semiconductors". Van Nostrand, New York.

Sutton, C. (1980). New Scientist, 11. September, pp. 786-789

Sørensen, B. (1979). Renewable Energy. Academic Press. London (2nd. ed., 2000)

Sørensen, B. (1979b). Nuclear power - the answer that became a question. Ambio, vol. 8, pp. 10-17

Sørensen, B. (1985). Securoty implications of alternative defence options for Western Europe. Journal of Peace Research, vol. 22, pp. 197-209

Sørensen, B. (1991). "Renewable energy - a technical overview", *Energy Policy* **19**, 386-391

Sørensen, B. (1992). "The future of renewable energy", *Ecodecision*, March, pp. 54-56

't Hooft, G. (1980). Scientific American, vol. 242, June, pp. 90-116

van den Bergh, S. (1981). Science, vol. 213, pp. 825-830

Vitale, B. (1982). END Papers One, pp. 54-67. Bertrand Russell Peace Foundation, Nottingham

Weinberg, S. (1972). Gravitation and cosmology: Principles and applications of the general theory of relativity. John Wiley & Sons, New York (p. 78)

Weinberg, S. (1977). The first three minutes. Basic Books, New York

Weyl, H. (1929). Zeitschrift der Physik, vol. 56, p. 330

Wilson, R. and Lissaman, P. (1974). "Applied Aerodynamics of Wind Power Machines". Oregon State University, Report No. NSF–RA–N–74–113.

Wit, J. (1981). Scientific American, February, pp. 27-37

Yang, J., S. Schramm, G. Steigman and R. Rood (1979). Astrophysical Journal, vol. 227, pp. 697-704

Yukawa, H. (1935). Proceedings of the Physical-Mathematical Society of Japan, vol.. 17, p. 48

Liste over tidligere udsendte tekster kan ses på IMFUFA's hjemmeside: http://mmf.ruc.dk eller rekvireres på sekretariatet, tlf. 46 74 22 63 eller e-mail: imfufa@ruc.dk.

332/97 ANOMAL SWELLING AF LIPIDE DOBBELTLAG
Specialerappoprt af: Stine Korremann
Vejleder: Dorthe Posselt

333/97 Biodiversity Matters
an extension of methods found in the literature on monetisation of biodiversity
by: Bernd Kuemmel

334/97 LIFE-CYCLE ANALYSIS OF THE TOTAL DANISH ENERGY SYSTEM
by: Bernd Kuemmel and Bent Sørensen

335/97 Dynamics of Amorphous Solids and Viscous Liquids
by: Jeppe C. Dyre

336/97 Problem-orientated Group Project Work at Roskilde University
by: Kathrine Legge

337/97 Verdensbankens globale befolkningsprognose
- et projekt om matematisk modellering
af: Jørn Chr. Bendtsen, Kurt Jensen, Per Pauli Petersen

338/97 Kvantisering af nanolederes elektriske ledningsevne
Første modul fysikprojekt
af: Søren Dam, Esben Danielsen, Martin Niss,
Esben Friis Pedersen, Frederik Resen Steenstrup
Vejleder: Tage Christensen

339/97 Defining Discipline
by: Wolfgang Coy

340/97 Prime ends revisited - a geometric point of view -
by: Carsten Lunde Petersen

341/97 Two chapters on the teaching, learning and assessment of geometry
by: Mogens Niss

342/97 A global clean fossil scenario DISCUSSION PAPER prepared by Bernd Kuemmel for the project LONG-TERM SCENARIOS FOR GLOBAL ENERGY DEMAND AND SUPPLY

343/97 IMPORT/EKSPORT-POLITIK SOM REDSKAB TIL OPTIMERET UDNYTTELSE AF EL PRODUCERET PÅ VE-ANLÆG
af: Peter Meibom, Torben Svendsen, Bent Sørensen

344/97 Puzzles and Siegel disks
by: Carsten Lunde-Petersen

345/98 Modeling the Arterial System with Reference to an Anesthesia Simulator
Ph.D. Thesis
by: Mette Sofie Olufsen

346/98 Klyngedannelse i en hulkatode-forstøvningsproces
af: Sebastian Horst
Vejledere: Jørn Borggren, NBI, Niels Boye Olsen

347/98 Verificering af Matematiske Modeller
- en analyse af Den Danske Eulerske Model
af: Jonas Blomqvist, Tom Pedersen, Karen Timmermann, Lisbet Øhlenschlæger
Vejleder: Bernhelm Booss-Bavnbek

348/98 Case study of the environmental permission procedure and the environmental impact assessment for power plants in Denmark
by: Stefan Krüger Nielsen
project leader: Bent Sørensen

349/98 Tre rapporter fra FAGMAT - et projekt om tal og faglig matematik i arbejdsmarkedsuddannelserne
af: Lena Lindenskov og Tine Wedege

350/98 OPGAVESAMLING - Bredde-Kursus i Fysik 1976 - 1998
Erstatter teksterne 3/78, 261/93 og 322/96

351/98 Aspects of the Nature and State of Research in Mathematics Education
by: Mogens Niss

352/98 The Herman-Swiatec Theorem with applications
by: Carsten Lunde Petersen

353/98 Problemløsning og modellering i en almendannende matematikundervisning
Specialerapport af: Per Gregersen og Tomas Højgaard Jensen

354/98 A Global Renewable Energy Scenario
by: Bent Sørensen and Peter Meibom

355/98 Convergence of rational rays in parameter spaces
by: Carsten Lunde Petersen and Gustav Ryd

356/98 Terrænmodellering
Analyse af en matematisk model til konstruktion af digitale terrænmodeller
Modelprojekt af: Thomas Frommelt, Hans Ravnkjær Larsen og Arnold Skimminge
Vejleder: Johnny Ottesen

357/98 Cayleys Problem
En historisk analyse af arbejdet med Cayleys problem fra 1870 til 1918
Et matematisk videnskabsfagsprojekt af: Rikke Degn, Bo Jakobsen, Bjarke K.W. Hansen, Jesper S. Hansen, Jesper Udesen, Peter C. Wulff
Vejleder: Jesper Larsen

358/98 Modeling of Feedback Mechanisms which Control the Heart Function in a View to an Implementation in Cardiovascular Models
Ph.D. Thesis by: Michael Danielsen

359/99 Long-Term Scenarios for Global Energy Demand and Supply
Four Global Greenhouse Mitigation Scenarios
by: Bent Sørensen (with contribution from Bernd Kuemmel and Peter Meibom)

360/99 SYMMETRI I FYSIK
En Meta-projektrapport af: Martin Niss, Bo Jakobsen & Tune Bjarke Bonné
Vejleder: Peder Voetmann Christiansen

361/99 Symplectic Functional Analysis and Spectral Invariants
by: Bernhelm Booß-Bavnbek, Kenro Furutani

362/99 Er matematik en naturvidenskab? - en udspænding af diskussionen
En videnskabsfagsprojekt-rapport af: Martin Niss
Vejleder: Mogens Nørgaard Olesen

363/99 EMERGENCE AND DOWNWARD CAUSATION
by: Donald T. Campell, Mark H. Bickhard, and Peder V. Christiansen

364/99 Illustrationens kraft - Visuel formidling af fysik
Integreret speciale i fysik og kommunikation
af Sebastian Horst
Vejledere: Karin Beyer, Søren Kjørup

365/99 To know - or not to know - mathematics, that is a question of context
by: Tine Wedege

366/99 LATEX FOR FORFATTERE - En introduktion til LATEX
og IMFUFA-LATEX
af: Jørgen Larsen

367/99 Boundary Reduction of Spectral Invariants and Unique Continuation Property
by: Bernhelm Booss-Bavnbek

368/99 Kvartvejsraport for projektet SCENARIER FOR SAMLET UDNYTTELSE AF BRINT SOM ENERGIBÆRER I DANMARKS FREMTIDIGE ENERGISYSTEM
Projektleder: Bent Sørensen

369/99 Dynamics of Complex Quadratic Correspondences
by: Jacob S. Jalving
Supervisor: Carsten Lunde Petersen

370/99 OPGAVESAMLING - Bredde-Kursus i Fysik 1976 - 1999
Eksamensopgaver fra perioden 1976 - 1999. Denne tekst erstatter tekst nr. 350/98

371/99 Bevisets stilling - beviser og bevisførelse i en gymnasial matematik undervisning
Et matematikspeciale af: Maria Hermannsson
Vejleder: Mogens Niss

372/99 En kontekstualiseret matematikhistorisk analyse af ikke-lineær programmering: Udviklingshistorie og multipel opdagelse
Ph.d.-afhandling af Tinne Hoff Kjeldsen

373/99 Criss-Cross Reduction of the Maslov Index and a Proof of the Yoshida-Nicolaescu Theorem
by: Bernhelm Booss-Bavnbek, Kenro Furutani and Nobukazu Otsuki

374/99 Det hydrauliske spring - Et eksperimentelt studie af polygoner og hastighedsprofiler
Specialeafhandling af: Anders Marcussen
Vejledere: Tomas Bohr, Clive Ellegaard, Bent C. Jørgensen

375/99 Begrundelser for Matematikundervisningen i den lærde skole hhv. gymnasiet 1884-1914
Historiespeciale af Henrik Andreassen, cand.mag. i Historie og Matematik

376/99 Universality of AC conduction in disordered solids
by: Jeppe C. Dyre, Thomas B. Schrøder

377/99 The Kuhn-Tucker Theorem in Nonlinear Programming: A Multiple Discovery?
by: Tinne Hoff Kjeldsen

378/00 Solar energy preprints:
1. Renewable energy sources and termal energy storage
2. Integration of photovoltaic cells into the global energy system
by: Bent Sørensen

379/00 EULERS DIFFERENTIALREGNING
Eulers indførelse af differentialregningen stillet over for den moderne
En tredjesemesters projektrapport på den naturvidenskabelige basisuddannelse
af: Uffe Thomas Volmer Jankvist, Rie Rose Møller Pedersen, Maja Bagge Pedersen
Vejleder: Jørgen Larsen

380/00 MATEMATISK MODELLERING AF HJERTEFUNKTIONEN
Isovolumetrisk ventrikulær kontraktion og udpumpning til det cardiovaskulære system
af: Gitte Andersen (3.moduls-rapport), Jakob Hilmer og Stine Weisbjerg (speciale)
Vejleder: Johnny Ottesen

381/00 Matematikviden og teknologiske kompetencer hos kortuddannede voksne
- Rekognosceringer og konstruktioner i grænselandet mellem matematikkens didaktik og forskning i voksenuddannelse
Ph. d.-afhandling af Tine Wedege

382/00 Den selvundvigende vandring
Et matematisk professionsprojekt
af: Martin Niss, Arnold Skimminge
Vejledere: Viggo Andreasen, John Villumsen

383/00 Beviser i matematik
af: Anne K.S.Jensen, Gitte M. Jensen, Jesper Thrane, Karen L.A.W. Wille, Peter Wulff
Vejleder: Mogens Niss

384/00 Hopping in Disordered Media: A Model Glass Former and A Hopping Model
Ph.D. thesis by: Thomas B. Schrøder
Supervisor: Jeppe C. Dyre

385/00 The Geometry of Cauchy Data Spaces
This report is dedicated to the memory of Jean Leray (1906-1998)
by: B. Booss-Bavnbek, K. Furutani, K. P. Wojciechowski

386/00 Neutrale mandatfordelingsmetoder – en illusion?
af: Hans Henrik Brok-Kristensen, Knud Dyrberg, Tove Oxager, Jens Sveistrup
Vejleder: Bernhelm Booss-Bavnbek

387/00 A History of the Minimax Theorem: von Neumann´s Conception of the Minimax Theorem - - a Journey Through Different Mathematical Contexts
by: Tinne Hoff Kjeldsen

388/00 Behandling af impuls ved kilder og dræn i C. S. Peskins 2D-hjertemodel
et 2. moduls matematik modelprojekt
af: Bo Jakobsen, Kristine Niss
Vejleder: Jesper Larsen

389/00 University mathematics based on problemoriented student projects: 25 years of experience with the Roskilde model
By: Mogens Niss
Do not ask what mathematics can do for modelling. Ask what modelling can do for mathematics!
by: Johnny Ottesen

390/01 Endnu ikke udkommet

391/01 Matematisk modelleringskompetence – et undervisningsforløb i gymnasiet
3. semesters Nat.Bas. projekt af: Jess Tolstrup Boye, Morten Bjørn-Mortensen, Sofie Inari Castella, Jan Lauridsen, Maria Gøtzsche, Ditte Mandøe Andreasen
Vejleder: Johnny Ottesen

392/01 "PHYSICS REVEALED" THE METHODS AND SUBJECT MATTER OF PHYSICS
an introduction to pedestrians (but not excluding cyclists)
PART III: PHYSICS IN PHILOSOPHICAL CONTEXT
by: Bent Sørensen.

393/01 Endnu ikke udkommet

394/01 "PHYSICS REVEALED" THE METHODS AND SUBJECT MATTER OF PHYSICS
an introduction to pedestrians (but not excluding cyclists)
PART II: PHYSICS PROPER
by: Bent Sørensen.

395/01 Menneskers forhold til matematik. Det har sine årsager!
Specialeafhandling af: Anita Stark, Agnete K. Ravnborg
Vejleder: Tine Wedege

396/01 2 bilag til tekst nr. 395: Menneskers forhold til matematik. Det har sine årsager!
Specialeafhandling af: Anita Stark, Agnete K. Ravnborg
Vejleder: Tine Wedege