Energy for the 21st Century

Professor George R. Tynan

Department of Mechanical and Aerospace Engineering
Jacobs School of Engineering
and
The Center for Energy Research

University of California San Diego

Table of Contents

Preface

These notes are provided to students of the MAE 118/119/120 sequence, and are not intended at this point in time for broader distribution, and represent an unpublished work by the author. As such, they are protected by the policies of the University of California as well as by the relevant portions of the U.S. Copyright protection.

Chapter 1: Fundamentals

Energy is often said to be the lifeblood of today's global human civilization. As we will see in our exploration of this subject, this aphorism is indeed true in so far that ready access to large quantities of energy allow human beings to grow enough food to feed nearly 7-8 billion people, provide clean drinking water, enable our health, education and other social services, and operate the balance of the world's economy. As we will also see, it is an uncomfortable fact that today the vast majority of the world's primary energy is obtained from inherently transient energy sources – namely fossil fuels – which are unsustainable in the long term and which are now leading to very serious global environmental degradation, including global climate change and local and regional air and water pollution. These two distinct issues motivate the development and deployment of energy sources and conversion technologies in the coming decades which can provide energy in relevant forms and at sufficient scale to power the lives of the 9-11 billion people that are expected to be living by the second half of the 21st century. We will find that the list of possible "primary" energy sources and technologies to convert these resources to useful form and that can scale to eventually replace fossil-fuels is disquietingly short. The goal of this book is to introduce students to the essential elements of these sources and conversion technologies in such a way as to highlight the potential contributions of the sources and technologies, and identify the fundamental scientific and/or technical issues that must be resolved before they can be deployed at scale around the globe.

The Basic Physics of Energy Systems

Before we can consider the issue in detail it is essential to first ensure that the fundamental definitions and physical constraints imposed upon the energy issues of interest here are first clearly defined and understood. Thus, let us first clearly define what we mean by "energy" and remind ourselves of the constraints that must always be obeyed by this quantity due to the laws of physics.

Definition of Energy

In physics, the term "energy" is used to denote a property which, when expended, provides the capacity to do work. This somewhat nebulous definition simply refers to another quanity, work, which we must define if we are to then understand what the term "energy" denotes. Work, in turn, is defined more clearly as a force imposed on an object which then moves over a distance. Mathematically the work, W, is given as the path integral of the force over the trajectory taken by the system as it moves from the starting position $\mathbf{x}_1$ to the end position $\mathbf{x}_2$

$$W = \int_{\mathbf{x}_1}^{\mathbf{x}_2} \mathbf{F}(\mathbf{x}) \cdot d\mathbf{x}. W = \int_{\mathbf{x}_1}^{\mathbf{x}_2} \mathbf{F}(\mathbf{x}) \cdot d\mathbf{x}.$$

This work could be associated e.g. with the expansion of a gas as it pushes and moves a piston against a resisting force, with the movement of a current-carrying wire across and magnetic field, or with the motion of an electrical charge against dissipating collisions within s conductor via the action of an applied electric field. Thus energy is a quantity that, when

expended, allows us to perform work. As we will see in a moment, energy is also a conserved quantity. Thus more precisely, work then performed as we transform energy from one form to another. In doing so, some of the energy is used to do work, while the balance of the energy is dissipated into a form that it can no longer readily be used to perform work. From this simple discussion we can see the essential elements emerge: we are interested in taking energy out of storage, converting it into a useable form, using that energy for the performance of work, and then having the energy dissipated into a form where it can no longer be used.

There are many different forms in which energy can be stored. For example, it can be stored in the random thermal motion of atoms in a system. If this collection of atoms is in thermal equilibrium then this motion can be described by a temperature, T, which is related to the average energy of the randomly moving particles. The energy can also be stored as internal energy, U, e.g. in the form of chemical or nuclear potential energy, or in the form of a potential energy due e.g. to a gravitational or other type of potential. This potential energy can then be converted into thermal energy in the form of hot matter, and this hot matter can then be used to generate useful work via some type of engineered system. Energy stored in one of these forms is then usually converted to some other useful form which is then used to perform a desired work task, thereby expending the energy. We can quantify these ideas with a simple example.

In a conservative potential field, U, the force acting on a body at a point can be expressed as $\mathbf{F} = -\nabla U$. Since the field is conservative, moving the body against this force through a displacement then results in work being done on the body and the subsequent accumulation of so-called potential energy by the body, and by definition there is no energy dissipated into heat. The amount of this energy, U_p, is given by

$$\Delta U_p = -\int_{\mathbf{r}_{ref}}^{\mathbf{r}} \mathbf{F}\cdot d\mathbf{r} = \int_{\mathbf{r}_{ref}}^{\mathbf{r}} \nabla U \cdot d\mathbf{r} = U(r) - U(r_{ref}).$$

Suppose then that a body has a potential energy U_p, and kinetic energy E_k. We can then think of the total energy of the body as being given by the sum of these two quantities, i.e. $E_{tot} = E_k + U_p$. In the absence of energy dissipation, then this total energy would be constant. Thus, potential energy and kinetic energy could be converted from one form into the other, but the sum would remain constant.

These simple considerations can be expanded to include more complex systems and conditions. For example, consider a material composed of many atoms or molecules. These particles have a variety of velocities and thus each has its own kinetic energy. When these particles are close to thermal equilibrium they can be described as having a temperature, T, which in turn describes how fast or slow the particles are moving. The average kinetic energy, E_{th}, of a particle in such a collection is then given by

$$E_{th} = \frac{3}{2} k_B T$$

where the constant $k_B = 1.38x10^{-23} J/degK$ is Boltzmann's constant and relates the microscopic kinetic energy of an atom or molecule to the temperature of a collection of particles. Thus, we can see that a collection of particles with a finite temperature T will have a thermal energy density U given by $U = nE_{th} = \frac{3}{2} nk_B T$. If the particles form an ideal gas such that $p = nk_B T$, then this energy density is related to the gas pressure p via the relation $U = \frac{3}{2} p$.

Energy can also be stored within the chemical or nuclear bonds of molecules or atomic nuclei. We can denote these types of stored energies by the symbols E_{chem} and E_{nuc}. When a

chemical reaction then occurs, energy that was stored in the chemical potential is then converted into another type of energy – usually into kinetic energy of the reactant particles. This energy can then be converted into thermal energy by the action of multiple collisions between the particles. In all cases, the total energy is conserved, and energy is simply reapportioned within the system.

Energy can also be stored within electric and magnetic fields E and B located in some volume V. The energy stored in this manner can be written as

$$E_{elec} = \int_V \frac{1}{2}\varepsilon(x)\left|E(x)\right|^2 d^3x$$

and

$$E_{mag} = \int_V \frac{1}{2}\mu(x)\left|B(x)\right|^2 d^3x$$

where denotes ε the dielectric permittivity and μ the magnetic permeability of the material filling the volume V and the integral is taken over the volume V. The total energy of a system that contains all of these components or types of energy then becomes the sum of all of these types of energy, i.e. we have

$$E_{tot} = E_k + U_p + U + E_{chem} + E_{nuc} + E_{elec} + E_{mag}\,.$$

The technologies that we have developed to perform useful work then convert from one or more of these forms, and transform some of it into mechanical work, and the balance is then rejected to the environment, usually in the form of heat at such a low temperature as to make further conversion into work impractical.

First & Second Laws of Thermodynamics

This property that we call energy obeys a conservation law (commonly known as the first law of Thermodynamics) which states that:

1. Energy is neither created or destroyed but, subject to the constrains imposed by the 2nd law of thermodynamics, can be converted from one form to another.

The implications of this physical law pervade the entire discussion of human energy demands. In particular, it implies that we must either seek out stored forms of energy which exist on the Earth and then convert them into useable form; once these resources have then been used, their energy content is then no longer available to be used again. Alternatively, we can devise mechanisms to capture energy that flows into the Earth system from external sources – e.g. from the sun – and then convert that energy into some useful form (e.g. into electricity). That energy is then used to perform some useful function and in the process is then dissipated into heat.

As alluded to in the above discussion, there is a second fundamental physical law that energy transformations must also follow. There are many different expressions of the second law of thermodynamics. For our purposes, we might consider the version put forward by Lord Kelvin:

2. It is impossible to convert heat completely into work in a cyclic process. In other words, the conversion of heat into work is never perfectly efficient; there is always some loss involved in the conversion process. The heat lost during the conversion process then generates entropy, which is a measure of the disorder of the system. These two fundamental physical laws provide strong constraints on how one can design and employ systems that produce useful work from energy inputs.

We will need more quantitative expressions of these principles and shall review their essential elements in a bit more detail later. However, for the present time these definitions suffice and allow us to consider next how human beings use energy to accomplish useful tasks.

Using Energy for Useful Purposes: Heating, Electricity, and Transportation

Energy is used by human beings to perform an extraordinarily large number of useful tasks. For example, the energy content of fossil fuels such as oil, coal, and natural gas is commonly released via combustion processes resulting in the heating of a stream of gas, usually at high pressure. If this resulting hot gas is then expanded within a device known as a heat engine, and some fraction of the thermal energy of the gas stream can be converted into mechanical work consisting of a force exterted over some distance traversed by a component of the heat engine. This work is then usually harnessed for a useful purpose – e.g. to apply a torque to wheels which propel a vehicle on a surface, to apply a torque to an electrical generator to thereby generate and electric voltage across a load and thereby drive a current through the load for a useful purpose, or to create a jet of high exhaust velocity gas which can then propel an aircraft through the air. In other industrial and domestic applications, the heat contained within the hot combustion exhaust is directly used e.g. for manufacturing, building heating, cooking, driving chemical reactions and other applications. Clearly improvements in the efficiency of the conversion of stored chemical energy into useful energy via a change in the conversion technology results in a reduction in use of stored energy for a given demanded useful energy. Thus we will need to understand what determines such conversion efficiencies, and what we can perhaps do to increase such

efficiencies. We will use these laws of thermodynamics to examine this issue of conversion efficiency and shall consider the efficiencies of these processes in more detail in a later chapter.

Implications of Thermodynamics: Energy Return on Energy Investment

In addition to impacting the thermal conversion efficiency, these basic physical laws impose significant impacts upon other aspects of energy systems as well. In particular, these laws imply that in order to acquire energy from some natural resource it takes energy (e.g. it takes some energy to drill an oil well and operate the equipment to extract the oil, it takes energy to mine and refine uranium ore into a form usable as fuel in a nuclear reactor, and so forth). The energy cost is paid in a number of forms. For example, usually this resource is not located immediately where the energy is needed; thus materials must be transported. There is always some energy cost to the ancilliary hardware and equipment that is required to make use of the resource. There may be a requirement to store the energy in some form prior to its practical use. There is also undoubtedly a physical infrastructure that had to be built in order to capture and convert the energy resource; this infrastructure has an energy cost. Finally to then convert the energy for use and deliver it to the location where it is need requires an energy investment. The sum of all of these energy requirements can be considered to be the energy invested. Once we have accounted for all of these energy costs, the energy remaining for practical use is the Energy Return. The ratio of the energy returned to the energy invested (EROEI) is then an important value for any prospective energy source/conversion technology combination. Clearly if this quantity is less than unity, then such a system cannot provide a net source of usable energy.

We can represent this problem schematically as follows:

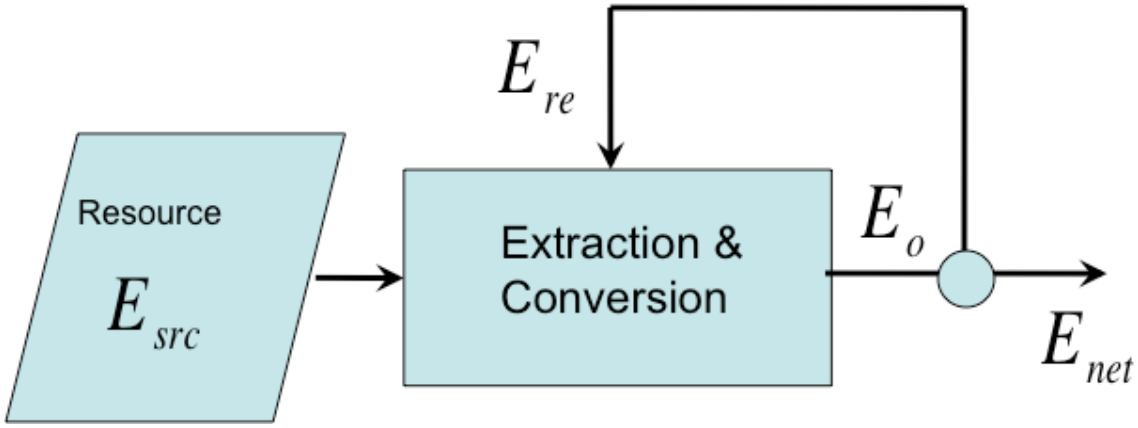


Figure 1.1: Schematic of the flow of energy from a natural source, through an engineered system which subsequently provides a net usable energy.

In this diagram, we imagine that a natural resource (e.g. petroleum, uranium, solar energy input) containing an energy resource denoted as E_{src} in the schematic above is somehow captured by an engineered system. This energy resource is then extracted and converted to usable form, yielding an energy output denoted as. E_o. The operation of the extraction and conversion (E&C) system requires some energy input denoted as. E_{re}. In this simple diagram, we are assuming that E_{re} $E_{re}$is provided by capturing some portion of E_o E_o and re-circulating it to operate the E&C system. The net energy available for other useful work is then denoted as. E_{net}. Obviously we have the requirement that

$$E_{net} = E_o - E_{re}$$

which must be positive. The ratio

$$E_R \equiv \frac{E_o}{E_{re}}$$

is defined as the EROEI, while the energy gain, G, is given by the ratio

$$G = \frac{E_{net}}{E_{re}} . G = \frac{E_{net}}{E_{re}}.$$

Obviously then the EROEI can be written in terms of G as

$$E_R = \frac{E_o}{E_{re}} = \frac{E_{re} + E_{net}}{E_{re}} = G + 1 \, E_R = \frac{E_o}{E_{re}} = \frac{E_{re} + E_{net}}{E_{re}} = G + 1.$$

Thus an energy gain G>0 (i.e. having positive net energy) then requires

$$E_R = \frac{E_{out}}{E_{in}} > 1 \, E_R = \frac{E_{out}}{E_{in}} > 1.$$

Let us now consider the impact that the EROEI has upon the required energy resource, E_{src} E_{src}, necessary to provide a given value of E_{net} E_{net}. Using the above definitions we can first write the ratio of net energy out to re-circulating energy as

$$\frac{E_{net}}{E_{\mathrm{Re}}} = E_R - 1$$

which then gives

$$E_{\mathrm{Re}} = \frac{E_{net}}{E_R - 1} \, E_{\mathrm{Re}} = \frac{E_{net}}{E_R - 1}.$$

We can also write an energy balance at the input to the extraction and conversion system as

$$E_{src} = E_{re} + E_o$$

Using the above expression for $E_{\mathrm{Re}}\, E_{\mathrm{Re}}$ and noting that $E_o = E_{net} + E_{re}\ E_o = E_{net} + E_{re}$ we can then write

$$\frac{E_{src}}{E_{net}} = \frac{E_R}{E_R - 1} + \frac{1}{G} = \frac{E_R + 1}{E_R - 1}$$

or equivalently

$$\frac{E_{src}}{E_{net}} = \frac{E_R + 1}{E_R - 1}\frac{E_{src}}{E_{net}} = \frac{E_R + 1}{E_R - 1}.$$

This is a simple but important result; Figure 1.2 below plots the variation of $E_{src}/E_{net}\ E_{src}/E_{net}$ with $E_R\, E_R$ and shows that for large values of E_R>>1, then $E_{src} \approx E_{net}\ E_{src} \approx E_{net}$ and the re-circulating energy is negligibly small. However, as the value of the EROEI falls towards unity (i.e. $E_R \to 1\ E_R \to 1$), very large amounts of energy must be extracted from the natural resource in order to provide a given demand of net energy (i.e. $E_{src} >> E_{net}\ E_{src} >> E_{net}$). *In such a case, large amounts of re-circulating energy are required simply to extract and convert the energy resource existing in nature to meet the net energy demand. As a result, the natural resource usage rate increases significantly, as will the net cost of the energy due to the requirement to handle large quantities of re-circulating energy within the extraction and conversion system.*

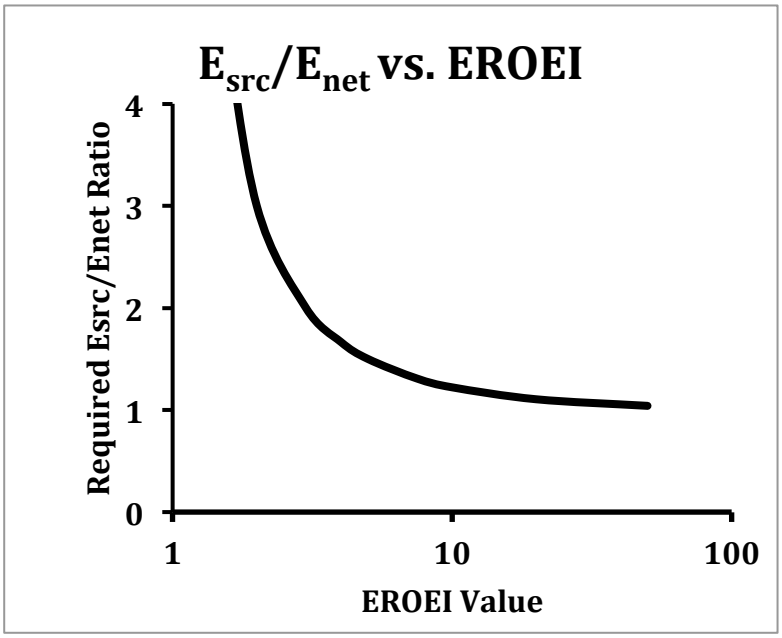


Figure 1.2: Required E_{src}/E_{net} ratio vs. EROEI for a hypothetical self-energized system. As the energy return of the source falls, the energy extraction from the natural source increases.

Minimum EROEI to Sustain a Civilization

It is clear from the above discussion that, in order for a set of prospective energy sources to be capable of providing useful work that can then be used to meet the needs of a population of human beings, at least some of these energy sources must have a sufficiently positive EROEI to provide energy input for energy systems requiring an external energy subsidy (e.g. the production of liquid fuels for transportation from either bio-fuel sources with negative EROEI by completely synthetic processes) and, in addition, meet the primary energy demands of the human population. For example, one could imagine a society that captures solar and wind energy resources, uses some of this energy to manufacture synthetic fuel for use in transportation vehicles, and uses the balance of this renewable energy to meet the other energy demands (e.g.

electricity, heat) of the population. Recent studies of this issue [C. Hall et. al., Energies 2009, 2, pp. 25-47] argue that there is therefore a minimum level of EROEI which is necessary to sustain organized human economies and civilizations, and that this minimum value of EROEI lies in the range of 3 or so; smaller values of EROEI tend to rapidly increase the amount of energy that must be consumed just to recover additional energy resources, thereby leaving fewer resources available for other necessary activities. Thus, as we consider the likely human energy demands in the coming decades, and we look at possible primary energy sources and the associated technologies that can be used to convert this energy into usable form, we must keep in mind the requirement that the EROEI must meet some minimum requirement. As we will also see, our recent experience with fossil fuels has provided experience with very high EROEI primary energy sources; if the new emerging energy technologies which are adopted have relatively smaller EROEI values, then the relative amount of re-circulating energy will increase correspondingly; it is also possible (and perhaps even likely) that as a result, the cost of these emerging energy sources will also be higher.

Chapter 2: Energy, Human Quality of Life and the Energy Status Quo

As we alluded in the introduction to this book, energy allows human beings to perform work on the natural environment, thereby manipulating it to meet our material and other needs. Thus access to adequate sources of energy is needed in order to generate the essential food, water and other physical elements needed to sustain the current human population. In addition, energy is required to drive the balance of the human economy to meet other real and perceived needs and wants. Thus it is natural for human populations that have inadequate access to energy to desire to change that situation, and acquire that access. It is this fact which in large part drives the demand for energy on the scale which is now being seen globally.

In this chapter, we examine the underlying drivers for increased demand for energy around the world. However, rather than building our analysis on purely economic metrics such as per-capita gross domestic product (GDP) or other explicitly economic measures (which are likely not universally valued across cultures, traditions and political organizations), we look at the linkages between energy access and other human quality of life measures such as literacy rates, child mortality rates, life space, and human population growth in order to understand the underlying drivers and implications that emerge from the increasing energy demand that is currently occurring around the globe.

Access to Energy and its Relationship to Human Quality of Life

Statistics such as infant or child mortality rates, literacy rates, average lifespan, education level, and population growth rate provide a quantifiable measure of human quality of life and are readily available at least for the last century or so for most regions and nations of the world. Furthermore, arguably these measures of quality of life are somewhat independent of culture (i.e. people in most cultures want a long lifespan, want reasonable health, want a good life for their children and so forth). These human quality of life measures have been found to be strongly correlated with the average annual per-capita energy use and, in particular, with access to electrical energy [see e.g. Pasternack (2000) and **Figure** below]. We emphasize that this is only a correlation; the cause of such improvements to human quality of life are obviously much more complex and involve many socio-economic variables and impacts that go well beyond the scope of this text. The point we are making here is simply that access to energy is needed to enable these causal agents to take hold. Thus clearly access to an adequate energy supply is correlated with significant improvements in the physical quality of life for human beings around the world. It would then seem reasonable that there is significant motivation for collections of individuals who lack access to adequate energy resources to wish to change that state of affairs. Fundamentally it is this human desire that then drives the need to provide an adequate quantity of energy in useful form to the human population. As we will see in more detail in Chapter 3, access to these energy resources is, as of this writing (2011-2012) highly non-uniform as we look around the globe.

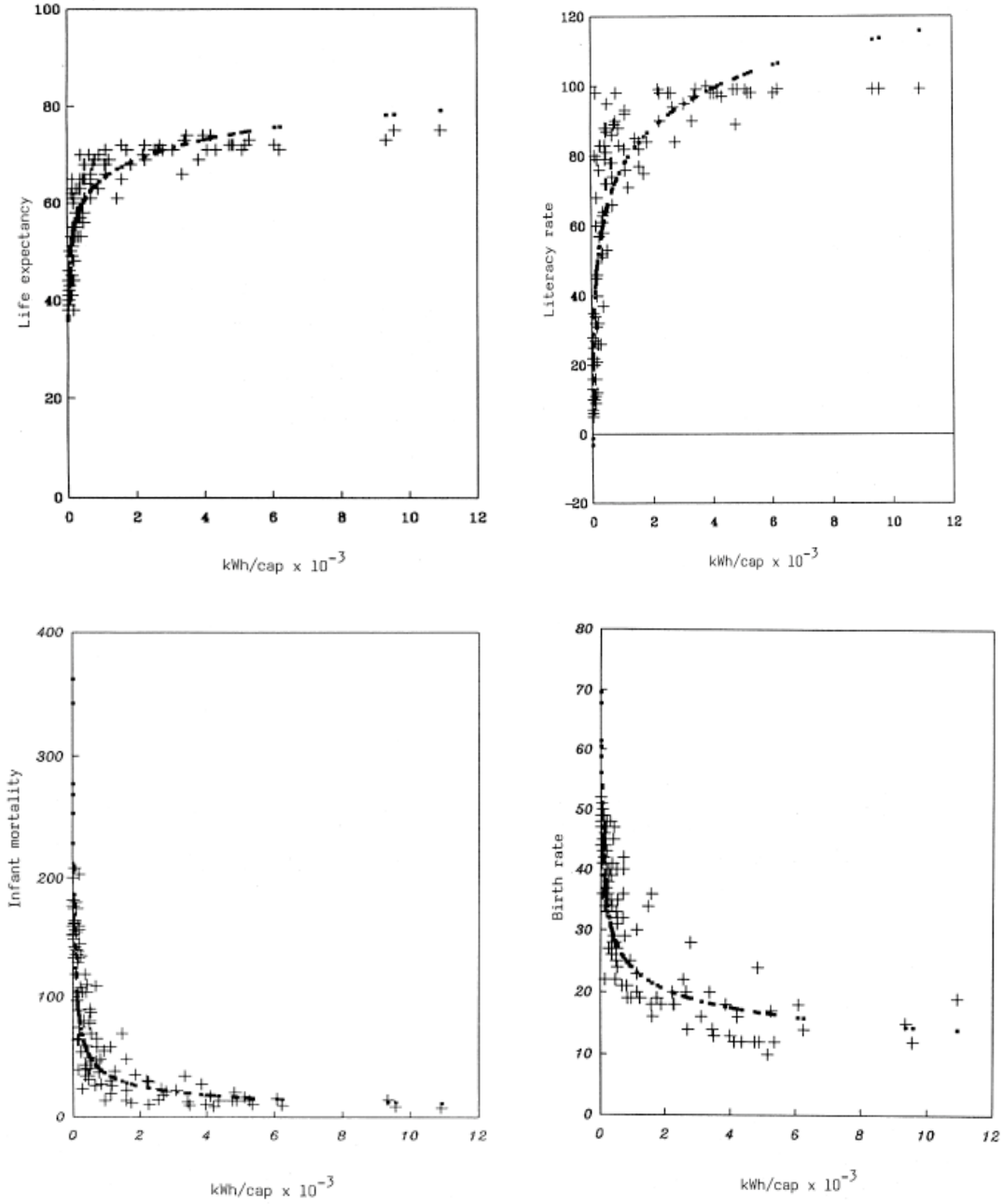


Figure 2.2: Upper left: Average life expectancy vs. energy access in kW-hr/person-year, Upper right: Literacy rate vs. energy access in kW-hr/person-year. Lower left: Infant mortality rate (deaths before 1st year per 1000 live births) vs. energy access, Lower right: Birth rate (births per year per 1000 women) vs. energy access (Alam et. al., 1998).

Next we examine historical data for population growth rates and per-capita energy access **Figure** 2.3 below shows the variation of the population growth rates for Brazil, South Africa, South Korea, India, and China plotted against the annual per-capita electrical energy usage in these nations. The trajectories in this growth rate vs. per-capita annual electrical energy clearly for the period from 1950-2005 clearly show a movement from high growth rate/low energy use to low growth rate/high energy use. Note that this trend is occurring across many different cultures and requires about 40-60 years (i.e. 2-3 generations) to occur. The lighter colored data points shown behind these highlighted trajectories show the current (2005) distribution of all UN-registered nation-states in this parameter space. If we were to examine the time history of each of these nations we would find that most if not all show the same trend towards lower growth rate/higher energy use as one moves forward in time. The primary difference between different nations is found in the starting time for this transition – but note that once the transition begins, many different nations and cultures then follow very similar growth rate/per capita energy trajectories, taking a few human generations to complete the transition.

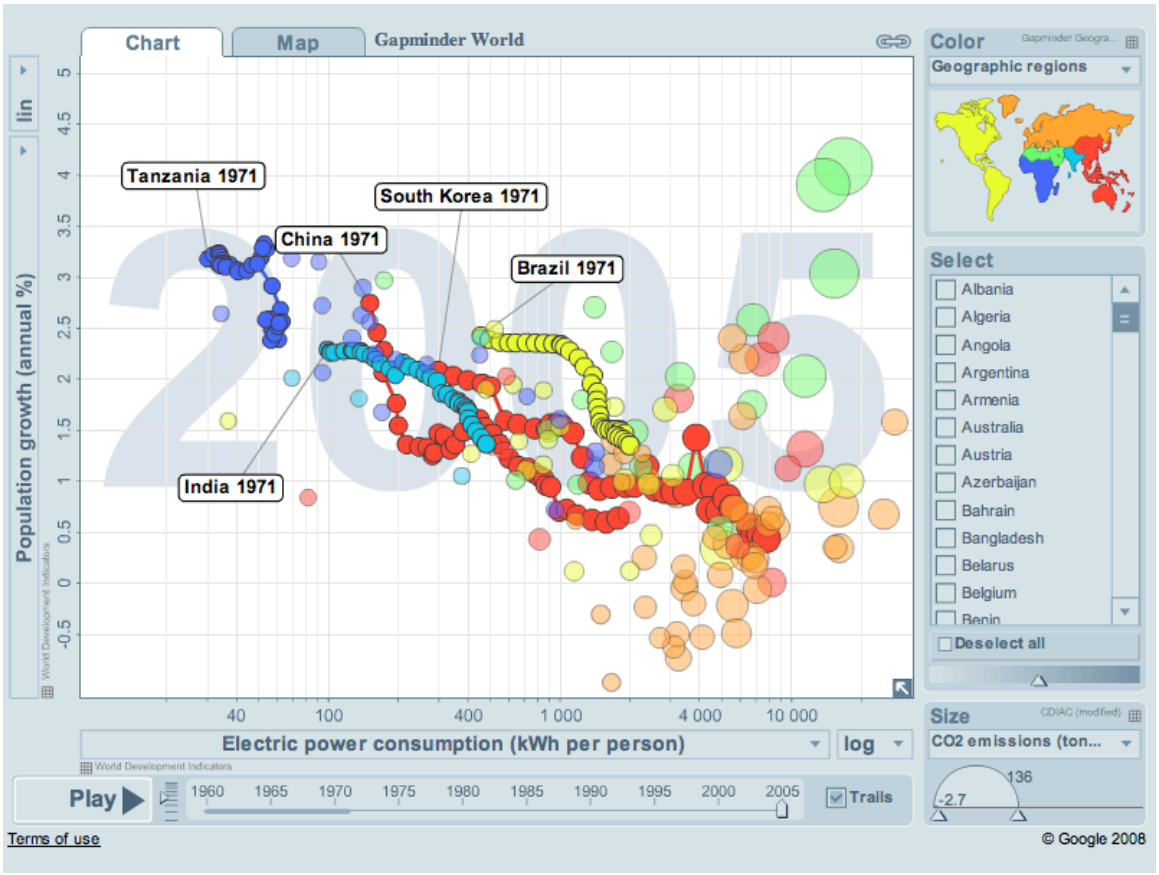


Figure 2.3: Population growth rates vs. annual per-capita electrical energy usage, with time as an independent parameter. Area of data points corresponds to per-capita CO2 emission. Figure obtained from www.gapminder.org, accessed September 2009.

Although these results make no statement on the casuality of these trends, the implication is clear: *reductions in population growth rates along with improvements in many other human quality of life measures are associated, or correlated, with increased access to energy resources*. It is this fact that motivates nations and peoples with little access to energy to desire to increase that access – because it will provide those adult populations and their children with a better life. There is also a corollary to this implication: *If humanity wishes to see the human population stabilize due to very real concerns about the human impact upon the global environment, then*

obviously the population growth rate must decrease; such decreases in population growth rates have historically been correlated with increased access to energy (and particularly to electrical energy). Thus, unless this linkage can be weakened or broken by other social developments, we can likely expect that future population stabilization that the majority, and eventually all, of the human population will require access to adequate level of energy resources. Historically this level of energy resource corresponds to roughly 4000-10,000 kW-hr electrical energy per person per year. Because this is such an important result, let us examine it in a bit more detail.

Energy and Human Population Growth

We can begin to understand the link between human population and access to energy by first considering the plot shown in Figure 2.1 below, which displays the historical human population as well as projections for the next several decades. For most of human history, the *per annum* population growth rate was small (<<1%) resulting in population doubling times (defined as the time required for the human population to double in number) that were measured in multiple centuries. Beginning about two centuries ago the human population growth rate began to increase; at roughly the beginning of the 20th century the population growth rate accelerated to values of ~2-3% per year, resulting in doubling times that within the last few decades have been as small as ~30 years. The result has been the well-known near-exponential growth rate of the human population in the 19th and 20th century. Of course, such growth cannot continue forever, and eventually the growth rate begins to slow down, resulting in an eventual saturation of the population. Indeed, recent studies of human population growth point towards evidence that such a decline in the growth rate is now underway [ref: U.S. Census Bureau].

These dynamics can be described reasonably well by the Verhults-Pearl model (also known as the logistics model) for population growth which posits that the time-dependent population, P(t), obeys an equation given as

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right)$$

where r denotes the exponential growth rate for periods when when P(t)<<K. Here the parameter K denotes the carrying capacity of the environment in which the population lives, equivalent to the maximum population that can be sustained or carried by the available resources without collapse. The solution to this nonlinear differential equation is given as

$$P(t) = \frac{KP_0}{P_0 + (K - P_0)e^{-rt}}$$

where $P_0 = P(t)$ $P_0 = P(t)$ denotes the initial population at t=0. Clearly, this model has limitations (e.g. it always predicts P->K for large t, excluding possibilities such as extinction and population collapse which are known to be possible); however for the purpose here it will suffice. We have taken the historical and near-future projected global human population, and have fitted that to this simple model. The results shown in Figure 2.1 suggest that many centuries ago the human population was (nearly) in an equilibrium (at least on a global scale) characterized by very slow population growth rates. Then, beginning sometime in the 19th century the growth rate began to increase significantly. Much more realistic demographic projections then suggest that human population will begin to level off in the late 21st century at

values in the range of 10 billion or so individuals. The question then arises: what allowed the rapid increase in population occur over the past two centuries, and how might this change be related to access to energy?

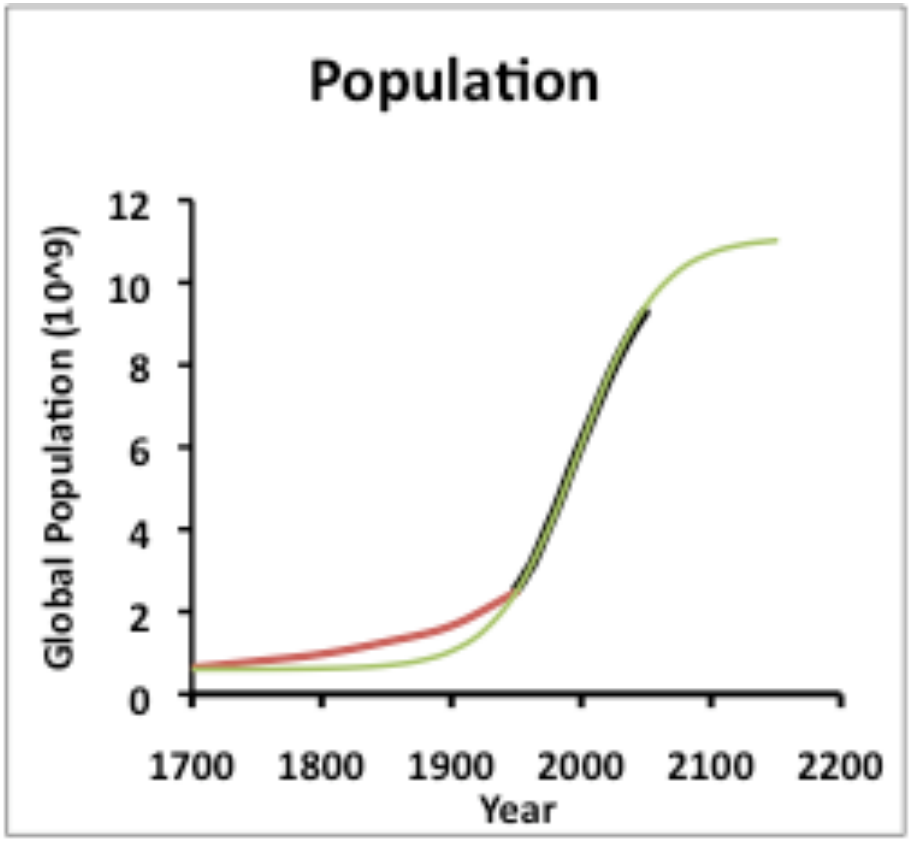


Figure 2.1: Human population for last ~300 years and projected to the year 2050 (brown and solid black lines) on UN demographic studies [US Census Bureau]. Green line: Verlhaus-Pearl model fit to the actual population, using t0=1910, r0=3.2%, 0.6 Billion initial population, and final carrying capacity of 11 billion.

We can begin to understand the population dynamics shown in Figure 2.1 by considering the possibility that the carrying capacity, K, could change in time due, e.g. to the introduction of a new source of food, water and so forth at some time t_0. Increased production or availability of these resources requires additional energy inputs into the physical environment (e.g. the

transportation of fresh water over long distances for irrigation purposes, or the production of synthetically produced fertilizer which in turn increases crop yields and therefore increases food production). In the context of this discussion, this increase in carrying capacity can be associated with the introduction of new sources of energy which can be used to increase the availability of food, water and other essential requirements for human life.

With these comments in mind, let us consider a simple model in which, for $t<t_0$ we denote the carrying capacity as K_1, while for $t>t_1$we denote the new carrying capacity as K_2. In this model it is understood that $K_2>K_1$, i.e. the carrying capacity is increased at $t=t_0$. Examining the solution P(t) above we see that for rt>>1 clearly P(t)=>K . Thus, for times prior to but approaching t_0, the population would be given as $P \approx K_1$ $P \approx K_1$. For the human population history shown above, it would appear that $K_1 \approx 0.3-0.5\times 10^9$ *individuals* $K_1 \approx 0.3-0.5\times 10^9$ *individuals*.

With the introduction of the additional food and water resources at $t=t_0$ (which are in turn correlated and perhaps even enabled by the introduction of energy technologies), then the carrying capacity would increase to a value $K_2>K_1$. The population then begins to undergo exponential growth at a rate r, and will increase from the equilibrium value K_1that held prior to t_0. Based upon the population history shown above, we estimate t_0 to be the year ~1900 or so, since that is when significant human population growth seems to have begun. From the data shown it appears that r0~0.03/year. A period of rapid population growth will then begin once the carrying capacity has changed. Then, once enough time passes such that $r(t-t_0)>>1$ the population will begin to gradually approach a value K_2 which we have assumed is larger than K. (the student should note that similarly a population decrease can occur is a change in the system occurs such

that $K_2<K_1$ due e.g. to a decrease in the natural carrying capacity of the environment, or if a reduction in the available food, water, and other critical components to sustain life were to occur). If we take r_0~3%, then we would estimate that in a period $t-t_0$~ 200 years or so the population saturation would occur. With a date t_0=1900, then this saturation would occur in the 2100 timeframe. The results of this simple model are shown as the green line in .

Projections made with much more sophisticated demographic models [U.S. Census Bureau] suggest that a value the human population will saturate in the range of 10-12 billion individuals in the late 21st century. We can therefore at least begin to understand the human population growth that has been occurring in the past ~200-300 years as being associated with the introduction of new physical and human resources such as additional food, improved sanitation, improved levels of education and medical care and so forth that could support an increased human population. Historically, fossil fuel consumption also began to grow rapidly during this same time period as technologies for converting the chemical energy stored within fossil fuels into useful purposes were developed in the early 19th century. These energy sources are inherently limited in their availability and, moreover, their widespread use has now begun to have very significant global environmental impacts.

The Link Between Future Population Growth and Future Energy Demand

We can now use these results to briefly examine the relationship between future human population growth and future energy demand. Here, we follow the development of Sheffield (1998) which provides a simple model of the linkage between future population growth and future energy demand.

In the year 2000 the global human population was about P_0~6.0 billion and the population growth rate r_0 = 1.6% [Sheffield, 1998]. If this growth rate were to continue unabated then after a time t the increment in population over a subsequent time interval Δt will be given as. $\Delta P = P_0\left[\left(1+r_0\right)^{t+\Delta t} - \left(1+r_0\right)^{t}\right]$. Next, we now suppose that for each succeeding year the population growth rate is decreased by a value 0<f<1 from the growth rate of the previous year, i.e. in the first year the growth rate $r_1 = fr_0$, $r_1 = fr_0$, in the second year the growth rate is $r_2 = f^2 r_0$ $r_2 = f^2 r_0$ and so forth. This reduction in the population growth rate as the population begins to approach the carrying capacity of the environment is simply a reflection of the forgoing discussion of population dynamics. Thus in the i-th year the population growth rate is given as $r_i = f^i r_0\, r_i = f^i r_0$ and we can write the population in the i-th year P_i as

$$P_i = P_0 + \Delta P_0 + f\Delta P_0 + f^2\Delta P_0 + \text{L} + f^{i-1}\Delta P_0$$

where

$$\begin{aligned}\Delta P_0 &\equiv \Delta P\big|_{t=0.\Delta t=1} \\ &= r_0 P_0\end{aligned}$$

denotes the increment in population during the first time interval. We can factor the above summation into the form

$$P_i = P_0 + \Delta P_0\left(1 + f + f^2 + ... + f^{i-1}\right)$$

which, using the definition above, can be written in summation form as

$$\begin{aligned}P_i &= P_0 + \Delta P_0\left(1 + f + f^2 + ... + f^{i-1}\right) \\ &= P_0 + r_0 P_0\left(1 + f + f^2 + ... + f^{i-1}\right) \\ &= P_0\left(1 + r_0 \sum_{j=0,i} f^j\right)\end{aligned}$$

If we now evaluate the summation in the limit that $i \to \infty$we can then write the final population for large times. This result can then be used to solve for the annual growth rate decrement, *f*, needed to arrive at a particular value of population. In order to do this, we first express the series term as

$$\lim_{i\to\infty} \sum_{j=0,i} f^j = \frac{1}{1-f}; \quad f<1$$

With this result, we can then solve for the annual growth rate reduction factor, f, in terms of the initial growth rate and the final target population:

$$f = 1 - r_0\left(\frac{P_\infty}{P_0} - 1\right)^{-1}.$$

If we take the final stable human population to be given as $P_\infty = 10^{10}$ individuals (consistent with more detailed demographic projections) and use the values given above in the year 2000, we then require f~0.976, i.e. we find that the population growth rate must decrease by a decrement of 2.4% each year.

In order to link this result to changes in future energy demand, we need to take into account the historical trends shown in **Figure** . Examining this figure, we surmise that, if these historical trends continue into the future, then such a growth rate decrement will be associated with an increase in per-capita annual electrical energy usage. For example, if f=0.976, then after ten years this would result in a decrease in population growth rate of f^{10}~0.7 or so. Taking a rough average of the trajectories of China, Brazil and India from Figure 2.3 we can construct a rough estimate of the globally averaged growth rate vs. per capita annual energy usage as shown in Figure 2.4 below. If we then chose the starting time such that the initial growth rate was 2%/year, then for f=0.976, after ten years we would expect that the growth rate would have

fallen to ~1.4%/year. Examining Figure 2.4, we then see that this corresponds to a 2-3x increase in per-capita annual electrical energy usage. Similarly, a transition to a steady state population requires a somewhat higher per-capita electrical energy access, and would likely take several generations to complete. Globally the human population is currently increasing at a rate of ~1.5%/year or so. The results of this analysis then suggest that a transition to a steady-state population would increase the average per-capita annual electrical energy usage from ~1000-2000 kW-hr/person/year up to 2-3x higher values of per capita electrical energy usage. If the final population then ends up in the range of ~10 billion or so as discussed above, then we might expect a 3-5x increase in global electrical energy demand by the time the human population saturates. There would likely be increased demand for energy used in other applications (heat, transportation). The question then becomes: how can such levels of energy demand be met indefinitely?

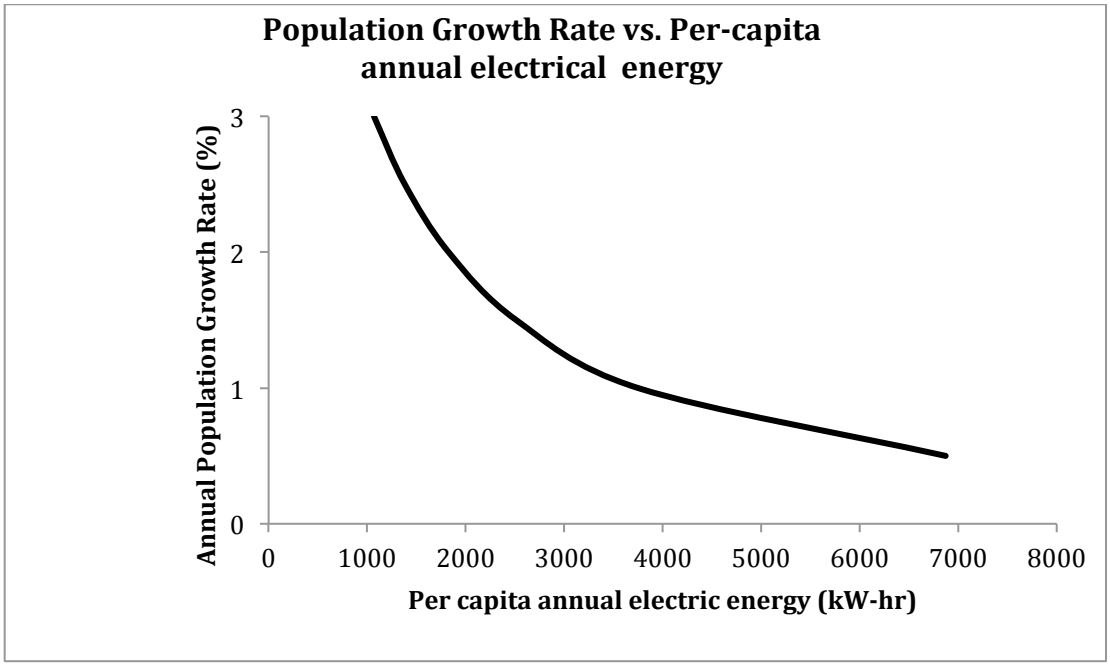


Figure 2.4: Rough fitted model to the growth rate-per capita annual electric energy shown earlier in Figure .

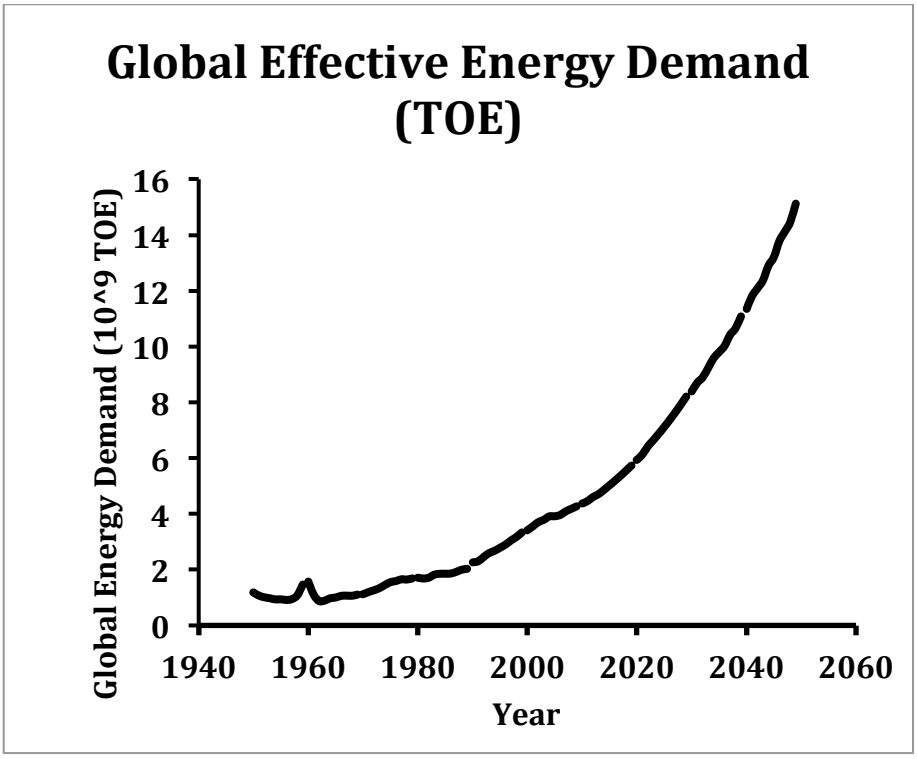


Figure 2.5: Global energy demand computed from current-year population and from the modeled growth rate-per capita annual energy required. Energy is computed in TOE, using a thermal efficiency of 33%.

Chapter 3: Historical and Current Primary Energy Sources

Human beings have used a variety of primary energy sources to meet their needs in the past. **Figure** below shows the historical and projected energy demand for all primary energy sources in the U.S., which for our purposes here we shall take as representative of the evolution of an industrialized economy over the last 150+ years. In the 19th century, nearly all of the energy used in the U.S. came from the combustion of wood and, to a lesser extent, coal. With the growth in the applications of heat engines for industrial and transportation applications, the consumption of coal increased substantially, and it became the dominant primary energy source by ~1880 or so. Coal consumption continued to grow for the remainder of the 19th century and into the first decades of the 20th century. With the advent of the automobile, consumption of petroleum became significant in the first decades of the 20th century. Oil and coal consumption then grew extremely rapidly in the period immediately following World War II and, at the same time, natural gas consumption began to grow rapidly. Finally, in the 1970s, nuclear fission begins to make a measurable contribution to the U.S. energy demand. The situation is similar for many industrialized economies and, as we will see in a moment, it is also reflected somewhat in the more recent history of the emerging economies of Asia and the Indian subcontinent. It is clear that for the last century coal, oil, and natural gas have provided the bulk of energy for human needs, and it is expected for the next several decades at least this trend will continue, as shown in the recent history and projections of Figure .

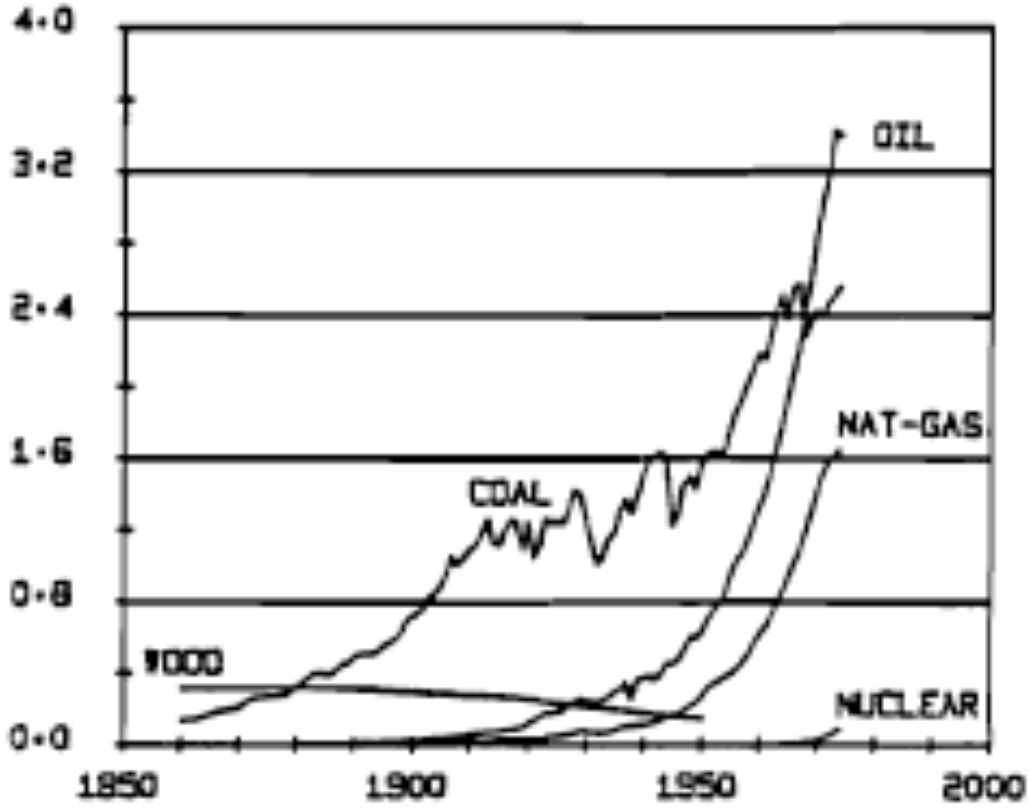


Figure 3.1: Absolute value of primary energy supplied in the US from various sources. Y-axis is in units of tones of coal equivalent (1 Tonne coal = 7 Gcalories). Figure taken from Marchetti & Nacinovik, IIASA RR-79-013.

It is also useful to consider the recent historical and projected total primary energy demand for the world (see Figure 3.2 below). Energy use in the OECD region has nearly saturated, and is expected to exhibit very small (<1%/year) growth rates going into the future. In contrast, energy use in China, India and the rest of the world (rest of non-OECD) is growing rapidly. As a result, the large variations in per-capita energy use that has existed for the last several generations as shown in Figure 3.3 below should be reduced by mid-century. Viewed in the context of the population dynamics discussion above, these trends suggest that the rich

regions of the world have already saturated their per-capita energy usage and their population growth, while the rapidly developing regions such as China and India are in the midst of the transition from a situation consisting of high population growth rate/low per-capita energy to a state of lower population growth rate/higher per-capita energy consumption. Indeed, an examination of demographic data suggests that this is nearly the case (e.g. if immigration is excluded, then the populations of Europe and the U.S. and Canada are nearly constant or even decreasing slightly; the populations of China and India are growing but the growth rates are becoming smaller with time consistent with the trends discussed in the previous chapter).

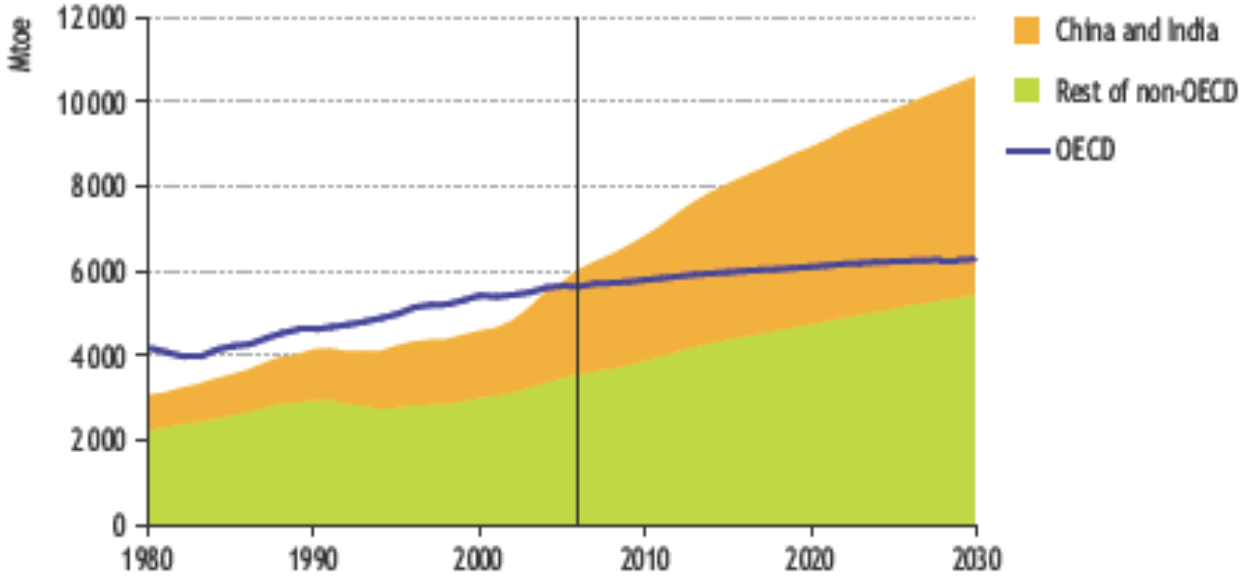


Figure 3.2: Historical and projected total primary energy demand for OECD, China and India, and remainder of non-OECD (i.e. rest of world). Figure taken from WEO 2008 Report by IEA.

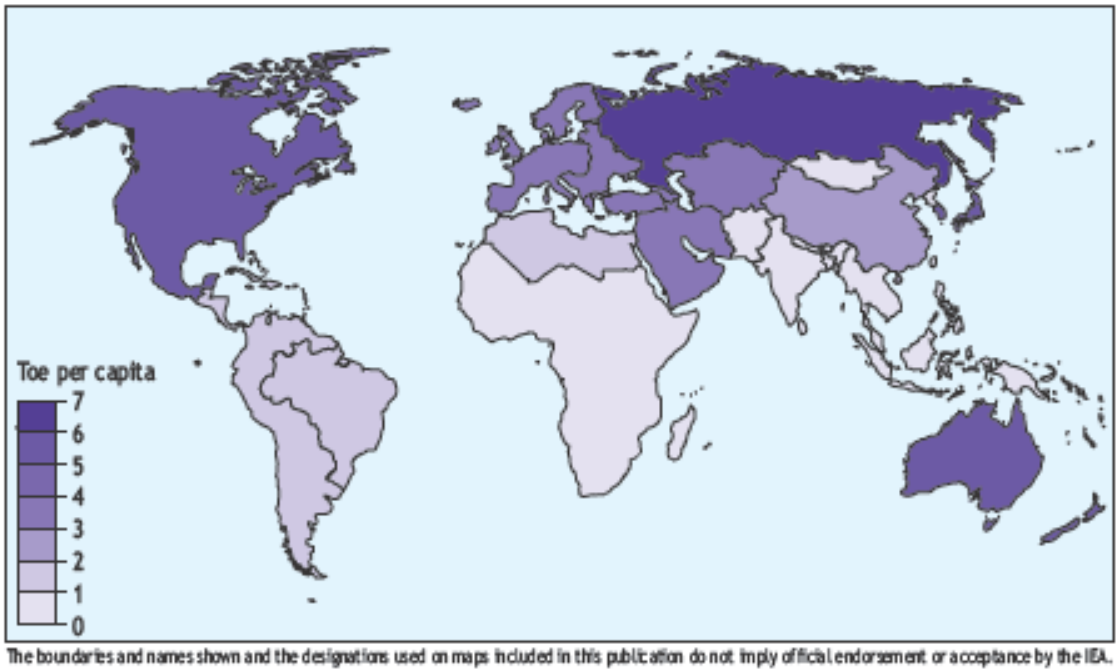


Figure 3.3: Per-capita annual energy use (in TOE equivalent) projected for 2030. Figure taken from World Energy Outlook 2008, published by International Energy Agency.

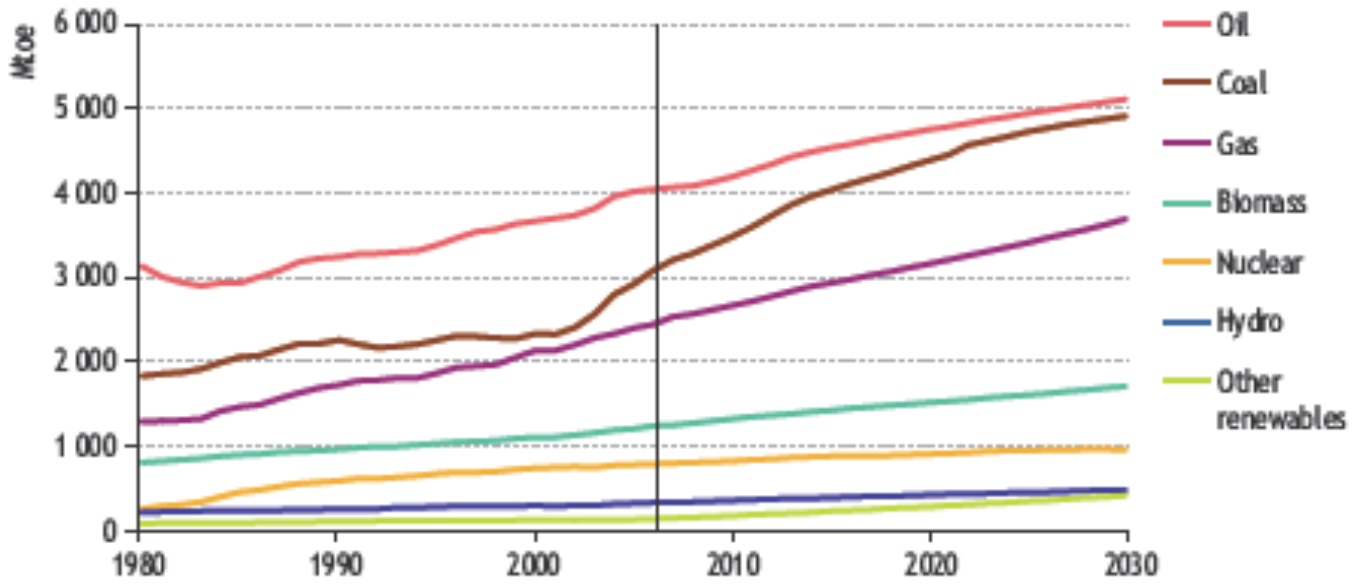


Figure 3.2: Near-term energy demand projections according to primary energy source. Figure taken from WEO 2008 Reprt by IEA.

This development, which is good news in the sense that a larger fraction of the human population is enjoying a better quality of life characterized by e.g. higher literacy rates, lower infant mortality rates, longer life spans and so forth, also carries with it the clear implication that the need for additional energy resources will only become larger as time goes on. The fact is that the majority (roughly 80%) of the current energy demand is met by fossil fuel consumption and that this dominance of the energy supply landscape by fossil fuels is expected to continue into the near future (see Figure 3.4). However, we know that fossil fuel resources are ultimately limited. Thus we are led to the deceptively simple question: roughly how long can we expect to be able to meet our energy demands by the combustion of fossil fuels?

Projections of Future Fossil Fuel Resources

Fossil fuels (i.e. petroleum, coal, and natural gas) represent finite resources that, once extracted and consumed, are not replaced on any timescale of interest for human beings. Thus, it stands to reason that at some point in time the rate of recovery of such finite resources must inevitably begin to decline. The question then becomes how to model and estimate these processes.

Let us denote the total quantity of a particular fossil fuel available for extraction as Q_0, and the rate at which these resources are extracted or produced is given as P(t). Note that these quantities are functions of both physical limits as well as economic factors such as the market price and recovery cost of the resource. Here, we will focus on the physical limitations and simply assume that if demand for the resource is high enough then the price would rise to the point where recovery would be economically feasible. We also let Q(t) denote the total cumulative production from very early time up until time t. From these definitions it is clear that $P(t) = \frac{dQ(t)}{dt}$ and then, of course, $Q(t) = \int_{t'=0}^{t} P(t')dt'$ $Q(t) = \int_{t'=0}^{t} P(t')dt'$ where Q(t) denotes the resource that has been recovered or extracted upon until time t. Because this finite resource is not replaced by nature during the extraction process, we can integrate this equation to write the remaining reserve at time t, $Q_{res}(t)$, as

.

$$Q_{res} = Q_0 - Q(t) = Q_0 - \int_{t'=0}^{t} P(t')dt'.$$

Obviously when $Q(t)=Q_0$ then the resource has been depleted and the production then ceases. These simple considerations are complicated by the fact that the value of Q_0 is *never* known

apriori and is, in fact, a complex function of the available extraction technologies, existing infrastructure for extraction and utilization, and of course, economic value or price. Keeping in mind these critical issues (which really go beyond the scope of our discussion here), we examine a widely used model for such resource extraction.

In 1959, Hubbert [REFERENCE] noted that production of such a finite resource would naturally start from a small value, grow to a peak value, and then decline again to small values. He hypothesized that Q(t) would follow a logistics law which, at early times when production rates are small and $\frac{Q(t)}{Q_0} \approx 0$ yields nearly exponential growth in production. Then, when $q(t) = \frac{Q(t)}{Q_0}$ is small but finite, the production rate will begin to saturate. This behavior can be described by the so-called logistics equation (which we have actually already seen in the context of population growth dynamics) for the evolution of q(t)

$$\frac{dq}{dt} = rq(t)\left(1 - q(t)\right)$$

where r denotes the growth rate at an early time where only a small fraction of the resource has been extracted, i.e. when0<q<<1.

The solution to this equation is given as

$$q(t) = \frac{1}{1 + \exp(-r(t - t_0))}$$
$$with$$
$$q(t_0) = 0.5$$

Which can be seen by taking the derivative of q(t) and substituting the result back into the original model equation.

We can now estimate the evolution of the production rate, P(t), by differentiating this expression:

$$P(t) = \frac{dQ(t)}{dt} = Q_0 \frac{dq(t)}{dt}$$
$$= rQ_0 \frac{\exp(-r(t-t_0))}{\left[1+\exp(-r(t-t_0))\right]^2}$$

Thus the long time behavior of resource production model is determined by the early growth rate, r, and the total reserve size, Q_0. As pointed out above, the total reserve size is never known until the resource is exhausted; at early times only poorly defined estimates of this value are known. Historically, the estimate of the total reserve tends to increase as resource exploration is carried out; however it eventually must approach the ultimate amount of the resource that can be economically extracted from the total reserve quantity that exists within Earth's geological structures. In other words, the production rate of a finite resource will initially grow, and then eventually peak and then begin to decline.

This type of analysis has been extensively applied to conventional petroleum resources. Estimates for the total recoverable reserve size Q_0 vary from $2\text{-}4\text{x}10^{12}$ barrels. The lower value is thought to be roughly the global conventional petroleum resource; the larger values take into account the recent economic recovery of so-called unconventional petroleum resources such as found in tar sands and shale deposits. These unconventional resources can be profitably extracted at currently prevailing prices using recently developed new technologies. Using historical growth rates for global petroleum consumption, r=4.5% [REFERENCE], this model

would then predict the production histories shown in below. Note that for Q_0 ranging from 2-4x10^{12} barrels the production rates would then be expected to peak between 2015 and 2025. The resource availability for coal and natural gas would eventually follow similar trajectories, although the consumption rates and total reserve size estimates put the peak production of these resources later in the 21st century. Despite the uncertainty over the exact timing of the peak production, the fact that at some point in the next few decades, the demand for these fossil fuels will begin to push up against the achievable production rate. We emphasize here that at that point we will *not* "run out" of petroleum. Instead, the result will be an increase in the price of the energy resource, which acts to depress demand and, at the same time, encourage movement to other sources of energy which, until that point in time, were more expensive that the conventional fossil fuels. The final result will be an eventual transition to some other replacement energy source that would then replace petroleum. The question of course then arises: what source(s) could provide this replacement? Furthermore, turning our attention to the other fossil fuels – namely natural gas and coal – a similar question arises. In that case, the fossil fuel is used primarily to produce energy in the form of electricity, whereas petroleum is used primary as the feedstock for producing liquid fuels for transportation systems.

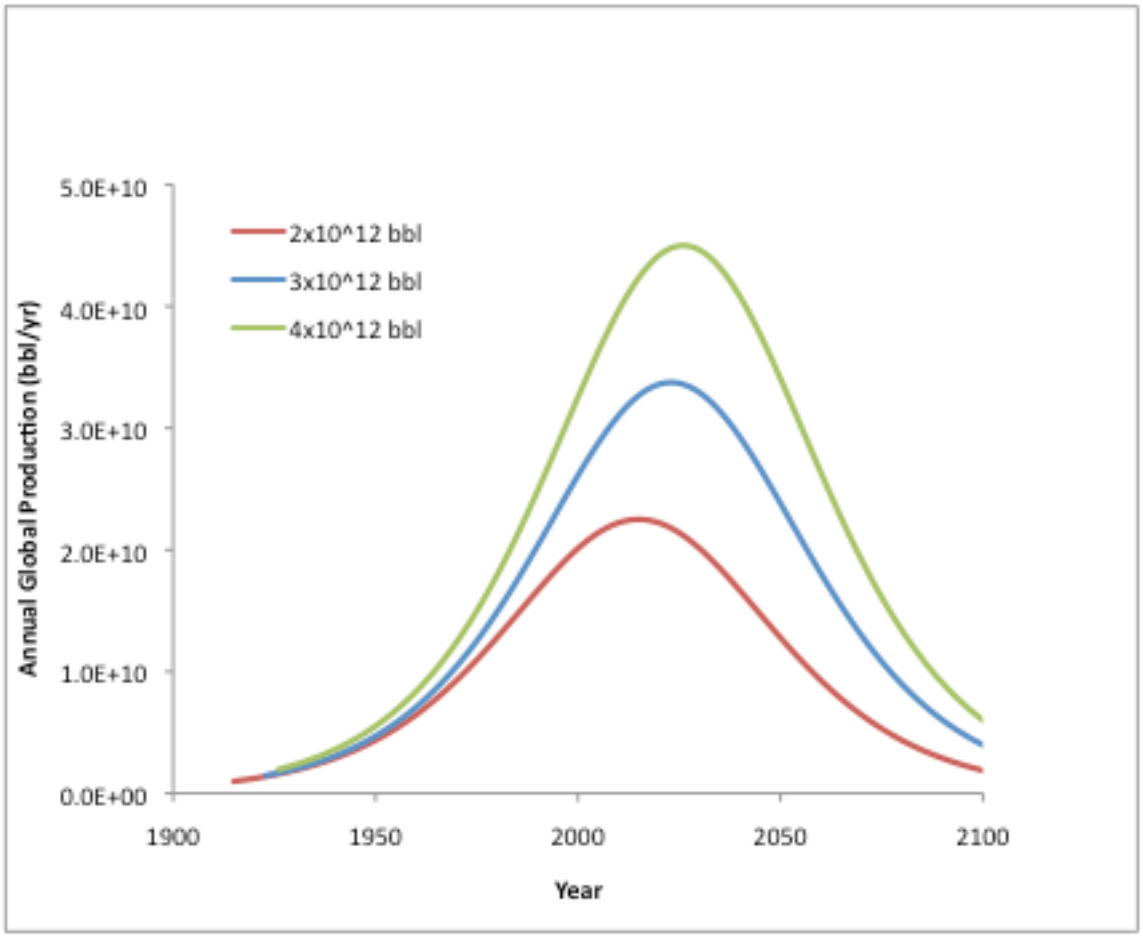


Figure 4.16. Estimated annual global production of petroleum based upon three different estimated ultimate reserves (2 x10 12 bbl, 3 x 10 12 bbl, and 4 x 10 12 bbl) corresponding to a range of estimated geologically available petroleum. R=4.5% for all cases. Models use historical data to establish exact year (e.g. P=2x10^9 bbl/yr in 1940). Peak production occurs in ~2015, 2023, and 2026 respectively.

Similar projections can be made for coal and natural gas; this is left as a homework problem for the students. The result of such a projection obviously depends upon the estimated reserve as well as upon the growth rate of consumption. Despite these factors, the implication is clear: fossil fuels are a finite resource and, given the historical growth rate in consumption of this finite resource, one can anticipate that at some point in the future (i.e. in a period ranging from a few decades to at most one century) the rate of extraction and conversion of these energy

resources will peak and then begin a decline. At that point, humans will either have to make a transition to other energy sources and conversion techniques, or be prepared to use less energy (and thus accept the consequences that accompany such a decision). If alternate sources are available and have been deployed on a sufficient scale, then the transition can happen in some sort of smooth (but perhaps inconvenient) manner. If such alternate energy sources have not been developed and deployed on a sufficient scale, then this transition is likely to be much more disruptive and difficult. In addition to these resource limitation considerations, there are also significant global climate change effects associated with the large-scale consumption of fossil fuels. These considerations also provide a strong motivation to make a transition away from fossil fuel energy sources.

Chapter 4: Review of Thermodynamics and Application to the Performance of Heat Engines

The previous discussion has focused on the large-scale demand for energy, the impact of this energy access on human quality of life and population growth. We then briefly examined where this energy comes from at present, and found that fossil fuels provide the large majority of this energy. We then make simple estimates of the duration of fossil fuel availability to meet this global demand. The ability of fossil fuels to meet human energy demand could be stretched out to longer times by improving the efficiency of our conversion technologies – i.e .by extracting more useful work out of a given quantity of fossil fuel. Thus we are led as ask the question: what determines the conversion efficiency of the devices (i.e. the heat engines) that we use today? To address this question, we must delve into the subject of thermodynamics, which we take up here.

Work and Heat Interaction between System and Environment

Let us now consider the exchange of work and heat between a system and the surrounding environment (see Figure 4.1 below). The boundary between the system and its environment are defined by the dashed line and we suppose that work and heat can move between these two regions. By convention, work that the system does on the environment is positive, i.e. a positive increment of work is given by

$$dW = \mathbf{F}_{env} \cdot d\mathbf{r}_{env}$$

where $\mathbf{F}_{env}$ denotes the force exerted by the system on the environment, and $d\mathbf{r}_{env}$ denotes the resulting displacement of the system-environment boundary.

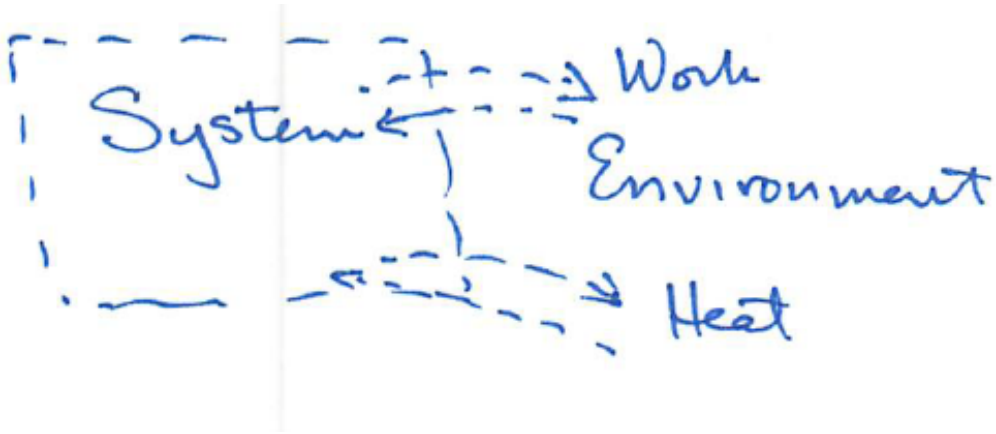


Figure 4.1: Schematic of the exchange of work and heat between a system and the surrounding environment.

We can illustrate these considerations by considering the work exerted by a high pressure gas on a piston that moves within a fixed cylinder as shown below in Figure 4.2. The system boundary in this case is formed by the walls of the cylinder and the face of the piston. The environment is then the piston itself and whatever other devices are connected to the piston. If the gas has a pressure p and the piston face an area A, and the gas pressure causes the piston to be displaced by a small distance dr, then the incremental work done by the system on the environment is given by

$$dW = Fdr = pAdr.$$

If the walls of the cylinder and the piston itself are perfectly insulating, then no heat can be exchanged between the environment and the system and, as a result of energy conservation, the gas must cool down. The amount of thermal energy loss would then precisely balance the amount of work done by the piston on the environment.

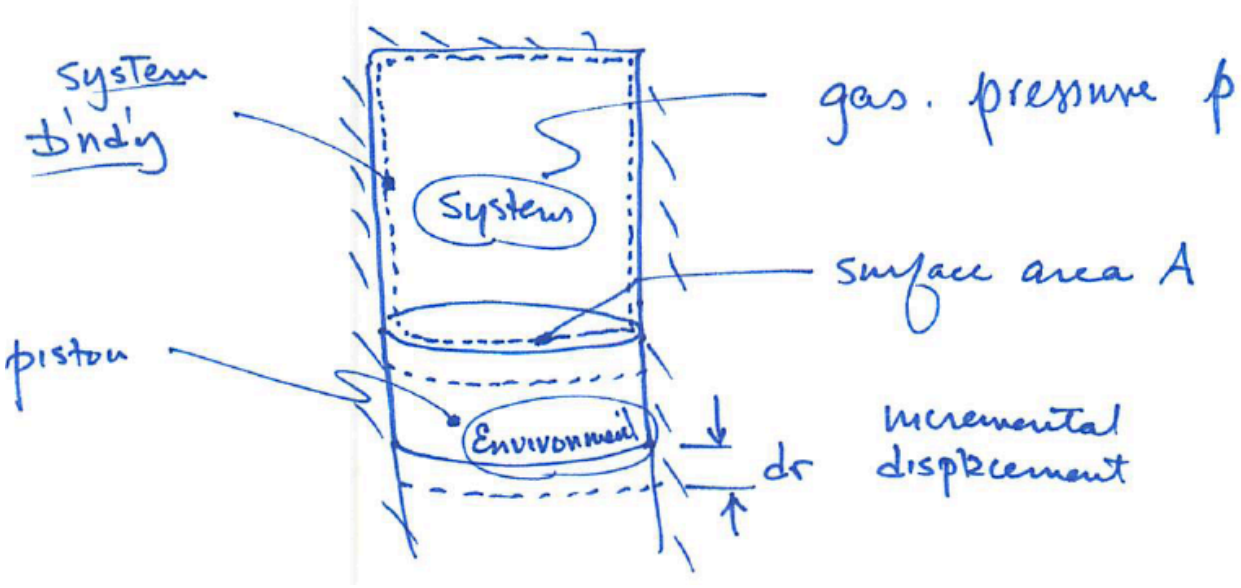


Figure 4.2: Schematic of a cylinder and movable piston that enclose a gas at pressure p.

These principles of energy conservation are universally observed in nature and thus are described as being governed by a physical law, commonly known as the first law of thermodynamics which is always true. This law states that the increment in system energy, dE, equals the increment in heat dQ which is transferred to the system from the environment minus the work, dW, done by the system on the environment. In symbols, this is expressed as

$$dE = dQ - dW\,.$$

If a sequence of such exchanges are considered to have occurred, then we can integrate this expression over all such exchanges to give

$$E_{final} - E_{init} = \int_{E_{init}}^{E_{final}} dE = \int_{init}^{final} dQ - \int_{init}^{final} dW$$

where here initial and final refer to the starting and ending conditions, or state, of the system during this exchange.

Let us now consider a question: What happens if the initial and ending states of the system are equal? To make progress on this question, we need to introduce several additional concepts. First, let us assume that a temperature scale, T, exists and that T is always positive, i.e. T>0. Furthermore, T is independent of the materials properties. Second, let us define a quantity known as the entropy S of a system. The increment in S is defined as

$$dS \equiv \left.\frac{dQ}{T}\right|_{rev}$$

Where dQ denotes an infinitesimal idealized reversible exchange of heat between the system and the environment which occurs at a system temperature T. This increment is so small that T does not change during this exchange and is reversible in the sense that the direction of heat transfer could reverse and all of the incremental heat addition could be returned to the environment.

The second law of thermodynamics then states that for any real incremental heat transfer process occurring between a system and the surrounding environment, the entropy must always increase or, at best, stay constant.

$$dS \geq \left.\frac{dQ}{T}\right|_{real}.$$

A real process where dQ=0 is referred to as an adiabatic process. In the ideal, or reversible, limit such processes have dS=0. In real adiabatic processes, we always have dS>0. Such processes are said to be irreversible. These concepts will play an important role in our evaluation of heat engine performance.

Thermodynamic properties

In order to analyze the exchange of heat between a system and its environment, we need to define several thermodynamic properties or quantities which will prove useful. These properties are usually classified as either being intensive properties – i.e. properties that do not depend upon system mass, and extensive properties, whose values do depend upon system mass. Examples of intensive properties include the kinetic pressure, p and the temperature T. Extensive properties include the total energy E, system volume V, and total entropy S. It is often useful to normalize these extensive properties by the system mass M, e.g.

Specific internal energy is defined as

$$e = \frac{E}{M}$$

Specific volume is defined as

$$v = \frac{V}{M}$$

and specific entropy is defined as

$$s = \frac{S}{M}.$$

These quantities can be used to re-express the first law of thermodynamics. To do so, it is useful to first define the specific enthalpy, h, as

$$h = e + pv.$$

Enthalpy is the sum of the specific internal energy (which is determined by the energy stored in chemical or nuclear bonds which can be liberated in suitable reactions) and the product of the kinetic pressure (which measured the kinetic energy per unit volume) and the specific volume (which determines the volume needed per unit mass. Thus, enthalpy is the sum of random thermal kinetic energy and internal stored energy.

To gain additional insight into the usefulness and meaning of the concept of enthalpy, consider the schematic shown in Figure below. A system with volume v is maintained at fixed kinetic pressure p_0. At t=0 an incremental amount of heat, dq, is added to the system from the environment, resulting in an increase in the temperature, i.e. T(t>0)>T(t<0). In order for pressure to remain constant the system volume increases by an amount dv. Applying the first law to this system allows us to write

$$de = dq - pdv.$$

We can re-arrange this to give

$$dq = de + pdv$$

which, since p=constant, can be re-written as

$$dq = de + d(pv)$$
$$= d(e + pv)$$

Using the definition of enthalpy then gives

$$dq = dh.$$

for a process that occurs at constant pressure. Thus we learn that at fixed pressure, the heat input into the system is equal to the increase in system enthalpy. This result will be important in our upcoming analysis of heat engines.

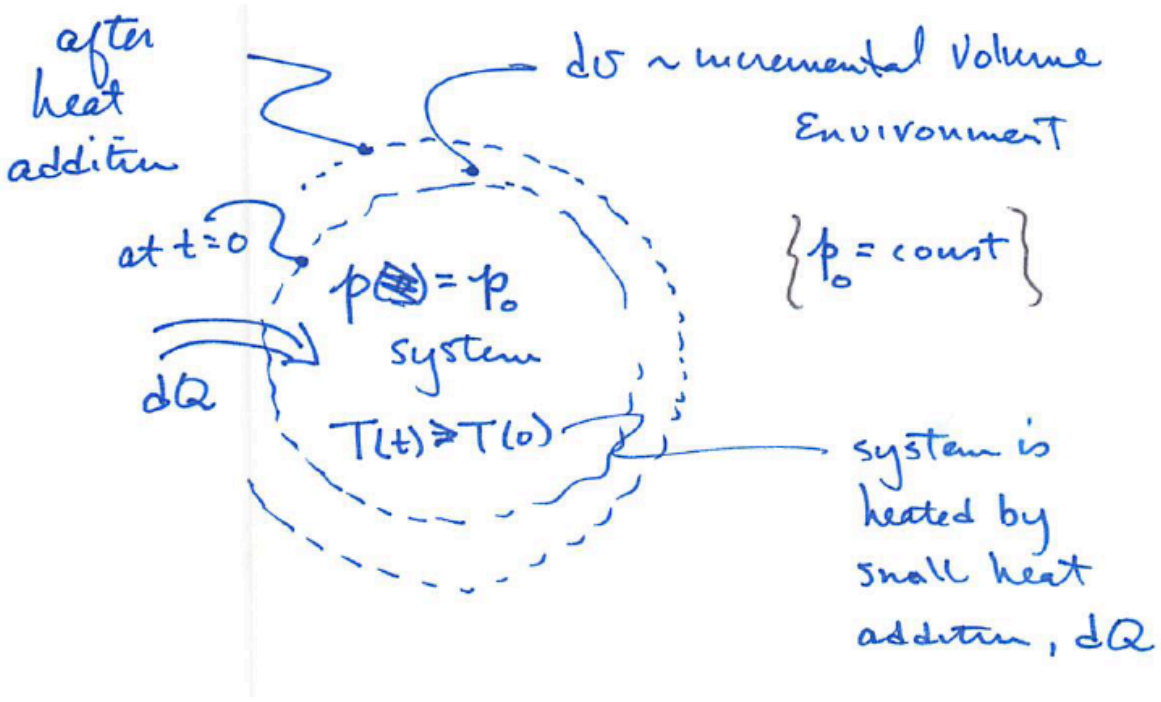


Figure 4.3: An incremental addition of heat, dq, to system at constant pressure p0 and temperature T(t) causes the system volume to increase by an amount dv.

Let us consider a second example which also serves to illustrate the concept of enthalpy and which will also be useful in examining the performance of heat engines which use combustion processes (see Figure below). Suppose that we have a unit mass of a substance that has an internal energy e at some time t_1. We then consider a process such that $e_1 \rightarrow e_2$ as time goes from $t_1 \rightarrow t_2$. Without loss of generality we can take $e_1 > e_2$. Furthermore, thinking ahead to consider an idealized combustion process, we shall assume that this process occurs at p=const, v=const. Applying the first law to this system, we find after a few manipulations that $dq = dh$. However, in this case we now find that the system must reject heat to the environment; the amount of heat rejected is precisely equal to the change in enthalpy which, since p=const and v=const, is equal to the change in internal energy of the system. This finding will also be useful in our analysis of combustion-based heat engines.

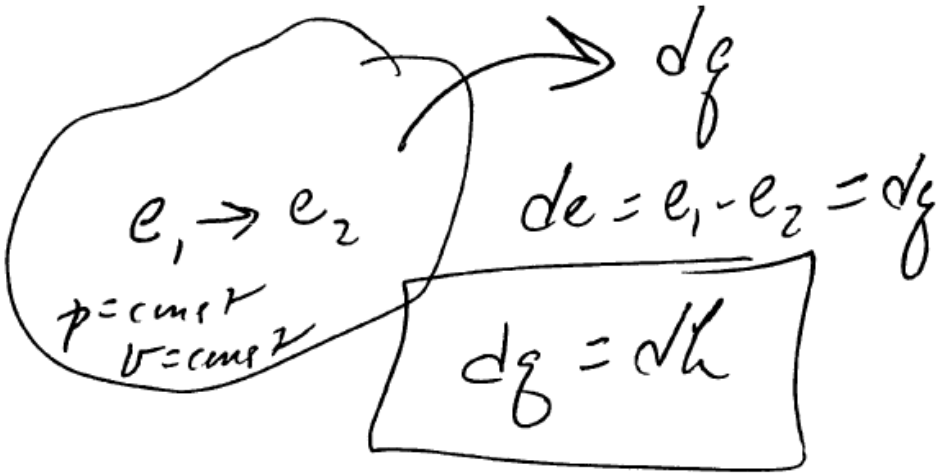


Figure 4.4: Schematic of system undergoing a change in internal energy at fixed pressure and specific volume.

The enthalphy can also be used to define several other quantities that are useful in the analysis of heat engines. For example, the specific heat at constant pressure, C_p, is defined as

$$C_p \equiv \left.\frac{\partial h}{\partial T}\right|_{p=const}$$

while the specific heat at constant specific volume is defined in terms of the change of internal energy as

$$C_v \equiv \left.\frac{\partial e}{\partial T}\right|_{v=const}.$$

The Gibbs free energy, f, is defined as

$$f = h - Ts = e + pv - Ts$$

This quantity proves useful particularly in the analysis of electrochemical devices such as fuel cells and batteries.

For processes with T=const, p=const, the second law of thermodynamics can be written as

$$dW \le df,$$

i.e. the amount of work done by such a process cannot exceed the decrease in Gibbs free energy. We can also relate these quantities for reversible processes. To see this first we use the definition of entropy to write

$$dq = Tds$$

which using the first law can be re-written for a constant pressure process as

$$Tds = de + pdv.$$

Using the definition of enthalpy this is equivalent to

$$Tds = dh - vdp.$$

Now, since f=h-Ts=e+pv-Ts we can then write the incremental change in Gibbs free energy as

$$df = de + d(pv) - d(Ts).$$

Using this result we can then write

$$Tds = dh + sdT - df.$$

We shall use these concepts in the evaluation of both idealized and real heat engine cycles in thefollowing sections.

Idealized Heat Engines

Let us now use these concepts to examine the performance of heat engines which convert heat into mechanical work. This topic is extremely important in so far as such devices form the vast majority of energy systems in use today. We will first examine two types of idealized engines, and then move on to consider several practical thermodynamic cycles that are the basis of the performance of real heat engines used e.g. for power generation (e.g. the steam cycle used in coal fired and nuclear power stations) and in transportation systems (e.g. automobiles and aircraft).

The Carnot Cycle

Let us now consider the simplest of the idealized heat engines, i.e. the Carnot cycle. We refer to the schematic shown in Fiogure 4.8 below to describe the key elements of this cycle. The T-s and p-v thermodynamic diagrams for the cycle are shown in Figures 4.8 and 4.9 below.

The engine consists of a hot reservoir held at a temperature Th. This reservoir transfers a quantity of heat, qh, at the hot reservoir temperature to the engine via an isothermal expansion, taking the system from state 1 to state 2. The system then undergoes an adiabatic (i.e. no heat

exchange) isentropic (no change in entropy) expansion, taking it from state 2 to state 3. During these two steps in the process, the pressure drops and specific volume increases, resulting in the exertion of mechanical work by the system on the environment. The cycle is then closed by rejecting heat to the cold reservoir at a fixed temperature T_C via an isothermal compression taking the system from state 3 to state 4, followed by an adiabatic isentropic compression taking the system from state 4 back to state 1.

For this process we can express the first law as

$$\oint dQ = \oint dW \Rightarrow q_h - q_c = w = \oint pdv$$

Furthermore since $\oint dS = 0$ we can then relate the change in entropy going from state 1-> 2 in terms, of the change in entropy from state 3 to state 4 as

$$\Delta s\big|_1^2 = -\Delta s\big|_3^4.$$

By assumption, the heat exchange steps occur at fixed temperature. Thus, we can use the definition of entropy to write

$$\int_1^2 dS = \int_1^2 \frac{dQ}{T_h} \Rightarrow \Delta s\big|_1^2 = s_2 - s_1 = \frac{q_h}{T_h}$$

and similarly we can write

$$\Delta s\big|_3^4 = -\frac{q_C}{T_C}$$

where the minus sign occurs due to the convention that a positive heat exchange event denotes an addition of heat from the environment to the system.

Using these results we then have

$$\frac{q_h}{T_h} = \frac{q_c}{T_c}.$$

If we then define the efficiency of the engine as the ratio of useful work to the heat input from the hot reservoir, we then can easily show that

$$\eta \equiv \frac{w}{q_h} = \frac{q_h - q_c}{q_h} = 1 - \frac{T_c}{T_h}.$$

this is the usual Carnot efficiency result, and indicates that in order to maximize the conversion efficiency for a fixed cold reservoir temperature, one then wishes to maximize the hot reservoir temperature. The question then becomes: what is this maximum temperature?

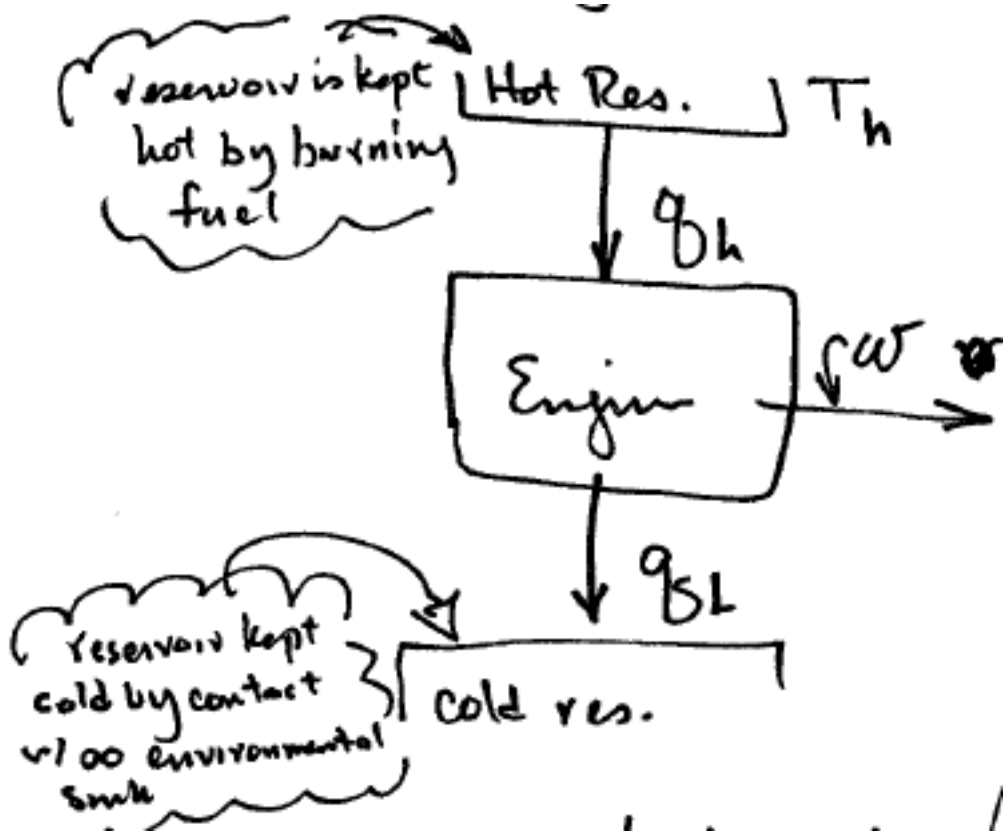


Figure 4.8: Schematic of idealized Carnot heat engine.

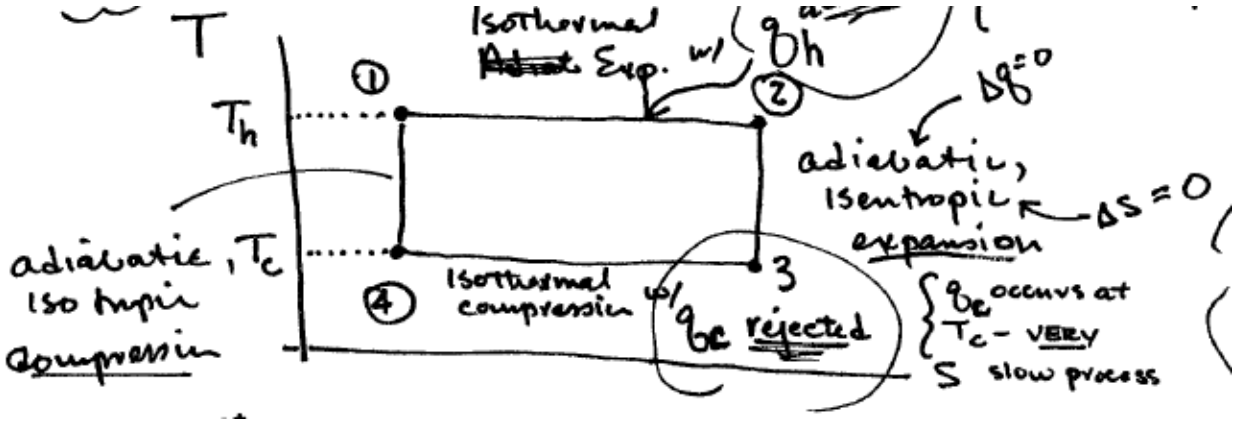


Figure 4.9: T-s thermodynamic diagram of idealized Carnot heat engine cycle.

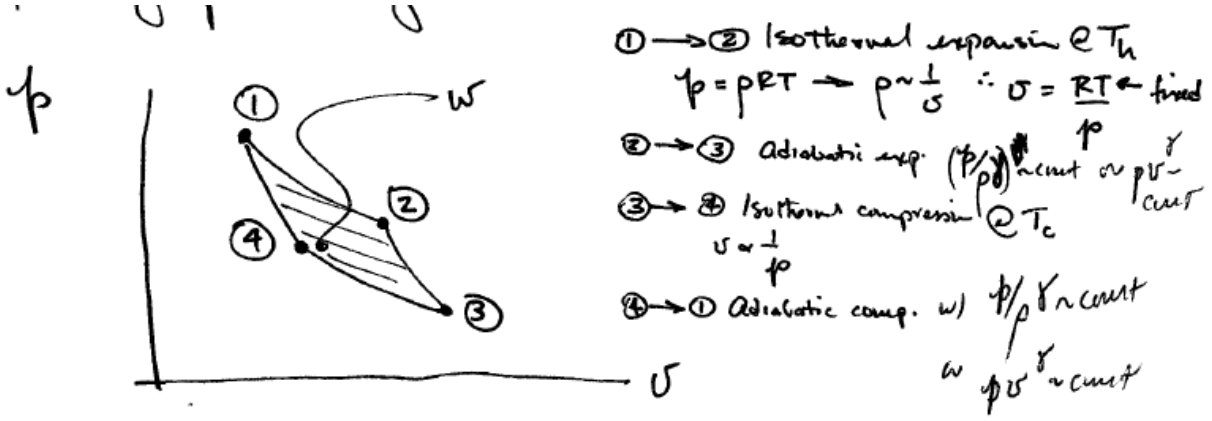


Figure 4.10: p-v thermodynamic diagram of idealized Carnot heat engine cycle.

Idealized heat engine with realistic constraints from combustion processes

Consider the idealized heat engine shown schematically in below. A fluid with steady-state mass flow rate $\dot{m}$ moves through the engine. The rate of heat input into the engine is given as $\dot{Q}$. The engine converts this heat input into work, and then performs work at a rate $\dot{W}$ on the outside environment. Applying the first law to the system boundary which operates at fixed pressure, and using the definition of enthalpy we can write that

$$\dot{m}\left(h_{out}-h_{in}\right)=\dot{Q}-\dot{W}\,.$$

Thus, without examining any of the details of how the engine works, but instead simply examining the heat and enthalpy of the products crossing the system boundary we can then determine the rate of work (i.e. the power) of the engine. This illustrates the usefulness of the foregoing concepts from thermodynamics. We then need to consider in more detail the heat transfer and combustion processes by which chemical potential energy is converted into thermal (i.e. heat) energy of a material medium. Thus we consider two key components in most heat engines: a) the heat transfer stage in which heat is exchanged between the system and the environment, and b) combustion process in which the energy stored within chemical bonds is released and converted into the thermal energy of a working fluid in the engine.

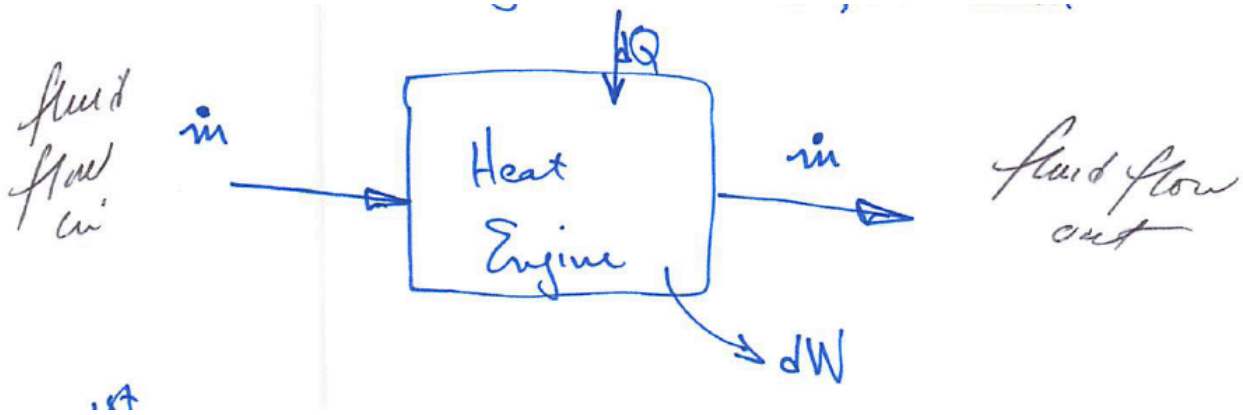


Figure 4.5: Schematic of a heat engine operating with a steady state fluid flow, heat input and work output.

In a heat engine, a heat exchanger is used to transfer heat between a working fluid at a temperature T_h and the environment which is at temperature T_c. Figure 4.6 shows a schematic representation of a surface of this heat exchanger. From studies of convective heat transfer, we know that this heat transfer rate $\dot{Q}$ is given by

$$\dot{Q} = \kappa A(T_h - T_c)$$

where A is the cross-sectional area of the heat exchanger and κ is the heat transfer coefficient which is a function of the geometry of the device, the conditions of the fluid flowing through the device, and so forth. Clearly from this general relation, if we wish to keep the heat transfer rate $\dot{Q}$ constant while decreasing the area A, then the temperature differential must increase or we must increase κ which in general requires a higher fluid flow rate through the system (and thus a larger pump which in turn requires more work to operate). Thus, there are clearly tradeoffs to be made in optimizing the performance of an engine that contains such a component (which includes nearly all heat engines in use around the world).

Next, let us consider the processes by which the stored chemical energy continued in fuels is released via a combustion process. To do so, we first recall a few elementary concepts from chemistry. Fossil fuels are composed primary of hydrocarbon molecules which react with oxygen in the atmosphere to form water vapor and carbon dioxide reaction products. We can write this reaction as

$$C_n H_m + \left(n + \frac{m}{4}\right) O_2 \rightarrow nCO_2 + \frac{m}{2} H_2O$$

Here we have taken m,n to be positive integers and require that the number of atoms of each species be conserved in the reaction (as is always the case in chemical reactions). Note that

for coal, n~m~1, while for natural gas (i.e. methane) we have n=1 and m=4. In general, higher values of m/n correspond to hydrogen-rich fuels which implies a higher chemical potential energy per unit reactant mass. It also, of course, implies a lower carbon content per unit of stored energy – an important implication for global climate change.

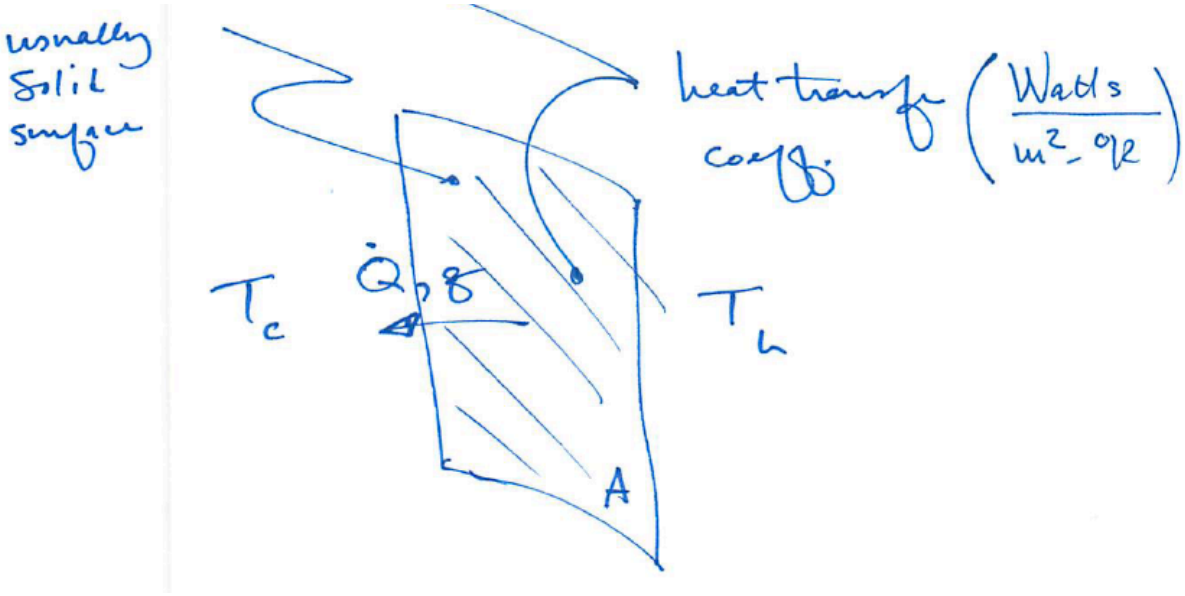


Figure 4.6 Schematic of a surface of area A separating two regions that are held at temperature $T_h > T_c$. A heat flux, q, moves across the surface and is proportional to the difference between the two temperatures.

Now consider an idealized combustion process; Figure 4.7 below shows a schematic of this process in an idealized perfectly insulating combustion chamber. Fuel and oxygen are introduced into the combustion chamber with mass flow rates $\dot{m}_{Fuel}$ and $\dot{m}_{O_2}$ as shown on the left of the schematic. These reactants are introduced at a temperature T_r and pressure p_r. The walls

of the chamber are taken to be perfectly insulating, and as a result there is no heat loss through the walls. Thus, $\dot{Q} = 0$ as shown.

We then take the combustion process to occur at the inlet pressure, i.e. $p_{comb} = p_r$. Finally, after the reaction occurs, the products are exhausted at a mass flow rate $\dot{m}_{ex}$ and a pressure equal to the combustion and inlet pressure, i.e. $p_{ex} = p_{comb} = p_r = const$. However, because the stored energy of the fuel and oxygen has been released and converted into thermal energy, we have an exhaust temperature $T_{ex} > T_r$ Given this idealized process, let us now consider the enthalpy through the process. From the first law we have

$$de = dq - pdv$$

which for isobaric processes gives

$$\begin{aligned} dq &= de + d(pv) \\ &= d(e + pv) \end{aligned}$$

or, using the definition of enthalpy, gives dq=dh. However, by assumption we have dq=0 and thus, as a result, we also have dh=0. Thus, the enthalpy is constant through this process. This finding is important, as it allows us to easily find the energy content of fuels by measuring the change in temperature of the mass flow through an idealized combustion chamber.

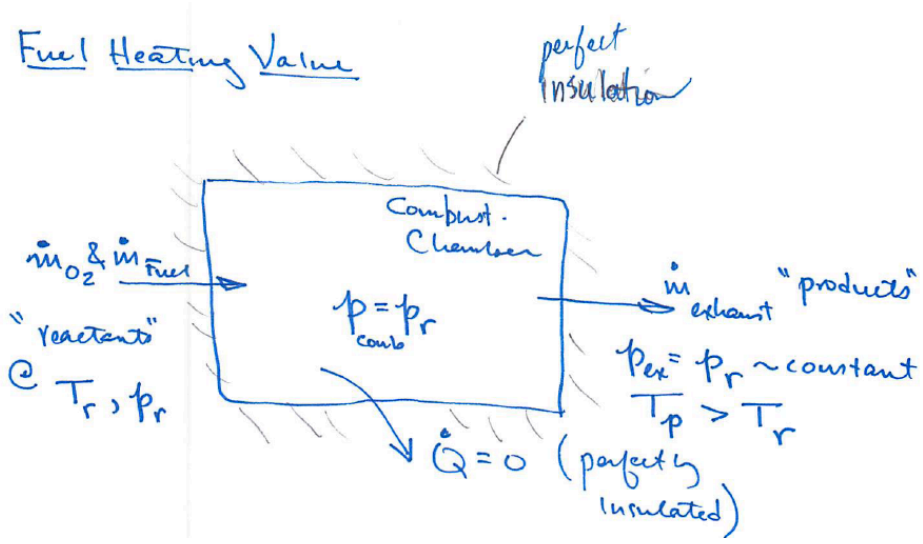


Figure 4.7 Schematic of an idealized combustion chamber in which combustion process occurs at fixed pressure. The chamber is considered to be perfectly insulated from the surrounding environment. Thus all energy release from combustion is captured by the change in the reactant and product temperatures exiting the region.

To see how this works, consider the idealized combustion process in the h-T thermodynamic space as shown in Figure 4.8 below. We draw two curves, $h_r(T)$ and $h_p(T)$ to denote the reactant and product enthalpies respectively. Furthermore, because the internal energy of the reactants is higher than the internal energy of the products, then for fixed T we have $h_r > h_p$. The reactants are then introduced at point 1 as shown on the diagram. The combustion process occurs during which there is no change in enthalpy (since in this step the internal energy is simply converted to thermal energy via an energy conserving process). Thus, this step results in a movement of the state from point 1 to point 2 along a line at h=constant. The reaction

products are then formed at a temperature T_p as shown, and we have $h_r(p_r, T_r) = h_p(p_p = p_r, T_p)$ at state 2.

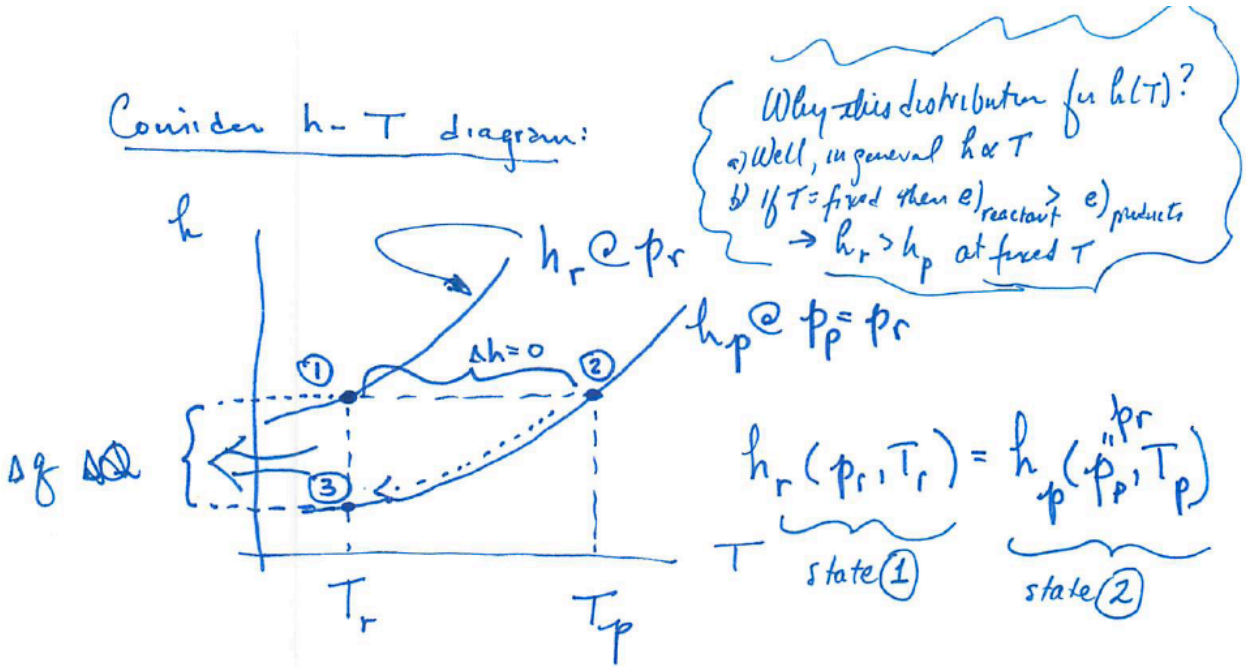


Figure 4.8 : h-T diagram of an idealized combustion process.

Now, consider what happens if we then capture these hot reaction products and cool them down until their temperature returns to the initial temperature of the reactants, i.e. until they reach a temperature T_r. This process will require the remove of some amount of heat from the products. Let us denote that quantity of heat as Δq. The products then arrive at state 3 as shown in the diagram in which their final temperature is equal to the temperature of the incoming reactant. From the first law, it follows that

$$\Delta q = h_p(T_p, p_r) - h_p(T_r, p_r)$$

Noting that $h_p(T_p, p_r) = h_r(T_r, p_r)$ we can then write the heat removed as

$$\Delta q = h_r(T_r, p_r) - h_p(T_r, p_r)$$

It is common to multiply this value by the ratio of product mass to fuel mass to yield the so-called Fuel Heating Value (FHV)

$$FHV|_{p_r,T_r} = \left(\frac{\dot{m}_f + \dot{m}_{O_2}}{\dot{m}_f}\right)\left(h_r(T_r, p_r) - h_p(T_r, p_r)\right).$$

This quantity provides the thermal energy content available per unit mass of fuel assuming a stochiometric idealized combustion process. For typical hydrocarbon fuels, the FHV~30-40 MJ/kg; For pure hydrogen, FHV~120 MJ/kg, and for lower quality coals and biomass, FHV~10-20 MJ/kg. These values assume that the water molecules contained in the exhaust products have not been condensed, which would then result in the release of the heat of vaporization of the water vapor contained in the combustion byproducts. In practice such condensation seldom occurs and thus the FHV expression given here is the most appropriate one for our purposes.

The specific enthalpy difference of the reactants and products denotes the energy available to heat the combustion products. Thus, we can find the maximum achievable combustion temperature, known as the adiabatic combustion temperature, T_{ad}, as

$$T_{ad} = T_r + C_p\Big|_{p_r,T_r}\left(h_r(p_r, T_r) - h_p(p_r, T_r)\right).$$

Typically for most hydrocarbon fuels, T_{ad}~1900 C (it is somewhat higher for pure hydrogen). This temperature is the maximum value achievable in a combustion process (actually in a realistic system the maximum temperature is somewhat less than this value) and thus it represents a maximum upper limit to the high temperature reservoir of the heat engine. When

we take up consideration of global climate change, it is useful to consider the amount of carbon contained per unit of stored chemical energy. This quantity, known as the carbon intensity, can be written as $C = \frac{m_f^C}{h_r - h_p}$. Fuels such as coal have a high carbon intensity while fuels such as methane have a somewhat lower carbon intensity. Thus, all other things being equal, switching from a high carbon intensity fuel to a lower carbon intensity fuel can reduce the quantity of CO_2 produced for a unit of energy release.

Idealized Heat Engine with an External Combustion-based Heat Source

Let us now use some of these concepts in the analysis of an idealized heat engine, now taking into account this idealized combustion process as the heat source. In the discussion above, we found that the highest combustion temperature possible was given by the so-called adiabatic temperature, T_{ad}, which was set by the stored energy density of the fuel and the heat capacity of the combustion by-products. Furthermore, the combustion process was taken to occur at fixed pressure, while each state in the cycle discussed above involves a change in pressure. In order to consider using such a combustion process to add heat to this idealized heat engine, clearly the combustion process must then occur outside of the heat engine. These considerations can also be applied to other heat source (e.g. a nuclear reactor heat source) which also adds heat to the working fluid at essentially a fixed pressure. Thus, the schematic of the heat engine, including the external heat source, would be as shown in Figure below.

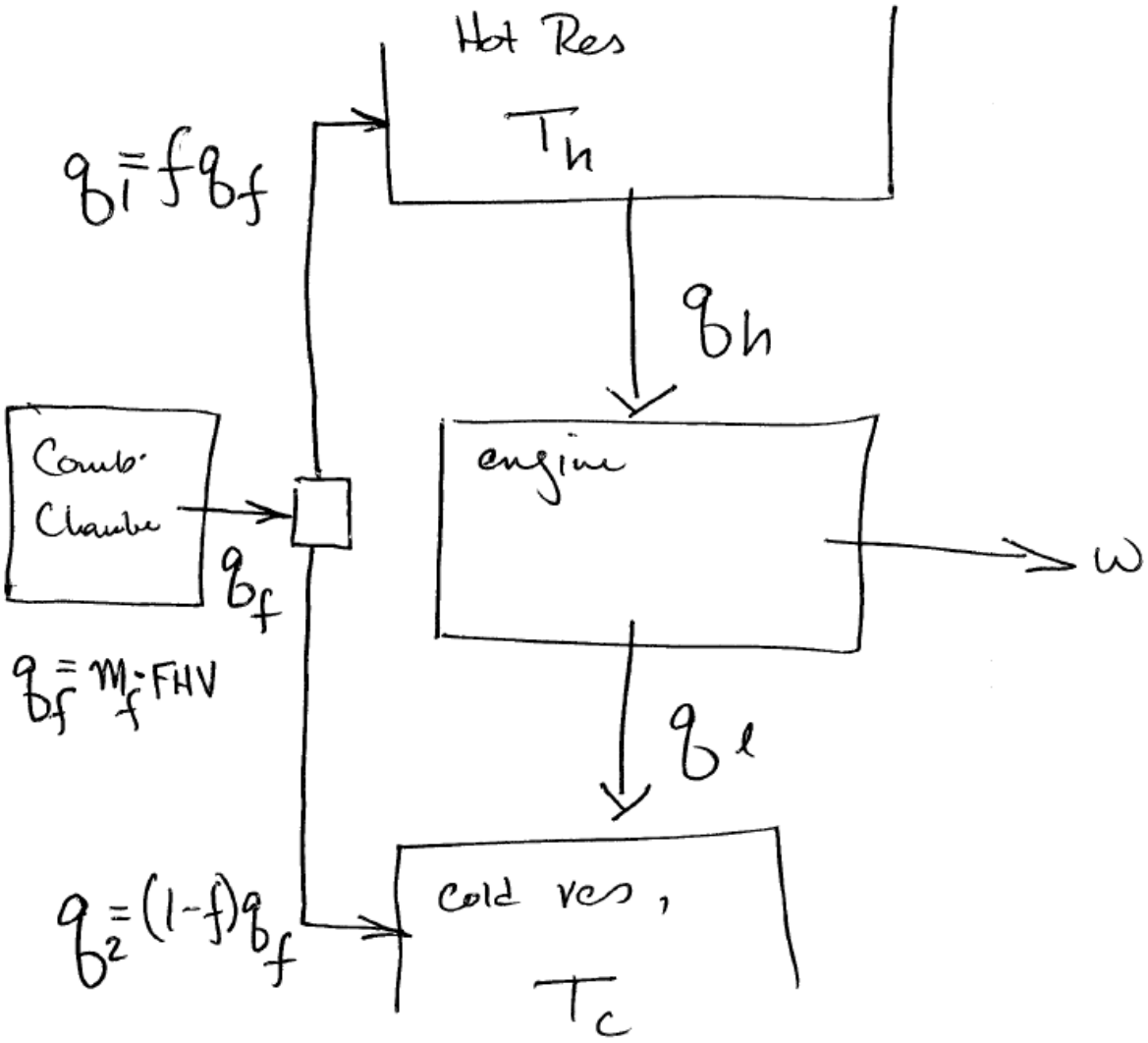


Figure 4.11: Idealized heat engine with external heat source.

Because it is located outside of the heat engine, the heat source is also in thermal contact with the environment and thus will only transfer some of its heat to the hot reservoir. Let us denote this fraction as f. Then the fraction 1-f of the heat will be transferred to the cold reservoir which is essentially at the same temperature as the environment. If the heat transfer between the

heat source and these two reservoirs occurs via convective heat transfer mechanisms (which will nearly always be the case), then the heat transfer rate is proportional to the temperature difference, and we can thus write

$$q_1 \propto \left(T_{ad} - T_h\right)$$

and

$$q_2 \propto \left(T_{ad} - T_c\right)$$

Furthermore we can write

$$q_f = q_1 + q_2,$$

$$q_1 = fq_f$$

and

$$q_2 = (1-f)q_f.$$

The usual definition of conversion efficiency is written as

$$\eta = \frac{w}{q_f}$$

Where *w* denotes the work done by the engine during the consumption of a unit mass of fuel containing fuel heating value q_f. This expression can be rewritten as

$$\eta = \frac{w}{q_h}\frac{q_h}{q_f}$$
$$= \eta_{Carnot} f$$

where we have used the results from the Carnot efficiency found earlier.

Let us now consider the implications of this simple result. If the two heat transfer steps occur via the transfer of heat via convective heat transfer through a heat exchanger of some sort, then we can write

$$\dot{q}_1 = k_1\left(T_{ad} - T_h\right)$$

and

$$\dot{q}_2 = k_2\left(T_{ad} - T_c\right)$$

Where k_1 and k_2 denote the heat transfer coefficients for these two processes. Now from the definitions given above we can write

$$f = \frac{q_1}{q_f} = \frac{q_1}{q_1 + q_2}.$$

This can be written as

$$f = \frac{k_1\left(T_{ad} - T_h\right)}{k_1\left(T_{ad} - T_h\right) + k_2\left(T_{ad} - T_c\right)}.$$

Thus the efficiency can be written as

$$\eta = \left(1 - \frac{T_c}{T_h}\right)\frac{k_1\left(T_{ad} - T_h\right)}{k_1\left(T_{ad} - T_h\right) + k_2\left(T_{ad} - T_c\right)}$$

and we require that $T_{ad} \geq T_h \geq T_c$.

Let us now consider two limits. First, consider the case when the hot reservoir temperature is maximized such that $T_h \to T_{ad}$. In this case the Carnot efficiency is maximized but $f \to 0$ and thus the actual engine efficiency $\eta \to 0$. Second, consider the case where f is maximized. This occurs when $T_h \to T_c$, but in this limit the efficiency again vanishes. Between these two limits the efficiency has a finite non-zero value. Thus, we conclude that for a given

adiabatic fuel temperature there must exist an optimum hot and cold reservoir temperature which maximizes the overall conversion efficiency. We can find a simple estimate for this optimum condition. To do so, let us assume that $(T_{ad} - T_h) << (T_{ad} - T_c)$ and for simplicity further assume that $k_1 \approx k_2 \approx k$. In this case, we can write

$$f \approx \frac{(T_{ad} - T_h)}{(T_{ad} - T_c)}$$

and thus can write the efficiency as

$$\eta \approx \left(1 - \frac{T_c}{T_h}\right)\frac{(T_{ad} - T_h)}{(T_{ad} - T_c)}.$$

It can be shown that when $T_h = \sqrt{T_{ad} T_c}$ the efficiency has a maximum value given as

$$\eta\big|_{max} \approx \frac{\left(\sqrt{T_{ad}/T_c} - 1\right)}{\left(\sqrt{T_{ad}/T_c} + 1\right)}.$$

Thus, for combustion of coal, with T_{ad}~2200 deg K and T_C= 300 deg K, we have an optimum hot reservoir temperature $T_h = \sqrt{T_{ad} T_c} \approx 800\ deg K$ and a maximum efficiency of about 45%.

Practical Heat Engine Cycles

The discussion above provided a discussion of idealized heat engines, irrespective of the details of the operation of the engine. In this section, we provide a short summary of several common heat engines. Taken together, these engines provide the large majority of useful energy in the world today. Thus, it is important to have an understanding of the basic principles and limitations of their operation.

Rankine Cycle

A schematic of an engine employing the Rankine cycle is shown in below. Such an engine has a pump which takes a liquid working fluid (step 1) and raises its pressure and exhausts the liquid (step 2) into an external heat source (step 3). An amount of heat q_1 is then added to the fluid either from combustion or from a source such as a fission reactor at constant pressure p_b; the liquid undergoes a phase change during this process, and then leaves the heating region as superheated vapor (step 4). This vapor is then sent to a turbine at a pressure equal to p_b (step 5) where it undergoes an ideal adiabatic expansion, converting some of the vapor energy content to mechanical work. The exhausted vapor (step 6) is then run through a heat exchanger where it is condensed. The excess heat of the vapor is rejected to the atmosphere, and the condensate is then pumped back to the heat source and the cycle is then repeated.

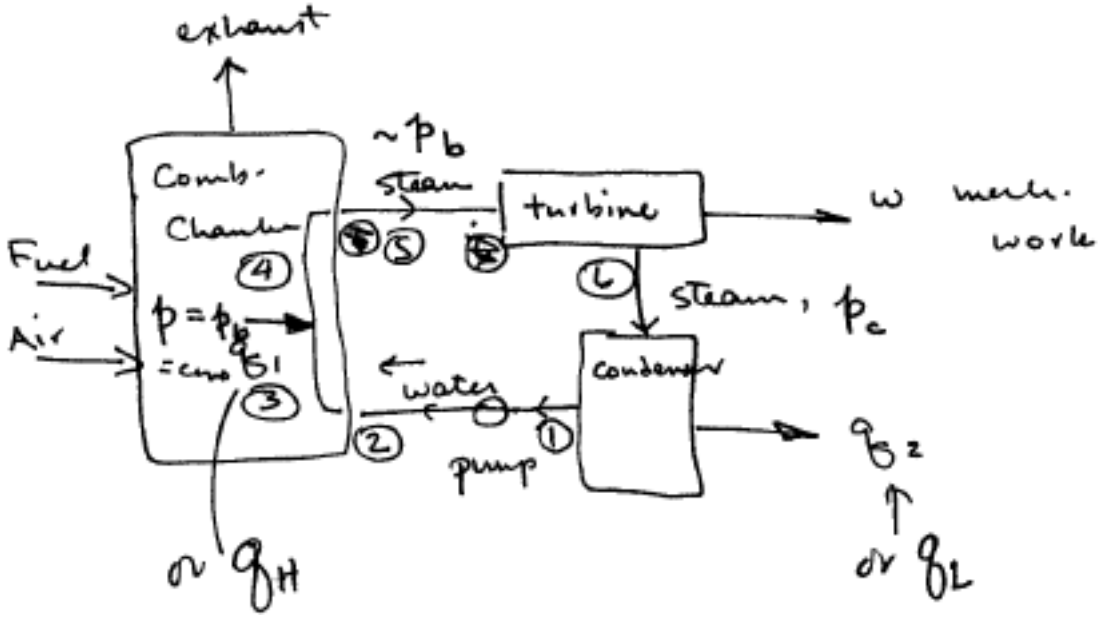


Figure 4.12: Schematic of a Rankine cycle engine identifying key steps in the cycle.

We can determine the overall functioning of such an engine by considering each component of the engine in turn using basic thermodynamics and referring to Figure 4.12 which shows a schematic of a Rankine cycle-based heat engine, and Figure 4.13 which shows the corresponding diagram of the engine state in the T-s space. We begin the analysis at the pump, step 1, where the working fluid has already been condensed back into liquid form. As such, we take the liquid to be incompressible and the pump provides an adiabatic isentropic process in which no heat addition occurs. From the first law, we can write the specific work done by the pump in terms of the change in enthalpy of the fluid moving through the pump

$$w_p = h_2 - h_1.$$

We can also write this work as

$$w_p = \int_1^2 v dp = v\left(p_2 - p_1\right)$$

Since the pump does not add any heat the working fluid, then from the second law we have

$$s_2 = s_1.$$

Thus we can write

$$\left(h_2 - h_1\right) = v\left(p_2 - p_1\right).$$

Next, let us examine the heat addition stage (e.g. the boiler in a coal fired power plant, the reactor core in a boiling water reactor, or the reactor core and primary heat exchanger together in a pressurized water reactor). No work is done on the fluid during this step, and thus w=0 for this stage. This is an idealized heat addition and thus from the first law we can write

$$q_h = h_5 - h_2.$$

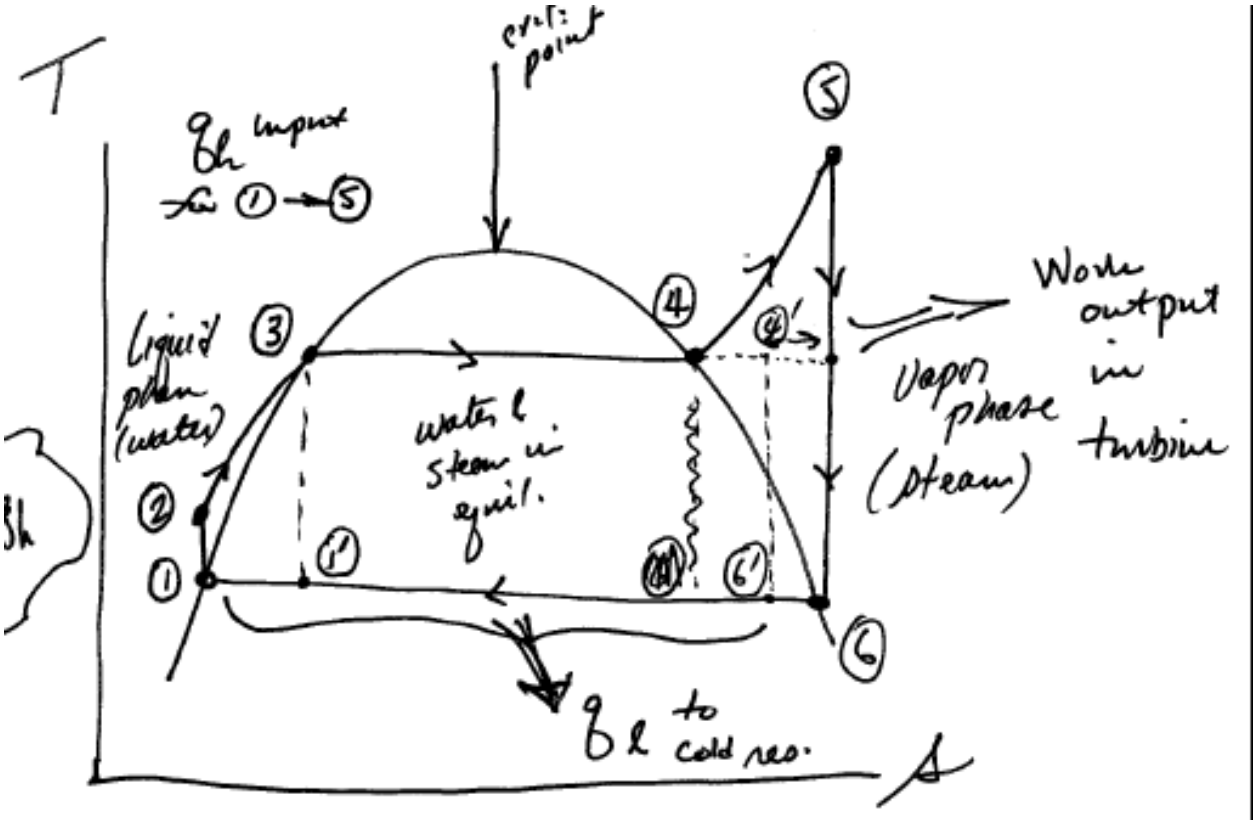


Figure 4.13 Thermodynamic cycle of a Rankine-cycle engine, plotted in T-s parameter space.

Next, we consider the turbine in which the fluid undergoes and isentropic adiabatic expansion. In this case we have from the first and second laws

$$w_t = h_5 - h_6$$

and

$$s_6 = s_5$$

respectively.

The condenser exerts no work on the fluid, and rejects an amount of heat to the cold reservoir given as

$$q_L = h_6 - h_1.$$

We can now use these results to determine the thermodynamic efficiency of such an engine. Writing as usual the efficiency as

$$\eta = \frac{w}{q_h} = \frac{q_h - q_l}{q_h}$$

and using the above results we can write

$$\eta = \frac{(h_5 - h_2) - (h_6 - h_1)}{(h_5 - h_2)}$$

which can be re-arranged to write

$$\eta = \frac{(h_5 - h_6) - (h_2 - h_1)}{(h_5 - h_2)}.$$

The first term in the numerator represents the change in enthalpy of the fluid as it passes through the turbine, while the second term in the numerator represents the enthalpy change in the fluid as it moves through the pump. The denominator represents the enthalpy change in the heat addition stage. It is difficult to design and build a pump that can handle vapor and liquid phases simultaneously. Furthermore, to avoid catastrophic damage to the turbine stage, it is important to avoid vapor condensation within the turbine. Now in Figure 4.13, the curve from states 1-5 shows the change in fluid state in the T-s diagram. In general, the working fluid will have a function $h=h(p)$ for a given T and s such that an increase in p will give an increase in h. Now, to maximize the efficiency of such a system we would wish to minimize the enthalpy at state 6 and maximize it at state 5. However, the corresponding temperature at state 5 is limited by the heat

source temperature (e.g. the adiabatic temperature of a combustion process, or the maximum operating temperature of a fission reactor). Furthermore, to avoid turbine damage, the temperature at state 6 cannot be below that shown in the figure; otherwise the working fluid will begin to condense back into liquid form within the turbine (which would then be destroyed by the liquid droplets). Likewise, the condensor operates at fixed T as the phase of the working fluid changes from 100% vapor to 100% liquid (states 6 to state 1), and the pump then operates on fluid which is in the liquid phase, taking the fluid from state 1 to state 2. Since the pressure is nearly constant from the turbine exhaust to the entrance to the pump (i.e. p~constant from state 6 to state 1) and pressure is nearly constant within the heat addition stage (i.e. p~const from state 2 to state 5). From these considerations, one can then see that the efficiency scales as $\eta_{Rankine} \propto \frac{P_{1-5}}{P_6} T_5$. Thus, a higher boiler pressure and higher maximum vapor temperature lead to an increase in conversion efficiency. The pressure p_6 corresponds to the point where saturated vapor is present.

Using thermodynamic property tables for the working fluid, and given the conditions at the exhaust of the boiler and the pressure at the condenser, one can then determine the efficiency of such an engine. For example, a Rankine engine operating with a steam boiler with a maximum steam temperature of 800 deg F and a gauge pressure of 40 atmospheres (600 psig) and a condensor operating at a gauge pressure of 1 psig the conversion efficiency will be ~32%. These are values close to what is found in a simple coal-fired power plant.

There are commonly used schemes to increase the efficiency of such engines. In one such scheme, known as the re-heated Rankine cycle (the schematic of such an engine is shown in

Figure 4.14 and the corresponding T-s diagram in Figure 4.15) the vapor is partially expanded through the turbine, and then sent back to the heat source where additional heat is added at fixed pressure. The re-heated vapor is then returned to the same point in the turbine where it undergoes further expansion. This Rankine cycle with re-heat can result in several percent increase (e.g. from ~32% to ~35-38%) conversion efficiency for the conditions of the example above. The analysis of this modified cycle is left an an exercise.

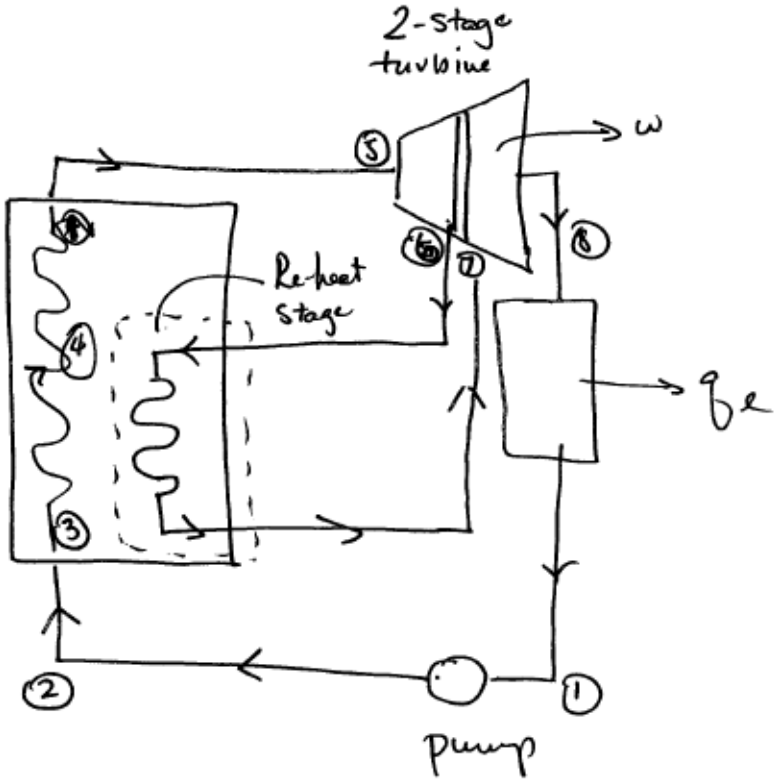


Figure 4.14 Schemetic of a Rankine cycle engine that employs a re-heating scheme to increase the overall thermal efficiency of the system.

With

$$\eta \equiv \frac{w_{net}}{q_h} = \frac{w - w_{pump}}{q_h}$$

$$= \frac{(h_5 - h_6) + (h_7 - h_8) - (h_2 - h_1)}{(h_5 - h_2) + (h_7 - h_6)}$$

Example: Use same values as from previous example:

- boiler pressure = 600 lb/in^2 800 °F) same as previous problem
- expands in turbine to 60 psi
- reheated to 800 °F
- expanded again in turbine to 1 psi) same as previous example

Can use steam tables ... find

$$\eta \simeq 38\%$$

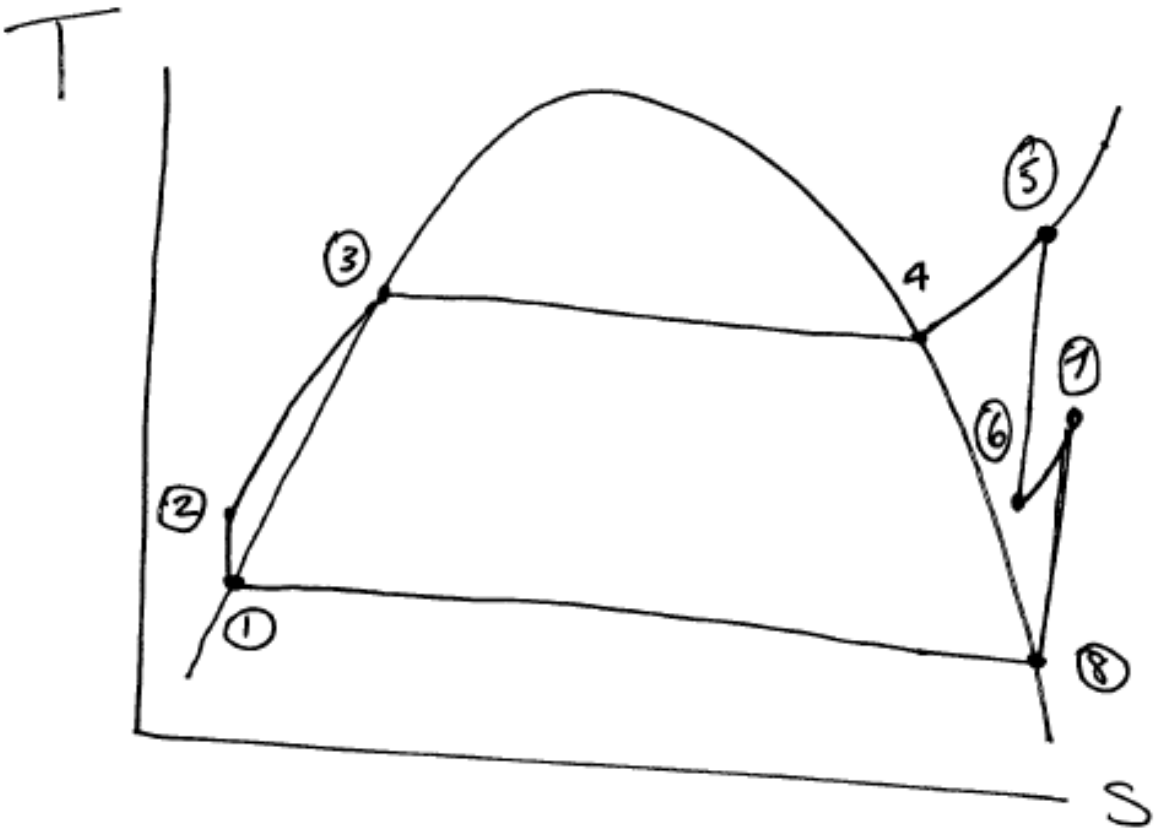


Figure 4.15 T-s thermodynamic diagram of a Rankine cycle engine that employs the reheating scheme to increase overall thermal efficiency.

Brayton Cycle

The Brayton cycle is the thermodynamic cycle used in the operation of gas turbines, which form the basis of jet engines and natural gas-fired power plants. This cycle has also been proposed for use in next-generation gas cooled nuclear fission power plants. Unlike the Rankine cycle discussed above, this cycle does not involve a phase change of the working fluid. A schematic view of such an engine, including a designation of different positions or states within the device,

and the corresponding T-s diagram are shown in

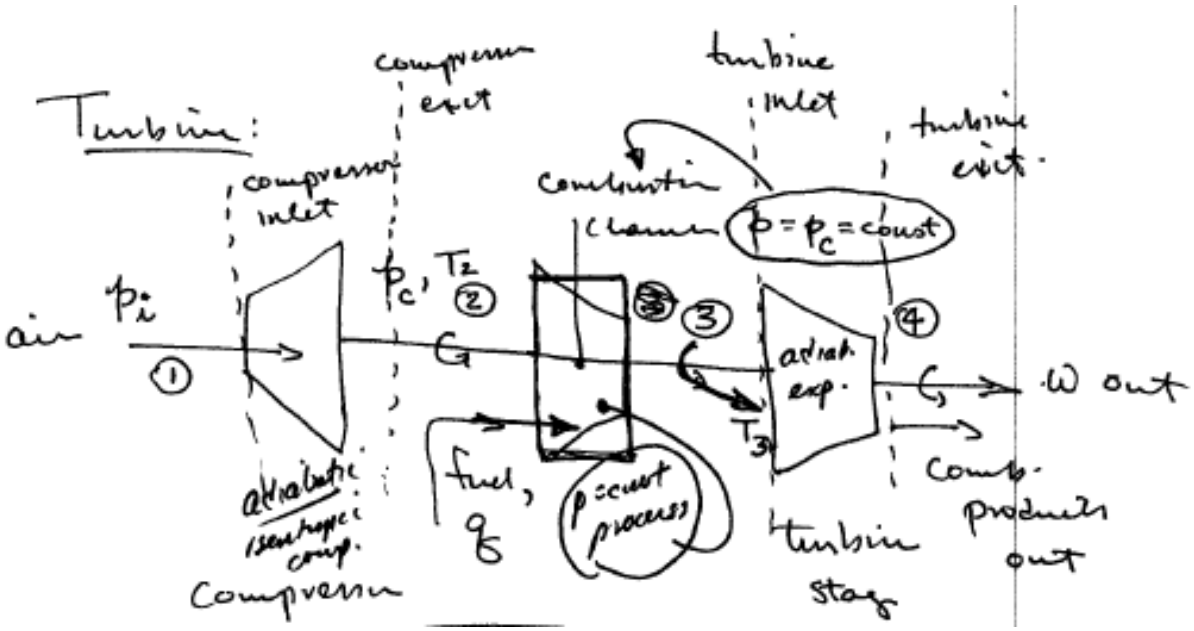


Figure and

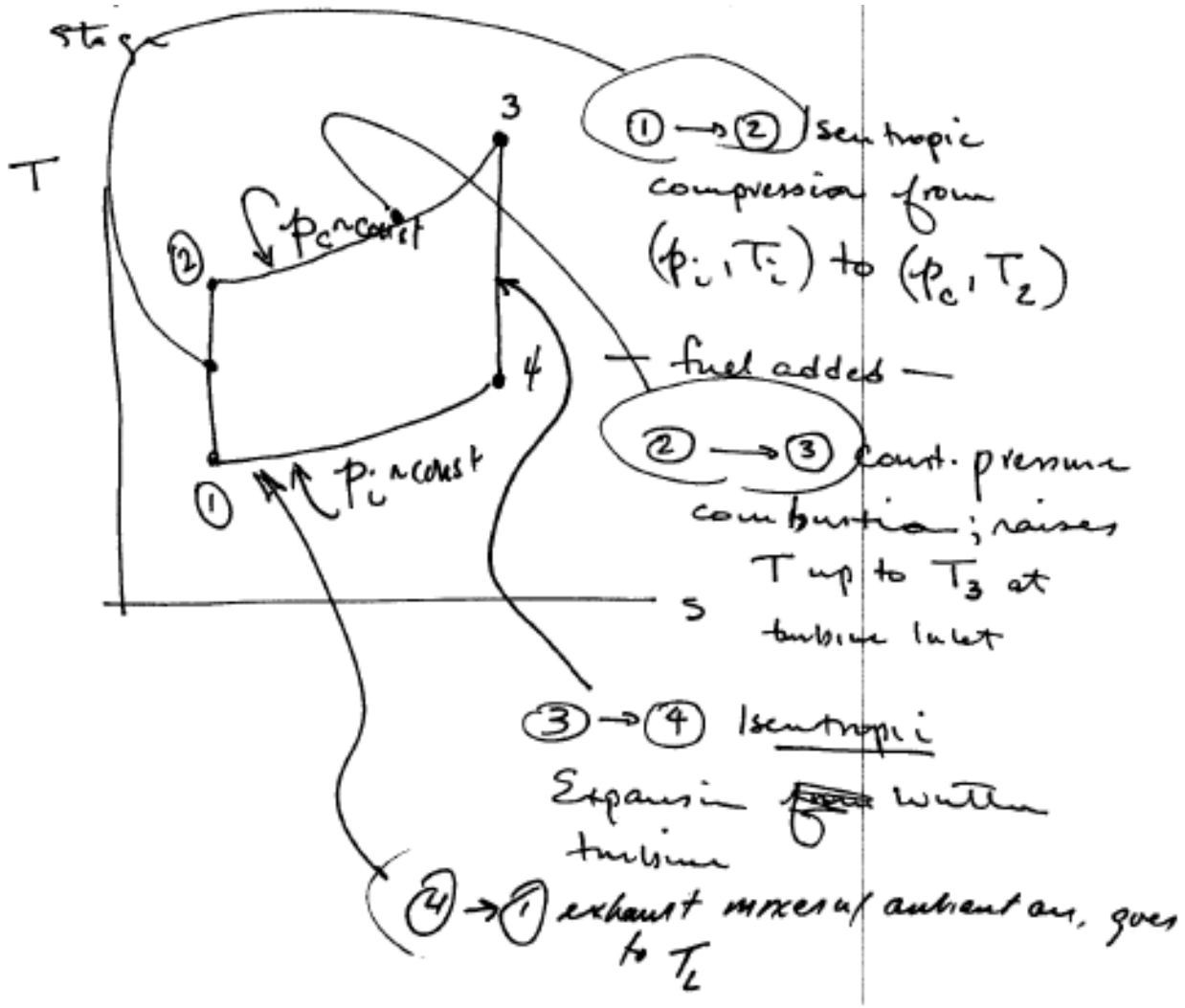


Figure below.

Working fluid (usually air) enters the engine at state 1 with an inlet pressure p_i and is ingested into a compressor, which performs an idealized adiabatic, isentropic compression on the fluid. The fluid at state 2 then enters an idealized constant pressure combustion stage where a quantity of heat, q, is added to the fluid. The resulting high pressure high temperature fluid at state 3 is then expanded through a turbine via an isentropic expansion. The turbine exhaust at

state 4 is then rejected to the atmosphere, where the excess heat is absorbed at the ambient temperature. Some of the work that is extracted by the turbine is returned to the compressor (usually the two are linked by a mechanical shaft which transmits torque from the turbine to the compressor), the remaining net work is then available for other useful purposes.

The analysis proceeds as follows. From our thermodynamic analysis above, we know that the work done by an isentropic process is related to the change in enthalpy by the relation $w|_{s=const} = \Delta h$. Since the net work available for useful purposes is given by

$$w_{net} = w_{turbine} - w_{comp}$$

we can then write

$$w_{net} = \Delta h_{turbine} - \Delta h_{comp}$$

or, writing the states explicitly

$$w_{net} = \left(h_3 - h_4\right) - \left(h_2 - h_1\right).$$

Note that here the turbine is doing work on the compressor which is reflected in the sign convention used here. For an idealized constant pressure combustion process that adds a quantity q of heat, we can write

$$q = h_3 - h_2$$

The efficiency of such an engine can then be written as

$$\eta = \frac{w_{net}}{q}$$

which, using the results above, can be written as

$$\eta = \frac{\left(h_3 - h_4\right) - \left(h_2 - h_1\right)}{\left(h_3 - h_2\right)}.$$

It is useful to re-arrange this expression to give

$$\eta = \frac{(h_3 - h_2) - (h_4 - h_1)}{(h_3 - h_2)}$$

or

$$\eta = 1 - \frac{(h_4 - h_1)}{(h_3 - h_2)}$$

For a perfect gas with constant specific heat, $C_p = const.$ we can then re-write this last expression as

$$\eta = 1 - \frac{(T_4 - T_1)}{(T_3 - T_2)}$$

which is equivalent to

$$\eta = 1 - \frac{T_1\left(T_4 / T_1 - 1\right)}{T_2\left(T_3 / T_2 - 1\right)}.$$

Now we note that for the combustion chamber section we can write $p_2 = p_3 = p_{comb} = const.$ and also since the turbine exhaust goes to the atmosphere we can write $p_4 = p_1$. Thus clearly the pressure satisfies the condition

$$\frac{p_3}{p_4} = \frac{p_2}{p_1}.$$

Noting that the compressor and turbine stages are isentropic thermodynamic processes then if the working fluid behaves as an ideal gas (which is a reasonably good approximation) we can relate the pressure and temperature via the expressions

$$\frac{p_2}{p_1} = \left(\frac{T_2}{T_1}\right)^{\gamma/\gamma - 1}$$

and

$$\frac{p_3}{p_4} = \left(\frac{T_3}{T_4}\right)^{\gamma/\gamma-1}.$$

Thus we can write

$$\frac{T_3}{T_4} = \frac{T_2}{T_1}$$

and also

$$\frac{T_3}{T_2} = \frac{T_4}{T_1}.$$

Thus we can write the efficiency in terms of the pressure ratio across the compressor

$$\eta = 1 - \frac{T_1}{T_2} = 1 - \frac{1}{\left(\frac{p_2}{p_1}\right)^{\gamma/\gamma-1}}$$

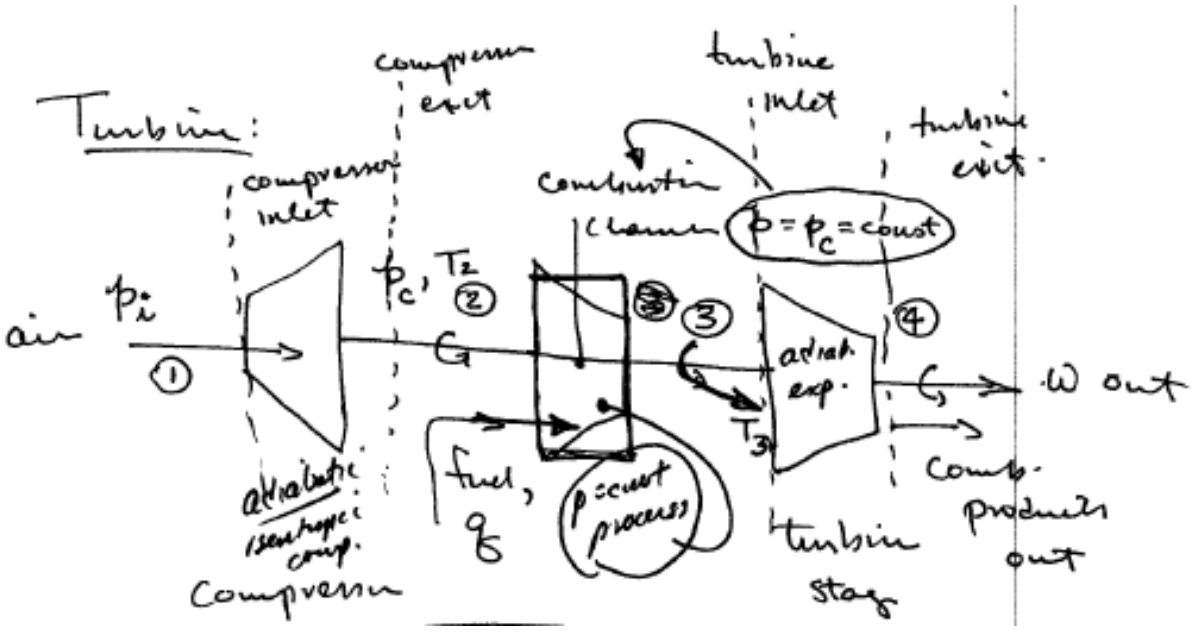


Figure 4.14: Schematic view of a Brayton cycle engine.

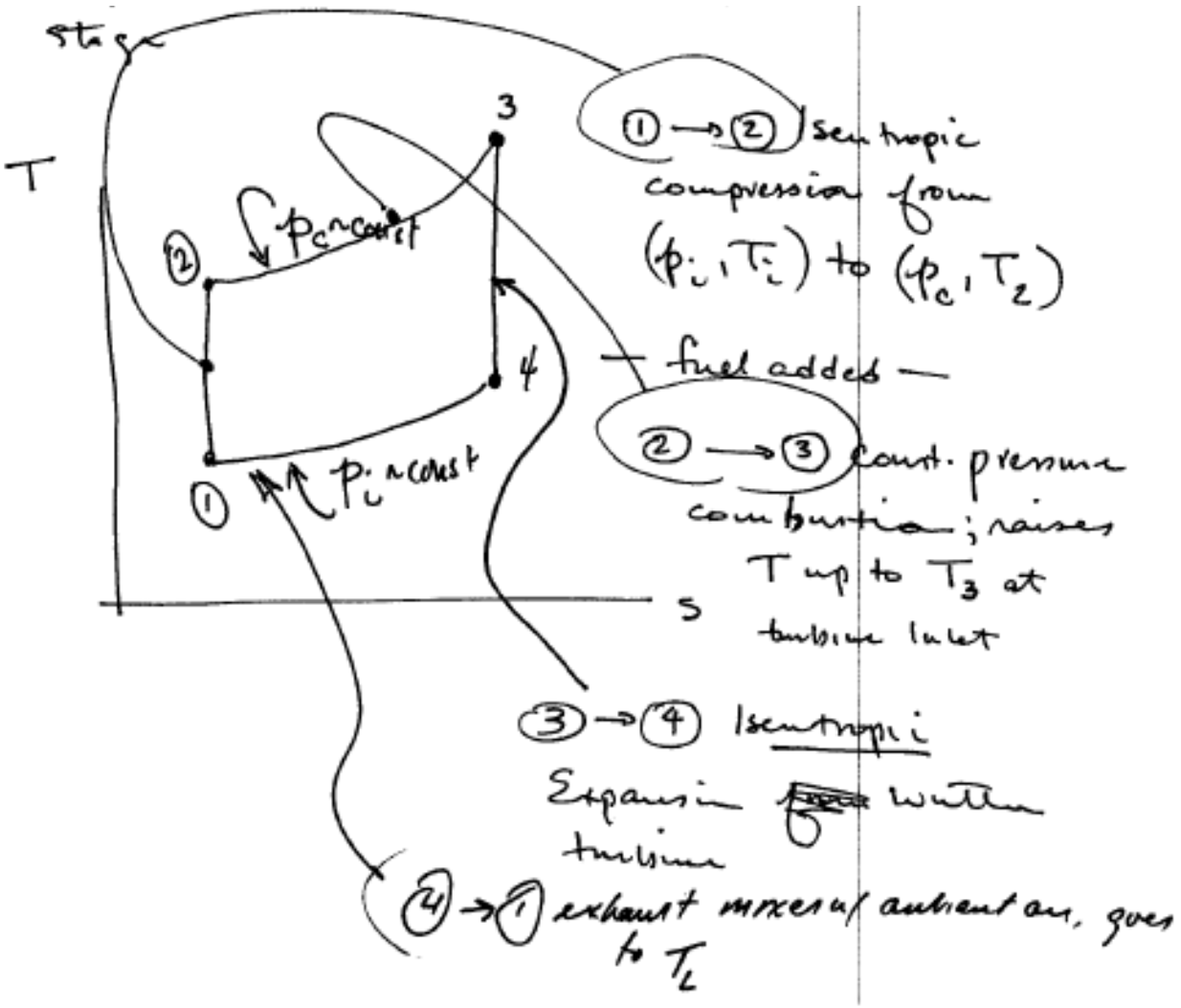


Figure 4.15: T-s diagram for the Brayton cycle, with states labeled according to their position within the engine schematic shown above.

We can now use these results to examine the performance of a typical gas turbine. Suppose the inlet pressure of a turbine p_1=1atm with T=20 deg C at the inlet. The pressure at the exit of the compressor p_2=5atm. These are typical values that might be found in an application. Since the compression is adiabatic, we can find from the ideal gas law that T_2/T_1~1.6. Materials

limits of the blades within the turbine section set an upper limit on the temperature that can be reasonably used in the combustion chamber. Currently this limit is in the range of 1200 deg K or so. The hot gas is then expanded in the turbine, which extracts some work and then exhausts the gas at roughly atmospheric pressure. For the values given here, the conversion efficiency will be approximately 36%.

The Combined Cycle Gas Turbine

The overall thermal efficiency of a heat engine can be increased from these values by combining the Brayton cycle-based gas turbine with a Rankine cycle-based steam cycle. In this way, conversion efficiencies approaching or even exceeding 50% can be achieved. A schematic of this approach is shown in Figure 4.16 below.

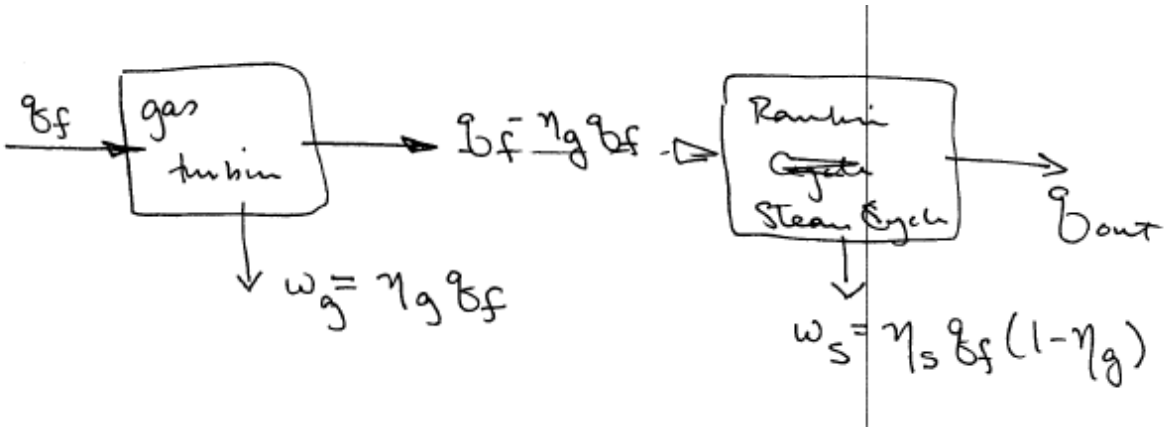


Figure 4.16: Schematic of a combined cycle heat engine that combines a Brayton-cycle based gas turbine with a Rankine-cycle based steam engine.

The analysis of this approach is fairly straightforward. Using the usual definition of thermal efficiency, we can write the overall thermal efficiency of the combined cycle as

$$\eta_{cc} = \frac{w_g + w_s}{q_f}.$$

We can then write the work done at the two different elements of the system as

$$w_g = \eta_g q_f$$
$$and \qquad .$$
$$w_s = \eta_s q_f (1 - \eta_g)$$

Combining these expressions then allows us to write the overall efficiency of such a system as

$$\eta_{cc} \sim 0.58 .$$

Thus for typical values of $\eta_g \sim 0.4$ and $+\eta_s \sim 0.3$ we find $. \eta_{cc} \sim 0.58$. Under some circumstances, the low grade heat exhausted by the Rankine cycle system can be used e.g. to heat steam for heating purposes, increasing the useful overall efficiency even further and extracting nearly all of the available energy from the chemical bonds of the fuel molecules. Thus by

NEED TO ADD DISCUSSION OF THE OTTO CYCLE FOR INTERNAL COMBUSTION ENGINES.

Chapter 5: Fundamentals of Climate Change

The combustion of fossil fuels results in the production of CO_2 which, as we will see below, can act to increase the absorption and trapping of infra-red radiation emitted from the Earth's surface due to heating from solar irradiation. Here, we develop a basic understanding of this phenomena with a view towards understanding the implications for future energy systems and sources. The aim here is not to provide a comprehensive introduction to the subject, but rather to sketch the essential element of the problem. We begin by considering distribution of the frequency and wavelength of light from a warm body which emits radiation in the form of light.

Blackbody Emission

A blackbody at a finite temperature emits a radiation intensity I that varies with frequency. The typical frequency-resolved intensity spectrum has the following shape:

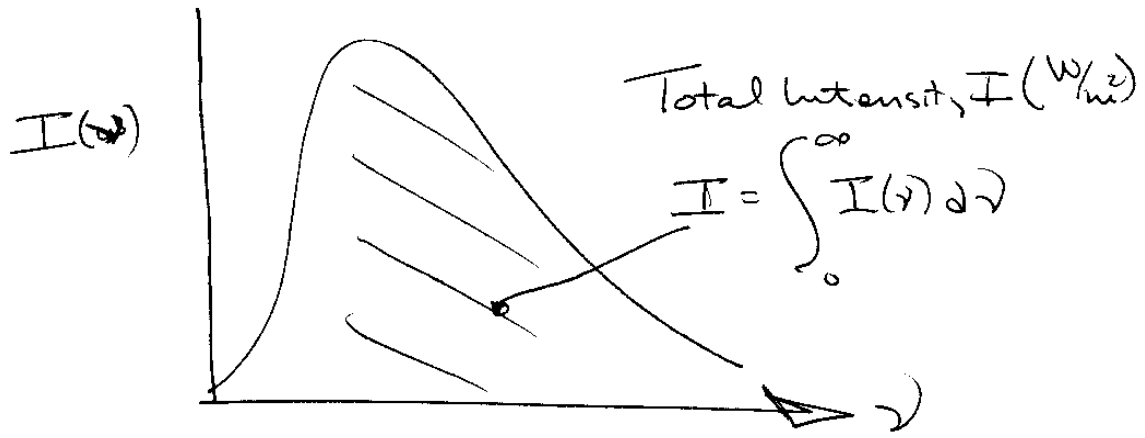


Figure 5.1: Schematic of the frequency distribution, $I(\omega)$, $I(\omega)$, of light emitted from a blackbody emitting radiation at a finite temperature, T.

Here the total intensity, I (W/m^2) is given by the integral of I(w) over all frequencies, i.e.

$$I = \int_0^\infty I(\nu) d\nu$$

For emission from a blackbody, quantum mechanics tells us that the frequency resolved emission intensity is given as

$$I(\nu) = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT} - 1}$$

where h ~ 6.6 x 10^{-34} J·s is Planck's constant (in MKS units), k ~ 1.38 x 10^{-23} J/°K is Boltzmann's Constant and T (in deg K) denotes the temperature of the emitting body.

Recalling the relation between frequency, wavelength, and the speed of light, we know that we can write $\nu\lambda = c$ which can be used to rewrite this expression in terms of the wavelength λ.

$$I(\lambda) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}} - 1}$$

For the climate change problem, we are interested in 2 different temperature ranges for the emitting bodies:

a) Sun's surface emission temperature, $T_e \approx T_e \approx 5800$ K

b) Earth Surface, $T_\oplus \approx T_\oplus \approx 300$ K

One can use these facts to show that the Sun's radiation intensity peaks at about $\lambda_e^{peak} \approx \lambda_e^{peak} \approx$ 500 nm, which corresponds to light particles (known as photons) with an energy of ~1-2eV,

while the Earth's radiation intensity peaks at $\lambda_{\oplus}^{peak} \approx \lambda_{\oplus}^{peak} \approx 10^4$ nm., which corresponds to a photon Energy ~ 0.03 eV. As we will see below, these two energy ranges play an important role in the basic physics of global climate change.

Radiation Transport in Gases

Next, let us develop a qualitative picture of light absorption by a molecular or atomic gas. Suppose we have a slab of gas, with a given density and some known thickness. A spectrum of radiation is incident upon the slab as shown in **Figure** below. The radiation moves through the slab, and some exits the slab. The incident radiation spectrum may look different than the exiting radiation spectrum due to absorption of radiation within the gas. This situation is shown schematically in Figure below. Radiation at discrete wavelengths is absorbed by the gas.

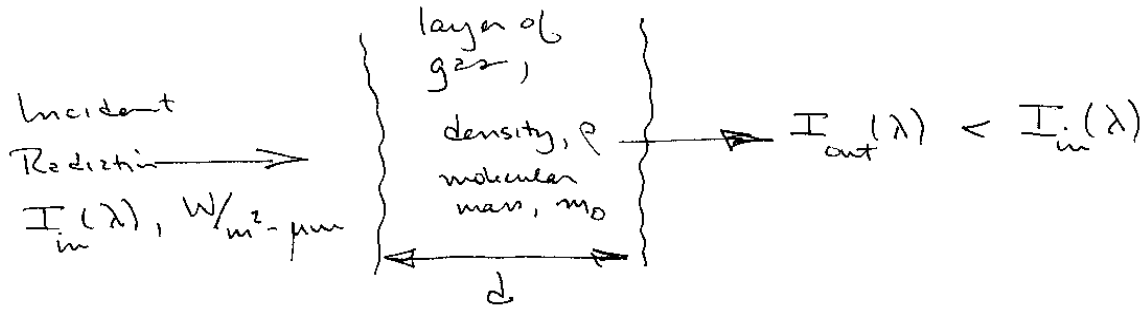


Figure 5.2: Schematic of incident and exiting radiation traversing a partially transparent slab of gas. The gas contains molecules or atoms that absorb radiation at selected wavelengths, and thereby modify the radiation spectrum.

Total incident radiation intensity, I_{tot} can then be written as the integral over the relevant wavelength range

$$I_{tot} \equiv \int_{\lambda_1}^{\lambda_2} I_{in}(\lambda) d\lambda .$$

A similar expression can be written for the radiation exiting the slab. Obviously, the total intensity of this un-interacted radiation could be lower than the incident total radiation intensity if there is partial absorption of the radiation by the gas.

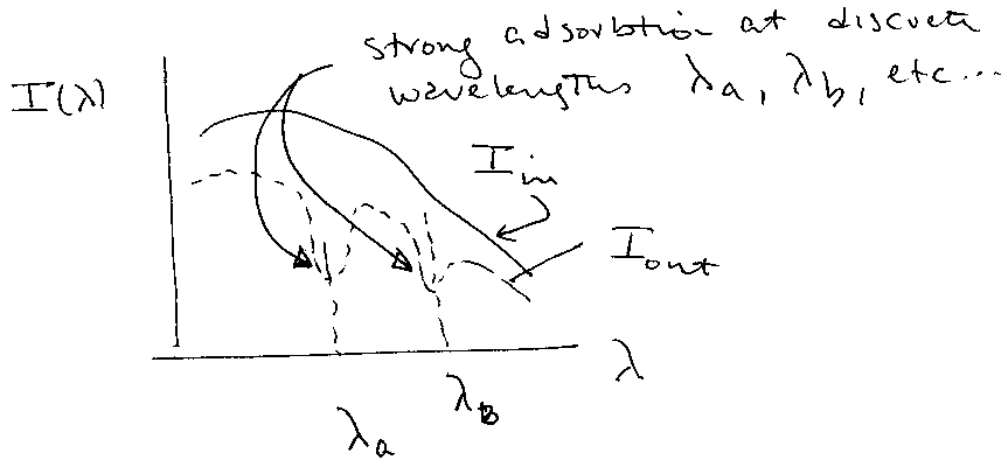


Figure 5.3: Schematic of incident blackbody spectrum, and spectrum of light exiting an absorbing medium. Absorption at λa, λb, etc… is due to resonant interactions between light waves and electrons in the molecules or atoms within the absorbing medium. (Put actual absorption spectrum for CO2 (and methane etc)?

We can write a (very) crude picture of the interaction of light with a molecule or atom by modeling these discrete particles using a classical harmonic oscillator as shown schematically in Figure below. We are essentially imagining a molecule as two vibrating masses m connected by a spring with constant k and damped at the rate v. If the two masses have a charge +q and –q and the oscillator is immersed within a fluctuating electric field, E, (which is related to the radiation intensity field I since from basic electromagnetic theory we know that $I(\omega) \propto E^2(\omega)$). $I(\omega) \propto E^2(\omega)$).

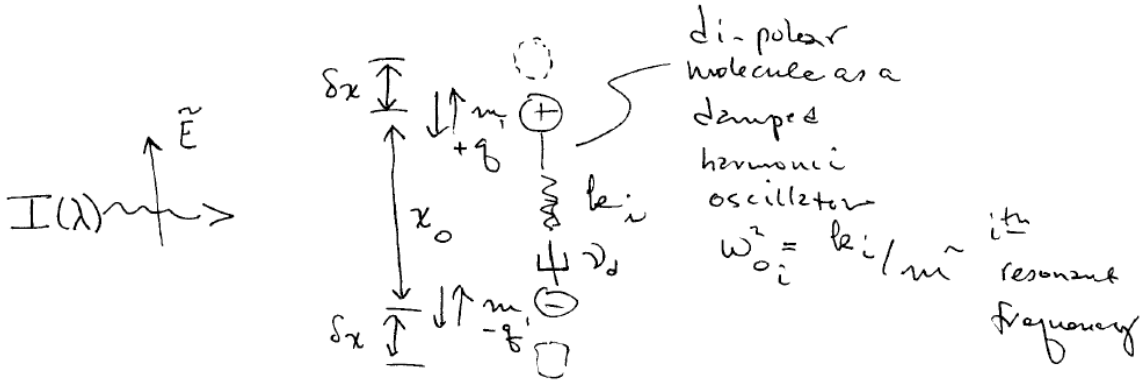


Figure 5.4: Schematic of a damped harmonic oscillator model of an atom or molecule undergoing excitation via irradiation from a light source.

Then the Force F on the dipole charges +q and –q is given as: $F = \pm q\tilde{E}(\omega) = \pm F(\omega)$. From elementary mechanics, we know that the equation of motion is given a $m\ddot{x} - vx + kx \doteq qE(w)$.

Suppose that all of the time-varying quantities behave as $e^{-\iota\omega t}$ – i.e. we expect oscillatory behaviors given by

$$x(t) = x_0 e^{-i\omega t},$$

$$F(t) = F_0 e^{-i\omega t}$$

and

$$E(t) = E_0 e^{-i\omega t}.$$

Then the time derivatives can be replaced by the frequency term, i.e. we can make the substitution $\frac{\partial}{\partial t} \to -i\omega \; \frac{\partial}{\partial t} \to -i\omega$ and the equation of motion then becomes:

$$\left[(-i\omega)^2 m x_0 - b(-i\omega)x_0 + kx_0\right] e^{-i\omega t} = qE_0 e^{-i\omega t}$$

We can now solve for the amplitude of the perturbation x_o

$$x_0\left\{-\omega^2 m + ib\omega + k\right\} = qE_0 \; x_0\left\{-\omega^2 m + ib\omega + k\right\} = qE_0$$

We can then re-arrange this expression to give the amplitude of the oscillation as

$$x_o = qE_o \frac{(\omega_o^2 - \omega^2) - i\nu\omega}{(\omega_o^2 - \omega^2)^2 + \omega^2\nu^2}$$

where $\nu \equiv b/m$ $\nu \equiv b/m$ and the undamped resonant frequency is given as $\omega_o^2 \equiv k/m$.

$\omega_o^2 \equiv k/m$.

and thus x(t) has the form:

$$x(t) = k_o e^{-i\omega t} = qE_o \frac{(\omega_o^2 - \omega^2) - i\nu\omega}{(\omega_o^2 - \omega^2)^2 + \omega^2\nu^2} e^{-i\omega t}$$

The real part of this expression then gives the physically meaningful oscillation amplitude.

For our purposes, we are most interested in the observation that the oscillation amplitude has a peak magnitude when the resonance condition $\omega = \omega_{res}$ is satisfied where the resonant frequency is given as $\omega_{res}^2 = \omega_o^2 - \frac{b^2}{2m^2}$ $\omega_{res}^2 = \omega_o^2 - \frac{b^2}{2m^2}$. A plot of the solution is given below.

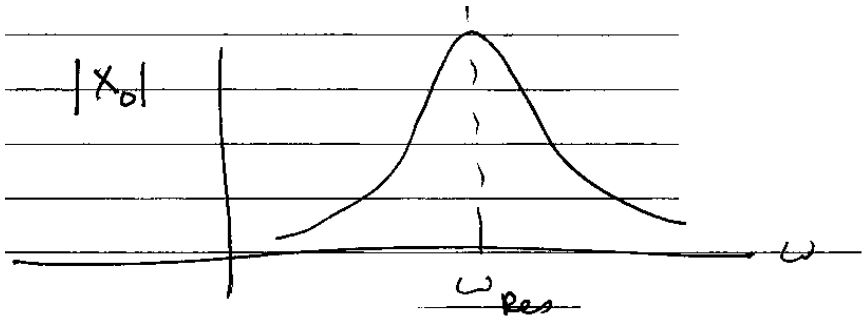


Figure 5.5: Amplitude of driven oscillator response near a resonant frequency.

The damping of the oscillator then causes the oscillator energy, which is driven by the work done by the external forcing term (in this case by the light wave) to be dissipated. As a result, the light wave which is moving past the oscillator will leave the interaction with a reduced amplitude. The dissipated energy will eventually be transformed into thermal motion of the atoms, resulting in a heating of the gas.

In order to understand how the two different ranges of light spectra (i.e. the visible spectrum emitted by the sun and IR emission from the Earth's surface) interact with the gas molecules in the atmosphere, we must consider the different types of excitation that these molecules can exhibit and the corresponding resonant frequency (or equivalently, the wavelength of light that results in resonant excitation) of the various types of molecular excitation that can occur.

Types of Molecular Excitation

Should we give quantitative equations for E_ij for carious excitation, so the magnitudes of the energies associated with the excitations are more apparent?

There are four type of atomic or molecular excitation that are of interest for the transport of radiation through the atmosphere:

Bound electron of transition from one orbital to another: In this class of dynamics an electron is bound to a nucleus, and moves from one orbital to another. Higher orbitals (e.g. j>i) correspond to a higher energy level. There is an energy associated with change in orbital, E_{ij} that is given by $E_{ij} = E_j - E_i$

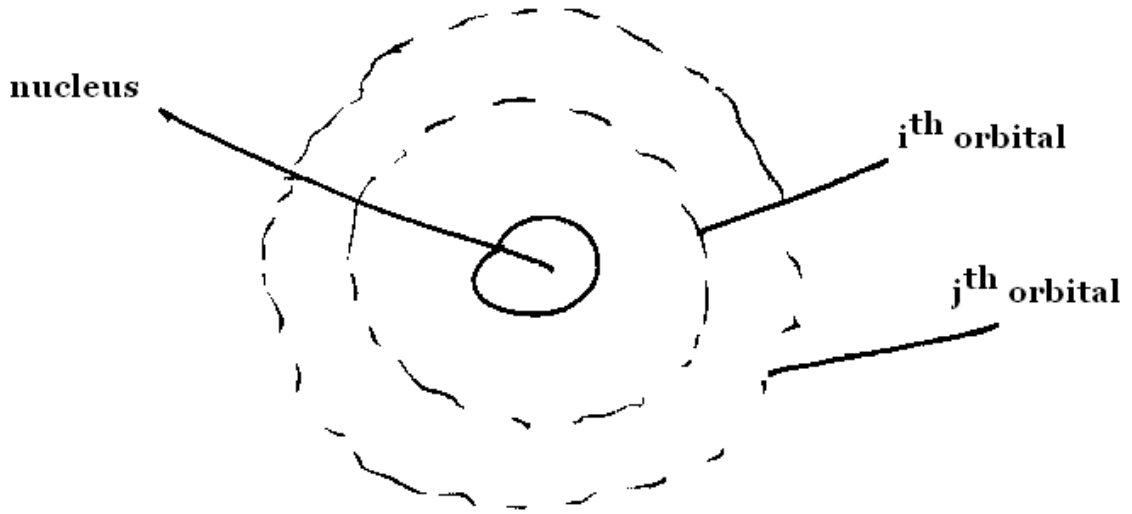


Figure 5.6: Schematic of bound electronic orbitals in a single atom.

Quantum mechanics tells us that this change in energy is associated with a characteristic frequency of light wave which can resonantly excite the transition (i.e. move an electron from the i-th orbital to the j-th orbital). This characteristic frequency is given as ω_{ij} [where ω is an angular frequency $\omega = 2\pi\nu$]

$$\frac{h}{2\pi}\omega_{ij} = E_{ij}$$

If the electron decays from the higher state, j, down to the lower state, I, then a quantum of light called a photon with be emitted. This photon will have an energy E_{ij} and a frequency given as in the above equation. Typically, these types of transitions correspond to photon energies in the range of several electron volts (eV) or more.

Molecules can also exhibit other types of dynamics. In particular, the nuclei (and also the electrons bound to those nuclei) can also vibrate relative to each other's position as shown schematically below.

Vibrational transitions: We can crudely think of these vibrations as being masses on a spring, much like the harmonic oscillator we just considered. Due to quantum mechanics, there are a series of natural resonant frequencies for this system; each of these resonant frequencies has a corresponding energy

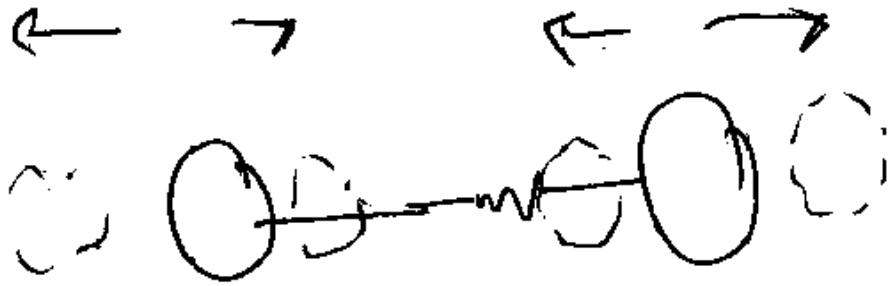


Figure 5.7: Schematic of vibrational excitation of a diatomic molecule.

Corresponding energy of frequency:

$$E_i^{vib} = \frac{h}{2\pi}\omega_i^{vib}$$

Molecular Rotations Molecules can also rotate, and can store energy in their rotation. This energy has a corresponding photon frequency that denotes the frequency of light that can resonantly excite this motion. If the motion is dissipated, then a photon of light with this energy will also be emitted by the molecule.

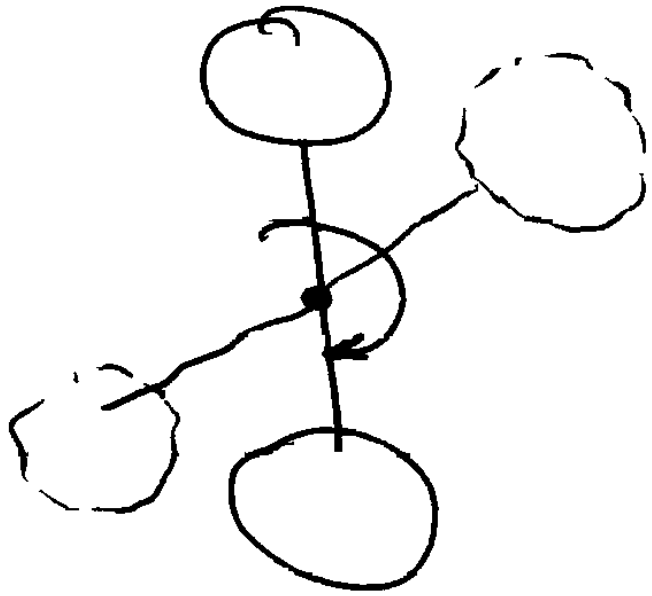


Figure 5.8: Schematic of rotational excitation of a diatomic molecule.

Bending Molecules can also bend, and can store energy in their bending. This energy has a corresponding photon frequency that denotes the frequency of light that can resonantly excite this motion. If the motion is dissipated, then a photon of light with this energy will also be emitted by the molecule.

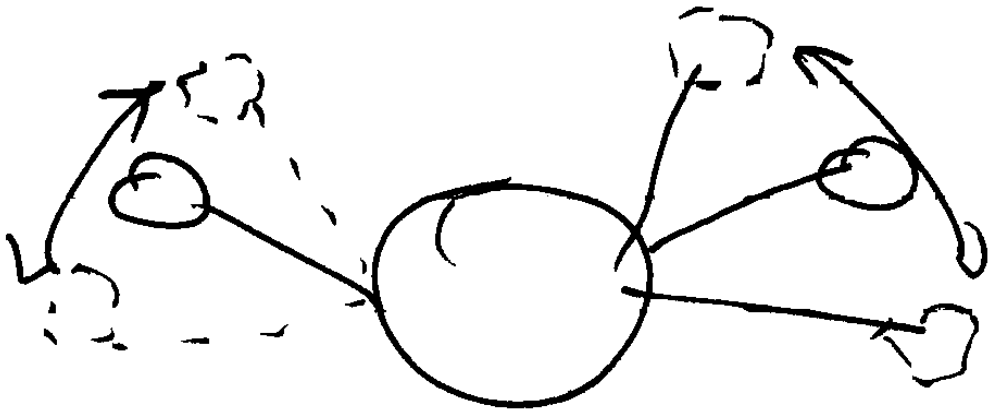


Figure 5.9: Schematic of molecular bending excitation.

$$E_i^{bend} = \frac{h}{2\pi}\omega_i^{bend}$$

Energy Ranges of Molecular Excitation. Now, for our purposes we note a key fact – namely that the energies that correspond to these four types of molecular motion have several different energy ranges. The electron transition energies are highest, the vibrational energies are next highest, and then the rotation and bending energies are considerably lower. Thus, these energy ranges will have the following rough ordering:

$E_{ij} \geq 10eV$

$E_j^{vib} < E_{ij} \geq 10eV$

$E_j^{vib} < 1$ eV or <u>less</u>

$E_i^{rot} \ : \ E_i^{bend} < 0.1$ eV or less.

We notice that there is an ordering to the energies of thee different processes such that $E_{ij} >> E_i^{vib} \geq E_i^{rot} \approx E_i^{bend}$.

Now, let us relate these facts back to the problem at hand – namely the transport of blackbody radiation through an absorbing gas. Let us first consider the typical energy of visible light photons which have a wavelength in the range of 400-800 nm. Using the previous expressions relating frequency, wavelength, and photon energy, we thus estimate that a typical visible light photon has an energy of roughly. $E_\nu^{vis} : 2eV$. These photons comprise the peak portion of the Sun's emission spectrum which to a good approximation is a blackbody spectrum from a body at ~5800 deg K. Next let us consider infra-red (IR) photons, which are emitted from the Earth's surface at a temperature of roughly 300 deg K. Thus photons will have a peak intensity at a wavelength of roughly 10 microns, which is about 20 – 40 times longer than the wavelength of visible light photons. This can easily be shown by evaluating the blackbody spectrum for T=300 deg K. The typically energy of these photons is . $E_\nu^{IR} \leq 0.1eV$.

Next, let us compare these energy ranges with the characteristic energies for the four different types of molecular dynamics summarized above. We note that visible light interactions with molecules will tend to be non-resonant since the photon energy doesn't correspond to the natural (or resonant) vibration energy. However, we also note that IR photons will be resonant with the molecular vibrations and bending. This suggests that the visible light photons will not interact nearly as strongly with the molecules as do the IR photons. This is indeed the case. We reach a key conclusion for this section: E_ν^{vis} interactions with molecules tend to be Non-Resonant

and E_v^{IR} interactions with molecules tend to be more nearly resonant. Of course, if other types of particles are present in the gas (for example, aerosol particles which form clouds) then these conclusions must be modified to include the interaction of the light with such particles.

Now suppose that light radiation passes through a gas, and the frequency of the radiation is such that it resonantly excites the molecules of the gas as discussed qualitatively above. This absorption creates a population of molecules that are in one of these excited states (let us for sake of discussion here assume that these are vibrationally excited molecules). Now suppose that, while they are in this excited state, these vibrating molecules collide with a neighboring molecule that is not so highly excited (i.e. it is in a lower energy state or is not "vibrating" using the oscillator model discussed above). In this case, the excited molecule can transfer some or all of its vibrational energy to the neighboring molecule, which will be kicked up into a more energetic state. In order to conserve the total energy, the original molecule will then lose a corresponding amount of energy. These vibrationally excited molecules can also transfer some of this energy into the translational kinetic energy of the molecules as well, resulting in an eventual heating of the gas via the absorption of radiation. To take this discussion further, let us consider the time scales for these excitation, collision, and decay processes to occur. Let us refer to Figure below, as we think about these three key processes.

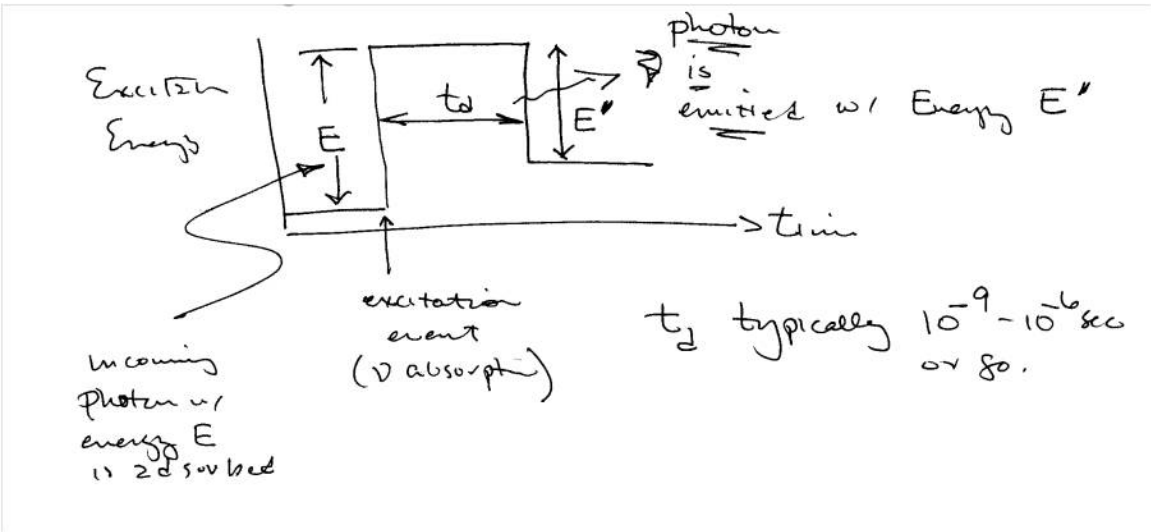


Figure 5.10: Schematic of the time evolution of a molecule or atom that undergoes an excitation event followed by a spontaneous decay event.

The vertical axis of the diagram corresponds to the energy state or energy level of a single molecule, while the x axis denotes time. When an incoming photon is absorbed, the molecule goes from a low energy state to a high energy state. This excited state has a finite lifetime, t_d. When t>td, the molecule has a high likelihood of decaying down into a lower energy state. This decay is accompanied by the emission of a photon whose energy E' corresponds to the change in energy of the molecule. For most of the decay processes we are concerned with, the lifetime t_d is typically in the range of nanoseconds to microseconds.

Now, suppose we have a collection of molecules (i.e. a gas) which can collide and interact with each other. Each individual molecule in this collection can undergo photon absorption and remission via the process just described. We can define an average time between collisions t_{col} by considering the average distance, l, between molecules, and the average thermal speed of the molecules, which we denote as $V_{thermal}$.

$t_{col}^{-1} \sim \frac{V_{thermal}}{l}$ where$V_{thermal}$ is typical thermal speed of molecule ($V_{thermal}$; $300m/s@300K$)($V_{thermal}$; $300m/s@300K$) and l is the distance between atoms or molecules:

We can write the number of molecules per unit volume, n, in terms of the mass density and the mass/molecule m using the expression $n \equiv \rho/m$. $n \equiv \rho/m$. Now clearly

n ~ l^{-3} ~ 10^{25} /m^3 or so for 1 atm

Therefore the typical distance between atoms at standard conditions in the Earth's atmosphere is then l ~ 200 Å. We can then estimate the typical time between collisions for molecules at typical temperatures of ~ 300 deg K to be given as $t_{col} \sim \frac{l}{V_{thermal}} \sim \frac{2 \cdot 10^{-8}}{300} \sim 10^{-10}$ sec,

$t_{col} \sim \frac{l}{V_{thermal}} \sim \frac{2 \cdot 10^{-8}}{300} \sim 10^{-10}$ sec,

i.e. we find that $t_{col} < t_d$ for most types of excitation. Now, note that when a molecule or atom in an excited state collides with a neighboring atom or molecule that is in a lower energy state, then some of the excitation energy can be transferred between the two colliding particles. If this energy transfer results in an increase in the translational kinetic energy of the lower energy particle, then we consider that particle to have been heated in the interaction. Because the collision frequency is often higher than the decay frequency, such energy transfer collisions are common. After many such interactions, the net effect is that the molecular gas is heated, and the light intensity decreases by a corresponding amount so that the overall energy conservation is maintained. We therefore conclude: *the absorption of the radiation by the gas leads to a heating*

of the gas and a reduction of the intensity of the radiation passing through the gas. There is an additional important implication of this analysis: the gas temperature will come into equilibrium with the radiation field temperature which is characterized by the width of the emission spectrum. This is by definition what we mean by a blackbody emitter, which has the atoms and molecules in thermal equilibrium with the radiation being emitted by the material. Let us now use these basic findings to build a simple model of the Earth's heat balance, and use that model to begin to understand how adding CO_2 molecules from fossil fuel combustion can impact this heat balance.

Simple Models of the Earth's Thermal Balance

Increasing CO_2 emissions from the consumption of fossil fuels has led to an increase in CO_2 concentrations in the Earth's atmosphere. These increased CO_2 concentrations can then lead to long-term global changes in the Earth's climate. As a result, there is interest in developing energy technologies which minimize or eliminate the production of CO_2. We can illustrate the physics origin of the problem and also illustrate the complex network of feedback loops in the Earth-Ocean-Atmosphere system with some simple considerations. Our discussion is motivated by the models found in Cushman-Roisin[1]. The interested reader is referred to the most recent Intergovernmental Panel on Climate Change (IPCC) reports for detailed discussions of this problem and of the status of predictions of future climate change.

[1] See e.g. B. Cushman-Roisin, Introduction to Geophysical Fluid Dynamics, 1994, Prentice-Hall, pp. 268-271 for an introductory discussion.

Let us now apply this to a simple thermal balance analysis of a slab of gas being illuminated by a visible light spectrum with total intensity $I_{in}(\lambda)$ as shown in Figure below. A small amount of this visible light is absorbed by the gas, which then re-radiates this energy as a black body.

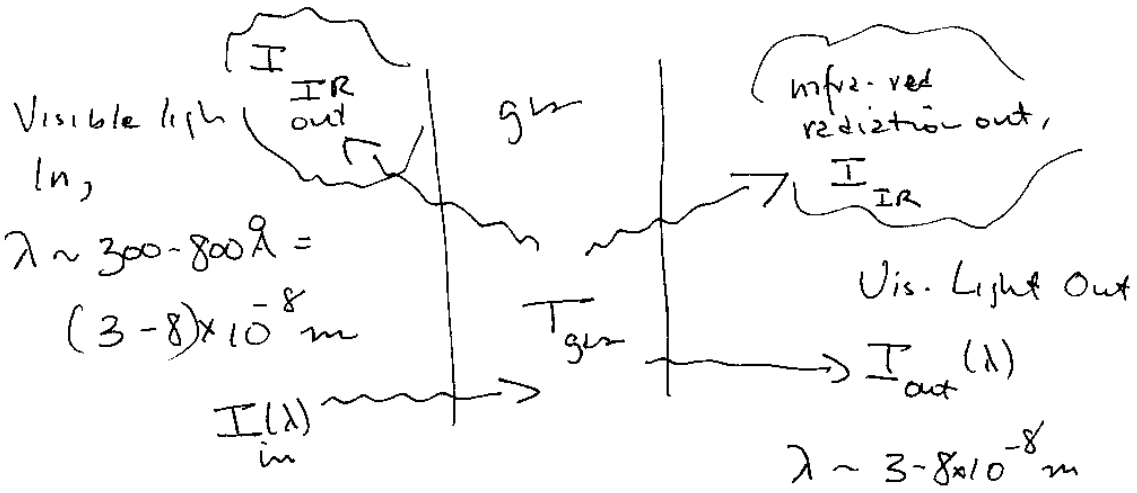


Figure 5.11: Schematic of visible light incident upon a slab of gas that re-emits as a blackbody at temperature Tgas.

If we apply a power balance to the slab we can then write:

$$\int [I_{in}(\lambda) - I_{out}(\lambda)] d\lambda = I_{IR} \quad \int [I_{in}(\lambda) - I_{out}(\lambda)] d\lambda = I_{IR}$$

Let us now introduce the transmission coefficient, β, to relate the incident and outgoing *visible* light radiation intensities $I_{out} = \beta\ I_{in}$; where β is the visible light transmission coefficient for the gas. From studies of heat transfer, we know that the infra-red radiation intensity is given as $I_{IR} = \sigma T_{gas}^4\ T_{gas}^4$ (from Stefan-Boltzmann Law) where σ denotes the Stefan-Boltzmann

emission constant. Here, we have implicitly assumed that the gas acts like a blackbody, which may not be precisely true, but for the purposes here it will suffice. We can now write the power balance by equating the absorbed and outgoing radiation and then solving for the resulting gas temperature:

$$(1-\beta_1)I_{in} = \sigma T_{gas}^4 \qquad \beta_1 \;;\; 0.49 \sim 0.5$$

$$\therefore \sqrt[4]{\frac{1}{2}\frac{I_{in}}{\sigma}} = T_{gas} \qquad I_{in} = 344 W/m^2$$

$$\therefore T_{gas} = 234K = -39^oC \qquad \sigma = 5.710^{-8} W/m^2 \text{ o}10^4$$

where we have taken $I_{in} \approx 340 W/m^2, \sigma = 5.7x10^{-8} W/m^2 - {}^\circ K^4, and\, \beta_1 \approx 0.5$. This temperature is much colder than the actual Earth's atmosphere, even though the values used above are roughly correct for Earth. Something is therefore clearly missing from this model.

Consider next the case of the Earth's surface with no atmosphere as shown schematically in Figure below. The incoming solar radiation flux, *I*, is incident on the Earth's surface. A fraction, α, (known as the albedo) is immediately reflected back into space primarily by water and ice surfaces. The balance of the radiation is completely absorbed by the Earth's surface which is consequently heated to a surface temperature T_0. In this model the Earth's surface is cooled only by blackbody radiation and consequently the power balance can be written simply as $(1-\alpha)I = \sigma T_o^4$ with the Stefan-Boltzmann constant given as $\sigma = 5.7\times10^{-8} W/m^2\, K^4$. Averaged over the entire surface, the Earth's albedo is α*~0.34.* Thus, in this model where there is no atmosphere to reabsorb radiation, and the earth absorbs all of the incident solar radiation, the equilibrium temperature of the Earth's surface is estimated to be 250 deg K – which is about 35 deg K lower than the actual value. The difference is accounted for by atmospheric absorption of

radiation. The emitted radiation has a peak intensity in the infra-red (IR) region of the spectrum, corresponding to blackbody emission at ~250 deg K. Thus, we must consider the transmission and adsorption of IR radiation through the atmosphere.

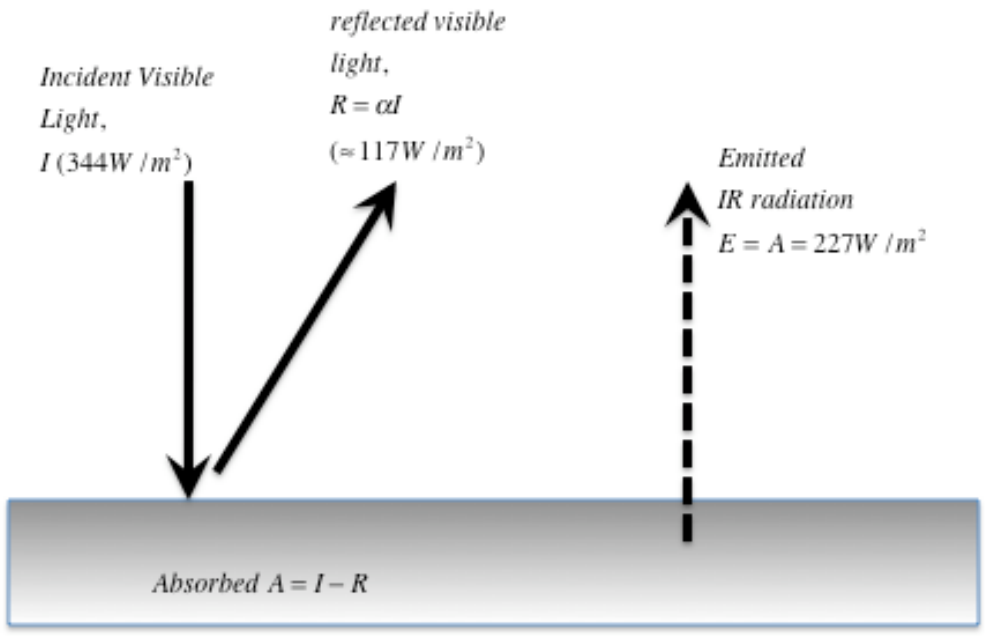


Figure 5.12: Schematic of Earth's surface thermal balance in the absence of an atmosphere.

Effect of Atmospheric Absorption on Earth's Thermal Balance

The Earth's atmosphere contains significant concentrations of water vapor, carbon dioxide, and methane. These molecules have a number of quantum mechanical rotational and vibrational states which have excitation energies of ~0.01-0.1 eV which correspond to the low energy photons which are emitted as blackbody radiation by the warm surface of the Earth. Thus, these

photons can be reabsorbed by the atmosphere and thus can lead to a warming of the atmosphere above the values predicted above. Because of the strong absorption of low energy photons and the relatively weaker absorption of shorter wavelength visible light by the atmosphere, we make a distinction between these two types of radiation: (a) short-wavelength radiation (consisting primarily of visible light) and (b) long-wavelength radiation (which lies in the infra-red portion of the spectrum). In order to see this process more clearly, consider the thermal budget model summarized in below.

Begin at the upper left-hand corner of the figure and proceed downwards towards the Earth's surface. The incident solar flux, averaged over the Earth's surface, is given as $I \sim 344\ W/m^2$ as discussed earlier. A fraction $R_1=\alpha_1 I$ (here $\alpha_1 \sim 0.33$ is the albedo of the atmosphere) is reflected from the atmosphere back into space without suffering absorption while the remainder of the flux passes through the atmosphere. This flux is still centered in the visible portion of the spectrum and thus a fraction $T_1=\beta_{vis} I$ is transmitted through the atmosphere and reaches the Earth's surface (the visible light transmission coefficient $\beta_{vis} \sim 0.49$). A portion $R_2=\alpha_2 T_1$ is reflected immediately back into the atmosphere as visible light, and the balance of the radiation A_2 is absorbed at the Earth's surface (here $\alpha_2 \sim 0.04$ is the albedo of the Earth's surface). The short wavelength radiation which is reflected from the Earth's surface passes back through the atmosphere, and a fraction of this radiation $T_2=\beta_{vis} R_2$ is transmitted back into space. There are two places where short-wavelength radiation is absorbed by the atmosphere. A fraction $(1-\beta_{vis})I$ of the incident light which is traveling towards the Earth is absorbed in the atmosphere. A

portion $(1-\beta_{vis})R_2$ of the visible light which is reflected from the Earth is also absorbed by the atmosphere. Thus, the total visible light flux which absorbed by the atmosphere, A_1, can be written as

$$A_1 = (I - R_1 - T_1) + (R_2 - T_2)$$
$$= [1 - \alpha_1 - \beta_{vis} + \beta_{vis}\alpha_2(1 - \beta_{vis})]I \quad .$$

with

$$E_1 = A_1 + E_2 - T_3$$
$$= A_1 + (1 - \beta_{IR})E_2 \quad . \qquad (2)$$

These last two equations can be solved together. Using the Stefan-Boltzmann constant given earlier along with the albedo of the atmosphere and surface, the visible light transmission coefficient $\beta_{vis} \sim 0.49$, and the IR transmission coefficient $\beta_{IR} \sim 0.04$, we can solve for the atmospheric IR emission $E_1 \sim 560$ W/m^2 and the surface IR emission $E_2 \sim 520$ W/m^2. Note in particular that the inclusion of atmospheric adsorption has resulted in a substantial increase in the IR emission from the surface. Finally, using the Stefan-Boltzmann law, we estimate the surface temperature to be ~310 deg K, which represents an increase of ~60 deg K from the earlier model discussed above. This result neglects the hydrological cycle which partially negates the IR adsorption of the atmosphere. However, the basic physics of this result is clear: strong IR absorption by the atmosphere leads to significant heating and results in a significantly increased surface temperature. This strong IR absorption is caused by the presence of water vapor, carbon dioxide, and methane in the atmosphere which have strong bands if IR absorption. Thus, if the concentration of these species were to increase (e.g. by the rapid combustion of fossil fuels) then

the IR transmission coefficient β_{IR} would decrease and the model would predict an increased surface and atmospheric temperatures.

We have now established a simple Earth-Atmosphere power balance model that includes transmission and absorption of visible light and IR radiation. In order to relate this model to the consumption of fossil fuels, we must now determine the relationship between IR Absorption and greenhouse gas concentration.

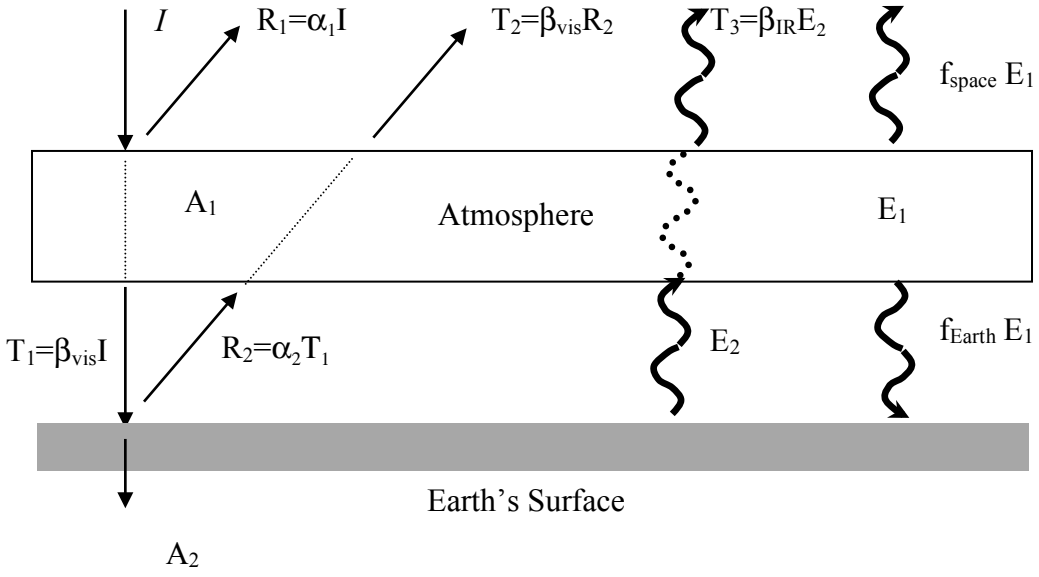


Figure 5.13: Schematic of Earth/Atmosphere Heat Budget, including IR absorption, but neglecting the hydrological cycle. Figure adapted from Cushman-Roisin

Linking Greenhouse Gas Concentration to IR Radiation Absorption

We now need to examine the quantitative link between the density of an absorbing species in the atmosphere and the resulting infra-red transmission coefficient. In order to do this, let us consider a simple 1-D model of the atmosphere as shown in Figure below. At x=0 we have IR

radiation incident upon this slab model, and the radiation moves through the slab. The slab has some number of gg molecules/unit volume, n_{gg}, as shown in the schematic.

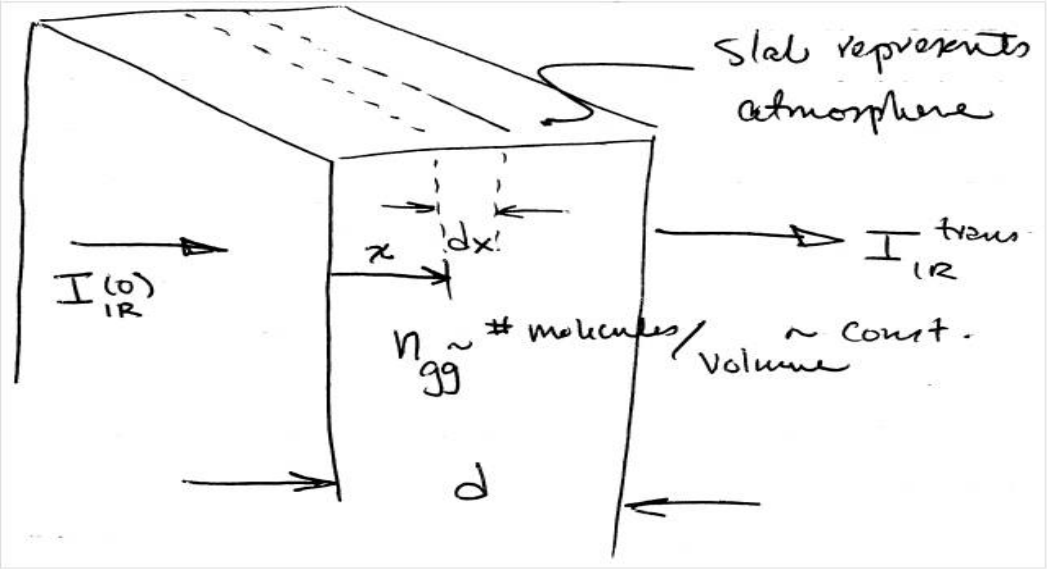


Figure 5.14: Schematic 1-D representation of the transport of infra-red radiation through an atmosphere containing absorbing greenhouse gas molecules.

At a depth, x, into the slab we can write the change in the IR radiation intensity, dI(x) as

$$-dI(x) \equiv I(x) - I(x+dx) \propto I(x) n_{gg}$$

where we have assumed that the change in intensity is proportional to n_{gg}. Let us define a constant of proportionality σ_{gg} ~ which quantifies the "absorption cross-section for n_{gg}" Then we can write the change in intensity as

$$dI(x) = -I(x)\sigma_{gg} n_{gg}$$

$$\int_0^d \frac{dI}{I} dx \equiv \int \sigma_{gg} n_{gg} dx \sim const \Rightarrow \ln I \Big|_0^d = \sigma_{gg} n_{gg} d$$

We can solve this as for the intensity at x=d and find

$$\ln\frac{I(d)}{I(0)} = -(\sigma n)_{gg} d \quad I(x=d) = I|_{x=0}\, exp(-n_{gg}\sigma d)$$

Rearranging this then gives

$$\frac{I(d)}{I(0)} = \exp\left[-(\sigma n)_{gg} d\right]$$

But $\frac{I(d)}{I(0)}$ is just the transmission coefficient for IR radiation, β_2, in our earlier analysis. Thus the infrared absorption fraction can be written as $A_{IR} = 1 - \beta_{IR} = 1 - exp(-n_{gg}\sigma d)$.

This analysis provides us with an important result: it shows that and increase in n_{gg} leads to a decrease in β_{IR} and an increase in A_{IR}. If we use this result in the 0-D Earth-Atmosphere power balance, we see that this then leads to an increase in the IR exchange between the Earth and the Atmosphere, which must be accompanied by an increase in the Earth's temperature. It is this link the gives rise to the origin of link between increased CO_2 concentration and concerns over global climate change. In the next chapter, we construct a simple model of the balance of C-containing molecules in the atmosphere and then in subsequent chapter, we combine these models to understand how much C-free energy might be needed to meet human energy demand while at the same time avoiding unacceptable climate change

Chapter 6: Fundamentals of the Carbon Cycle

The results of the previous chapter indicate that if the CO_2 content of the atmosphere (or indeed the content of other greenhouse gases that have a long lifetime in the atmosphere) is increased, more of the IR radiation emitted by the Earth's warm surface will be trapped within the atmosphere, resulting in a heating of the atmosphere and an increase in the rate of IR radiation transfer between the surface and the atmosphere. As a result, the overall temperature of the surface and atmosphere will increase. One substantial source of CO_2 is the combustion of fossil fuels. Therefore, to complete the link between the consumption of fossil fuels and the global climate change problem, we must also develop a basic understanding of how CO_2 moves between the atmosphere, ocean, and land which each act as reservoirs for carbon. Obviously the transport and exchange of carbon between and within these reservoirs is a complex problem involving chemistry, biological processes, and physical transport and, as a result, a detailed study of the problem is well beyond the scope of this book. Our purpose here it to develop a very basic, low-level description of this complex problem that is sufficient to gain an understanding of the linkage between energy consumption, (and in particular fossil fuel energy consumption), climate change, and carbon content (in the form of CO_2) in the atmosphere.

A Simplified Carbon Balance Model

As was discussed above, CO_2 is exchanged between the atmosphere, ocean, and land which each can act as reservoirs for carbon. The precise mechanism by which C is stored in each of these reservoirs can vary – e.g. C can be contained in the atmosphere in the form of CO_2 and CH4, in the land in the form of pure C (e.g. coal) or within organic material usually in the form of

hydrocarbon or carbon-oxygen-hydrogen molecules, and in the ocean in the form of CO_2 as well as in carbonic acid and in solid form within the skeletal structures of sea creatures. Here, we are not interested in such a detailed description of the storage locations and mechanisms for carbon. Instead we are interested in developing a very simple mass balance for carbon in the atmosphere. We can represent these three key reservoirs of carbon via three control volumes as shown below:

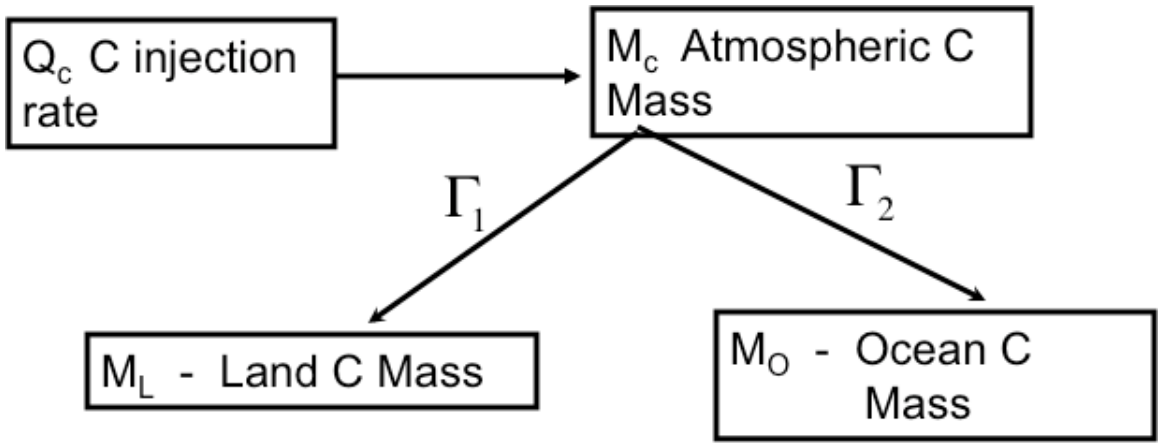


Figure 6.1: Schematic of C storage within atmospheric, land based, and ocean based reservoirs. A source of C injection, produced by the combustion of fossil fuels, injects C into the atmosphere in the form of a source QC(t) of CO_2.

Clearly the mass balance for atmospheric carbon is then given as

$$\frac{dM_c}{dt} = Q_c - \Gamma_1 - \Gamma_2$$

Where

Q_c ~ rate of input of C into the atmosphere,

Γ_1 ~ Net rate of C absorption by the Earth Land Area, and

Γ_2 ~ Net rate of C absorption Earth's Oceans.

We can integrate this expression to then find the atmospheric C mass:

$$M_c(t) = \int [Q_c(t') - \Gamma_1(t') - \Gamma_2(t')]dt + M_{c_0}$$

To make further progress we need to first consider the source injection rate and the exchange fluxes. In the earlier chapters, we saw that there are a number of motivations that drive human beings to desire access to some minimal level of per-capita energy resources. These considerations are often described in simple economic terms (which really only serve to quantify the more important human quality of life drivers discussed earlier). For example, the C emission rate $Q_C(t)$ can be linked to energy production by the so-called Kaya Identity:

$$Q_c(t) = N(t)G(t)E(t)C(t)$$

where

N(t) ~ population at time t
G(t) ~ average GDP per person per year
E(t) ~ average energy intensity per unit GDP
C(t) ~ average intensity of energy sources

This relation simply expresses the C carbon emission rate $Q_c(t)$ in terms of the human population, the average economic activity per person, the energy intensity required to sustain the economic activity, and the carbon content of the fuel used to produce that energy.

As we saw earlier, the population has been growing at a rate, r(t). Recent demographic data show that for most of the 20th century r~const at a few percent/year, yielding exponential growth in the human population. Very recent data in the past decade or so suggest, and modeling also suggests, that the growth rate may be starting to decrease. If that trend continues, it could result in a saturation of the human population at values estimated to range from 9-11 billion people in the late 21st century. The quantity G(t) has also been growing at a rate r1(t), the energy intensity has been decreasing at a rate r2(t), and the average carbon intensity of primary energy sources has been decreasing at a rate r3(t). Thus, defining a net emission growth/decay rate, rnet, as $r_{net} = r + r_1 - r_2 - r_3$ $r_{net} = r + r_1 - r_2 - r_3$ we write the net C source rate, Qc(t), as

$$Q_C = Q_{C_0} \exp(r_{net} t).$$

Where we need to recognize that the emission rate can vary in time, i.e. that $r_{net}=r_{net}(t)$. Clearly if $r_{net}>0$ then the atmospheric C concentration can increase and, since carbon in the atmosphere is contained primarily in the form of CO_2, and since CO_2 is a greenhouse gas, then IR transmission will decrease and thus the Earth will warm. This is the key physics behind the concerns over global climate change.

To gain further insight into the implications of this C balance model, let us consider a few limiting cases to the solution to the carbon mass balance. We will consider the following models for the C flux Γ_1 and Γ_2:

a) $\Gamma_{1,2}$ = constant

b) Diffusion Limited.

Let us first consider the case where the fluxes Γ_1 and Γ_2 are fixed, which is akin to assuming that the land and ocean C uptake does not change as the C concentration in the atmosphere changes due to human fossil fuel combustion. In this case,

$$M_C(t) = M_C(t=0) - (\Gamma_1 + \Gamma_2)t + \int_0^t Q_{C_0} \exp(r_{net} t)$$

i.e. the atmosphere simply integrates the C emissions over time. Now, presumably before the introduction of large-scale fossil fuel combustion, the C balance in the atmosphere-land-ocean system was in equilibrium, that is to say, there were no *net* exchanges of C between these three reservoirs (clearly there are exchanges between these reservoirs; however in the absence of anthropogenic sources presumably these reservoirs are in equilibrium). In this case, Γ_1 and Γ_2 both are equal to zero since they denote the net flux between these reservoirs. Thus, since we have assumed that the fluxes do not change in response to human injection of C into the atmosphere, we can then assume that these fluxes remain at zero after the introduction of C injection into the atmosphere. If the growth rate r_{net} does not change in time, then we can solve for M_C simply as

$$M_C(t) = \frac{Q_C(0)}{r_{net}} \exp(r_{net} t) + M_C(0).$$

Let us now evaluate this with current values. In 2008, we have roughly that $Q_c^{(0)} \sim 7 \frac{GTonnes}{year}$ $Q_C(0) \sim 7x10^9\ Tonnes/year$ (Note that is this GTonnes of C per year, not GTonnes of CO_2 per year). Current trends suggest that the time to double the C concentration in the atmosphere, i.e. the so-called doubling time for Q_c is ~ 20 years, i.e.

$$\frac{Q_c(20)}{Q_c(0)} = 2 = e^{20r}$$

which gives the growth rate, r, for the carbon inventory as approximately $r = \frac{ln(2)}{20} \approx 3.5\%/yr$

Thus, in 20 years we expect the total, integrated C mass in the atmosphere to be approximately

$$\left.\frac{M_c(t)}{M_c(0)}\right|_{t=20} = \frac{Q_c^{(0)}}{M_c^o r} e^{rt} + 1 .$$

$$\left.\frac{M_C(t)}{M_C(0)}\right|_{t=20} = \left.\frac{Q_C(0)}{M_C(0)r} e^{rt}\right|_{t=20} + 1.$$

With the values given above, we find $\frac{Q_{C_0}}{r} \sim \frac{7 GTonnes/yr}{0.035/yr} = 150\, GTonnes \approx 1.5 \times 10^{11}\, Tonnes$. To determine the relative increase in C inventory that this increment corresponds to, we must find the current C carbon mass, $M_C(0)$ in the atmosphere. The total atmospheric mass is roughly ~ 5 x 10^{18} kg ~ 5 x 10^{15}Tonnes ~ 5 x 10^6GTonnes [REF], and the current CO_2 concentration is roughly 380ppmv ~ 4 x 10^{-4}. We can therefore estimate the current C mass in the atmosphere as

$$M_C(0) = \frac{12}{44} 400(CO_2 ppmv) 5x10^6 (GTonnes) \frac{44(CO2\, density)}{29(air\, density)} \approx 790 GTonnes$$

Note that this expression takes account of the molecular weight of C and CO_2 and the relative density of air and CO_2 at standard conditions. We can now find the relative increase in carbon mass that current growth rates will yield.

$$\frac{M_c}{M_c^{(0)}} = \frac{Q_c^{(0)}}{rM_c^{(0)}} + 1 = \frac{7}{0.035(5 \cdot 10^2)} + 1;\ 1.4.$$

.

Thus, at current growth rates for C emission, we expect that in the 2028-2030 time frame the C content of the atmosphere will have grown by 40% and the CO_2 fraction in the atmosphere would then be approximately 1.4*380~550 ppmv. This value is, of course, only valid in the limit that the uptake of C by the land and oceans does not change (which it most certainly will in fact change).

Now, let us estimate the effect that this increase in C content will have on the infrared transmission coefficient, β_2 using our earlier model linking greenhouse gas content to IR transmission We use these earlier results to write the relative change in transmission as

$$\frac{\beta_2(t=20)}{\beta_2(t=0)} = \frac{\exp\left[-\sigma_{gg}\frac{M_C(t=20)d}{V}\right]}{\exp\left[-\sigma_{gg}\frac{M_C(t=0)d}{V}\right]}$$

where we have expressed the number density of CO_2 molecules in terms of the C mass as n=MC/V where V is the volume of the atmosphere. Now, noting that the cross section, thickness, and volume do not change, we can re-write this expression as

$$\frac{\beta_2(t=20)}{\beta_2(t=0)} = \exp\left[-\frac{\sigma_{gg}d}{V}\left(M_C(t=20)-M_C(t=0)\right)\right].$$

In our earlier 0-D analysis of the Earth's thermal balance, we stated that β_2 ~0.05 and from above we found that $M_C(t=0)\approx 500\ Gtonnes$. $M_C(t=0)\approx 500\ Gtonnes$. We can then use these current conditions to estimate that the constant term in the exponential has a value of approximately

$$\frac{\sigma_{gg}d}{V} = -\frac{1}{M_C}\ln(\beta) \approx \frac{3}{500} = 0.006\ Gtonnes^{-1}.$$

Now, the C mass balance estimate above shows that we expect a change in carbon inventory over this period of time of about 200 Gtonnes. We can now estimate the change in IR transmission to be given by

$$\frac{\beta_2(t=20)}{\beta_2(t=0)} = exp\left[-\frac{3*200}{500}\right] \approx 0.3$$

This simple model therefore would suggest that, if current trends continue and CO_2 content increases to ~550ppmv in the next 20-30 years, then the IR transmission coefficient of the atmosphere would decrease to a value of ~1/3 of the current value (i.e. it would fall from ~5% to about 2% or so). Of course, this simple estimate neglects many potential complicating factors such as changes in Earth's albedo, cloud cover, and so forth. Nonetheless, from the heat balance analysis presented earlier, we can now use this in the Earth's heat balance model to find E_1, E_2, and then use these to solve for $T_{1,2}$. Clearly, this will result in an increase in global temperature. Estimating this increase is left as an exercise for the students.

This model of course assumes no change in C uptake by land and oceans. This is probably a poor assumption. Let us now consider the second model that attempts to rectify this oversimplification. We make one crucial (and seemingly reasonable) assumption in order to keep this simple model tractable. Let us assume that the *net* flux of C between the atmosphere, ocean and land is proportional to the *change* in C concentration in the atmosphere. In other words, let us define $M_{C_0}\, M_{C_0}$ to be the equilibrium C mass in the atmosphere at which the net fluxes, Γ_1 and Γ_2 vanish. Presumably $M_{C_0}\, M_{C_0}$ corresponds to the atmospheric C mass prior to the introduction of large scale fossil fuel combustion in the early 19th century. Now, with

$Q_C(t) > 0$ $Q_C(t) > 0$ we have an increase in atmospheric C mass, i.e. we have the deviation of atmospheric C mass away from equilibrium given as

$$\delta M_C(t) \equiv M_C(t) - M_{C0} > 0.$$

We can now use our assumption about the form for the C fluxes to write

$$\Gamma_1 \propto \delta M_C(t)$$
$$\Gamma_2 \propto \delta M_C(t)$$

We can turn this proportionality expression into an equation by introducing the carbon exchange time scales, t_1 and $t_{2\,w}$ which relate the fluxes of carbon between the atmosphere and either the land or ocean reservoirs to the change in atmospheric carbon inventory via the equations

$$\Gamma_1 = \frac{\delta M_C(t)}{t_1}$$

$$\Gamma_2 = \frac{\delta M_C(t)}{t_2}$$

These exchange time scales correspond to the average time it takes for a carbon atom to be exchanged between the atmosphere and land, or between the atmosphere and the ocean. One of the simplifying assumptions we make is that the relaxation timescales $\tau_{L,O}$ τ_L and τ_O are independent of the C concentration in the atmosphere and of the mean temperature as well. Obviously, if changes in these quantities lead to a large change in the land/ocean system (e.g. suppose that changes in temperature or rainfall patterns lead to a large change in plant

distribution around the globe). In that case, the uptake of C from the atmosphere would certainly be changed and thus would lead to a change in the value of τ_L). τ_L).

In this case, the carbon balance model can then be written as

$$\frac{\partial}{\partial t}\delta M_C(t) = Q_C(t) - \frac{\delta M_C(t)}{t_{net}}$$

where $\frac{1}{t_{net}} = \frac{1}{t_1} + \frac{1}{t_2}$ is assumed constant.

This is a linear inhomogenous ordinary differential equation which can be solved in general for a given Q(t) if $t_{net}$is fixed. Let us consider several simple cases to illustrate the key behavior we are interested in.

Case 1*:* $Q_c = 0$ First, consider the special case that $Q_c = 0$, i.e. that there is no source of anthropogenic carbon injection into the atmosphere. In this case, it is easy to show that $\delta M_C(t) = 0$ always, i.e. the system is always in equilibrium as would be expected. The details of this are left as an exercise for the student.

Case 2*:* Step-wise decrease in Q_C: The second case is the response of the system to a step-wise change in the carbon injection rate. In particular, we assume that the carbon injection rate is held constant for t<0 and then at t=0 the carbon source is instantaneously turned off for all time t>0, i.e. the carbon source term is given as

$$Q_C(t) = \begin{cases} Q_C(0) = Q_{C_0} = const & t < 0 \\ 0 & t > 0 \end{cases}$$

We then want to know $\delta M_C(t)\, \delta M_C(t)$ for all time. Let us first consider times such that t<0. The mass balance equation then reads

$$\frac{\partial}{\partial t}\delta M_C(t) + \frac{\delta M_C(t)}{t_{net}} = Q_{C_0}.$$

If the C injection is constant for all time t<0, then the system must be in an equilibrium and thus clearly we have $\frac{\partial}{\partial t}\delta M_C(t) = 0$ and we immediately write the perturbed carbon mass as

$$\delta M_C = Q_{C0} t_{net}.$$

Notice that, if the net C uptake time scalet_{net}, is increased (decreased) for a fixed injection rate, then the departure from equilibrium will also increase (decrease). Now, let us consider the response for times t>0. In this case, the mass balance equation reads

$$\frac{\partial}{\partial t}\delta M_C(t) + \frac{\delta M_C}{t_{net}} = 0$$

since we have already assumed that the source vanishes for t>0 in this case. The solution to this equation is given as

$$\delta M_C(t) = A \exp(-t / t_{net}).$$

We solve for the constant A by noting that $\delta M_C(t)\big|_{t=0-\varepsilon} = \delta M_C(t)\big|_{t=0+\varepsilon}$ to find that

$$A = Q_{C0} t_{net}.$$

Thus for t>0 the solution in this case is given as

$$\delta M_C(t) = Q_{C_0} t_{net} \exp(-t/t_{net})$$

and thus after the source is reduced to zero, the deviation from equilibrium decays on a timescale t_{net}. Much more complex modeling, as well as measurements in the atmosphere tell us that the estimated value for t_{net} is of the order of 100's of years [REFERENCE]. Using this fact, our simple analysis here then tells us that, even if human sources of CO_2 injection were instantly turned off (something that will clearly not happen) it will still take several C exchange times (i.e. many centuries) for the C inventory in the atmosphere to return to undisturbed levels.

***Case 3**:* Let us consider the response of the atmosphere to a step-wise increase in $Q_C(t)$ Let us suppose that for $t<0$ $Q_C=0$ and then at $t=0$ the source is instantly turned on such that $Q_C(t)=Q_{C0}$=const for all $t>0$. The deviation of the atmospheric C inventory from equilibrium is then given as

$$\delta M_C(t) = Q_{C_0} t_{net} \left[1 - e^{-t/t_{net}}\right]$$

The relative time behavior is given qualitatively in **Figure** below. Examining this solution note a crucial result: For times such that t/t_{eff}, we see that $\delta M_C \propto t$, $\delta M_C \propto t$, i.e. *for timescales that are short compared to the C exchange timescales, the atmosphere simply integrates, or collects and stores, all of the* CO_2 *that human beings inject.* This is a truly crucial result, as it helps understand why very large reductions in C injection rates appear necessary to successfully avoid significant changes in climate in the coming decades.

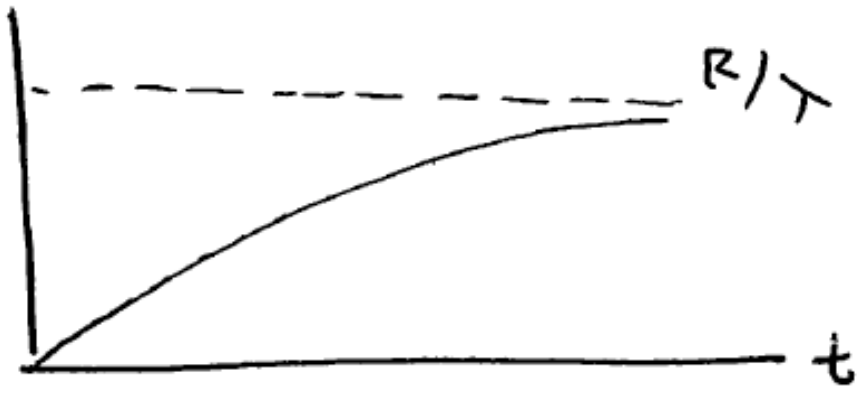


Figure 6.2: Solution of the perturbed atmospheric C inventory for the case QC(t)=const for t>0.

In general Q_C (t) is not constant and is, in general, a function of time, i.e. $Q_C=Q_C(t)$. The carbon balance equation in this case is an inhomogenous linear ODE with constant coefficients. In this case the solution becomes a bit more complex:

$$\Delta M_C(t) = \Delta M_C(0) + e^{-t/\tau_{ex}} \int_0^t e^{t'/\tau_{ex}} Q_C(t')dt'$$

The student should confirm that if Qc(t)=const in this case then the solution reverts to the earlier solution discussed above. This model can also be used with historical data for energy growth rates and estimated values of the C exchange timescales to make estimates of atmospheric C inventory, and this model can then be compared with actual measurements to gain an idea of how accurately this model represents the much more complex behavior of the actual system.

Summary:

Let us summarize what we have learned up until this point. First, we have seen that there are strong correlations between the various social, non-economic measures of human quality of life (e.g. average lifespan, infant mortality rates, literacy rates, etc.) and per capita energy usage. These measures would seem to be fairly independent of particular culture and country and thus may be fairly commonly valued across the human population. The correlations are strongest when examining access to electrical energy in particular. This correlation between quality of life and energy provides a strong motivation for human beings who do not have access to adequate energy resources to increase that access and thus increase their energy use. This motivation lies at the heart of increasing global energy demand. Furthermore, we also found that increased access to energy was also correlated to decreases in population growth rate. Again, this finding seems to hold across a variety of cultures and countries, suggesting that achieving a stable human population will also be linked to increased global energy consumption.

Second, we examined the current sources of energy used around the world. We found that the large majority (about 75-80%) of the human energy demand is provided by the combustion of fossil fuels. This process releases CO_2 into the atmosphere on a fairly large scale. As a result, the atmospheric CO_2 concentration has been steadily rising in the past 100 years. Furthermore, these fossil fuel energy resources are finite, which implies that at some point become more and more difficult to access at adequate rates. EROEI considerations will then force more and more energy resources to focus simply upon the task of acquiring adequate energy resources, increasing the cost of acquiring and using these energy resources.

Third, we examined the effect that this CO_2 emission has on the Earth's heat balance. We considered a very simplified model in which we carefully tracked heat input in both the visible light and infra-red portions of the spectrum. These two spectral ranges are relevant because the Sun's radiation input to the Earth is peaked in the visible portion of the spectrum, while the emission of heat from the Earth occurs primarily in the infrared part of the spectrum. We also examined how molecules interact with these types of radiation, and found that the typical molecule will strongly absorb radiation at certain discrete resonant frequencies that correspond to characteristic excitation modes of the molecule. We also found that the visible light interaction is typically weaker than the infrared interaction due to the fact that the excitation frequencies tend to occur more often, and with stronger intensity, in the infrared part of the spectrum. We also made a simple model for radiation transport in the atmosphere and found that the transmission of radiation decreases exponentially with an increase in resonant molecular density. Thus, an increase in the concentration of a resonant absorber will lead to an exponential decrease in the transmission of resonantly interacting radiation. Finally, we made a very simple model of the carbon balance in the atmosphere, and we learned that, at a crude level, the atmosphere simply accumulates the human-generated CO_2 since the response times for CO_2 uptake from the land and oceans is relatively long (many decades). Thus, continued reliance upon fossil fuel combustion and injection of the byproducts into the atmosphere will lead to a decrease in IR radiation transport through the atmosphere, and will therefore result in an increase in the Earth's surface temperature. Feedback effects can act to either reinforce and dampen these effects (and indeed accurately modeling these types of complexities is the subject of current research), but the basic physics linking CO_2 injection into the atmosphere and Earth's thermal balance is clear.

Clearly, this situation is unsustainable. Humans are demanding access to increasing amounts of finite resources, and the consumption of these resources has a significant impact upon the global climate. There is, therefore, a clear and compelling need for new energy sources that do not inject C into the atmosphere. These energy sources must be capable of providing an adequate amount of energy to the majority and, eventually, all of the human population – otherwise unsustainable population growth ensues. The question is then obvious: What sources can be used at sufficient scale to meet this demand and not cause irreparable harm to the Earth's environment? We turn our attention to this issue in the next chapters.

Chapter 7: Estimating Future Carbon Free Energy Requirements

Introduction

In this chapter, we seek to estimate the required C-free energy that will be needed to simultaneously meet human energy demands and limit CO_2 concentrations to some desired level due to constraints imposed by global climate change concerns. Note that this discussion assumes that adequate fossil fuel resources would be available to meet any prospective human demand. If this is not in fact the case (and our brief considerations of this issue earlier in the book suggest that there may not in fact be adequate fossil fuel resources at some point), then the C-free energy requirements will be increased from the estimates made here.

To proceed, we first consider the carbon emission trajectories that might be required. We then revisit our earlier studies of future human energy demand and examine the difference between the energy release from the constrainted carbon emission trajectory this future projected energy demand. This difference must then be provided by some new C-free energy source – or the demand must go unmet. Obviously, such projections have significant uncertainties associated with them, but they serve a very useful purpose in that they give us a rough (within factor or 2 or so) estimate of the scale and scope of the required C-free energy sources. As we will see, humans will require adoption of such new energy sources on a vast scale. It is the magnitude of this scale that then quickly limits the potential new energy sources that can actually be scaled up to meet this demand.

Carbon Emission Trajectories

The simple carbon balance model described in the previous chapter provided a critical result: For timescales of a century or less, the atmosphere simply integrates the injected CO_2 from fossil fuel combustion. Therefore, we can see that if CO_2 concentrations are desired to not exceed some maximum value which is decided upon by humans, then the net injection rate of CO_2 into the atmosphere must be decreased and eventually brought to zero. Precisely how the C-emission trajectory needs to be engineered depends upon the final desired target value of CO_2 that is deemed acceptable in the atmosphere, but regardless of this precise value, the C-emission rate must at some point decrease to small (or zero) values.

The precise trajectory of C-emission that results in a given CO_2 concentration late in the 21st century has been studied in great detail using much more sophisticated versions of the C-balance model described in the previous chapter. These models include many of the complicating factors of the actual C-balance that exists in nature – e.g. spatial variations in the atmosphere, ocean, and land, feedback loops involving capture of carbon in living organisms (particularly in plants) and changes in the plant growth rate due to changes in climate, changes in ocean chemistry and so forth. A typical set of such C(t) trajectories is found below [source: IPCC, J. Holdgren, AAAS, 2007, Plenary Talk]; each emission trajectory corresponds to the best estimate for the ultimate atmospheric CO_2 concentration.

The figure shows the same essential feature that we discovered in our simple considerations of this problem: As the ultimate allowable atmospheric CO_2 concentration is lowered, the CO_2 emission rate must fall at earlier and earlier times, and must also eventually go to zero at earlier times. For example, if the ultimate CO_2 concentration is desired to lie in the

450-550 ppm range, then carbon emissions must peak in the 2020-2030 timeframe, and then must approach zero in the 2060-2100 timeframe. Keep in mind that, at the same time, the human population is expected to climb to ~9-10 billion in the same time frame (this of course assumes that the energy and other resources needed to support this population are indeed available), and that currently humanity emitted ~8 Gigatonnes of C into the atmosphere each year.

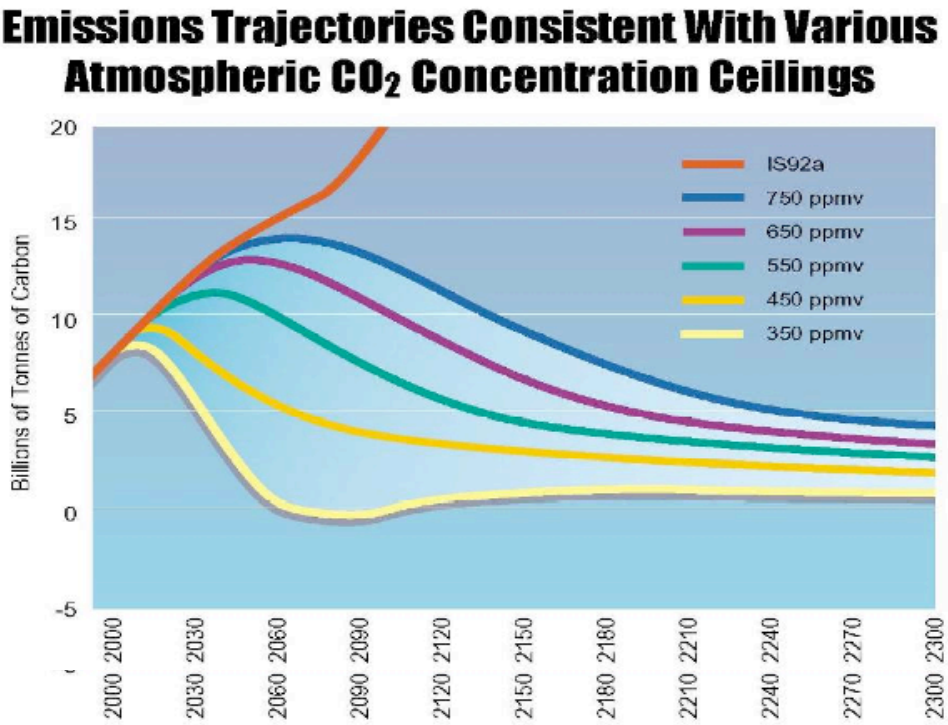


Figure 7.1: C- emission trajectories, C(t), for various CO2 stabilization scenarios. Source: IPCC, J. Holdgren AAAS 2007 Plenary Talk.

Estimating Future C-free energy demands

Once a particular C-emission trajectory has been identified, then an estimation of the future C-free energy demand can be made by finding the difference between the total future human demand and the energy content of the allowable C-emissions. Here, we summarize the results of

such an analysis that was made in 1999 by Hoffert [Hoffert, et. al., Nature, 1999] then determines the amount of net C-free power needed in the future for any given final CO_2 stabilization value. In this work, the future energy demand was estimated by first starting with the Kaya identity, which we introduced earlier, and which simply relates the energy demand to the product of the human population, N(t), the per-capita economic intensity (a.k.a. the per-capita GDP), the energy intensity E(t) which gives the energy required to produce a given unit of economic activity, and the carbon intensity, C(t), which denotes the mass of C produced per unit of energy release. Energy intensity is a rough measure of the type of economic activity that is pursued (e.g. a unit of economic output resulting from steel making will likely have a higher energy intensity than the same unit of economic output produced by a service industry such as music, accounting, and so forth, and coal has a higher C intensity that petroleum, which in turn is higher than the C intensity of natural gas combustion. C-free sources such as renewable and nuclear energy have very small C intensities arising only from secondary C emission from services provided in support of these sources). Hoffert and coworkers then examined historical data from the 20th century to estimate the variation of these parameters, and then assumed that these historical trends would continue into the future in order to estimate the total human energy demand in the 21st century. Figure shows the historical trends and projected future values for these quantities.

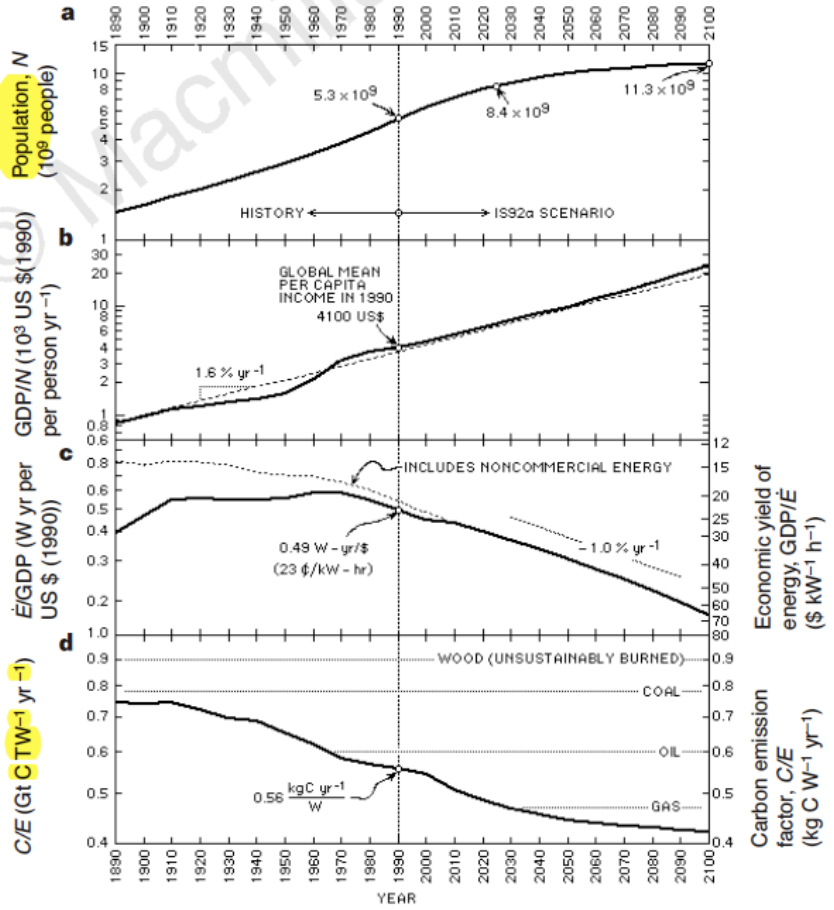
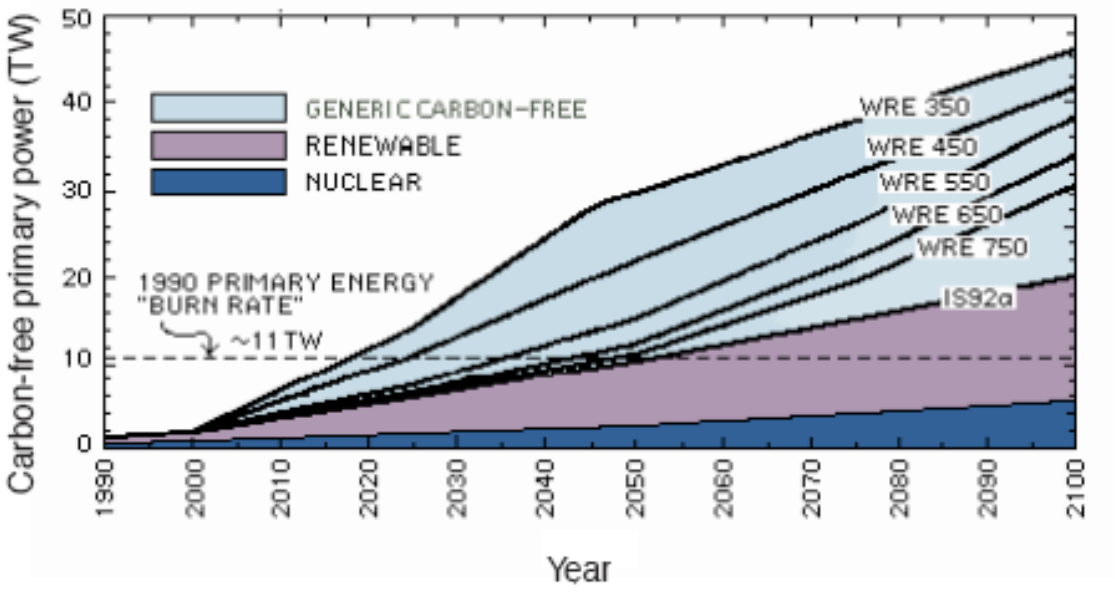


Figure 7.2: Historical and projected variation of (a) human population, (b) globally averaged per-capita annual GDP, (c) globally averaged energy intensity, and (d) globally averaged carbon intensity. Population growth is expected to saturation in the late 21st century; historical trends for per-capita GDP, energy intensity, and carbon intensity are used to project these quantities into the 21st century. Figure taken from Hoffert et. al., Nature, 1998.

These projections – which are based upon historical trends from the past 100 years – resulting in a projected 3x increase in C emission by the end of the 21st century – i.e. C emissions growth from `~6 Gigatonnes/year in 2000 to ~20 Gigatonnes/year in the year 2100. Again, this scenario assumes that adequate fossil fuel resources are located to provide for this energy demand using current technologies and sources. When such an emission scenario is compared to the results of the balance modeling shown in **Figure** , we would then expect that CO_2

concentrations would exceed 1000ppm by the late 21st century. Such values are thought to lead to unacceptable and perhaps catastrophic climate change. Hoffert et. al. then looked at the question of how much C-free power would be needed to meet future demand while simultaneously limiting C-emissions to a trajectory that results in a CO_2 concentration that does not exceed a proscribed value. No change in energy intensity or per capita economic activity was assumed; instead some hypothetical C-free energy sources were assumed to become available as necessary to meet these two constraints.



(a)

Figure 7.3: Estimated C-free power required vs. year for various values of the final stabilized CO_2 concentration in the atmosphere. For example, stabilization of CO_2 at twice pre-industrial levels (i.e. at 500 ppm CO_2) requires about 10 TW of C-free power in the year 2035. Figure taken from Hoffert et. al., Nature, 1998.

The results of the analysis are shown in Figure , which shows the estimated future C-free power requirement (which is simply the time-rate of delivery of energy) needed to meet projected human energy demand for various values of ultimate atmospheric CO_2 concentration. For example, the analysis shows that, in order to limit the ultimate CO_2 concentration to 550 ppm

(i.e. roughly double the preindustrial value) then by 2050 the world would need to produce ~14-16 TW of C-free power, and in 2100 this demand would have grown to something in the neighborhood of 30 TW of C-free power. To put these estimates in perspective, the total global power demand today is approximately 14-15TW, of which ~80% is provided by fossil fuel consumption.

Clearly, reductions in the energy intensity, which reflect a mixture of energy efficiency measures as well as the type of economic activity has a significant impact on these estimates. Hoffert et. al. also examined the effect that variations in the rate of decrease in energy intensity (which correspond to an increase in energy efficiency) have on these estimates. They found that if the historical rate of decrease in energy intensity could be doubled (i.e. the energy intensity decrease could go from a value of -1%/year (which is close to the historical record as seen above) to a decrease of -2%/year, then the c-free power required to maintain 550ppm CO_2 concentration could be decreased from 14-16TW in 2050 to values of 3-34 TW in 2050. Clearly then, efficiency gains are critical to this problem as well. It is unclear if such energy intensity decreases could be obtained on a global scale. Certainly in highly developed economies where energy consumption is very large (e.g. in the U.S.) such rapid energy intensity decreases have been obtained [REFERENCE]. However, as we have seen earlier, the majority of the current human population lives with access to little or no energy resources to speak of and thus it seems unlikely to expect that those regions of the world would experience such a rapid decrease in energy intensity.

Clearly the estimates for the C-free energy and power requirements in the middle of the 21st century vary widely depending upon the assumptions used for economic growth and

improvements in human quality of life, as well as on assumptions about energy intensity and efficiency gains. However, it is clear from the considerations discussed here that if we are to have the energy necessary to drive these changes and, at the same time, avoid large changes in Earth's climate, then by the middle of the 21st century the world will need to have in place C-free energy sources and conversion technologies that are capable of producing something in the range of ~10-15 TW of C-free power. These sources will need to meet transportation, electricity production, and heat sources for industrial, commercial and residential demands. These requirements will then increase even further later in the 21st century, and by that point in time will likely exceed current-day (2009) energy demand from all sources.

As we begin to consider what sources and conversion technologies might be capable of meeting this demand, it is essential to bear in mind the expected scale and scope of this demand. As we will see next, there are many potential C-free energy sources to consider; however when we impose the requirement that these new sources be capable of meeting a significant fraction (say ~10% or so) of this expected new demand, we find that the available choices become much more limited.

Chapter 8: Overview of Primary Energy Sources

The previous discussion has shown that there is a clear link between how human beings live their lives and their access to energy, which enables them to undertake activities which would otherwise be difficult or impossible to do. At present only a fraction of the current human population has access to a reasonable level of energy (in particular, about a third of the human population has access to more than adequate energy supplies, another roughly one third are increasing their energy access and are approaching adequate supply, and then the final one third have nearly no access to energy supplies other than what they can acquire from the consumption of biomass (mostly dung and scavenged wood supplies). We then found that even if those regions which currently dominate the world's energy usage were to decrease their per capita energy use, we can still expect significant (e.g. factors of 2 or more) increase in global energy use in the 21st century. Given the fact that the current world energy economy is largely based upon fossil fuels, which results in the emission of significant quantities of CO_2, and that this emission has serious environmental impacts, it seems likely that the world will need to transition to net carbon-free energy sources. Furthermore, the fact that these fossil fuel resources are finite implies that (eventually) the rate of extraction of the fossil fuels will first peak and then begin to decline. When the rate of extraction of these resources begin to fall below demand, serious economic and social impacts would then occur. This fact also forces us to consider what alternative energy sources can be considered to power human civilization into the latter part of this century and beyond. Lastly, our considerations provided an estimate of the scale of this required energy. Any proposed new energy source must be capable of providing some significant fraction of this energy if it is to be material to the problem at hand.

Summary of Potential Carbon-free Energy Sources

In this text, we are interested in primary energy sources that can be scaled to the requisite magnitude to meet human energy demand through the 21st century while not emitting significant CO_2 into the environment. Thus even assuming widespread use of carbon capture and sequestration, large scale use of fossil fuels will eventually be unable to meet this demand due to

the fact that these resources are finite and thus unlikely to be able to meet human energy demands into the indefinite future. Furthermore, technologies such as hydrogen-powered fuel cells are not primary energy sources, but are rather energy carriers and energy conversion technologies that might offer replacements for current technologies. Thus, as such, they fall outside of the scope of this text.

Thermodynamics indicates that primary energy sources must either receive energy input from outside of the Earth (e.g. solar energy input), or must be based upon converting energy stored within some material that is available on Earth. This energy must be converted into usable form and delivered to its end use point. The quantity of energy available must be sufficient to meet the human demand, taking into account the conversion efficiencies, geographic and temporal availability of the resource and demand. If the energy is provided by capturing energy stored within a resource on the Earth, then clearly this will be a finite resource and thus subject to the limitations inherent in such resources. Thus any such new energy source must have sufficient widely available resources to meet human energy needs for a long enough period of time (likely centuries or more) to be of interest. If the energy source is a renewable resource, i.e. it regenerated by the input of energy from outside of the Earth's boundaries, then the rate of input of energy must be adequate to meet the anticipated human energy demand.

With these elements in mind, there are a number of potential primary energy sources to consider. They include:

- Wave and tidal power
- Ocean currents
- Ocean thermal gradients
- Hydropower
- Wind power
- Solar power
- Geothermal energy
- Nuclear energy

Some of these potential energy sources scale sufficiently; as we shall see, others do not. Let us briefly examine them to determine which scale and which do not. We will then focus our

more detailed study on the energy sources that appear to scale to meet a significant fraction of human energy demand.

Wave Power

Wave action on the surface of the ocean is driven by wind. The velocity gradient that results in the region just above the water results in a shearing force on the water surface. This force in turn can generate an instability which leads to the generation of surface waves on the water, resulting in periodic oscillation in the surface height of the ocean. Typically, the period of the oscillation is in the range of 5-20 seconds or so, and the amplitude of the oscillation is of the order of 1 meter. We are interested in estimating the energy resource that might be extracted from these waves. Figure 8.1 below provides a schematic of an idealized sinusoidal surface wave.

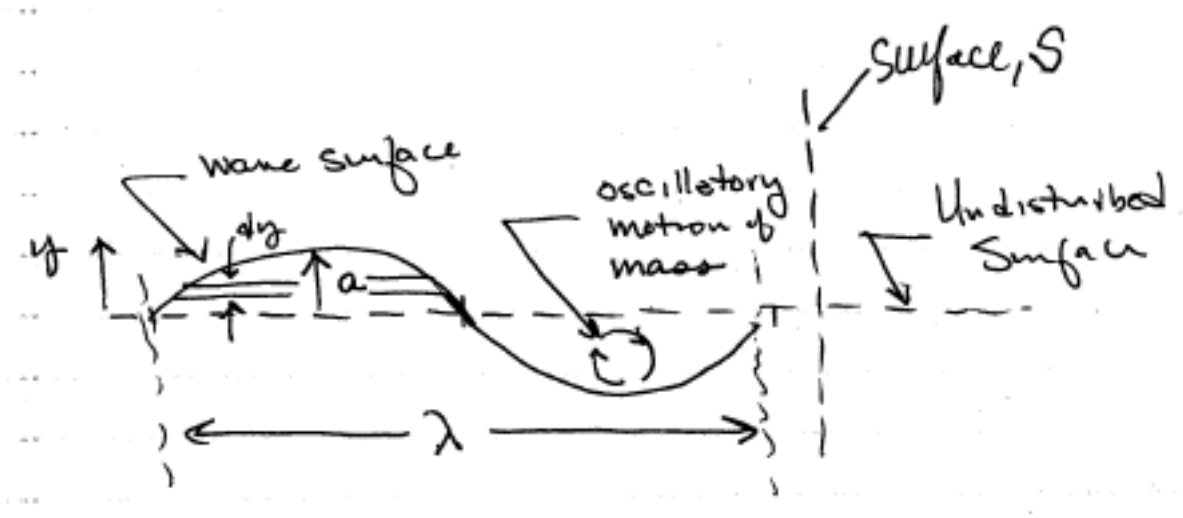


Figure 8.1 Schematic of a surface wave, showing the perturbed sea surface, the wavelength the oscillatory motion of the water associated with the passage of the wave, and the vertical displacement.

We will examine an idealized sinusoidal surface wave to determine an estimate of the energy potential of this resource; the results will not be terribly different if multiple wavelengths were to be included.

The vertical perturbation of the sea surface can be taken to be given as $y(t) = a\sin(\omega t - kx)$ where a denotes the magnitude of the wave perturbation, ω denotes the angular frequency of the wave, $k = \frac{2\pi}{\lambda}$ denotes the wavenumber of the wave with λ giving the wavelength, and x denoting the direction of wave propagation. This simple description of the wave will suffice as long as the amplitude, a, is much smaller than the water depth, d.

If one were to identify a position on the wave of fixed phase, and then follow that point as the wave propagates, the speed of propagation would be the phase velocity of the wave which is given as $v_{ph} = \frac{\omega}{k}$. It can be shown for these types of small amplitude surface waves that $v_{ph} = \frac{g}{\omega}$ where here g denotes the Earth's gravitational acceleration. Notice that $v_{ph} = v_{ph}(\omega)$ and thus the speed of propagation depends on the frequency; such waves are said to be "dispersive" since a tightly spaced set of waves (known as a wavepacket) will tend to breakup or disperse as they propagate in space. The speed at which such a localized wave packet will move is called the group velocity, and is given as $v_g = \frac{\partial \omega}{\partial k}$, and plays an important role in determining the energy that can be extracted from waves. Using the relations above, we can write the so-

called dispersion relation for surface waves as $\omega^2 = gk$. We can then easily find that the group velocity is given as $v_g = \frac{1}{2}\frac{g}{\omega} = \frac{1}{2}v_{ph}$.

Now lets denote the total energy contained within one wavelength, per unit depth, to be given as $U_{tot} = U_{pot} + U_{KE}$, and is composed of a potential energy component U_{pot} that originates from the vertical perturbation of the water surface within the Earth's gravitational potential, and an kinetic energy component U_{KE} that arises from the periodic circular motion of a fluid element as shown in Figure 8.1. It can be shown that these two components are equal to each other. Thus if we determine the contribution of one of these components to the wave energy, we have then determined the total wave energy.

Let us therefore examine the potential energy contribution since this evaluation is straightforward. Considering the small incremental vertical displacement, *dy,* as shown in Figure 8.1, we can write the incremental perturbation to the potential energy arising from this element as

$dU_{pot}(y) = \rho g y x dy$.

And the distance x denotes the horizontal distance of the upward going portion of the wave. At the end of our analysis we shall then need to account for the fact that there is also an equally long component of the wave in which the water surface is displaced downwards. The resulting combination of downward fluid element motion and a decrease in the local fluid density (i.e. the density goes from that of water to that of air for this element) then results in an equivalent contribution to the wave energy.

For simplicity we evalution the energy at t=0. We then have $y(t) = a\sin(kx) = a\sin\left(\frac{2\pi}{\lambda}x\right)$ which can be used to find $x = \frac{\lambda}{2\pi}\sin^{-1}\left(\frac{y}{a}\right)$. We can now write the incremental potential energy as $dU_{pot}(y) = \rho g \frac{\lambda}{2\pi} y \sin^{-1}\left(\frac{y}{a}\right) dy$.

We can then find the total potential energy by integrating this over the range [0,a]. We then find that $U_{pot} = \rho g \frac{\lambda a^2}{16}$ for the region with $0 < y < \frac{\lambda}{2}$. As mentioned above, we must also account for the potential energy of the wave for the region $\frac{\lambda}{2} < y < \lambda$. Taking this into account, and then also accounting for the kinetic energy of the oscillatory motion of the water, we find the total wave energy per unit length parallel to the face of the wave to be given as $U_{tot} = \rho g \frac{\lambda a^2}{4}$. Finally we complete our analysis by noting that this energy is stored across a wavelength, which passes a point of observation in a wave period $T = \frac{2\pi}{\omega}$. Taking into account the wave dispersion relation, we can then find the wave power per unit length to be given as

$$P_{tot} = \frac{U_{tot}}{T} = \rho g^{3/2} \frac{\lambda^{1/2} a^2}{4\sqrt{2\pi}}.$$

If the expression on the right is evaluated in MKS units, then P will have units of Watts/meter. Let us now evaluate this expression for typical values. We take T~10 seconds, and a~1m. With a water density of 1000 kg/m^3 and g~10 m/s^2 we then find P~40kW/m. The actual conversion efficiency of wave energy devices has been in the range of a few 10%, say 20-40% or so. Thus

we see that actual power delivery would lie in the range of 10 kW/m of length, or 10 MW/km of coastline. As a rough estimate, we can estimate each inhabited continent on the Earth has a coastline of a few 1000s km. Taking a typical value of 3000km for a continent, we could estimate a maximum possible upper limit of global wave power generated from near-coastal region stations to be of the order $10\text{-}15 \times 10^3$ km X 10 MW/km = 100-150 GW of power. While this is quite substantial, it is less than 1% of the estimated global power demand today, and will be even a smaller fraction at mid-century due to the anticipated growth in global demand for energy. Thus these elementary considerations show that wave power will not contribute in a major way to global energy demand, and thus we shall not consider it further here. This is *not* to say that wave power will not, perhaps, be an important contributor to local and regional energy demand in areas where weather, ocean and geography conspire to make wave energy a serious contender. However it is clear that wave power cannot play a globally significant role in the global energy economy.

Tidal Power

Tidal variations are caused by the diurnal variation of the position of the moon and sun, whose gravity perturbs the distribution of the ocean water. Again, the displacement of the water due to tidal effects is highly predictable and can be (and in some places is) captured and converted for useful purposes. Let us estimate here this resource, and compare this estimate against current and expected global demand for energy in order to determine if tidal power can play a materially significant role in future global energy economy. Our analysis will make reference to Figrues 8.2 and 8.3, which provide a schematic view of the small variation in net gravitational

acceleration due to the combined effect of the Earth and moon (Figure 8.2) and the resulting impact that this has on the surface of the ocean (Figure 8.3). As the Earth rotates, the locally observed water height will oscillate as the Earth turns beneath the perturbed ocean surface. The goal of the analysis is to first estimate the height of this oscillation based on elementary considerations, and then use this estimate to determine how much power could potentially be extracted from the resource and compare this to anticipated future human energy demand. The key to the approach is to note that the surface of the ocean will sit on an equipotential surface. The gravitational potential of a parcel of fluid is determined by the fluid density (which is a constant), the gravitional acceleration, and the height of the parcel relative to an arbitrary reference position. Thus if we can determine the acceleration we can solve the problem.

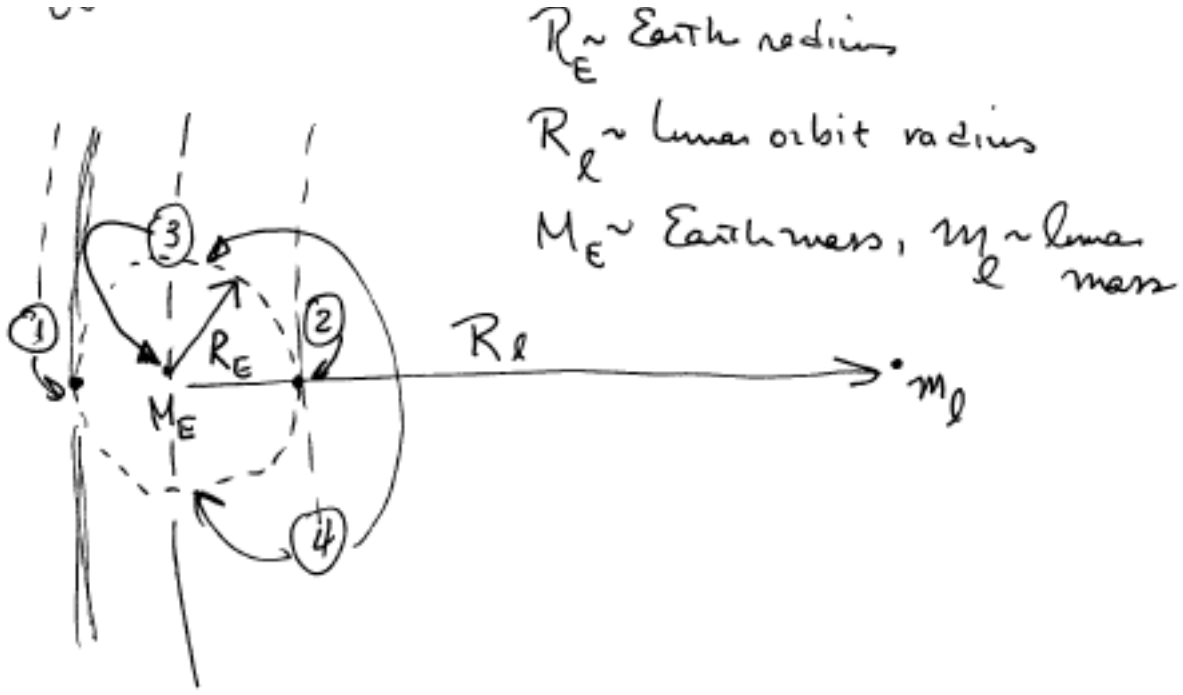


Figure 8.2: Schematic of the Earth-moon system showing the point away from the moon (point 1) which experiences a net gravitational acceleration that is slightly less than that experiences at the point closest to the moon (point 2). Points 3 and 4 experience a net gravity acceleration that is intermediate to that of points 1 and 2.

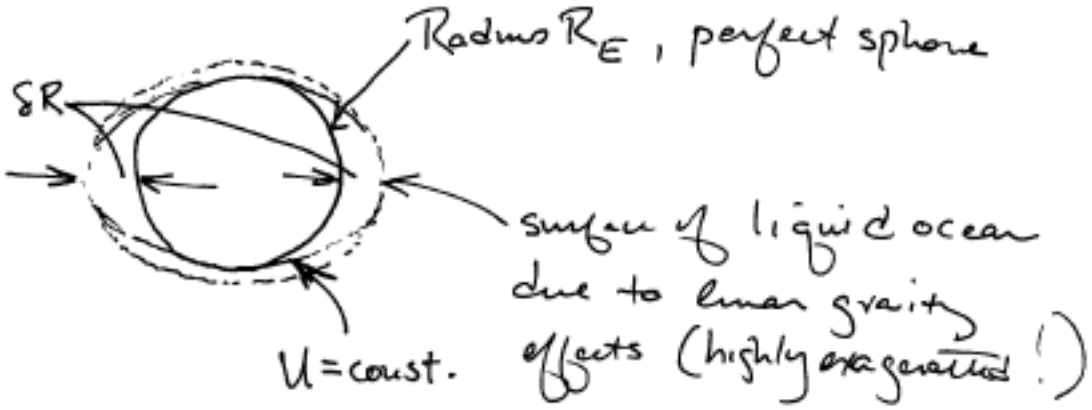


Figure 8.3: The surface of a liquid within a gravitational potential well will lie at an equipotential. For the Earth's oceans, the liquid water surface will acquire a slight deviation away from spherical due to the variation in the net gravitational acceleration arising from the moon (and to a lesser extent, the sun).

Refering to Figure 8.2, the difference in the force on a fluid parcel of mass m due to the lunar mass, m_l, at positions 1 and 2 on the Earth is given as

$$F_2 - F_1 = \frac{mm_l G}{\left(R_l - R_E\right)^2} - \frac{mm_l G}{\left(R_l + R_E\right)^2}.$$

Normalizing this difference to the lunar gravitational force at position 3 we find the relative variation in the lunar gravitational acceleration on the parcel of mass m to be given as

$$\frac{F_2 - F_1}{F_3} = \frac{1}{\left(R_l - R_E\right)^2} - \frac{1}{\left(R_l + R_E\right)^2}.$$

Since we know that $\frac{R_E}{R_l} \sim \frac{1}{60}$ we then can see that the lunar gravity varies by about +/-3% from point 1 to point 2. The ratio of the average lunar gravitational force on mass m to the Earth's gravitational acceleration is given simply as

$\frac{F_E}{F_3} = \frac{M_E}{m_l}\left(\frac{R_l}{R_E}\right)^2 \sim 80 \times 60^2$. Thus we can find the deviation of the net (i.e. Earth + Moon) gravitational acceleration across points 1-2, normalized to Earth's gravitational acceleration, as

$\delta F_{net} = \frac{F_2 - F_1}{F_3}\frac{F_3}{F_E} \approx 1 \times 10^{-7}$. Finally, we note that the potential energy, U, of a unit mass parcel of fluid located at the Earth's surface, is given as $U = F \cdot R$ where F is the acceleration on the unit mass and R is a vertical displacement relative to an arbitrary reference plane. Let us take the reference plane to be the surface of spherical ocean that would occur if there were no lunar tides. We can differentiate this expression to write

$\partial U = \partial F \cdot R + F \cdot \partial R$.

However, we now note that in the presence of the lunar tidal effect, the surface of the ocean will still lie on an equipotential, and thus we will have $\partial U = 0$. Thus we can find the displacement of the ocean surface due to the tidal effect as

$\partial R = -\frac{\partial F}{F} \cdot R$.

Using the result above, we then find that $\partial R = -\frac{\partial F}{F} \cdot R \approx 1 \times 10^{-7} R_E \sim 1m$, which is consistent with typical tidal heights of ~1-2m.

This tidal variation occurs on a timescale corresponding to the time needed for the Earth to rotation approximately ¼ of a turn, i.e. about 6 hours. In some special geographic regions, the ocean water is partially enclosed in a basin. If the period of oscillation, or "sloshing" of the water within this basin is close to this period, then a resonant oscillation can be setup. In this case, much larger tidal oscillations (sometimes up to 10m in a few locations on the Earth) can develop. However, typically the values are of order of a few meters at most.

Tidal power systems are then arranged as follows. A basin region is enclosed from the open sea by the construction of a barrier between the open sea and the basin. This barrier is designed to allow the incoming tidal waters to flow into the basin. Then, when the tidal reverses and the flow begins to move out to sea, the basin traps or retains the water within the basin. As the open sea level then recedes, the resulting height differential can be used to create a potential energy difference. The trapped basin water can then flow through a suitable mechanism that extracts the potential energy and converts it into useful form (usually electricity). Obviously this scheme has a temporal variation that oscillates. If the basin has a surface area, A, and we have a tidal height, h, then the basin volume is obviously V=Ah. This mass has a potential energy, U, given as $U_{pot} = \rho g A h^2$. This potential energy can be extracted on a tidal period T~6 hours. Taking into account the oscillatory nature of the process then gives an average power extraction rate P as

$$P_{ave} \approx \frac{1}{2} \frac{\rho g A h^2}{T}.$$

Let us now make a numerical estimate of the average power that is available from this scheme. Let us assume that the basin has a surface area of 1 km^2. For tidal heights in the range of [1,10] meters, we then estimate $P_{ave} \approx 3\times10^5 - 3\times10^7\ MW/km^2$. The larger values would only hold in few special geographic locations around the world; the intermediate and smaller values could be achieved in most coastal regions. Keeping in mind the anticipated future world energy demand, we estimate that to produce 1 TW of average power we would then require tidal basin areas ranging from $3\times10^6\ km^2$ (for height h=1m) to $3\times10^5\ km^2$ (for tidal height h=3m). We note that the largest proposed tidal power system (proposed for the Kamchatka peninsula area in the Far East of Russia) has an area of ~20,000 km^2 with a tidal height h=9m, and an estimated average power of ~100 GW (i.e. 1% of estimated future carbon free power demands). The largest existing tidal power system, located in France, has an average power capacity of about 200 MW. Thus it would seem that tidal power, while using well established physical principles and technologies, and which produces power on a highly predictable schedule, nonetheless is limited to contributing small fractions of future global demand. Thus, like wave power, we shall not consider it further here.

Ocean Currents

The world's oceans exhibit steady, large-scale currents driven by a combination of wind surface stresses, variations in density caused by salinity and thermal effects, and the Earth's rotation. The kinetic energy contained within these flows is much larger than the annual human energy demand, and is replenished continually by the input of energy from the sun. However, to capture this energy, large arrays of submerged turbines would need to be deployed in suitable locations.

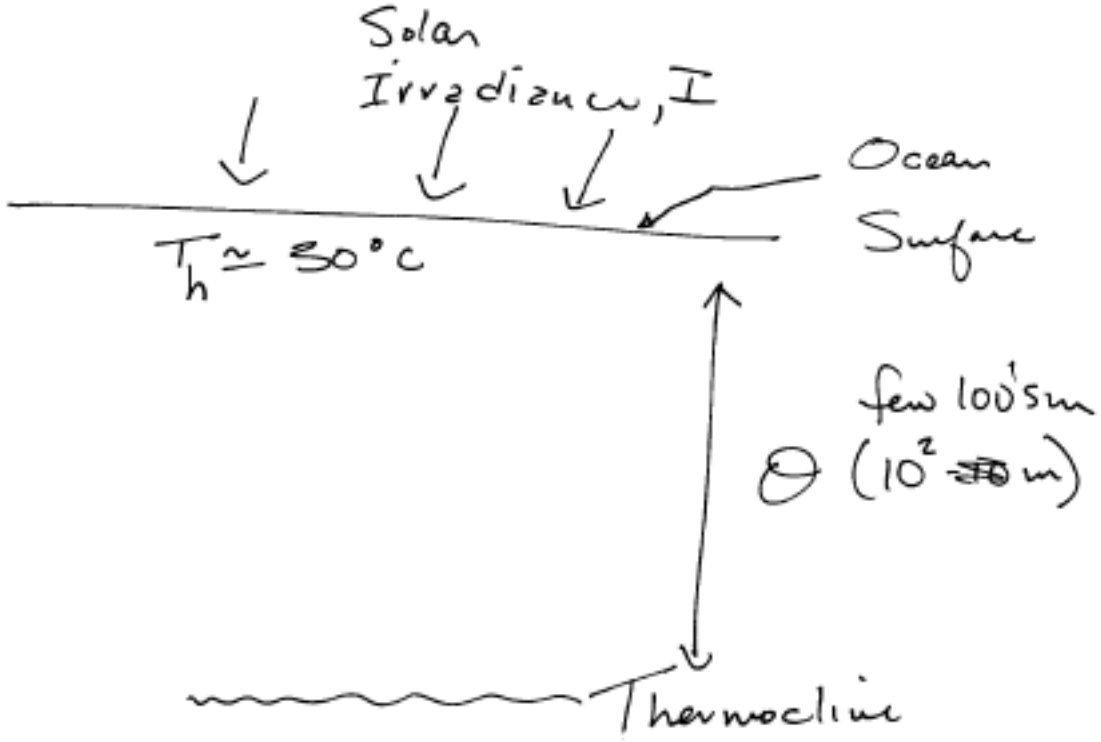
Although there are plans underway to do so in regions where suitable currents flow in the region near the shore, it appears unlikely that this potential energy source can be scaled to meet even a small fraction of future human energy demand. Thus, we do not consider it further here.

Ocean Thermal Conversion

Ocean thermal energy conversion (OTEC) uses the temperature difference between the tropical region surface waters and the deeper (~1km or so) waters. This temperature difference can be of order 10-20 deg C, and can be used in a thermal power conversion cycle to produce electricity. The heat capacity of the ocean in this surface layer is huge compared to annual human energy demand. Thus on the face of it, one might conclude that this resource presents a large opportunity. Let us consider it then in more detail.

Figure 8.4 shows a schematic view of the near (1km) ocean surface. The incident solar irradiance, I (with an average value of about 250 W/m^2) is absorbed within the first 10 or so m of the surface of the ocean. This results in a surface temperature of as much as 30 deg C. The deep ocean is much colder, with a temperature near 0 deg C. Since colder water is more dense than warm water, this results in the formation of a stably stratified layer in the upper region of the ocean. This layer is known as the thermocline, and has a depth of a few 100s meters up to ~ 1km. The resulting temperature differential can in principle be used by a suitably designed heat engine to perform useful work. OTEC technology proposed to do this to provide a renewable source of electricity. The concept would have a long pipe extend from the surface down to the region of the thermocline. This pipe would pump cold water up to the heat engine, which is located on the surface. The warm surface water then acts as the heat source for the hot reservoir

of the engine, and the cold water acts as the heat sink. For this scheme to be sustainable, the rate of power extraction cannot then exceed the rate of heat input into a region with ocean surface area A.



Figrue 8.4: Schematic of near ocean surface showing surface and deep (~1km) water temperature differences across the ocean thermocline.

To evaluate this concept, we take the idealized Carnot conversion efficiency as the upper limit on thermal conversion efficiency. We have $\eta_C = \frac{T_h - T_c}{T_h} \approx \frac{303K - 273K}{303K} \approx 5\%$; actual values will be about half of this upper limit. For such a system to operate in a sustainable manner (i.e. so

that we do not cool off the upper ocean surface region too much), the system must satisfy a power balance given as

$$P_{out} = \eta\left(I - I_{IR}\right)A_{surf}$$

where $P_{out} = \eta\left(I - I_{IR}\right)A_{surf}$ denotes the output power of the system, I_{IR} denotes the infra-red emission of the warm ocean surface back towards the atmosphere and space, and A_{surf} denotes the effective surface area of the ocean affected by the OTEC power plant. Typically I_{IR} ~ 150 W/m^2 to a clear sky in the warm tropical regions of the ocean. Thus for a 1 GW class OTEC power system operating at the maximum upper limit of conversion efficiency the area affected would be about 14 km^2. If the thermocline is located at a depth of 1km, then this corresponds to ~14 km^3 volume of water.

The mass flow rate through the 1GW 12OTEC system would be given as

$$P_{out} = \eta \dot{m} C_p \Delta T$$

Taking $\Delta T = 30K$ and $C_p = 4kJ/kG$ we then find a mass flow rate through this system of ~160 Tonnes/second, which corresponds to a volumetric flowrate of 160 m^3/sec of water. For the given volume affected by the plant, it would then take ~30 years to move the 14km^3 of water through the OTEC system. This timescale is considerably longer than the time needed to reheat the upper layer of the ocean due to turbulent mixing. If this upper layer has a depth, *d*, and due to wave action at the surface there is a turbulent mixing process that occurs within the upper regions of the thermocline, resulting in a turbulent diffusion coefficient, D (with units of m^2/sec)we can then estimate the turbulent mixing time in the upper layer as

$$\tau_{mix} \sim \frac{d^2}{D}.$$

Usually turbulent mixing occurs via a random-walk like process occurring due to the existence of multiple turbulent eddies with a scale size, l_{turb}, and an overturning timescale τ_{eddy}. This then results in a turbulent diffusion coefficient that scales like $D \sim \frac{l_{turb}^2}{\tau_{eddy}}$. Thus the mixing timescale in the upper regions of the thermocline due to turbulence will be given as

$$\tau_{mix} \sim \frac{d^2}{l_{turb}^2}\tau_{eddy}.$$

With $l_{turb} \sim few\, m$ (corresponding to the vertical displacements from surface waves) and $\tau_{eddy} \sim few\, hours \sim 10^4$ sec and d~100-1000m, we can then estimate that the mixing time is of the order of weeks to a few months. The key point then is that the rate of heat extraction from the OTEC system could then be slow enough that the surface can be successfully reheated by the sun, and turbulent mixing can then carry this heat down to depths of ~100m at a rate that is much faster than the rate of heat extraction. Thus the system could in principle be sustainable and operated on an indefinite basis (at least until some equipment failure forces the system to cease for maintenance or repair).

However, note that in this simple conceptual analysis a 1GW system requires a surface area of 14 km^2. The question then becomes: how many such systems could in principle be stationed in a near-off shore area, produce useful power, and then transmit this power to the shore for human use? To get a crude estimate, we estimate the coastal shoreline length in regions with warm surface waters. A quick look at the globe might suggest that this length is of the order of 10,000

km or so across the whole globe. If such stations could be positioned off shore to a maximum distance of about 20-50 km, then the total available surface area in these near coastal regions is quite large, ranging from values of 200,000 – 500,000 km. If a reasonably small fraction (1% so as to not cause significant environmental impacts in the coastal regions?) of this area were devoted to OTEC systems, then we could expect to produce between 200GW and 500GW of power from OTEC systems distributed globally. If the relative surface fraction devoted to OTEC could be increased further, say to 10%, then these values could increase by a factor of 10 or so and begin to approach a significant fraction of anticipated global power demand. However this would require the use of several 1000s of large OTEC power plants operating in near coastal waters. Given that there are no such large systems in operation today, and that more detailed engineering evaluations of the costs, lifetimes, and reliability of such systems have not been encouraging, OTEC systems are not under large scale development and deployment today. Thus we will not consider them in detail in this text.

Hydropower

Hydropower refers to the capture and conversion of the gravitational potential energy available in water flowing from elevated regions down to lower regions and, eventually, to the oceans where the hydrological cycle (powered by the sun of course) then returns water to the higher elevations. Estimates of the world's hydropower potential [REFERENCE] show that this resource can meet ~3-10% of the world's energy demand, but cannot be increased substantially above this range of values. Certainly, there are localized regions whose geography permits much higher fractions of electrical energy production via hydropower (e.g. the country of Norway

provides the majority of its electricity needs using hydropower). However, there are many more regions where the geography is not so suitable to widespread use of this technology. Furthermore, the technology of hydropower conversion is well developed and is unlikely to undergo significant advances that will permit wider use of this energy resource. Thus, we do not consider hydropower in this text.

Remaining Options

This leaves only four primary energy sources and conversion technologies to examine for meeting future world energy demand:

- *Wind energy*
- *Solar energy*
- *Biologically or synthetically produced fuels*
- *Nuclear energy*

In the following chapters, we take up each of these sources in turn, examining the magnitude of the available resource and identifying the essential physical mechanisms by which this resource can be converted into useful form.

Chapter 9: Wind Energy

Introduction

The kinetic energy contained in the wind can be captured and converted into mechanical work and from that to electrical energy. The use of this resource for electricity production is becoming more commonplace, and as of this writing wind generated electrical power generation capacity is growing at 20-30% per year and currently represents a few percent of the U.S. electrical power generation capacity. Relevant questions for us to consider here include:

- How does a wind-turbine function, and what is the maximum possible energy conversion efficiency?
- How does wind variability affect power production?
- What is the potential wind energy resource, and is it sufficient to meet a significant fraction of the world energy needs?
- What considerations impact the ultimate potential for wind generated electrical power?

Estimating Maximum Theoretical Resource

The Earth's wind is driven by heat input from the sun, and thus operates as a heat engine. Thus one way to estimate the energy resource that might be available from capturing wind energy is treat the atmosphere as a heat engine, estimate how much of the heat input from the sun can theoretically be converted into mechanical motion of the atmosphere, and then estimate what fraction of this kinetic energy could be captured and converted by turbines located in the lower reaches of the atmosphere. This top-down approach provides a simple, rough estimate of the potential of wind power, and provides a useful comparison for more detailed, bottom-up approaches to the same estimation.

Figure 9.0 provides a highly simplified schematic view of a one-dimensional slab-like model of the atmosphere operating as a heat engine which converts the temperature gradient within such an atmosphere into mechanical motion of the atmospheric fluid. There in heat input from the sun , Qin. This radiation is in the form mostly of visible light, to which the atmosphere is reasonably transparent as we saw in earlier discussions. Thus this radiation heats the deeper regions of the

atmosphere as well as the surface of the Earth, which is heated to an average temperature Th~290 K or so. The upper reaches of the atmosphere are at a significantly lower temperature, Tc~230K or so. Taking these two temperatures as the hot and cold reservoir temperature, we estimate the maximum conversion wefficiency to lie in the range of 20% or so; in reality the conversion efficiency will lie in the rnage of 5-10% of the thermal power input.

A significant portion of the incident solar radiation is directly re-emitted to space as infra-red emission from the atmosphere itself; this radiation is lost of the "engine" and is not available to drive mechanical motion. In the average input radiation flux is ~250 W/m^2 and about half of this incident radiation is immediately re-radiated to space, then we have a net radiation flux of about 120 W/m2 available to drive the engine. With a ~10% conversion efficiency, this implies that there is about 10 W/m^2 of mechanical power being dissipated per unit surface area. This power is then distributed across the vertical depth of the atmosphere. If we take the atmosphere to be about 10 km deep, and our wind turbine systems have a height of order 100 m, then if the mechanical work in the atmosphere is uniformly distributed and dissipated within the atmosphere, then a few percent, or about 0.2-0.5 W/m2 of mechanical power is available per unit surface area to be captured in the first ~100 m thick layer of the atmosphere. Studies of the wind speed distribution across the globe indicate that about 10^6 km^2 = 10^{12} m^2 of land area and near-coastal ocean area could potentially be made available for use as wind farms. These simple estimates indicate that or order 100-1000GW of power could be estracted from the wind if a sufficient number of turbines were distributed in these wind-rich regions. This is not insignificant, but at the same time the larger values of these estimates approach perhaps 10% of anticipated global demand for power. Thus wind seems to be capable of providing a significant portion of future human energy demand. Furthermore, of all the renewable technologies, wind power seems to have the best economics (at least at the moment) and thus has enjoyed a healthy growth rate and market adoption in recent years. Thus the resource and technology is worth a closer examination. In this chapter, we provide an introduction to the aerodynamics and mechanics of wind turbines, we examine the theoretical and practical conversion efficiencies of these systems, estimate the wind power potential using several other approaches, and summarize key issues that must be overcome in order for wind power to achieve its full potential.

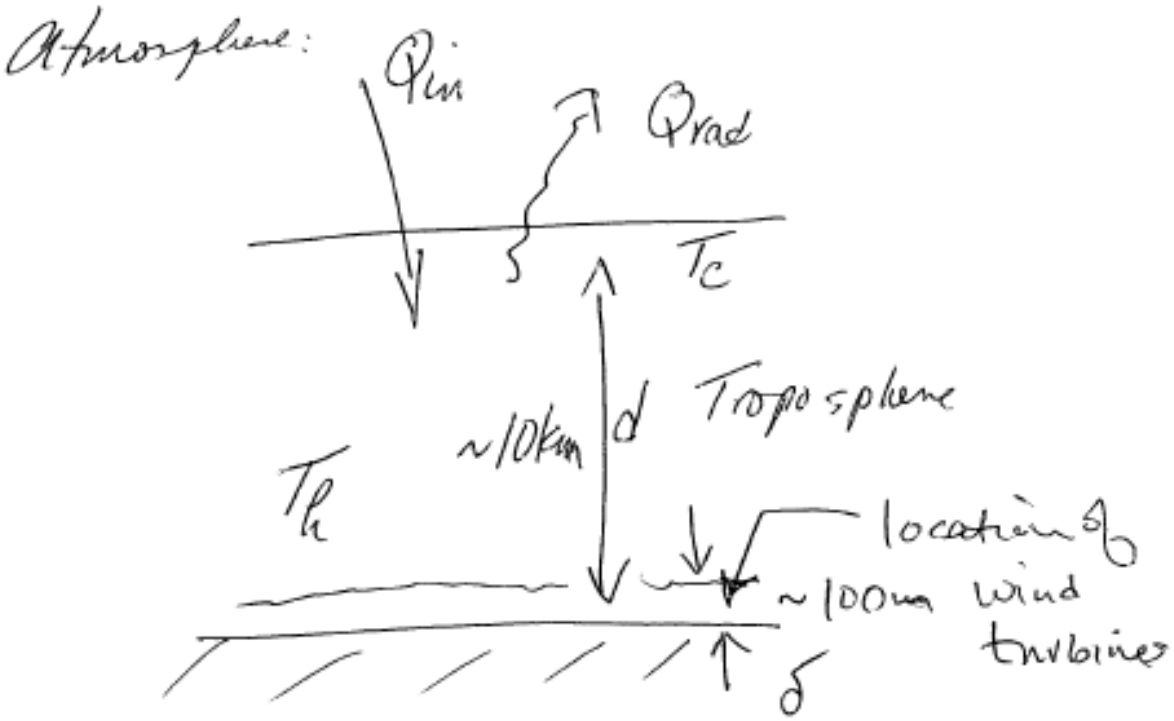
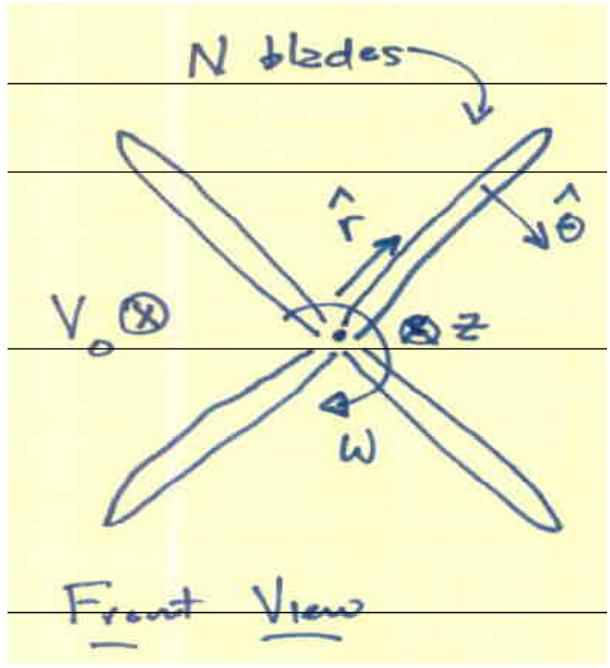


Figure 9.0 Schematic view of a one-dimensional atmosphere as a heat engine.

Wind Turbine Mechanics and Aerodynamics

We first introduce the basic operational principles of a wind turbine by considering the aerodynamics of the turbine blades immersed within a uniform flow of air. The geometry is presented in the figure below.



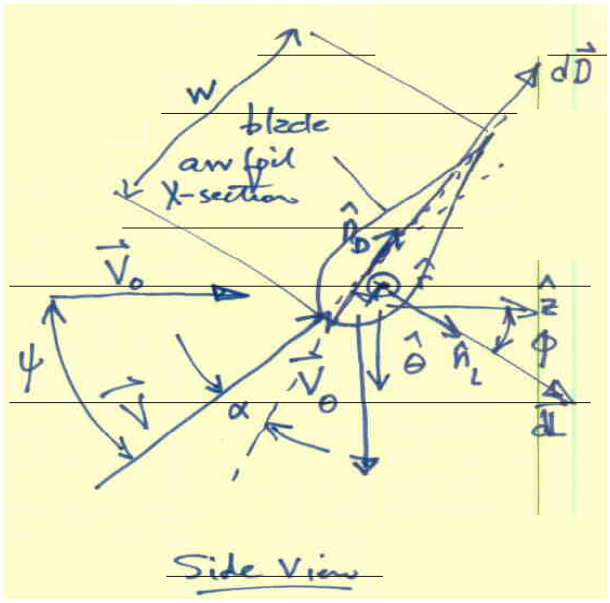


Figure 9.1: Left panel: Front view of a wind turbine immersed within a uniform freestream flow. Right panel: Side view of a turbine blade cross-section showing the free-stream velocity vector component V0, the azimuthal velocity component Vq due to rotation, the total velocity vector V, as well as the local normal and axial unit vectors along with the incremental lift and drag forces.

The total velocity vector of the wind relative to a position at radius r on a blade spinning with an angular rotation frequency ω is given as

$$\vec{V} = \vec{V}_\circ - \vec{V}_\theta = V_\circ \hat{z} + V_\theta \hat{\theta}$$
$$V_\theta = \omega r$$
$$\therefore \vec{V} = V_\circ \hat{z} + \omega r \hat{\theta}$$

From aerodynamics, we know that the incremental lift and drag on an airfoil with an incremental span *dr* are given in terms of the lift and drag coefficients as:

$$\vec{L} = \frac{1}{2}\rho V^2 C_L \cdot w \cdot dr \cdot \hat{n}_L$$
$$d\vec{D} = \frac{1}{2}\rho V^2 C_D \cdot w \cdot dr \cdot \hat{n}_D$$

where w denotes the chord length of the turbine blade.

The net force in the $\hat{\theta}$ direction is then given by the scalar product of these incremental forces with the azimuthal unit vector

$$d\vec{F}_\theta = \left(d\vec{L} + d\vec{D}\right) \cdot \hat{\theta}\hat{\theta}$$
$$= \frac{1}{2}\rho V^2 w dr \cdot (C_L \hat{n}_L + C_D \hat{n}_D) \cdot \hat{\theta}\hat{\theta}$$

From geometry, we can write the unit vectors as

$$\hat{n}_L = \cos\phi \hat{z} + \sin\phi \hat{\theta}$$

and

$$\hat{n}_D = \sin\phi \hat{z} - \cos\phi \hat{\theta}.$$

We then can write the incremental force in the azimuthal direction as

$$d\vec{F}_\theta = \frac{1}{2}\rho V^2 W dr \cdot (C_L \sin\varphi - C_D \cos\varphi)\hat{\theta}$$

This force then exerts a moment about the z axis given as :

$$d\vec{M} = \vec{r} \times d\vec{F}_\theta = \frac{1}{2}\rho V^2 W r \cdot (C_L \sin\varphi - C_D \cos\varphi) dr\, \hat{z}$$

The total moment exerted by N blades on the shaft is found by integrating this expression over the entire span of the blades

$$\vec{M} = \int d\vec{M} = \int \vec{r} \times d\vec{F}_\theta = \int_{r=0}^{r_0} \frac{1}{2}\rho V^2 W r \cdot (C_L \sin\varphi - C_D \cos\varphi) dr\, \hat{z}$$

To proceed further we now write the square of the velocity as

$$V^2 = \vec{V} \cdot \vec{V} = \left(V_\circ \hat{z} + V_\theta \theta^2\right) \cdot \left(V_\circ \hat{z} + V_\theta \theta^2\right)$$

$$= V_\circ^2 + V_\theta^2 = V_\circ^2 \left(1 + V_\theta^2 / V_\circ^2\right)$$

Now let us assume that the blade width is uniform and the lift and drag coefficients are also uniform along the radius of the blade. This latter assumption translates into a requirement that the local angle of attack between the blade chord vector and the local flow velocity vector is constant. From the geometry above, this will then translate into a requirement that the blade pitch changes with radius, i.e. that $\phi = \phi(r)$ such that the angle of attack is constant. This assumption then implies that

$w(r) = w \sim$ constant

$C_L(r) = C_L \sim$ constant

$C_D(r) = C_D \sim$ constant

$V_\theta = \omega r$

ρ ~ constant

With these assumptions the total moment is then given as

$$\vec{M} = \frac{1}{2}\rho V_0^2 W \int_{r=0}^{r_0} r \cdot \left(1 + \frac{\omega^2 r^2}{V_0^2}\right)(C_L \sin\varphi(r) - C_D \cos\varphi(r))\, dr\, \hat{z}$$

For a typical airfoil the lift and drag coefficients, and the lift to drag ratio vary with angle of attack as shown below. We note in particular that there is a value for the angle of attack, given as $\alpha_{\max}$, where the lift is maximized. This condition will result in a maximum torque generation

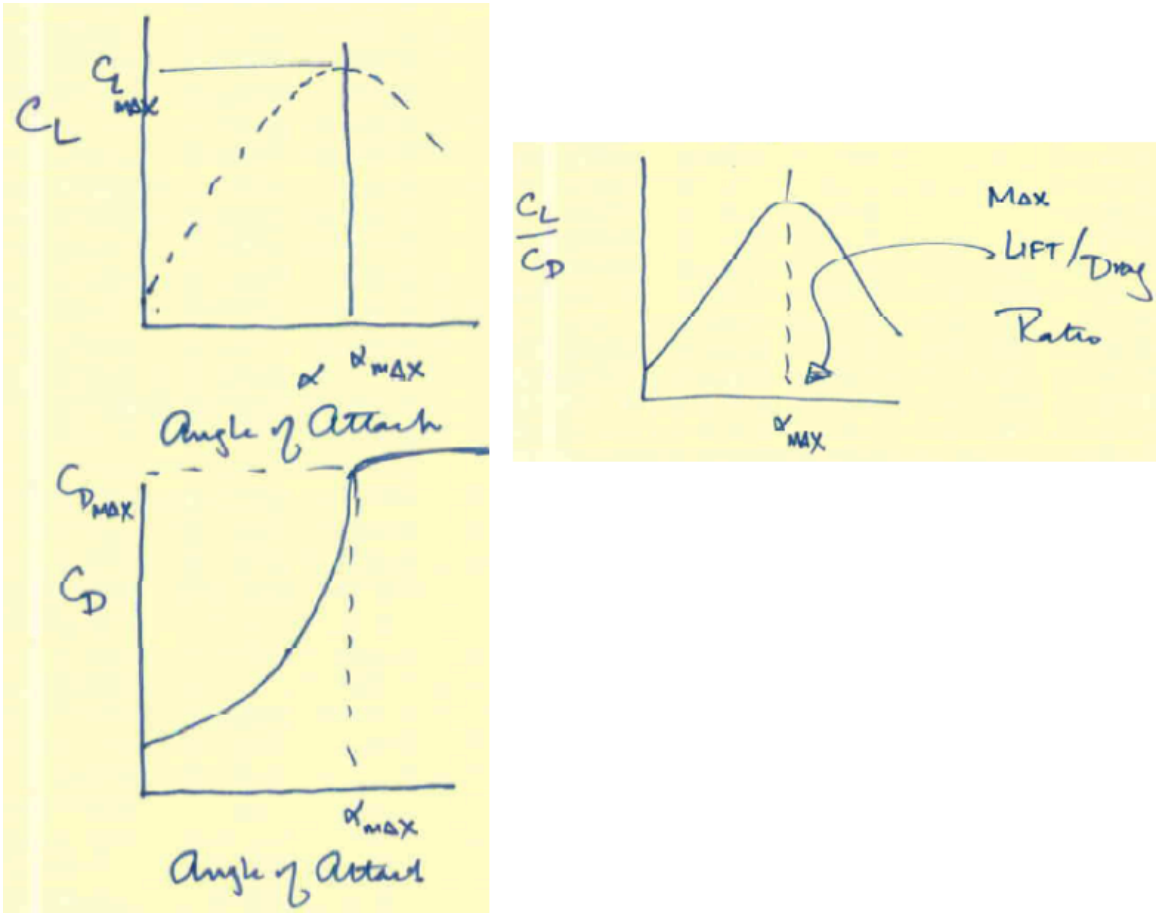


Figure 9.2: Typical variations of lift coefficient, drag coefficient, and lift/drag ratio verses angle of attack.

If one designed the turbine such that $\phi \sim \phi_o \sim$ constant and we have $V_\theta = \omega r = f(r)$, then $\alpha = \alpha(r)$ and the moment exerted on the shaft would not be maximized. Thus we see that in general a turbine would be designed with a variation $\phi(r)$ such that $\alpha(r) \sim \alpha_{max}$ for all r.

Now from the geometry shown in the figure above we can write the following relations

$$\psi = \tan^{-1}\left(\frac{V_\theta}{V_o}\right)$$

$$\psi + \alpha + \frac{\pi}{2} + \phi = \pi$$

$$\phi = \frac{\pi}{2} - (\psi + \alpha)$$

$$\phi = \frac{\pi}{2} - \tan^{-1}\left(\frac{V_\theta}{V_o}\right)$$

And if $\alpha = \alpha_{max} \sim const$ we can then find the radial variation of the blade pitch required to maintain constant angle of attack:

$$\phi = \frac{\pi}{2} - \tan^{-1}\left(\frac{\omega r}{V_o}\right) - \alpha_{max}$$

A rough design procedure for the turbine would then have the designer first choose $\phi = \phi(r)$ for maximum torque generation at the optimum wind speed. One could then evaluate the integrated moment produced by the turbine as a function of wind speed and then finally determine the power extraction for a given wind speed:

$$P = \vec{\omega} \cdot \vec{M} = \omega M$$

Clearly, the optimum blade pitch depends upon the angular rotation frequency, the free stream wind speed, and the optimum angle of attack. By incorporating the ability to change the blade pitch via a mechanism within the wind turbine, a range of values for the blade pitch can then be accommodated. This approach is used in current turbine designs. However, this capability cannot accommodate an arbitrarily wide range of free

stream wind speeds. Thus, there will be a minimum and maximum wind speed over which the turbine can operate due to a combination of effects arising from the turbine blade aerodynamics and structural considerations. Furthermore there will be a maximum rotation frequency which the turbine can accommodate safely and there is a limited range for the blade pitch control. Thus, if the free stream wind speed becomes too high, the local angle of attack can exceed $\alpha_{\max}$ and the airfoil can stall, resulting in a decrease in applied torque (and thus, power production); the resulting separated flow can also then exert time-varying forces on the turbine blade due to the unsteady generation of vortices behind the airfoil. This in turn can generate blade vibrations, which can present a dangerous operating condition in high speed flow conditions. As a result there is a maximum wind speed at which turbine operations can be permitted. This speed is known as the cutout speed; when the wind speed exceeds this value the turbine blades must then be rotated such that no torque is applied (i.e. they are "feathered") and turbine operations cease until the wind speed decreases into a safe range.

Maximum conversion efficiency

We are interested in determining the maximum possible conversion efficiency of a wind turbine. We shall refer to the schematic shown in Figure . First, let us first define ε as the kinetic energy of a column of air. Referring to the figure below, we can see that it is given as

$$\varepsilon = \frac{1}{2}\rho V_0^2 A V_0 \delta t$$

where dt denotes an arbitrary time interval.

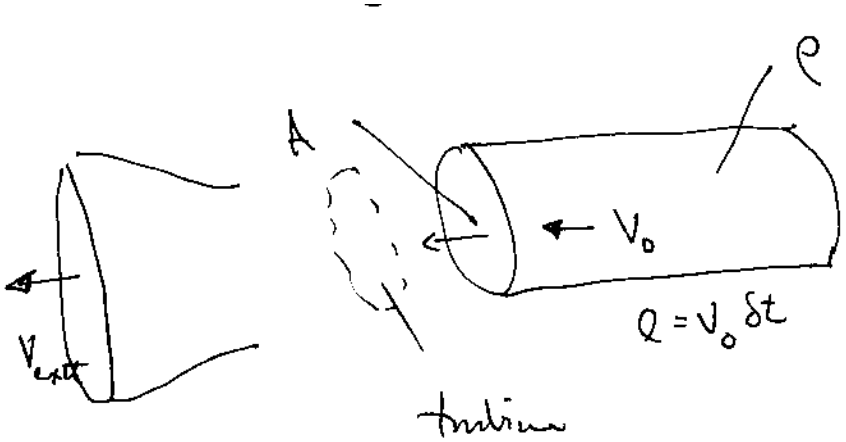


Figure 9.3: Streamtube passing through a disk defined by the rotation of a horizontal axis wind turbine.

Suppose now that an amount of energy δε is extracted from the column during this time interval dt. Since the power is then given simply as $P = \frac{\partial \varepsilon}{\partial t}$ it is then clear that the power available scales like $P_{out} \propto V^3$. Clearly then, one wishes to site wind turbines in regions where higher wind speeds prevail in order to maximize the power output of the system (obviously if the wind speed cannot exceed some maximum design limit; otherwise the turbine could be destroyed by the forces of such a high speed wind). The question then becomes: what is the maximum theoretical efficiency at which we can extract wind power? To answer this question, consider Figure below which shows a wind turbine and the streamtube associated with that turbine.

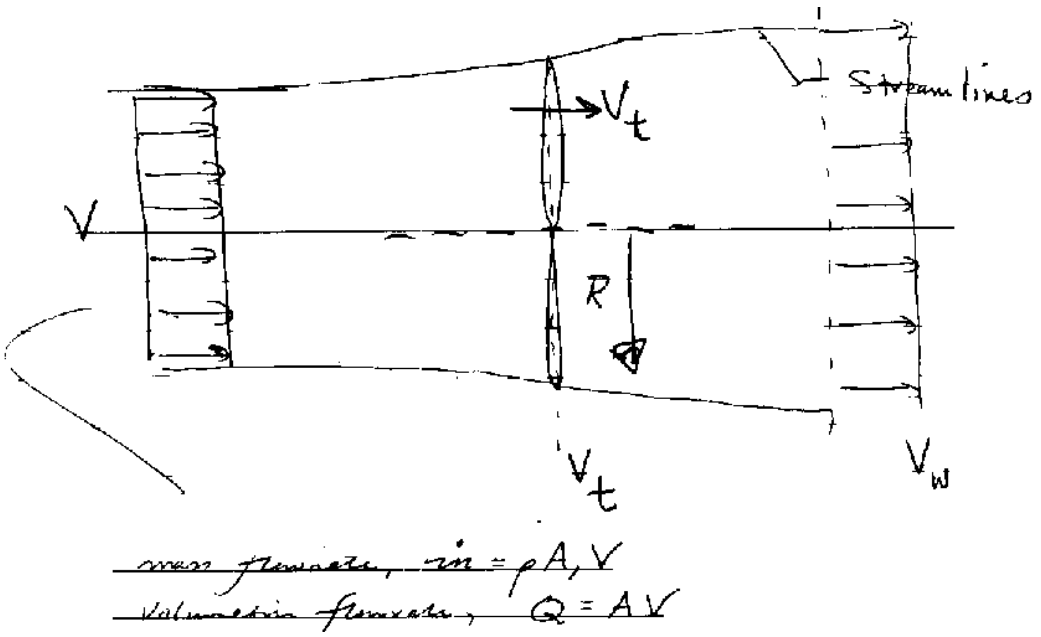


Figure 9.4: Sideview of streamtube passing through a turbine.

The kinetic energy per unit volume of air, $p = \frac{1}{2}\rho V^2$. If the turbine extracts power from the wind one can then clearly write the wind speed ordering as $V > V_t > V_w$, where V denotes the free-stream wind speed, Vt denotes the wind speed at some intermediate location, and Vw denotes the wind speed downstream of the turbine. Assuming the flow is incompressible, then we can write the change in the kinetic energy of the wind in terms of an identity:

$$\frac{1}{2}\cancel{\rho}\left(V^2 - V_w^2\right)\cancel{A} = \cancel{\rho}\cancel{A}V_t\left(V - V_w\right)$$

$$\frac{1}{2}\left(V - V_w\right)\left(V + V_w\right) = V_t\left(V - V_w\right)$$

$$V_t = \frac{1}{2}\left(V + V_w\right)$$

where Vt is just the average between the upstream and downstream velocities. The turbine power *P* is then:

$$P = \frac{1}{2}\rho V_t A\left(V^2 - V_w^2\right)$$

$$P = \frac{1}{2}\rho A\left(V^2 - V_w^2\right)\cdot\frac{1}{2}\left(V + V_w\right)$$

We see that the turbine power, P, is given by the mass flow rate $\dot{m}$ through the turbines: $\dot{m} = \rho V_t A$ multiplied by change in kinetic energy per unit mass, i.e.

$$P = \frac{1}{2}\rho V_t A\left(V^2 - V_w^2\right).$$

We can re-write this result in terms of a dimensionless variable $x = V_w/V$ and then find the value of x that maximizes the power production:

$$P = \frac{1}{2}\rho A\left[\frac{(V-V_w)(V+V_w)^2}{2}\right]$$

$$P = \frac{1}{2}\rho AV^3\left[\frac{\left(1-V_w/V\right)\left(1+V_w/V\right)^2}{2}\right]$$

$$P(x) = \frac{1}{2}\rho AV^3\left[\frac{(1-x)(1+x)^2}{2}\right]$$

$$\frac{\partial P}{\partial x} = \frac{K}{2}\left[-(1+x)^2 + 2(1-x)(1+x)\right] = 0$$

Solving this expression for x, we find one physically plausible solution (the other solution is non-physical)

$$-\left(1+2x+x^2\right)+2\left(1-x^2\right)=0$$

$$3x^2+2x-1=0$$

$$x = \frac{-2+\sqrt{4+12}}{6} = 1/3$$

$$\therefore P_{\max} = P(x)\big|_{x=1/3} = K\left[\frac{\left(1-1/3\right)\left(1+1/3\right)^2}{2}\right]$$

$$= K\left(1/3\right)\left(4/3\right)^2 = K\frac{16}{27}$$

$$P_{\max} = \frac{16}{27}\cdot\frac{1}{2}\rho AV^3$$

The power conversion efficiency can then be written as

$$\eta = \frac{P_{\max}}{\frac{1}{2}\rho A V^3} = \frac{16}{27} \approx 0.59$$

This result, called Betz's Law after the engineer who first derived it, gives the maximum possible power conversion efficiency of a turbine, independent of the details of the turbine design and thus it represents a maximum possible value which can be approached - but never exceeded - by an actual turbine design.

The actual power conversion efficiency of a turbine is usually quantified with the power coefficient, C_P, which is defined as the ratio of the actual power output to the power content of the wind that sweeps through the wind turbine, and is given as

$$C_P = \frac{P_{out}}{\frac{1}{2}\rho V^3 A_t}.$$

Due to the aerodynamic considerations discussed above, the power coefficient is a strong function of the ratio of the tip speed of the turbine blades to the incident wind speed, otherwise known as the tip-speed ratio, V_{tip}/V_∞ as shown in Figure below. The power coefficient has a small value when $V_{tip}/V_\infty \sim 1$, rises to a maximum value for $V_{tip}/V_\infty \sim 10$ or so, and then falls again. The power coefficient is always limited by the maximum theoretical conversion efficiency discussed above.

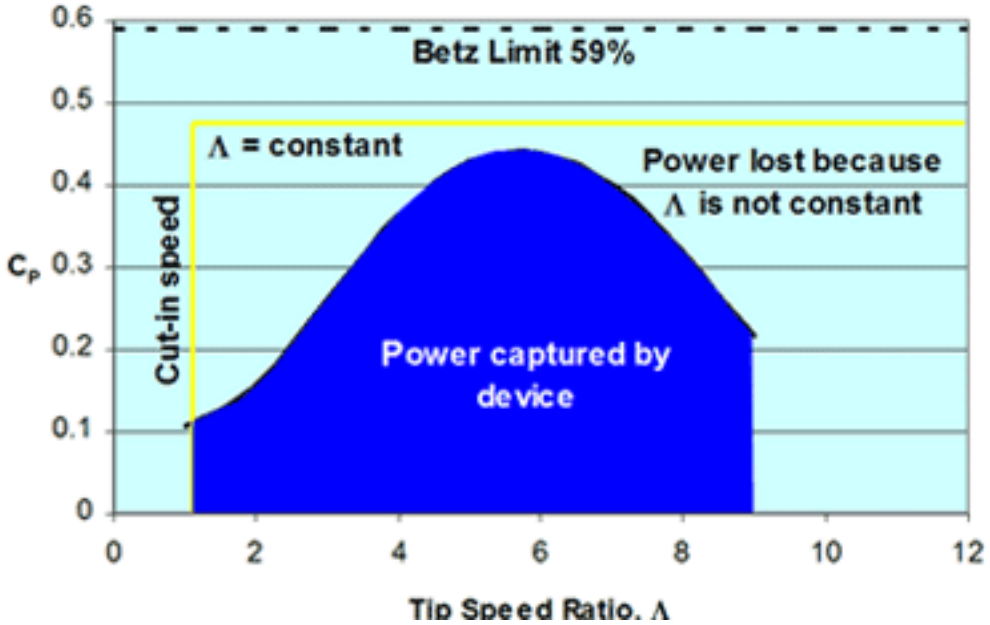


Figure 9.5: Power coefficient verses tip speed ratio. Figure from http://www.reuk.co.uk/Wind-Turbine-Tip-Speed-Ratio.htm, accessed 7 October 2009.

Relating Windspeed Variability to Power Output Variability

Clearly, a variation in wind speed will result in a very significant change in turbine power output, since the power output varies with the third power of the wind speed. Since the wind velocity is (nearly) a two-dimensional quantity localized to the horizontal plane and the speed is bounded by zero from below, the wind speed distribution follows the Rayleigh probability distribution which has a mathematical form given by

$$f(V) = \frac{V}{V_0^2} \exp\left(-\frac{V^2}{2V_0^2}\right)$$

where the parameter V_0 is related to the average wind speed V_{ave} via the relation

$$V_{ave} = V_0\sqrt{\frac{\pi}{2}}.$$

and the average wind speed is given by the integral

$$\begin{aligned} V_{ave} &= \int_{V=0}^{\infty} Vf(V)dV \\ &= \int_{V=0}^{\infty} \frac{V^2}{V_0^2}\exp\left(-\frac{V^2}{2V_0^2}\right)dV \end{aligned}.$$

The function *f(V)* is a probability distribution function; the quantity *f(V)dV* denotes the probability that the wind speed will lie within the range of *(V, V+dV),* with *V>=0.* Now, because the wind speed is related to the power density P via the relation $P = \frac{1}{2}\rho V^3$, we can find the power density probability distribution function, g(P), by defining the change in power density *dP* in terms of the change in wind speed *dV* via the relation

$$dP = \frac{3}{2}\rho V^2 dV.$$

We can then use the fact that the probability of a wind speed between *(V,V+dV)* will give rise to a corresponding probability that the power density will lie in the range (P,P+dP) to then write

$$g(P)dP=f(V)dV$$

which we can then solve for the power density probability distribution, g(P), which is given as

$$g(P) = f(V)\frac{dV}{dP}$$
$$= \frac{2f(V)}{3V^2}$$
$$= \frac{2}{3VV_0^2}\exp\left(-\frac{V^2}{2V_0^2}\right)$$

The average power output from this power distribution function is then given by the average over the power distribution function

$$P_{ave} = \int_{V=0}^{\infty} Pg(P)dP.$$

Using the definition of power density and the result for dP above we can then recast this in terms of an integral over the wind speed:

$$P_{ave} = \int_{V=0}^{\infty} \frac{1}{2}\rho V^3 \frac{2}{3VV_0^2}\exp\left(-\frac{V^2}{2V_0^2}\right)\frac{3}{2}\rho V^2 dV$$
$$= \int_{V=0}^{\infty} \frac{1}{2}\rho^2 V^2 \frac{V^2}{V_0^2}\exp\left(-\frac{V^2}{2V_0^2}\right)dV$$

Comparing this expression for the average power against the expression for the average wind speed, we can see that the power density corresponding to the average wind speed is significantly lower than the average power, i.e. we can see that

$$\frac{1}{2}\rho V_{ave}^3 < P_{ave}.$$

Figure 9.6 below illustrates f(V) along with actual wind speed histogram data taken at a representative location; the power distribution g(P) is also shown and clearly illustrates that the average power is highly weighted by the higher wind speeds.

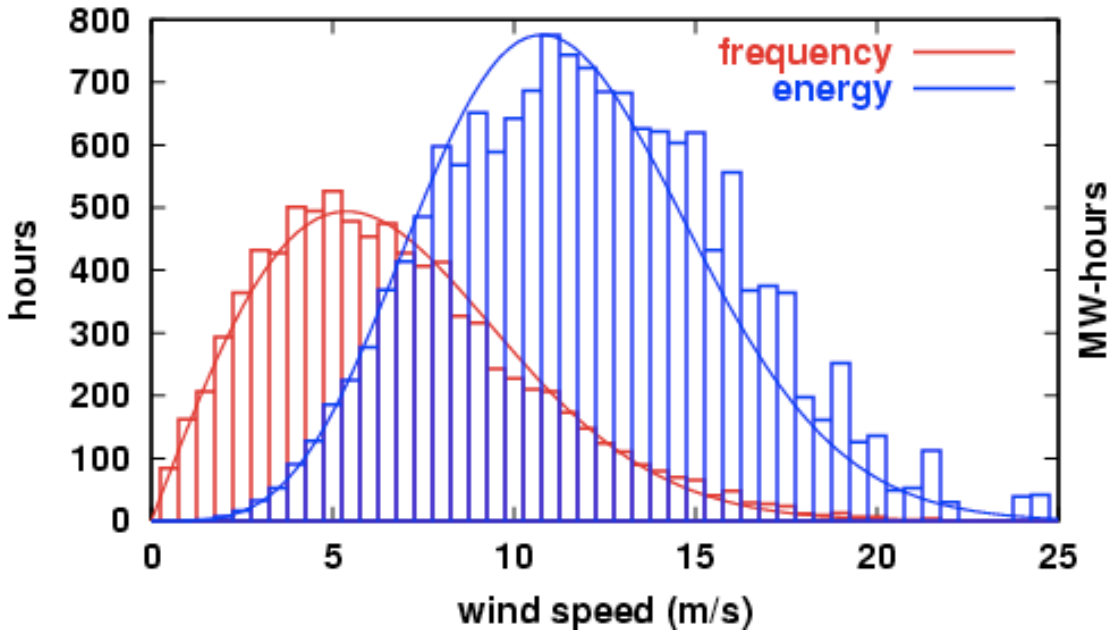


Figure 9.6: Wind-speed and power output histogram for a typical wind turbine site.
Source: http://en.wikipedia.org/wiki/Image:Lee_Ranch_Wind_Speed_Frequency.png

These results have implications for both the economics of wind power generation as well as for the time variation of the power output of the turbine (which imposes technical issues for incorporating large quantities of time varying wind power into a large

scale electrical power grid). Let us examine the cost issue first and then comment briefly on the technical issues associated with power intermittency.

In an idealized world where the cost to borrow money is zero and the turbine lifetime is T, then at a minimum the turbine revenue from power generation must recoup the initial turbine cost C in time T (absolute Minimum requirement). Thus, the turbine must produce revenue at a rate given by C/T. Suppose that the power P produced is $P = C_P A \frac{1}{2} \rho V_o^3$ and this wind speed is steady state, and power sells for K $ / Watt. Thus, we must satisfy the requirement $KP \geq \frac{C}{T}$, or $K \varepsilon A \frac{1}{2} \rho V_o^3 \geq \frac{C}{T}$ in order to recoup the initial investment. Thus, there exists a minimum wind speed below which the turbine will not pay for itself. Of course, real world effects like a finite borrowing cost, finite operations and maintenance costs and so forth will increase the total costs further, and therefore increase the minimum allowable wind speed.

Of course, wind speed is not constant but instead has the Rayleigh distribution as discussed above. Changes in wind speed can therefore have a significant effect on the wind power output of a turbine or turbine array. This significance of this effect can be appreciated by defining the wind power capacity factor, C_{WP}, in terms of the ratio of the average power produced by a turbine (or perhaps an array of wind turbines) (the average is taken over some suitably long period of time to account for short term and seasonal

variations) to the maximum rated power production possible by the turbine or turbine array, i.e.

$$C_{WP} \equiv \frac{P_{ave}}{P_{rated}}.$$

Operational experience in a number of locations around the world suggest that the capacity factor lies in the range of $C_{WP} \approx 0.25 - 0.4$ [REFERENCE].

As mentioned earlier, the variability of wind speed and the associated power output introduces the potential to degrade the electrical power quality on a large power grid. These quality issues typically include voltage regulation, phase control, harmonic content and so forth, and present serious power engineering issues that must be carefully considered, particularly when the wind power input to the grid becomes a significant fraction (say 10% or so) of the total grid power capacity [REFERENCE]. It is thought that introduction of energy storage solutions and some degree of demand management techniques can increase the maximum wind power fraction on a large scale grid to values of 20-30% [REFERENCE]. Further increases likely would require more substantial grid architecture changes or other wholesale changes in the power grid design. Current research is focused on identifying what such architechture changes look like and how they would impact the maximum feasible wind energy that can be incorporated into a power grid.

Wind turbine arrays: Interference between turbines

The blades of a wind turbine shed vortices from the tips and also induce turbulence along the trailing edge of the blade chord. The resulting air disturbances then propagate downstream with the free-stream wind flow. As a result, a turbine located upstream of another turbine will interfere with the performance of the downstream turbine; such disturbed airflows can reduce the power coefficient of the downstream turbine and could, if strong enough, also reduce the lifetime of the downstream turbine due to materials fatigue issues caused by the disturbed airstream. As a result, there is typically a minimum packing density for an array of turbines to the next limits spacing. Typically, the flow-wise spacing needs to be roughly 10 times the upstream turbine diameter; the lateral spacing can be smaller (if the prevailing wind direction does not change that often); values of 5 times the upstream turbine diameter are often found in discussions of the subject. Thus, for a turbine diameter d, a minimum land area of approximately 50-100d^2 is needed to avoid these issues. Figure illustrates this result.

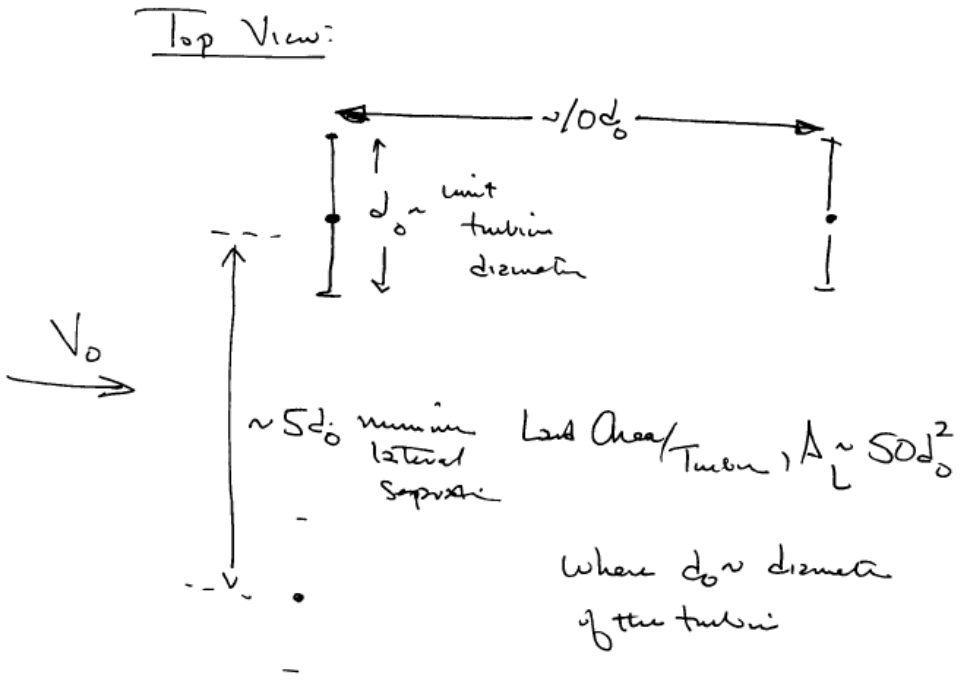


Figure 9.7: Top view of wind turbine array showing spacing parallel and transverse to the prevailing wind direction.

Example: Suppose we wish to build a wind farm with P_{out} = 1GW in an area with a wind turbine energy density, $\frac{1}{2}\rho V_0^3$, of 250 W/m^2 and a power coefficient of 40%. We then seek to answer two questions:

- How many turbines are needed?
- What land area is needed?

Let us denote P_T ~ output of a single turbine and N_t ~ number of turbines. From the results above we know

$$P_T = \frac{\pi}{4} d_o^2 \cdot \eta \cdot \frac{1}{2} \rho V_o^2$$

Where $\eta \frac{1}{2}\rho V_0^3$ = 100 W/m^2

With the spacing restrictions given above we then have

$$P_{tot} = N_t D_T = N_t \cdot 25 \cdot \pi d_o^2$$
$$A_{land}^{tot} = N_t \cdot 50 d_o^2$$
$$P_{tot} = 10^9 W = N_t \cdot 25 \cdot \pi d_o^2$$
$$\therefore N_t d_o^2 = \frac{10^9}{25\pi} \sim 10^7 m^2$$
$$\therefore A_{land}^{tot} \; ; \; 5 \cdot 10^8 m^2 = 500 km^2$$

If $d_o \sim 50$ m

$$P_T \sim \frac{3}{4} 50^2 \cdot 100 \sim 200 kN$$
$$N_t \sim \frac{10^7}{d_o^2} \sim \frac{10^7}{2.5 \cdot 10^3} \sim \frac{10^4}{2.5} \; ; \; 4000$$

i.e. 1 GW output would require roughly 4000 turbines @ 200kW / turbine on 500km^2 of land area.

Example 2: Naïve Estimate of Potential of Wind Power

Let us use the above results to estimate the maximum possible power output from wind energy in the U.S. According to the U.S. Department of Energy NREL, there is approximately 6x10^5 km^2 of land with >300 W/m^2 wind power density located within 10km of transmission line in the U.S. Assuming a turbine Spacing (with d_o being the rotor diameter) consistent with the above discussion, i.e. assuming $10d_o$ ~ parallel to wind and spacing $5d_o$ ~ perpendicular to wind, and taking the current largest wind

turbine to have a rated power of ~ 3MW in such a wind, we estimate the required turbine area is then given by 3MW/300W/m^2 ~ 10,000 m^2. This then gives a rotor diameter of about 120m. Therefore, the land area per turbine is approximately $5x10^5$ m^2. With the available land, we would then expect to be able to pack approximately 10^6 such large turbines on suitable land in the U.S. These turbines would have a peak rated power of ~3TW. However, taking into account the typical capacity factor of ~0.3 or so, the average power output of such an array of turbines would be ~1TW – approximately equal to the current electrical power consumption in the U.S.

However, this analysis – which mirrors many published estimates [see e.g. PNAS, 2009] – neglects the fact that the wind turbines actually slow down the wind speed. As a result, the downstream turbines do not see the original wind speed – they see this reduced value. In a large enough array, the reduction in wind kinetic energy will then be balanced by the turbulent transport of wind kinetic energy from the regions immediately above the turbine array downwards through the atmospheric boundary layer into the near-ground region where the turbine array is located. It is this effect that ultimately will limit the amount of wind energy that can be produced from large turbine arrays. We thus take up this important physical effect and examine the impact on the maximum possible power generation from wind.

Ultimate Physical Limit of Wind Power Generation

We have examined the theoretical efficiency and performance of a wind turbine, and found that such devices can extract a significant fraction (~0.5) of the available kinetic energy contained in the wind, assuming that the device is operating close to the design optimum. Given that such wind turbines are immersed within the boundary layer that exists near the Earth's surface, and given the fact that the placement of a large number of turbines in an array will therefore extract a significant fraction of the kinetic energy contained within the boundary, the question then arises: *How closely can an array of such turbines be placed together before the existence of the turbine array disturbs the boundary layer to such a degree that the power output of the array is degraded?* In order to answer this question, we must examine aspects of turbulent boundary layers on a flat plate, which we take to approximate the Earth's surface. This discussion is based upon the paper by Best[2].

---

[2] R.W.B. Best, Energy Conversion, v. 19, pp. 71-72, (1977).

Turbulent Boundary Layer Analysis[3]

The flow of the wind over the Earth's surface forms a turbulent boundary layer as shown schematically in the figure below. Because of the gradient in wind speed at the surface there will exist a shear stress, τ, at the Earth's surface as shown in the subsequent figure.

The *average* wind velocity *u* will vary with height, *h*, via the function *u(h)* given as

$$\frac{u(h)}{w} = 2.5\ln(h/r) + 5.5$$

where w denotes the so-called "skin friction velocity", defined in terms of the shear stress, τ, according to $\tau = \rho w^2$ where ρ denotes the fluid density and *r* denotes the "surface roughness height" and is given by $r \approx \nu / w$ with ν denoting the viscosity of air. It is important to remember: the boundary layer is turbulent, and here *u(h)* denotes the average flow speed, after the turbulent velocity fluctuations have been averaged away.

---

[3] See e.g. Landau and Lifschitz, *Fluid Mechanics*, section 42 or any other fluid mechanics textbook discussion of turbulent boundary layers on a flat plate.

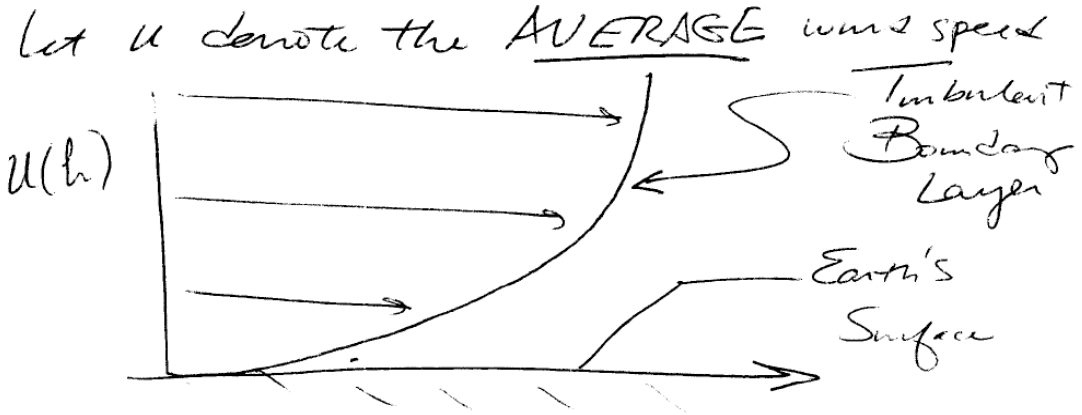


Figure 9.8: Schematic of the average wind speed distribution within the turbulent boundary layer.

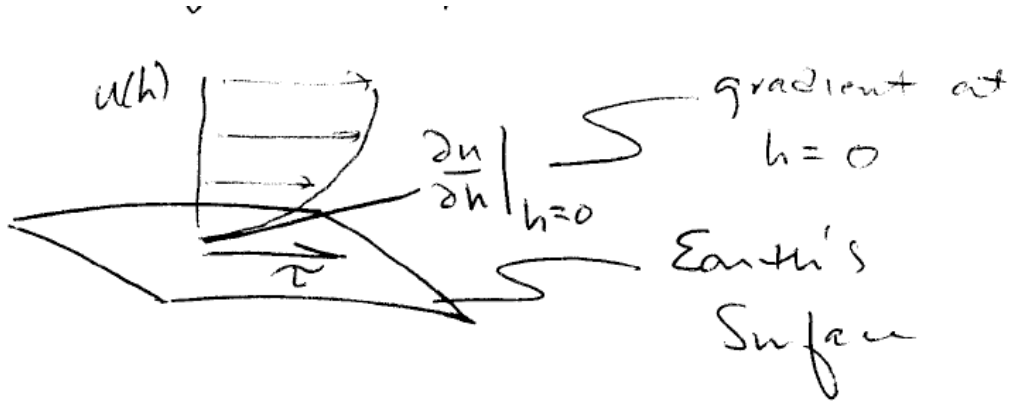


Figure 9.9: A shear stress, t, exists at the Earth's surface due to the exchange of momentum between the flowing air mass and the surface.

Now the wind speed will of course vary with time. Let us denote the instantaneous wind speed to be given as *v*. Then the probability F(v) that v>*u* is given for the Rayleigh distribution of wind-speed (which is the usual model for wind speeds) is given as

$$F(v) = \exp\left(-\frac{\pi}{4}\frac{v^2}{u^2}\right).$$

A sketch of F(v) is shown below.

We can now use this to calculate the average power density contained within a wind flow. We know for a wind speed v that this power density is given as $P(v) = \frac{1}{2}\rho v^3$. We can then find the *average power* from the expression $P = \int_{v=0}^{\infty} P(v)dF$. Using the expression for F(v) above we can write $dF = -\frac{\pi}{2}\frac{vdv}{u^2}$. Using P(v) and performing the integral, we find that the average power, P, is related to the average wind speed, u, by the relation

$$P = \frac{3}{\pi}\rho u^3.$$

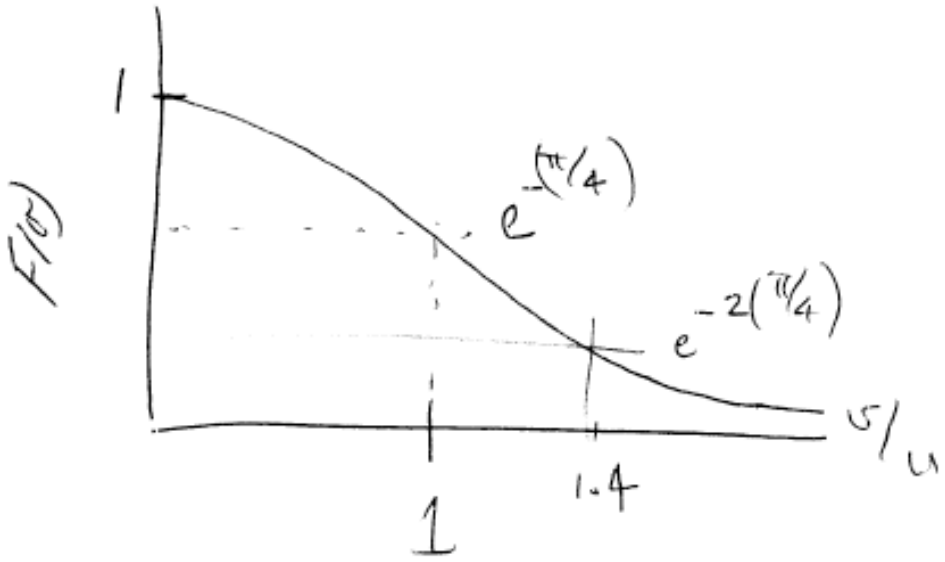


Figure 9.10: Sketch of the probability that the wind speed v will exceed the average wind speed u, assuming that the wind speed obeys the Rayleigh distribution.

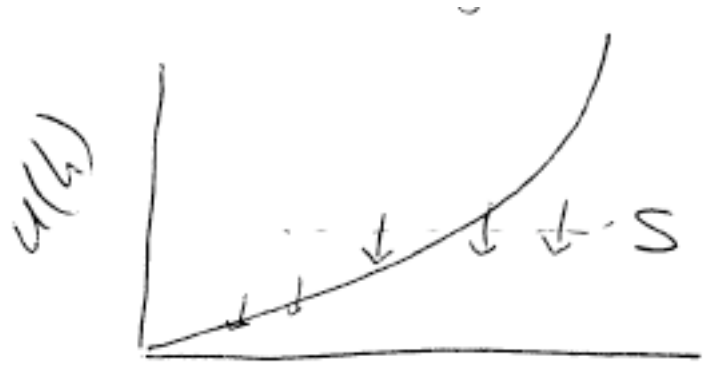


Figure 9.11: Sketch of the vertical flux S of horizontal kinetic energy

Now within the boundary layer, there will exist a vertical transport of horizontal wind kinetic energy. Let us denotes this flux of energy in the vertical direction as S. It can be shown (see Landau & Lifschitz) that this flux is given by

$$S \approx \frac{6}{\pi}\rho w^2 u.$$

Let us define the ratio of the frontal area of an array of N turbines, each with area At, to the ratio of the terrain surface area A as the ratio

$$\lambda = \frac{NA_t}{A}.$$

Let us suppose that the number of turbines is large enough (question: how large would this number actually need to be?) such that the presence of this large number of turbines makes a substantive change in the turbulent boundary layer. In particular, we suppose that there is a change in the shear stress that is proportional to this area ratio; it will also be proportional to the wind kinetic energy, i.e.

$$\delta\tau = \lambda\frac{1}{2}\rho u^2.$$

Using the definition of the unperturbed shear stress, we can also write

$$\delta\tau = 2\rho w\delta w.$$

Equating these two we can then find that

$$\frac{\delta w}{w} = \frac{\lambda}{4}\frac{u^2}{w^2}..$$

Now the average wind energy flux through the turbine array is given by λP per unit area of terrain where P is the average wind power density given above. We can write the ratio of this power flux to the vertical flux of horizontal kinetic energy, S, as

$$\frac{\lambda P}{S} = \frac{\lambda}{2}\frac{u^2}{w^2} = 2\frac{\delta w}{w}.$$

It is reasonable to assume that the height of the boundary layer, H, is proportional to w. Thus we can relate a change in H to a change in w by the relation

$$\frac{\delta H}{H} = \frac{\delta w}{w}..$$

The so-called “gradient wind”, U, is defined in terms of H, r, and w as

$$\frac{U}{w} = 2.5\ln\left(\frac{H}{r}\right) + 5.5$$

and represents the characteristic wind speed at the top of the boundary layer associated with a particular shear stress and surface roughness at the bottom of the boundary layer. Now, the introduction of the wind turbine array will disturb r and w, but U will not change U. Thus, we can perturb this expression to find

$$\delta\left(\frac{U}{w}\right) = \delta\left[2.5\ln\left(\frac{H}{r}\right) + 5.5\right]$$

or

$$-\frac{U}{w^2}\delta w = 2.5\left(\frac{\delta H}{H}\right) - \delta\ln r$$

or

$$\delta\ln r = \left(1 + 0.4\frac{U}{w}\right)\frac{\delta w}{w}$$

where in the last expression we used the result above for the change in H in terms of the change in w.

Now, from the boundary layer velocity equation given above, we can write the perturbed average velocity as

$$\begin{aligned}
\delta u(h) &= \delta\{w[2.5\ln(h/r)+5.5]\} \\
&= \delta w[2.5\ln(h/r)+5.5] + 2.5w\delta\left(\ln(\frac{h}{r})\right) \\
&= \delta w\frac{u}{w} - 2.5w\delta(\ln(r)) \\
&= \delta w\frac{u}{w} - 2.5w\left[\left(1+0.4\frac{U}{w}\right)\frac{\delta w}{w}\right] \\
&= \delta w\left(\frac{u-U}{w} - 2.5\right) \\
&= -2.5\delta w\left(1+\ln\frac{H}{h}\right)
\end{aligned}$$

Based upon this result, we can then write the change in the average speed due to the turbine array, normalized to the original average wind speed as

$$\frac{\delta u}{u} = -\frac{2.5}{4}\frac{u}{w}\left(1+\ln\left(\frac{H}{h}\right)\right)\lambda.$$

Now from the definition of the average power we can write

$$\frac{\delta P}{P} = 3\frac{\delta u}{u}.$$

Similarly the perturbation in S due to the turbine array is given as

$$\frac{\delta S}{S} = 2\frac{\delta w}{w} + \frac{\delta u}{u}.$$

We can now use these perturbation analysis results to estimate the maximum theoretical wind power extraction per unit of land area. The analysis works as follows. We first must specify the parameters r and w, which are determined by the regional geometry (for r) and by the shear stress in the near-ground area. Once these parameters are specified, then we chose a turbine height, h. From this choice, we can then calculate

u/w, $\delta w/\lambda w$, $\delta u/\lambda u$, and $\delta S/\lambda S$. The choice of w determined U. We can then calculate u at the height h and from this, find the average power P.

We then specify the maximum allowable decrease in average power, i.e. we specify $\delta P/P$. These quantities then can be used to find the area ratio, λ. That value, combined with P then determine how much average power per unit land area can be extracted from a given wind flow.

Results from Turbulent Boundary Layer Analysis: Maximum Theoretical Wind Power Extraction per unit Terrain area.

Let us now use this analysis to estimate the power extraction per unit land area. We follow the analysis of Best, Energy Conversion v. 19, pp. 71-72, (1977). We look at coastal and inland areas. The table below summarized these two cases for a 100m high wind turbine. If we set a 20% extraction fraction of kinetic energy as an upper limit (this will give about a 7% reduction in wind speed averaged over the entire area of the wind turbine array), then the frontal projected area of the turbine array, normalized to the terrain area occupied by the array, can only be ~0.2%. The resulting power extraction density (MW power/terrain area) is then shown in the figure below for both coastal and inland regions.

Table 1: Calculation of maximum theoretical power extraction capability for 100m high wind turbine in Class 7 coastal and inland areas.

These calcuations based upon analysis contained in the paper

"Limits to Wind Power", R.W. B. Best Energy Conversion v.19 pp.71-72 (1977)

turbine axis height,		
h	100	m
Gradient Wind, U	10	m/sec

	Coastal region analysis		Inland region analysis
INPUT QNTY'S			
roughness, r	0.6	m	10
Boundary layer thickness, H	400	m	1000
skin friction velocity, w	0.5	m/sec	0.64
CALCULATED QNTY'S			
u/w	18.28998952		11.25646273
dw/(lambda*w)	83.6309292		31.67698831
du/(lambda/u)	-27.27831179		-23.23464126
dS/lambda*S)	139.9835466		40.11933536
u	9.144994762		7.204136149
U	13.62786271		12.8682723

for 20% loss in wind power density, dP/P=0.2 we then find lambda:

dP/P		-0.2	-0.2
allowable du/u		-0.066666667	-0.066666667
Lambda		0.002443944	0.002869279
Average wind power density, P	watts/m^2	730.3344136	357.0402003

maximum power/unit terrain area, lambda*P	1.784896414	1.024447925
Maximum power for 1 km^2	1784896.414	1024447.925

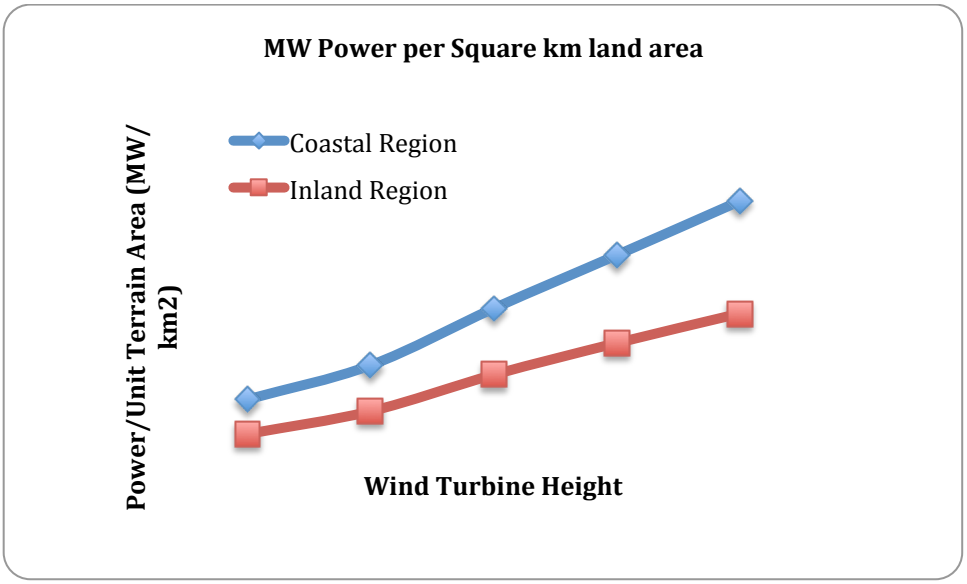


Figure 9.12: Maximum power extraction per unit land area vs. wind turbine height for coastal and inland regions. Based upon analysis technique of Best, Energy Conversion v19, pp. 71-72, (1977).

Estimates of Maximum Possible Wind Power Resources – U.S. Focus

We are now in a position to make some rough estimates of the maximum theoretically available amount of wind-power that could be produced in a given region. Because the data are widely available, let us focus on the lower 48 states contained within the U.S. Wind power is usually classified according to the power density. Class 1 and 2 power densities lie in the 100-200W/m2 range and are not thought to be commercially viable on

a large, utility scale. Class 3 (300 W/m2) and higher sites are thought to be commercially viable on a large scale.

The figure below shows the regions in the USA which have Class 3 and above wind resources [ref: http://rredc.nrel.gov/wind/pubs/atlas/chp2.html#areal accessed 26 Feb 2009].

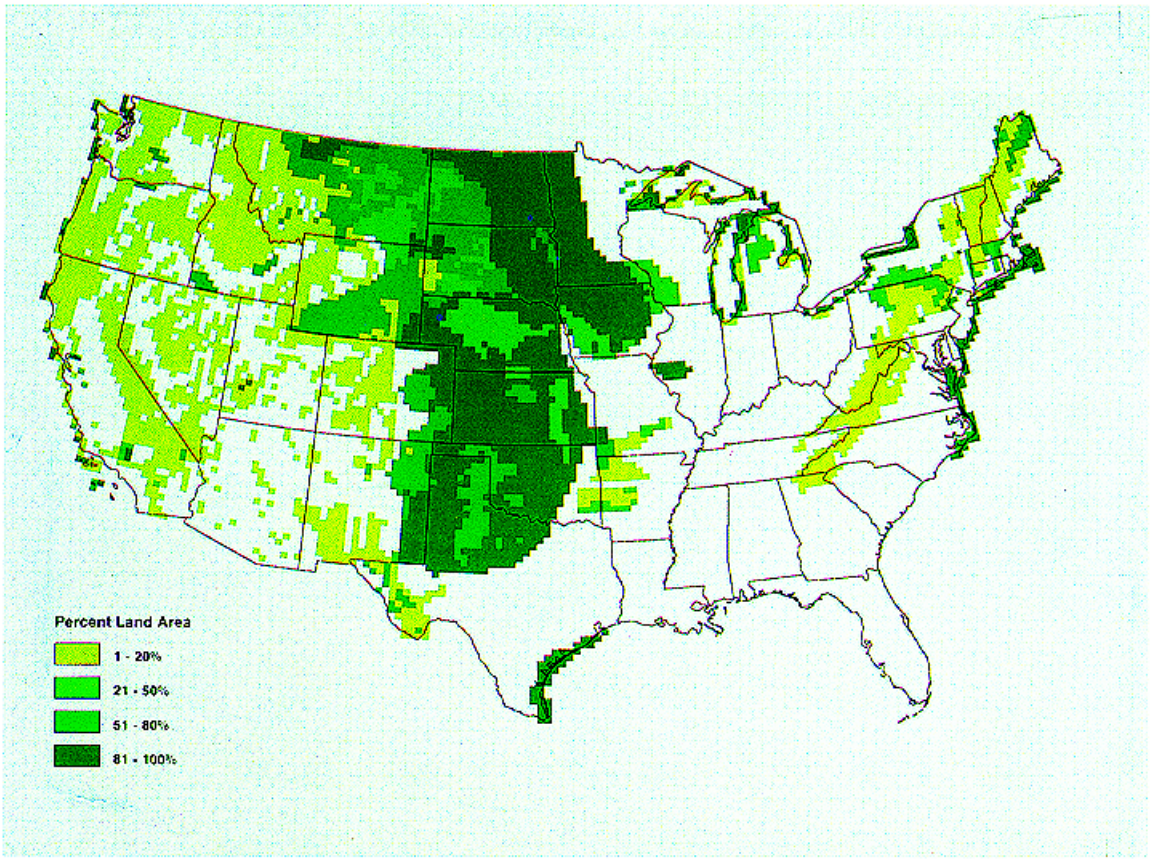


Figure 9.13: Class 3 and higher areas within the 48 U.S. states. Wind equals or exceeds specified class rating at least 80% of time.

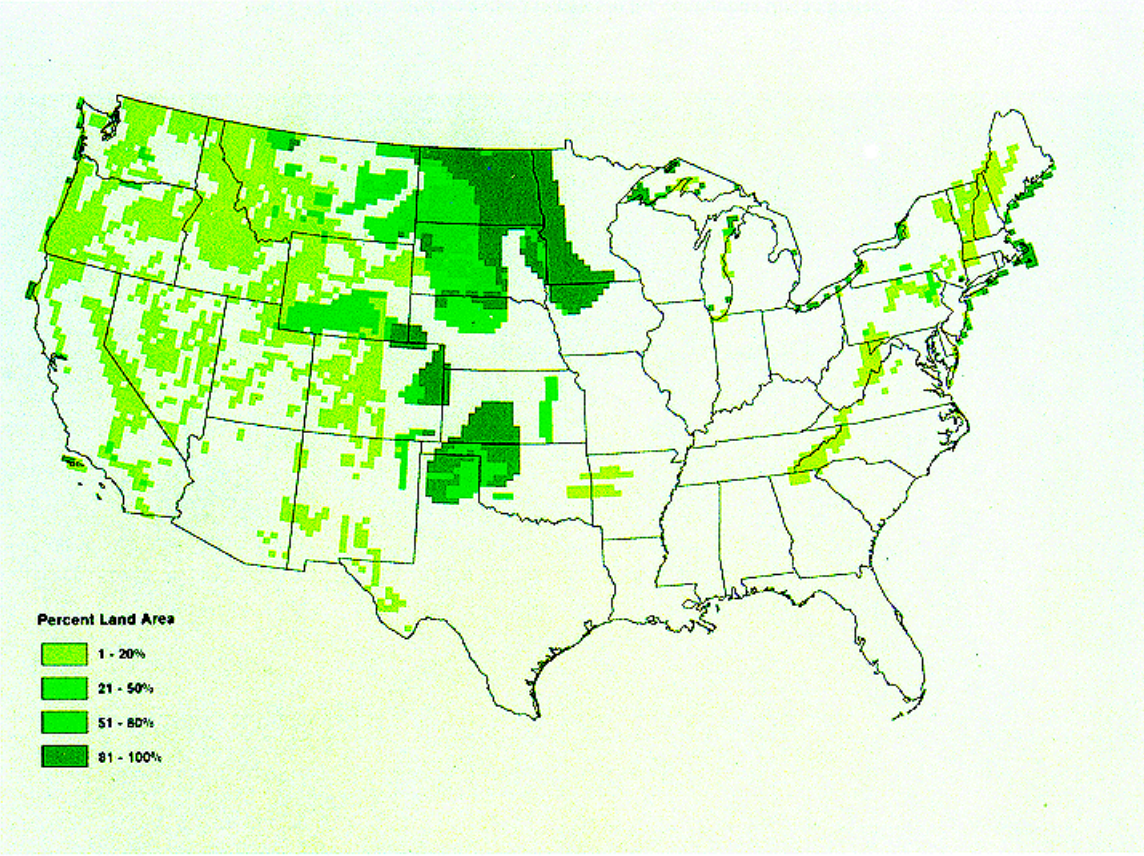


Figure 9.14: Distribution of Class 4 and higher areas within the 48 U.S. states. Wind equals or exceeds specified class rating at least 80% of time.

We can roughly estimate that for Class 3 winds, there are of order 3x10^6 sq km of land area available, while for Class 4 winds, there are about 5-6x10^5 sq km of land area available.

Our turbulent boundary layer analysis tells us that, for 100m height turbine axes in Class 3 wind areas (i.e. with mean wind speeds of about 5-6msec), with large turbine arrays extracting 20% of the available wind kinetic energy (this will correspond to a decrease of about 6% in the average wind speed in the areas covered by the arrays), we

can extract about 1MW/km^2 of terrain area. Increasing the maximum power extraction fraction to say 30% increases this maximum power extraction density to about 1.5-1.7MW/km^2. Thus, we can estimate the range of theoretically maximum wind power.

Let us assume that we utilize all Class 3 areas with 20% power extraction. The boundary layer analysis above then indicates that the maximum power would then result in 2-3 TW. If we extract 30% of the available kinetic energy, we can increase this maximum power to the range 3-4TW – roughly equal to the entire U.S. power consumption (integrated over all primary energy sources) today. A large wind turbine today has a maximum power rating of ~3-4MW. Thus, this system would correspond to a turbine spacing density of about one turbine/2-4 sq km making for a total system of about ~1 million such large turbines.

If we use only Class 4 and above areas (i.e. with average wind speeds exceeding 7 m/sec) the turbulent boundary layer analysis tells us that we can expect to extract slightly more than 1MW/sq km of terrain area – perhaps in the range of 1.2-1.4MW/sq km. If we used all of the available Class 4 land area this would then correspond to a maximum theoretical power production of about 0.6-1TW of power.

Of course, these estimates represent a maximum upper limit. In practice, it is unlikely that 100% of the available land area would be made available for such use; instead some fraction perhaps ranging from a few % to 10-20% of the land area might be used for wind power generation. In that case, the wind power generation resource estimates would lie somewhere between ~50GW on the lower end to something

approaching 500-600GW of power. For comparison, total U.S. electrical power consumption is in the range of 1TW today; nuclear fission (currently the largest C-free power source in the U.S. today) provides about 150-200GW of electrical power. Thus, it seems theoretically possible that wind energy could provide an electrical power contribution roughly equal to or perhaps somewhat larger than the current install nuclear fission capacity in the U.S., but will require installation of ~100,000s to ~1,000,000 large wind turbines over a very large (10^5-10^6 sq km) area. These values are reduced by factors of ~2-10 from the estimates made earlier that neglected the boundary layer effects.

This analysis examines the peak power production. In reality, historical experience shows that the average power produced from a turbine is only about 20-25% of the peak power rating of the turbine. The values above discuss the peak turbine power rating. If this experience held for such large arrays envisioned above, then the average power production would be reduced from the peak values listed above by about factor of 4-5.

Chapter 10: Geothermal Energy

The interior of the Earth is maintained at high temperatures relative to the Earth's surface via the decay of naturally occurring radioactive materials, resulting a transport of heat that makes it way towards the surface. In isolated surface regions, the molten rock from the interior approach or even reaches the surface, forming geologically active volcanoes and isolated features that are characterized by easily accessible geothermal sources. These localized regions (e.g. in Northern California, Iceland, and other similar geologically active regions) geothermal energy systems in which the near-surface heat content is converted to electricity and/or used for industrial and space heating, is already being exploited. However, due to the limited regions in which these near-surface geothermal sources are available, the relative contribution to the global energy demand is quite limited (i.e. power levels of GW's to 10's of GW's are involved), and it is difficult to see how these resources can be scaled up by several orders of magnitude to provide a materially significant energy resource in the 21st century. However, there are other approaches to consider that do take advantage of the available geothermal energy resource. In this chapter, we summarize the basic technical considerations involved in exploiting this resource, and we estimate the scale of this geothermal energy resource.

Steady-state Geothermal Heat Flux

The Earth's central temperature is maintained at high values relative to the surface temperature due to the energy release of radioactive materials. As a result, a heat flux directed outwards towards the Earth's surface exists and can, in principle, be captured

and converted to useful form. Let us first make an estimate of this flux and of the resulting useful work that could be extracted for human consumption. Figure below shows a schematic of the distribution of the temperature within the Earth.

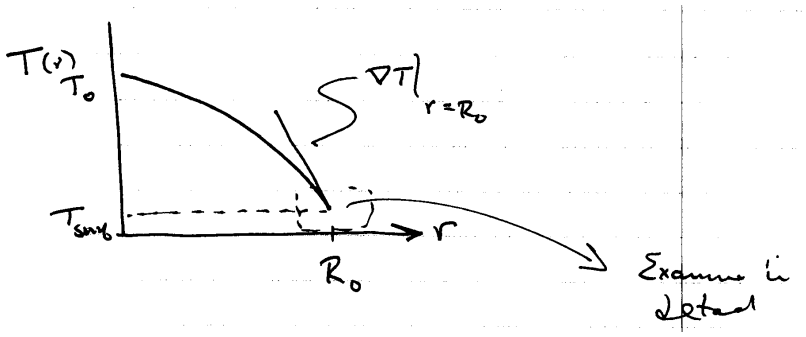


Figure 10.1: Schematic of the Earth's temperature vs. radius. A gaident exists and carries a heat flux via thermal conduction to the surface.

Earth has a geothermal gradient, $\nabla T \sim 20-25\frac{\deg K}{km}$ at depths of ~ 10^3 – 10^4m. Because of Earth's gravity, some of this gradient is due to a change in gravitational potential energy. To account for this effect, we relate the change in gravitational potential energy to the change in thermal energy via the relation

$$\cancel{\rho} g\Delta h = \cancel{\rho} c_p \Delta T$$

$$\therefore \frac{\Delta T}{\Delta h} = \frac{g}{c_p} = \frac{9.8\,{}^{m}\!/\!_{s^2}}{800\,{}^{J}\!/\!_{kg\cdot K}};\ 12\frac{K}{km}$$

This value is known as the geopotential thermal gradient and has a value ${}^{\nabla T}\!/\!_{g}$, i.e. ${}^{\nabla T}\!/\!_{g} = 12\frac{K}{km}$. This thermal gradient exists due to the gravitational potential alone, and is not available to transport heat via diffusion. Thus thermal diffusion will require a gradient that exceeds this value.

Assuming that the heat transport occurs via thermal diffusion, we can estimate the heat flux at the surface: In the presence of this gradient, the Fourier heat conduction then gives:

$$q_r = -K\left(\nabla_r T - \nabla T \big/ g\right)$$

For most near-surface material the typical value of thermal conduction lies in the range $K \sim 2-4\frac{W}{m\cdot K}$.

Thus we estimate the available heat flux in the near-surface region is given as

$$q \sim (2-4)\cdot(10-15)\cdot 10^{-3}\frac{W}{m^2}$$

$$q \sim (20-60)\frac{mW}{m^2}$$

The surface area of Earth is given as $4\pi R_o^2$ where

$$R_o \sim 6\cdot 10^6 m$$

$$A;\ 12(36)\cdot 10^{12};\ 4\cdot 10^{14} m^2$$

Thus, the total thermal power escaping from Earth is estimated to lie in the range

$$Q = qA \sim (20-60)\cdot 10^{-3}\frac{W}{m^2}\cdot 4\cdot 10^{14} m^2$$
$$\sim (80-240)\cdot 10^{11} W$$
$$\sim (8-24) TW$$

Suppose a fraction, *f*, of the Earth's surface was tapped for geothermal energy.
Maybe $f \sim 10^{-3}$ - 10^{-2} (Note: this is still 4x10^{11} – 4x10^{12} m^2), i.e. 6x10^5 meters on a side, or:

$$\frac{f \sim 10^{-3}}{f \sim 10^{-2}} \text{ or } \frac{600\cdot 600 km(+)}{1800\cdot 1800 km}$$

If the maximum Practical Depth is ~10km or so, then the temperature of the hot side, T_H:

$$T_H \sim 10 \cdot \nabla T \sim 200K$$

For simplicity, let us assume that the heat escaping at this temperature can be converted at the Carnot Efficiency:

$$\eta_{\max} \; ; \; 1 - \frac{T_C}{T_H} = 1 - \frac{300}{500} = 0.4$$

The maximum steady-state power that could be extracted using the steady-state heat flux is then given as:

$$\begin{aligned} P_{\max} &= f \eta_{\max} Q \\ &\sim (10^{-3} - 10^{-2}) \cdot 0.4 \cdot (8 - 20) TW \\ &\sim (3 - 8) \cdot (10^{9} - 10^{10}) W \end{aligned}$$

This is much less than anticipated carbon-free energy demand discussed earlier. Thus, we conclude that the steady-state extraction of geothermal power does not play a significant role in fueling carbon-free power needs. However, there is another approach which can potentially provide significant energy resources, which we take up next.

Hot Rock Geothermal Energy: Heat Mining

The preceding analysis is assuming that the geothermal resource is used in a steady-state, and leads to a conclusion that capturing and converting this steady-state heat flux probably does not lead to a significant new primary energy resource simply due to the small value of the steady state heat flux in the near surface region. However, a reasonable question is: Could one extract the thermal energy form a volume rock at a rate that is fast

enough to be economically viable? In other words, could one "mine" the heat content of a volume of rock for a reasonably long period of time and then, once the rock volume has cooled to the point where the energy capturing process was no longer cost competitive, then move the mine to a new region where the process would be replicated? If so, then what is the potential resource provided by this approach?

The schematic view of such a heat mine would be as shown in Figure below. Consider a volume of rock, located at an average depth, d, below the Earth's surface. Cold working fluid at a temperature T_c is injected through a well into this volume of rock. The rock is either naturally porous or has been artificially fractured such that the working fluid can move through the volume. At some other location within the volume, the working fluid then is removed at a temperature T_h. For our initial purposes, we have taken the volume $V \sim l^3$ where l is the characteristic length of the heat mine.

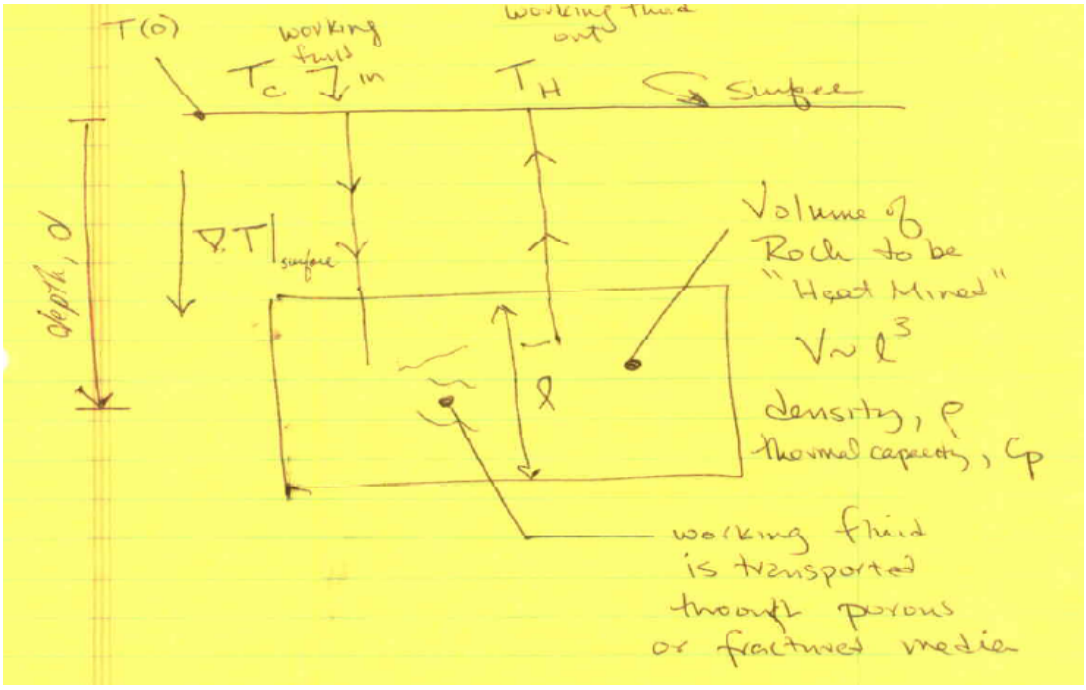


Figure 10.2: Schematic of a heat mine.

The temperature at depth *d* would be roughly: $T(d)\,;\; T(0)+d\nabla T\big|_{surface}$. As we discussed above, the Earth's near surface temperature gradient has values in the range: $\nabla T\big|_{surface}\,;\; 20-25\frac{K}{km}$. If *d* was ~10km then the hot reservoir temperature would be roughly $T_H \approx d\nabla T\big|_{surf} \sim 200-250\,C^{\circ}$ - i.e. high enough such that a reasonably efficient heat engine could be constructed and operated.

If volume is 14km x 14km x 1km, and if $\rho\,;\; 10^3\,{}^{J}\!/_{kg\cdot K}$, the thermal energy W_{th} that must be extracted in order to cool this amount of rock from 250 deg C to 200 deg C (ΔT=50degC) is estimated simply as

$$W_{th} \; ; \; V \cdot C_P \cdot \rho \cdot \Delta T$$
$$; \; 14^2 \cdot (10^3)^2 \cdot 10^3 \cdot 10^3 \cdot (3 \cdot 10^3) \cdot 50$$
$$; \; 200 \cdot 10^9 \cdot 3 \cdot 10^6 \cdot 50$$
$$; \; 6 \cdot 10^{17} \cdot 50$$
$$; \; 3 \cdot 10^{19} J$$

Compare this to the annual energy use in the U.S.:

$$W_{US} \; ; \; 4TW \cdot 10^4 hrs \cdot 3.6 \cdot 10^3 \frac{\text{sec}}{hr}$$
$$= 4 \cdot 10^{12} \cdot 4 \cdot 10^7$$
$$; \; 16 \cdot 10^{19} J$$

i.e. it represents 20% of the U.S. energy use. Thus, although this geothermal energy cannot be used in steady state, this simple example shows that there is sufficient heat content in deeper rocks to be of potential interest for energy applications. Of course, only a small fraction of this heat content could be tapped; however we also note that the volume of rock considered in this hypothetical example is relatively small compared to the possible rock volume that could possibly be used.

Let us now examine a second question: If we had such a geothermal heat mines, what would the lifetime of the mine be? To answer this question, let us assume that the hot and cold reservoir temperatures are given as

$$T(0) \approx T_C$$
$$T_H \approx T(d)$$
$$T_H - T_C \approx d \nabla T\big|_{surface}$$

where ∇T is the earth's temperature gradient at depths $d \to 10km$. Now, let us consider the heat mine volume to be a control volume, and apply a time-dependent energy conservation analysis.

The time-dependant energy conservation equation is:

$$\frac{\partial}{\partial t}\left(\rho C_p T\right) - \nabla \cdot Q \; ; \; S_{in} - S_{out}$$

where Q denotes the heat flux and the terms on the RHS denote the volumetric sources and sinks of heat. We can safely neglect S_{in} if we are extracting heat out of the volume quickly enough. The heat flux Q is the diffusive heat flux from the background (i.e. coming from deeper within the Earth). Let us integrate this expression over the volume V as denoted in the schematic above:

$$\rho C_p \int_V \frac{\partial T}{\partial t} dV - \int_V \nabla \cdot Q dV = \int_V \left(S_{in} - S_{out}\right) dV$$

Using Gauss' theorem we can re-write the second term on the left in terms of a surface integral:

$$\rho C_p \int_V \frac{\partial T}{\partial t} dV - \int_S Q \cdot dA = \int_V \left(S_{in} - S_{out}\right) dV$$

where S denotes the surface of the volume V. Let us now define the volume averaged quantities

$$\bar{T} = \int T dV \Big/ V$$

$$S_{in}^{tot} = \int_V S_{in} dV$$

$$S_{out}^{tot} = \int_V S_{out} dV$$

.

We can then re-write the energy balance equation in terms of these averaged quantities as

$$V\rho C_p \frac{\partial \bar{T}}{\partial t} - \int_S Q \cdot dA = \bar{S}_{in} - \bar{S}_{out}$$

We will also assume that $S_{out}^{tot} >> \int_A \underline{q} \cdot d\underline{a}$. To see the justification for this assumption, suppose that **Q** is mostly directed in the radial direction due to $\nabla T\big|_{surface}$, and has a value $q_r \sim 50 mW/m^2$ or so as we estimated in the previous section. If the projected surface area of the volume V in the $\hat{r}$ direction A_r ~10 x 10km = $10^8 m^2$ and we then estimate this area integral has a value

$$\int_A \vec{q} \cdot d\vec{A} \sim q_r A_r \sim 50 x 10^{-3} \cdot 10^8 = 5 MW$$

.

Note that the heat flux through the sides of the volume are small since Q is assumed to be in the radial direction. Thus, for our 10 x 10 km^2 heat mine, if $S_{out}^{tot} >> 5MW$ then we can safely neglect $\int_A q \cdot da$.

Now if S_{in}^{tot} is also negligibly small (which it will be unless the volume has an extraordinary concentration of radioactive material acting as a heat source within it), then our integral heat conservation equation becomes…

$$\frac{\partial}{\partial t}\int_V \rho C_p T dV \;;\; -\int_V S_{out} dV$$

$$\bar{T} = \int T dV$$

$$S_{out}^{tot} = \int S_{out} dV$$

$$V\rho C_p \frac{\partial}{\partial t}\bar{T} = -S_{out}^{tot}$$

Let us approximate $\frac{\partial \bar{T}}{\partial t}$; $\frac{\Delta \bar{T}}{\tau}$ where τ denotes a characteristic temperature decay time. Solving for τ we then find

$$|\tau| = \frac{\rho C_p \Delta \bar{T} V}{S_{out}^{tot}}$$

The thermodynamic conversion efficiency is limited by Carnot Value: $\eta \le 1 - \frac{T_C}{T_H}$. Thus, if T_H gets too low then η will get too small. Since it is likely that at the start of operations of the heat mine T_H ; $T(d)$, then it is likely that the initial conversion efficiency η will be determined by well depth:

$$\eta_{\max} = 1 - \frac{T_C}{T_o(d)}$$

Where again $T_o(d) = T(d)$ at the start of operations. The efficiency η will then decrease as the heat is mined and T_h falls. The question then arises: how long will the mine operate before the efficiency falls to some unacceptably low value?

To answer this question, let us define some minimum acceptable η, η_{min}. Obviously η_{min} corresponds to a minimum value for T_H, which we shall denote as $T_H^{\min}$. If T_H=T(d) always (working fluid comes to temp of the rock) then we can estimate the lifetime of the heat mine as follows:

Knowing that $|\tau| = \frac{\rho C_p \Delta \bar{T} V}{S_{out}^{tot}}$, let the total allowable change in rock temperature be given as:

$$\Delta \bar{T} = T_H\big|_{t=0} - T_H^{\min}$$
$$T_H\big|_{t=0} = T(d)$$
$$\Delta \bar{T} = T(d) - T_H^{\min}$$
$$T_H^{\min} = \frac{T_C}{1-\eta_{\min}}$$
$$\Delta \bar{T} = T(d) - \frac{T_C}{1-\eta_{\min}}$$
$$\therefore \Delta \bar{T} = \frac{(1-\eta_{\min})T(d) - T_C}{(1-\eta_{\min})}$$

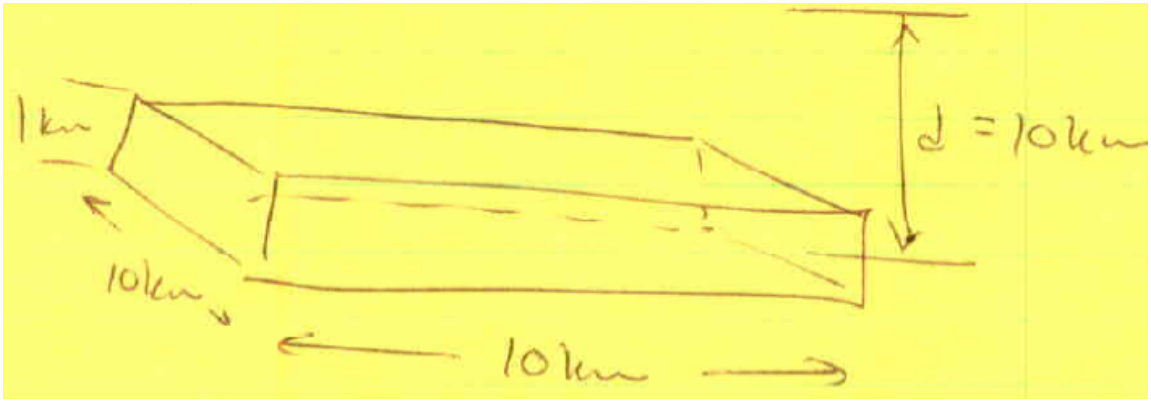
Thus the lifetime of the mine is $\tau = \frac{\rho C_p V}{S_{out}^{tot}} \cdot \frac{(1-\eta_{\min})T(d) - T_C}{1-\eta_{\min}}$, where we take S_{out}^{tot} to be fixed (it will actually decrease over time). The useable power P_{out} will be: $P_{out} = \eta S_{out}^{tot} \le \eta_{\max} S_{out}^{tot}$

Max $E_{electricity} = \int \rho C_p V dt \cdot \eta(T) = \int \rho C_p V \left(1 - \frac{T_c}{T}\right) dT$

$\eta(T)$ is T dependent

Example: let the depth d=10km, then $T(d)\big|_{t=0} \approx d\nabla T\big|_{surf} \sim 200 - 250\,C^c$. If T_C=$T_{surface}$=20°C and $\eta_{\max} = 1 - \frac{T_C}{T_H} = 1 - \frac{300K}{500K}$; 0.4 (though the actual efficiency will be lower, say 0.2 or so). Suppose we can afford $\eta_{\min} = \frac{1}{2}\eta_{\max}$.

$$\eta_{\min} = 0.2$$



V = 10^{11}m^3, ρ = 3 x 10^3 kg/m^3, $C_p \sim 10^3$ J/kg*K, and suppose $S_{out}^{tot} = 10^{10}$ W.

Then the lifetime estimate looks like:

$$\tau = \frac{3\cdot 10^3 \cdot 10^3 \cdot 10^{11}}{10^{10}} \frac{(0.8)500 - 300}{0.8}$$

$$\tau \simeq 360 \cdot 10^7 \simeq 4 \cdot 10^9 \text{ sec}$$

$$1 year \simeq 400 days \cdot 24 \cdot 3600 \simeq 3.6 \cdot 10^7 \text{ sec}$$

$$\therefore \tau \sim 100 years$$

The power output would start out at $P_{out} \le \eta_{\max} S_{out}^{tot} = 4GW$ and slowly decay to ~1GW over the timescale of ~100 years.

This simple analysis indicates that heat mining of significant volumes of rock located at depths of a few kilometers or greater could potentially provide GW-class sources of heat which could be converted to electricity or used directly for applications requiring low-to-medium temperature heat sources. We note that recent studies have explored this concept in more detail, and have also concluded that hot rock mining represents a potentially significant new primary energy resource. Although it is not renewable in the strict sense of the word, the resource is at the same time very large, and is widely distributed.

Put some real examples of the EGS:? (e.g. **www.altarockenergy.com**)

EERE.energy.gov/geothermal

Highlight the high capacity factor for geothermal (73%). (Compared to PV (14%) and wind (21%))

Source: 2007 survey of energy resources

Chapter 11: Solar Energy from Photovoltaic Cells

Introduction

Solar photovoltaic (PV) cells are composed of a junction of n-type and p-type doped semiconducting materials and electrical contacts to carry electrons to an external circuit. A simple schematic of such a cell is shown in Figure below. Photons, which are particles of light, impinge upon the PV cell and, if the photons have sufficient energy, can be absorbed and create an electron-hole pair within the cell. These charges can then move through the device, and the electrons can then move through an external circuit. Key questions that we would like to understand include:

- What mechanisms determine the maximum possible efficiency of these devices?
- How do the charge carriers move through the device?
- What does the current-voltage response of the device look like?
- What mechanisms determine the actual efficiency of such a device, and how can this efficiency be improved?

In this chapter, we introduce the reader to the essential elements, processes, and mechanisms that determine the answers to these questions.

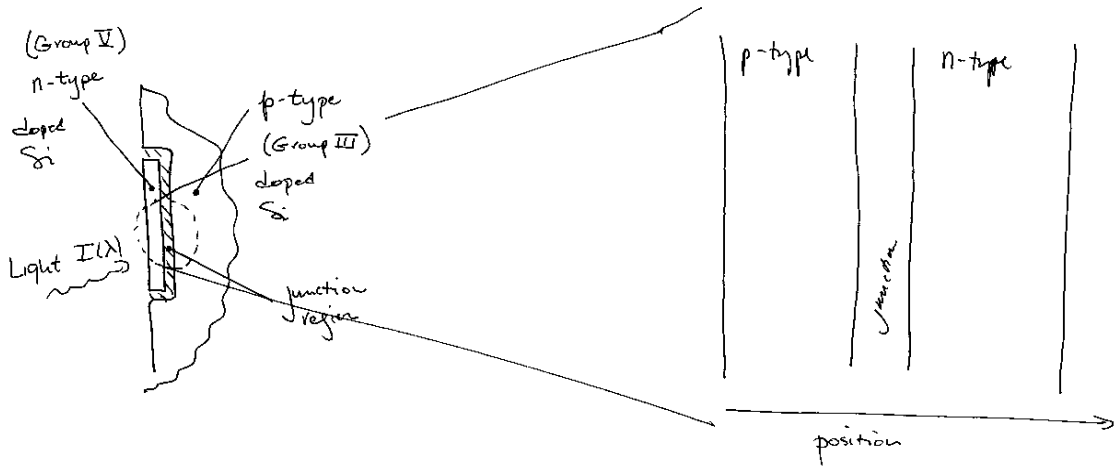


Figure 11.1: Schematic of p-n junction acting as a solar cell.

To proceed, we must first consider a few aspects of the behavior of solid state materials in general, and then specifically focus upon several important properties of semiconducting materials.

Fundamental aspects of solid state materials

We know from quantum mechanisms that individual atoms have distinct, discrete energy levels that are available to the electrons that are bound to the atomic nuclei. This is shown schematically in Figure below.

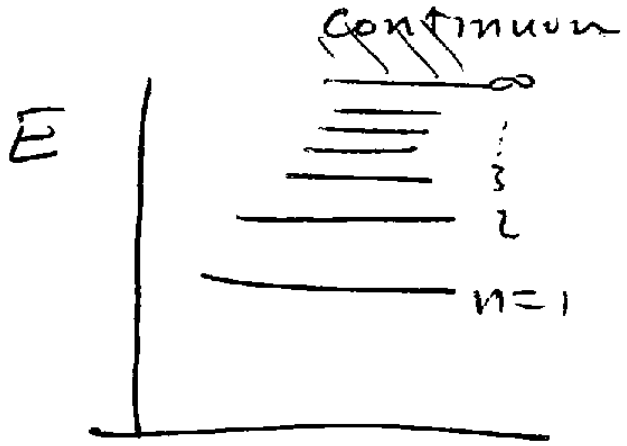


Figure 11.2: Schematic of allowed principle quantum energy levels within an individual atom

Let us now consider, on a qualitative sense at least, what happens to these discrete energy levels when we bring together many identical atoms. As the inter-atomic spacing becomes comparable in size to the electronic wavelengths associated with the bound electrons, then the electrons can begin to interact with each other, resulting in a merging or overlapping of the energy levels of the allowed electronic energy levels, again as shown schematically in Figure below. The quantum mechanical exclusion principle applies, which states that no two electrons can occupy the same state (i.e. be at the same location and same energy). As a result, when the atoms are at a finite temperature, then

the electrons follow Fermi-Dirac statistics, and have an energy distribution function given as

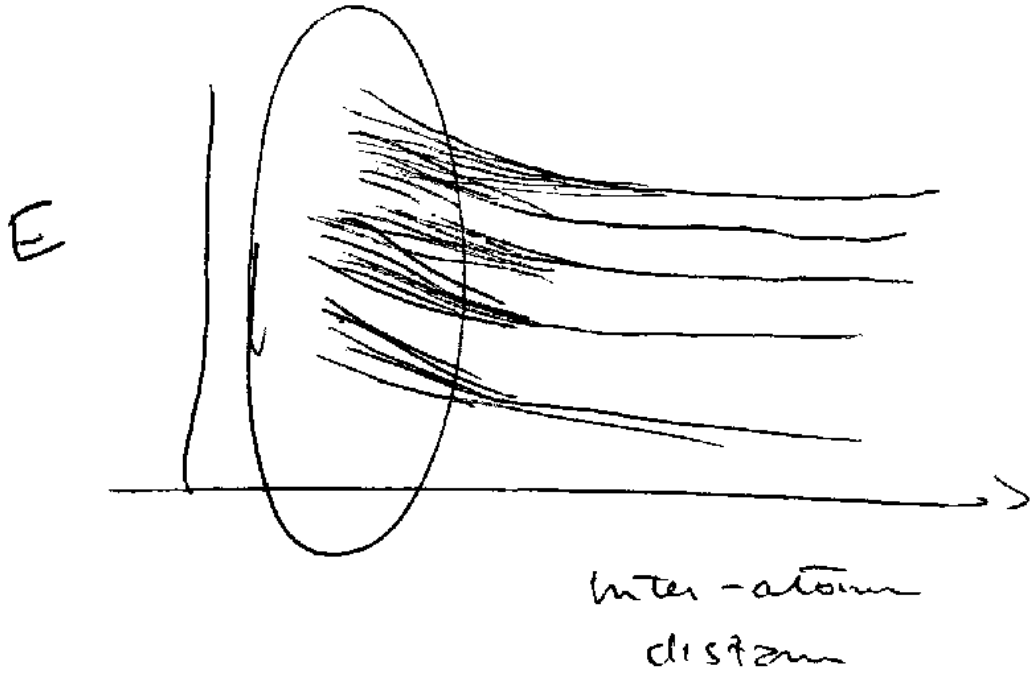


Figure 11.3: Schematic of evolution of allowed energy levels as the distance between atoms is decreased to the values encountered in solid materials.

In solid material, the envelopes of the electrons blend together and one has a distribution of energies within the solid.

Estimating Available Charge Carrier Densities

In a classical system in which particles are thought of as mathematical points, no two particles can occupy the same position. When the particles have quantum mechanical properties (such as the charged particles in a solid material), the equivalent principle states that the particles (e.g. electrons) cannot occupy the same quantum mechanical state

where the term State refers to the particle energy, momentum, spin, and position. This principle results in the Fermi-Dirac statistics for electrons in solids.

$$f(E) = \frac{1}{(exp^{\frac{E-E_F}{kT}}) + 1}$$

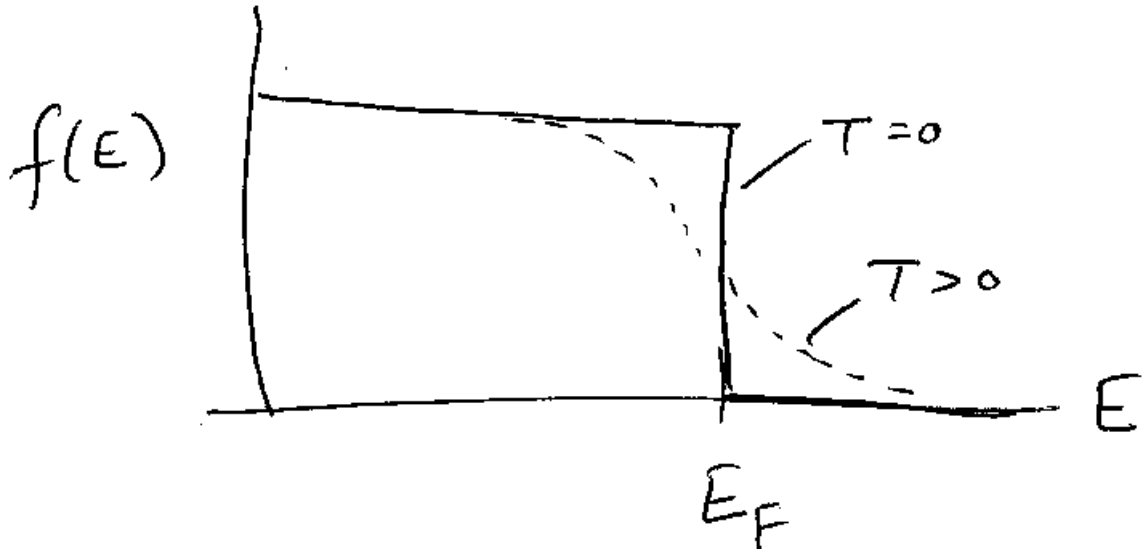


Figure 11.4: Schematic of the Fermi-Dirac probability distribution function for T=0 and T>0 systems.

The probability that particles have energy between E and E+dE is given as *f*(E)de. When one accounts for the fact that solid materials of interest here generally have a crystalline structure – i.e. the atoms of the solid are arranged in a periodic lattice which occupies three spatial dimensions – and that this spacing is comparable to the quantum mechanical wavelength of the electrons in the material, it turns out that the lattice can act somewhat like a diffraction grating acts with light, and can create constructive and destructive interference effects depending upon the wavelength of the electrons (which is

related in turn to the electron energy via the uncertainty principle). As a result, there exist a range of allowed electron energies (which correspond to constructive interference of the wave functions) and a range of forbidden electron energies (which correspond to the destructive interference of the wave functions).

Figure below shows these bands schematically for metals, insulators, and semiconducting materials. Metals, which area generally good electrical conductors, have the lower allowed energy range (denoted as the valence band) only partially filled with electrons, and have a Fermi energy, E_F that is below the maximum allowed valance band energy, E_V (i.e. $E_F<E_V$). The gap between E_V and the lowest allowed value E_C of the conduction band electrons, defined as $E_{gap}=E_C-E_V$, is not too large relative to the thermal energy, kT, and thus a few electrons can make their way into the conduction band, which are electrons whose energy is high enough to allow them to move through the material. Insulators have the lower valence band filled and have a relatively high band-gap energy relative to the thermal energy, i.e. $E_{gap}>>kT$, and thus have few electrons in the valence band and thus are poor conductors. Semiconducting materials have the Fermi energy lying in between EV and EC, and also have a smaller Egap such that $E_{gap}\sim kT$. As a result, their conducting properties lie somewhat between metals and insulators.

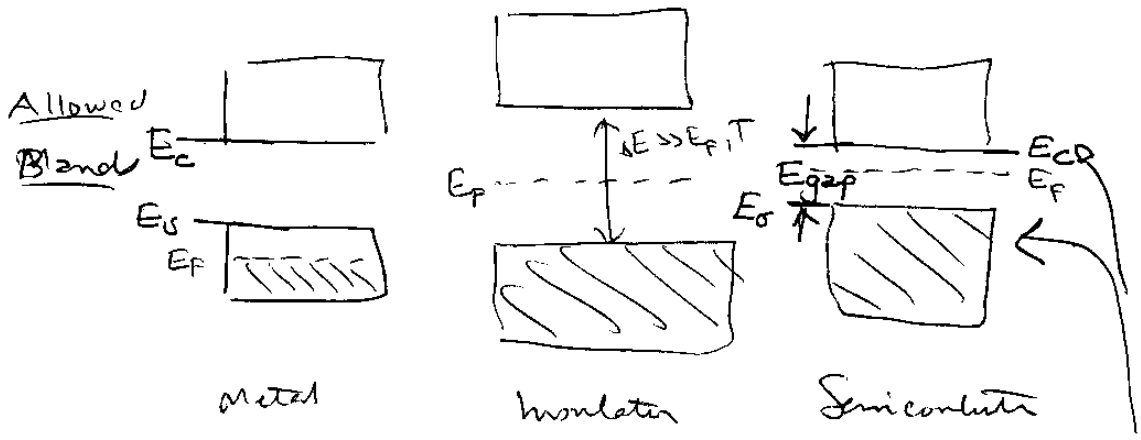


Figure 11.5: Schematic of allowed energy levels in metallic, insulating, and semiconducting materials.

Estimate of Maximum Theoretical Conversion Efficiency of a Solar Cell

The above result indicates that it requires an energy E_{gap} to move an electron from the valence band (where the electrons are immobile and tied to a particular atom in the crystal lattice) into the conduction band (where they are mobile in the semiconducting material). We can use this result to estimate the maximum possible conversion efficiency of a solar cell. First, we note that the flux of photons from the sun follow the blackbody distribution vs. frequency, f, which is given as

$$P(f) = A\frac{f^3}{\exp(hf/k_BT)-1}$$

where here A is a constant (which will end up canceling out in the end of the analysis). Now the energy of a photon of frequency f is given as Ephoton=hf, where h is Planck's constant. If a photon with energy Ephoton>Egap is absorbed in the material, then a

mobile electron (and a corresponding mobile hole) (i.e. an electron-hole pair) is created. The photon energy that is in excess of Egap then appears as the kinetic energy of the charge carrier pair. If the photon energy is such that $E_{photon}<E_{gap}$ then the photon can be absorbed, but this absorption simply heats the material via the generation of lattice vibrations (called phonons in the solid state literature) which have much lower energy than Egap, and does not produce a charge carrier pair. In the process, the photon energy is no longer available to generate charge carrier pairs and thus is lost from the system.

To quantify this efficiency, we first write the fraction of the photon spectrum that has $E_{photon}<E_{gap}$ as

$$G_L = \frac{1}{P_{tot}} \int_{f=0}^{f=f_{gap}} P(f)df$$

where we have defined the photon frequency corresponding to the band gap energy as $f_{gap} = E_{gap}/h$ and P(f) corresponds to the blackbody photon spectrum. This fraction G_L of the photon power is immediately lost. To proceed further, let us now define the flux of photons (i.e. number of photons incident on a surface per unit time) that have $f>f_{gap}$ to be given as

$$\phi_{gap} = \frac{1}{h} \int_{f_{gap}}^{\infty} \frac{1}{f} P(f)df$$

.

When a photon in this population is absorbed in the material we will then assume that it creates a charge-carrier pair that have a kinetic energy given as

$$E_{pair} = E_{photon} - E_{gap}$$
$$= hf - E_{gap}$$

i.e. we assume that the excess photon energy is converted into kinetic energy. This energy is then dissipated as heat in the material via a series of collisions. Thus, the maximum useful power that can be extracted from the photon flux ϕ_{gap} is then given as $P_{\max} = \phi_{gap} E_{gap}$. We can then define a maximum possible conversion efficiency in terms of the ratio of this maximum possible power to the total power flux contained in the incident radiation

$$\eta_{\max} = \frac{P_L}{P_{tot}} = f_{gap} \frac{\int_{f_{gap}}^{\infty} \frac{1}{f} P(f) df}{\int_0^{f_{gap}} P(f) df}.$$

This can be put into a convenient form for evaluation using the expression for the blackbody emission above:

$$\eta_{\max} = \frac{15}{\pi^4} \xi_0 \int_{\xi_0}^{\infty} \frac{x^2}{e^x - 1} dx$$
$$where \quad \xi_0 = \frac{qE_{gap}}{k_B T_{bb}}$$

For Egap=1.1eV (energy gap for silicon) and Tbb=6000 deg K (i.e. the temperature of the Sun) we find that $\eta_{\max} \approx 0.44$. This is an intrinsic limit to the conversion efficiency, set only by the assumption that photons must have an energy that equals or exceeds the energy cost of producing a charge carrier pair in the material. Losses of these charge

carrier pairs subsequent to their production, other non-charge carrier producing absorption processes, and so forth will then all conspire to reduce the efficiency from this upper limit. Of course, the value of the intrinsic efficiency can also be increased by decreasing the band gap energy – which is accomplished by changing materials for example.

Governing Equations of a p-n junction PV cell

The above model provides an estimate of the maximum conversion efficiency of a PV cell made of material with a band gap energy, E_{gap}. However, it neglects important aspects that tend to reduce the actual efficiency to values that are smaller than this maximum estimated efficiency. In order to quantify the actual efficiency of a PV cell, we need to develop a mathematical model for the p-n junction that includes the following:

- The density of charge carriers in the p-type, n-type, and junction regions.
- An understanding of how these charge carriers move in these 3 regions
- Understand how charge carriers are generated (photon absorption) and lost.
- Relate charge carrier distribution to the electric field in the cell.
- Put it all together.

We can use this picture to begin to understand then the density of electrons that are in the conduction band of a particular material. Referring to Figure below, let us denote N(E) as the number of allowed quantum mechanical states per unit volume per unit energy in the material. The quantity f(E) then denotes the probability that a particular energy level is occupied. Thus, it stands to reason that the product of these two quantities will then give the number of electrons in the conduction band per unit energy per unit

volume. Integrating will then give the number of conduction band electrons per unit volume.

Charge Carrier Densities

For the conduction and valence band regions, we can write the number of allowed (not necessarily occupied) states / unit volume / unit energy

Cond. Band: $N_C(E) = \frac{8\sqrt{2}\pi m_e^{*3/2}}{h^3}(E - E_C)^{1/2}$; $E \geq E_C$

Valence Band: $N_c(E) = \frac{8\sqrt{2}\pi m_h^{*3/2}}{h^3}(E_V - E)^{1/2}$; $E \leq E_V$

Where E_c, E_V are energy of conduction band and valence band

Where m_e^* ~ effective mass of an electron $(\frac{m_e^*}{m_e} \cong 0.2mSi) \Leftarrow$ and m_h^* is the effective mass of a hole. More complex density-of-states calculations show that

$m_e^*/m_e^o = 1.08$

$m_h^*/m_e^o = 0.81$

where m_e^o is the mass of a bare electron.

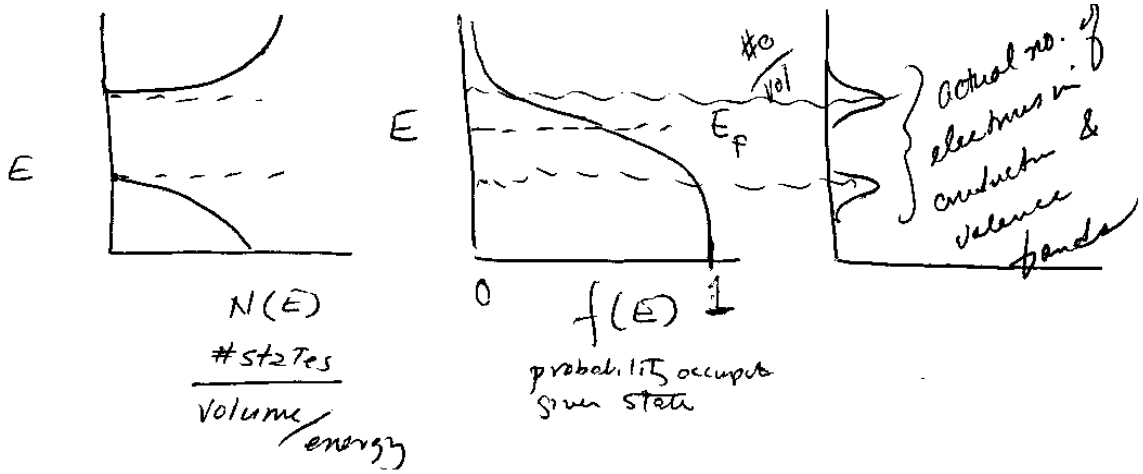


Figure 11.6: (a) N(E), Number of allowed states/unit volume/unit energy, (b) Fermi-Dirac Distribution function, f(E), given by number of particles/unit energy, (c) Number of occupied states given by N(E)f(E).

The number of electrons in conduction band per unit volume is then given as:

$$n = \int_{E_C}^{E_c^{\max}} f\left(E\right) N\left(E\right) dE$$

For $E_C >> kT$, and approximating the maximum conduction band energyas $E_C^{max} \approx \infty$, we can then expand *f*(E) and then integrate analytically. The result for the number of electrons in conduction band per unit volume, i.e. for the conduction band electron density, n, is then given as

$$n = N_C \exp\left[\left(E_F - E_C\right)/kT\right]$$

$$N_C = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2}$$

$$n = N_C \, exp\left[\left(E_F - E_C\right)/kT\right]$$

where

$$N_C = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2}$$

One can go through a similar procedure to track the density of holes (which are essentially equivalent to a missing electron – i.e. a net positive charge) in the valence band:

$$p = N_V \exp\left[\left(E_V - E_F\right)/kT\right]$$

$$N_V = 2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{3/2}$$

For a pure semiconductor (no dopant) we have: $n = p = n_i$ since the material has no net electrical charge. It is common to define the "intrinsic concentration", n_i. in terms of the product of n and p:

Now since n=p in a pure semiconductor we can write:

$$N_C \exp\left[\left(E_F - E_C\right)/kT\right] = N_V \exp\left[\left(E_V - E_F\right)/kT\right]$$

This expression can then be solved for the Fermi energy E_F:

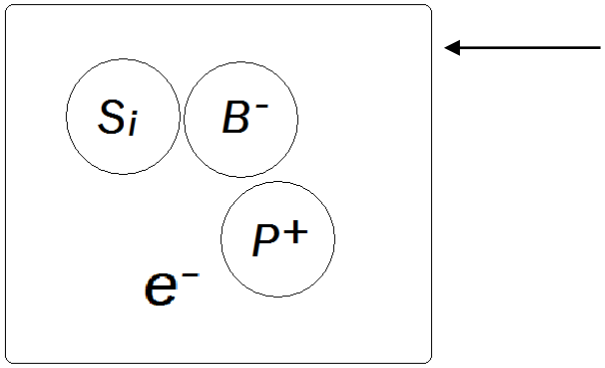
$$E_F = \frac{E_c + E_V}{2} + \frac{kT}{2}\ln\left(\frac{N_V}{N_C}\right)$$

The above considerations hold for pure semiconducting materials. However, the PV cell is composed of a junction of so-called doped semiconducting materials, which have had carefully chosen impurities added in extremely dilute quantities. Dopants which have a single weakly bound electron in the outer valence orbital can easily give up this electron to the surrounding material; semiconductors that have been doped with these

types of dopants are referred to as n-type materials. Similarly, dopants that are missing a single valence electron can easily acquire an extra electron; semiconducting materials that have been doped with these materials are referred to as p-type materials. Let us now assume that we have an n-type material that has a donor dopant density N_D which givens the number of dopant atoms per unit volume. Charge neutrality give us

Better to give a figure?

No net Charge



$$p - n + N_D^+ - N_A^- = 0$$

where here we denote the hole and electron density as p and n respectively and we have assumed that all of the dopant atoms have lost their electron and hence have formed a positive ion. Since all of the mobile electrons, n, have come from the valence electrons of the dopant atoms we can then write that $n = N_D^+$ and, since we expect that $N_D \approx N_D^+$ (i.e. most of the dopant atoms are ionized), we can then write from charge neutrality that

$$p = n - N_D^+ << n$$

in other words the hole density is much smaller than the electron density in n-type materials. In p-type Si with a dopant density N_A (called the acceptor density), the carrier densities satisfy an inverse ordering as compared to the n-type densities:

$$N_A^- \approx N_A$$
$$p \approx N_A$$
$$n \approx \frac{n_i^2}{N_A} << p$$

In doped semiconductors, we can then approximate the electron and hole densities as:

$$n \approx N_D \approx N_C \, exp\left[\left(E_F - E_C\right)/kT\right]$$
and
$$p \approx N_A \approx N_V \, exp\left[\left(E_V - E_F\right)/kT\right]$$

Solving for $E_F - E_C$ and $E_V - E_F$ for the two types of materials then gives

$$E_F - E_C = kT \ln\left(\frac{N_D}{N_C}\right)$$
$$E_V - E_F = kT \ln\left(\frac{N_A}{N_V}\right)$$

for n-type and p-type materials respectively and where E_C and E_V are fixed by the material properties of the semiconducting material. These expressions imply that as the n-type (p-type) dopant level is increased, the Fermi level moves towards E_C (E_V). The results also imply that if we form a junction consisting of an n-type region in contact with a p-type region, then the hole and electron density distribution will look something like that shown schematically in Figure below.

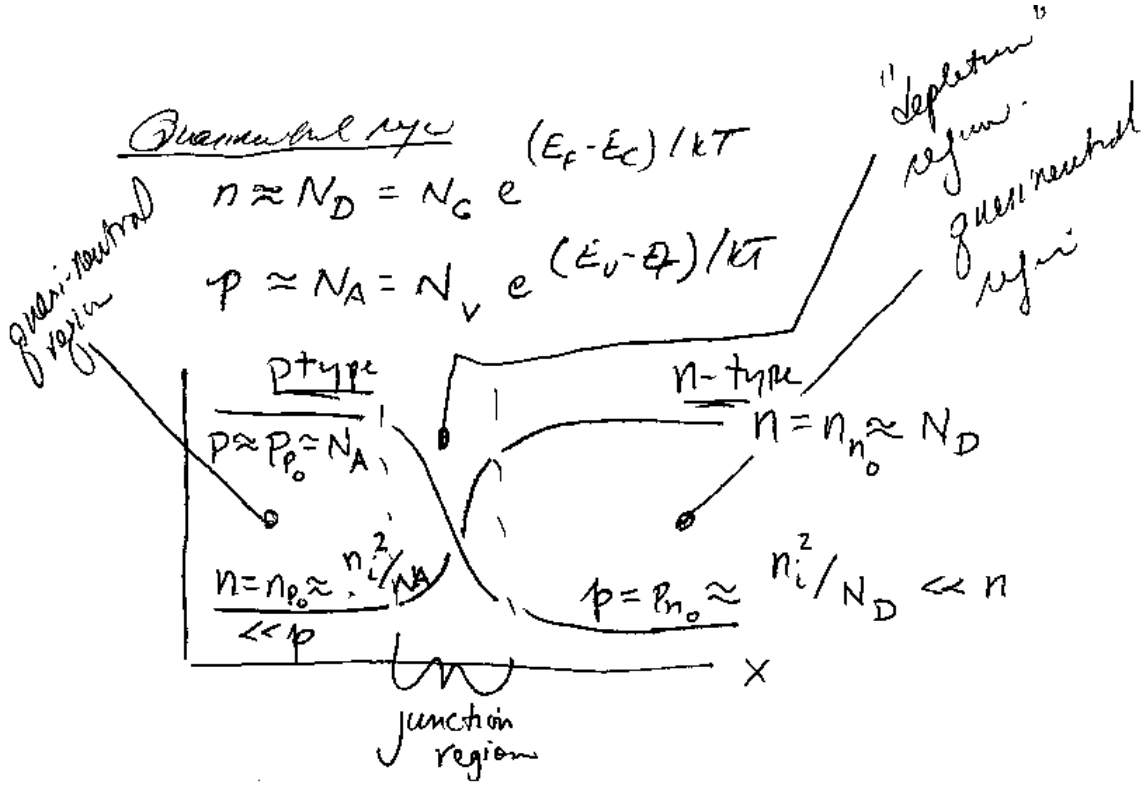


Figure 11.7: Schematic of hole density, p, and electron density, n, across a p-n junction.

Relating Charge Density to Electric Field

The change distribution in this device is related to the divergence of the electric field via the Poisson's equation, which is given as

$$\nabla \cdot E = \frac{1}{\varepsilon_s}\rho$$

Where ε_s is the permittivity of the semiconductor/solid, and ρ is the charge density given as

$$\rho = q(p - n + N_D^+ - N_A^-)$$

where q = 1.6 x 10^{-19} C is the (absolute magnitude) charge of an electron

p ~ the number of holes per unit volume [units: m^{-3} or cm^{-3}]

n ~ the number of electrons per unit volume [units: m^{-3} or cm^{-3}]

N_D^+ ~ the density of donor impurities in the semiconductor

N_A^- ~ the density of acceptor impurities in the semiconductor

$\xi_s = \xi \cdot \xi_o$

ξ_o = **permittivity of vacuum=8.85·10^{-12} F/m**

ξ = **relative permittivity (17.7 for Si)**

In one dimension, Poisson's equation then reads $\frac{\partial E}{\partial x} = \frac{\rho}{\varepsilon_S}$ where the electric field E is now understood to point in the x direction.

.

Motion of charges

The electrons and holes can move through the solid material under the action of an electric field and/or a concentration gradient at finite temperature. This motion forms a current density, *J*, with units of Amperes per unit area. *J* is composed of the motion of electrons and holes, i.e. $J = J_e + J_h$ where J_e and J_h, are related to *n* and *p* by the equations:

$$J_e = q_e \mu_e n E + q D_e \frac{dn}{dx}$$ ← **Current due to electron flow (el⁻ are negatively charged)**

and

$$J_h = q_h \mu_h p E + q D_h \frac{dp}{dx}$$

Here, μ_e and μ_h are the mobilities **(in cm^2/v-s)** of the electrons and holes which determines the average drift speed of the electrons and holes under the action of the electric field, i.e.

$$u_e = \mu_e \cdot E$$
$$u_h = \mu_h \cdot E$$

and D_e and D_h are the diffusion coefficients for the electrons and holes. In other words, there is a convective portion of the charged particle motion ($\propto E$) and a diffusive part proportional to the spatial gradient of the electron or hole density. The coefficients D and μ are related by the Einstein relations:

$$D_e = \frac{kT}{q}\mu_e$$
$$D_h = \frac{kT}{q}\mu_h$$

here k ~ Boltzmann's constant and T refers to the temperature of the semiconductor.

Charge Conservation:

Because the electrons and holes can move and can be created by photon absorption, and destroyed by recombination within the PV cell, the conservation law for their density must include suitable volumetric source and sink terms that correspond to the rate of production or destruction of the charges per unit volume per unit time.

In analogy to the fluid mass conservation law we can then write:

$$\frac{\partial n}{\partial t} + \frac{1}{q}\nabla \cdot J_e = U - G$$

$$\frac{\partial p}{\partial t} + \frac{1}{q}\nabla \cdot J_h = -(U - G)$$

here U denotes the net recombination rate for electrons and holes

Units: [U] ~ $\frac{\#}{volume \cdot time} \sim \frac{\#}{m^3 \cdot \sec} or \frac{\#}{cm^3 \cdot \sec}$

And G denotes the net generation rate of electrons and holes and has the same units as U.

Here, we are only interested in steady-state, 1-D version of this equation. Thus, the equations for charge conservation in this case then become:

$$\frac{1}{q}\frac{\partial J_e}{\partial x} = U - G$$

$$\frac{1}{q}\frac{\partial J_h}{\partial x} = -(U - G)$$

For convenience, let us now recap the key equations that will govern the behavior of the charge-carriers in a simple one-dimensional p-n junction device, which when illuminated with a flux of photons with energy above the band gap energy will then produce a current in the device.

$$\frac{\partial E}{\partial x} = \frac{\rho}{\varepsilon}$$

$$\rho = q(p - n + N_D^+ - N_A^-)$$

$$J = J_e + J_h$$

$$J_e = q\mu_e nE + qD_e \frac{dn}{dx}$$

$$J_h = q\mu_h pE - qD_h \frac{dp}{dx}$$

$$\frac{1}{q}\frac{\partial J_e}{\partial x} = U - G$$

$$\frac{1}{q}\frac{\partial J_h}{\partial x} = -(U - G)$$

Next, we develop an approximate solution to these equations, which gives an understanding of how a PV cell operates.

The Ideal Diode as a Solar Cell

The previous set of equations has an approximate solution which allows us to understand the essential elements of PV cell operation. We must make several approximations to make this solution tractable.

The density of electrons and holes in a semiconductor is given as

$$n = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2} \exp\left[\left(E_F - E_C\right)/kT\right]$$

$$2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2} = N_C$$

$$p = 2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{3/2} \exp\left[\left(E_F - E_V\right)/kT\right]$$

$$2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{3/2} = N_V$$

Where m_e^* and m_h^* are the effective mass of an electron and a hole, and E_F is the "Fermi Energy", E_C is the minimum energy of the conduction band and E_V is the maximum energy of the valence band as we have discussed earlier.

As discussed above, for a pure semiconductor, n=p since creation of each electron leaves a vacancy or hole in the conduction band. Thus, in this case

$$np \approx N_C N_V \, exp\left[-E_g/kT\right]$$

Where $n_i^2 = np$ is the squared intrinsic concentration of charge and $E_{gap} = \left(E_C - E_V\right)$is the band gap energy. The Fermi Energy, E_F, was found earlier to be given as

$$E_F = \frac{E_c + E_V}{2} + \frac{kT}{2}\ln\left(\frac{N_V}{N_C}\right)$$

For weakly doped n-type material we found

$$n \approx N_D$$

$$N_D^+ \approx N_D$$

$$p \approx \frac{n_i^2}{N_d} << n$$

with

$$n = N_C \, exp\left[\left(E_F - E_C\right)/kT\right]$$

and

$$E_F - E_C = kT \, ln\left(\frac{N_D}{N_C}\right).$$

$$n = N_C \exp\left[\left(E_F - E_C\right)/kT\right]$$

$$E_F - E_C = kT \ln\left(\frac{N_D}{N_C}\right)$$

Similarly, for weakly doped p-type material we found

$$p \approx N_A$$

$$N_A^- \approx N_A$$

$$n \approx \frac{n_i^2}{N_A} << p$$

with

$$p = N_V \, exp\left[\left(E_V - E_F\right)/kT\right]$$

and

$$E_V - E_F = kT \, ln\left(\frac{N_A}{N_V}\right).$$

So what happens to the spatial distribution of n(x) and p(x) if we place a piece of p-type and n-type Si together? This is the configuration of a "p-n diode" which forms the basis of the simplest and most common type of solar PV cell. Figure below illustrates the result. Suppose the interface layer lies between two locations, a and b, which lie within the p-type and n-type materials respectively. At the instant that the two materials are joined, there will be a step=like distribution to the hole and electron densities. As a result, rapid diffusion will cause this step-like density distribution to relax into a smoother distribution, as shown in the figure. Eventually the diffusion process will reach a new equilibrium in which the hole and electron densities smoothly vary as shown in the

schematic. We now seek to learn what the distribution looks like quantitatively; in particular we wish to determine the hole density at point b (denoted as p_{nb} in the figure) and the electron density at point a (denoted as n_{pa}).

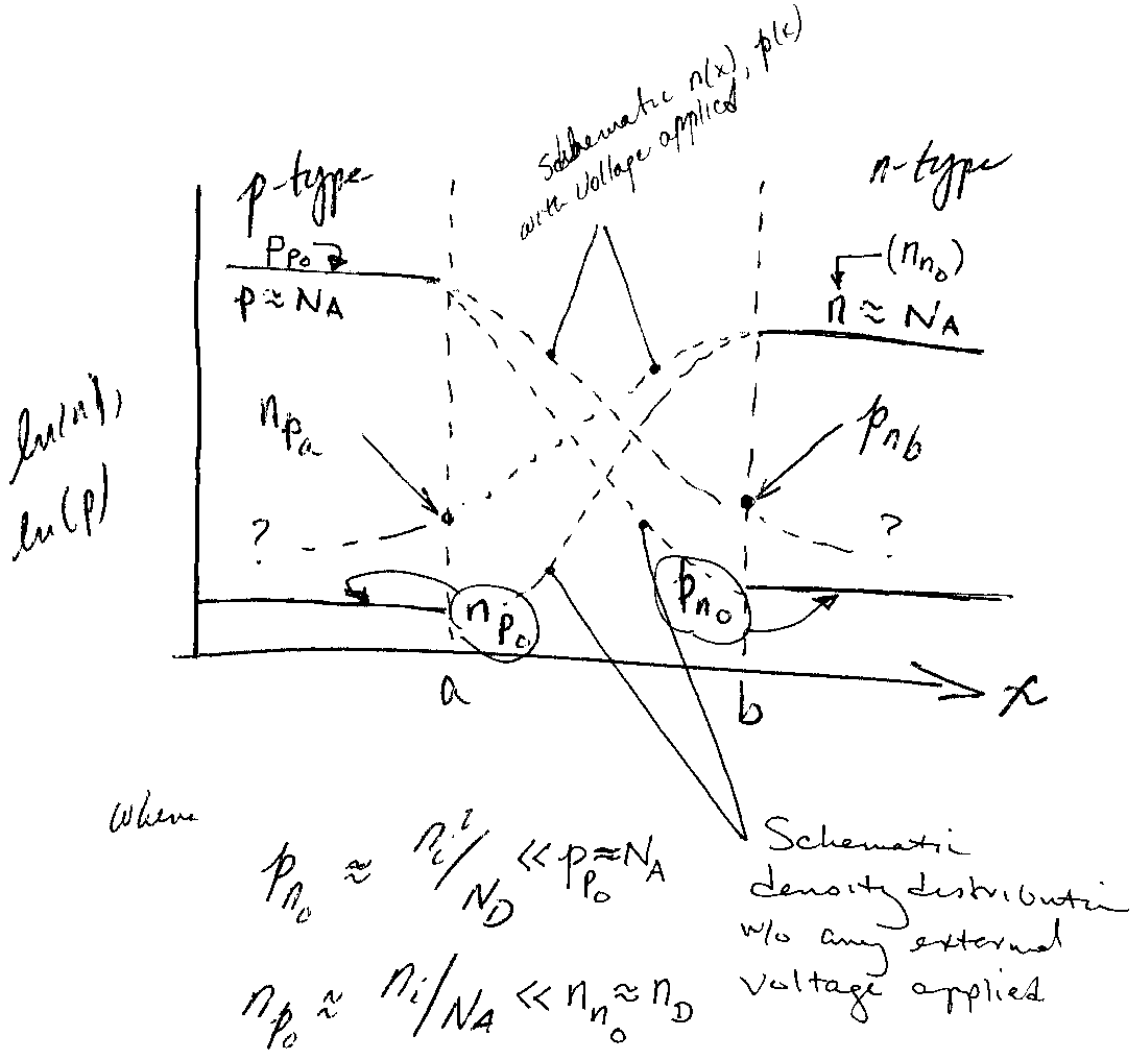


Figure 11.8: Schematic of a p-n diode showing the relative densities of holes and electrons. Note the development of a gradient in the densities in the region near the interface, which is also known as the junction.

To determine the exact charge carrier distribution across the junction, we must carry out some additional analysis. First, we note that in general, a system in thermal equilibrium can have only one E_F. Thus, when we bring the two sides of the p-n junction together, the energy levels have to adjust so that there is a single Fermi energy across the whole device. For a p-n junction the energy levels will then look like:

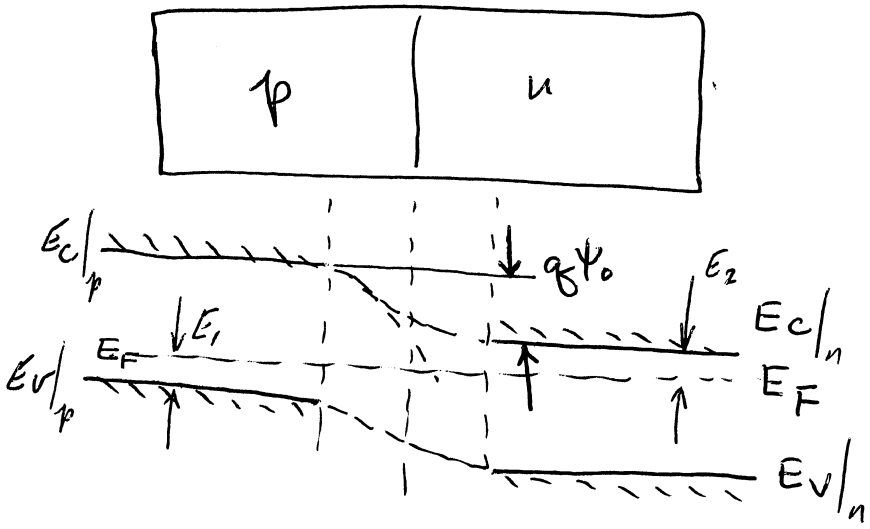


Figure 11.9: Energy levels within a p-n junction

Referring to Figure above, we define ψ_o , as the change in potential energy across the diode. It can be written as

$$\psi_o = E_g - E_1 - E_2$$

where

$$E_1 = E_F - E_V = kT \ln\left(\frac{N_V}{N_A}\right)$$

$$E_2 = E_C - E_F = kT \ln\left(\frac{N_C}{N_D}\right)$$

Combining expressions for E_1 and E_2 gives their difference as

$$E_F - E_V = \frac{E_C - E_V}{2} + \frac{kT}{2}\ln\left(N_V / N_C\right)$$

and we can then solve for the potential change across the diode as

$$q\psi_o = E_g - kTln(\frac{N_C N_V}{N_A N_D})$$

$$n_i^2 = N_C N_V \exp\left(-E_g / kT\right)$$

$$E_g = kTln(\frac{N_C N_V}{n_i^2})$$

$$q\psi_0 = kTln\left(\frac{N_C N_V}{n_i^2} \cdot \frac{N_A N_D}{N_C N_V}\right)$$

$$\psi_0 = \frac{kT}{q}\ln\left(\frac{N_C N_V}{n_i^2}\right)$$

In the p-type and n-type materials far away from the junction, we have already found the hole and electron densities and, presumably, far away from the interface these densities have not changed. In particular, we found earlier that the minority carrier densities (i.e. hole density in the n-type material and electron density in the p-type material) are given as

$$p_{n_o} \approx \frac{n_i^2}{N_D} \text{ and } n_{p_o} \approx \frac{n_i^2}{N_A}$$

while the majority carrier densities (i.e. the hole density in the p-type region and the electron density in the n-type region)

$$p_{p_o} \approx N_A + n_{p_o} \; ; \; N_A$$
$$n_{n_o} \approx N_D + p_{n_o} \approx N_D$$

We can now use these relations to find the ratio of majority to minority carrier density across the junction

$$\frac{p_{p_o}}{p_{n_o}} = \exp\left[q\psi_o \Big/ kT \right]$$

and

$$\frac{n_{n_0}}{n_{p_0}} \approx exp\left[q\psi_0 \Big/ kT \right]$$

or

$$\frac{n_{p_0}}{n_{n_0}} \approx exp\left[-q\psi_0 \Big/ kT \right]$$

and

$$\frac{p_{n_0}}{p_{p_0}} \approx exp\left[-q\psi_0 \Big/ kT \right]$$

Thus, the minority carrier density is decreased exponential from the majority carrier density located on the opposite side of the junction.

Next, let us now consider the charge distribution in the junction region. We refer to the Figure below.

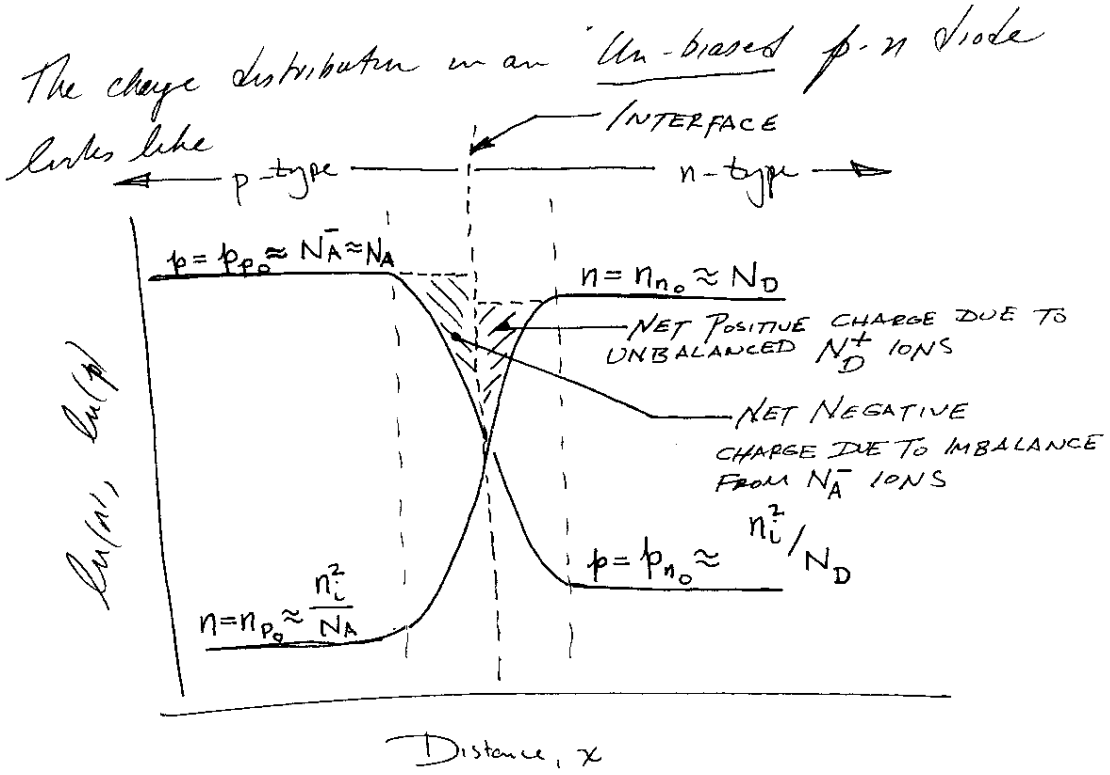


Figure 11.10: Spatial distribution of mobile charge carrier densities and net uncovered ion densities across a p-n junction

In the p-type region, located to the left of the Interface, the hole density is depleted from the original value p_{p_0} over some distance into the p region. As a result, the acceptor ions, with a density N_A^- no longer have their charge completely cancelled out by the holes and, as a result, this region acquires a negative charge density. Similarly, in the n-type region, located to the right of the interface, the electron density is depleted from the original value n_{n_0} over still-to-be determined distance. Again, this region has a net

charge density, but the sign of this charge is now positive since the donor ions with a density N_D^+ have been "uncovered" by the diffusion of the electrons across the interface.

To make further progress, we must now make the first approximation: the so-called Depletion Approximation which assumes that the p-n device connects different regions:

a) The "Quasi-neutral" regions where the charge density is zero, and

b) The Depletion region where the electron and hole densities are so small that they can safely been taken to vanish, i.e. n≈0, p≈0.

This crude model for the charge distribution then gives a charge distribution as shown in Figure below.

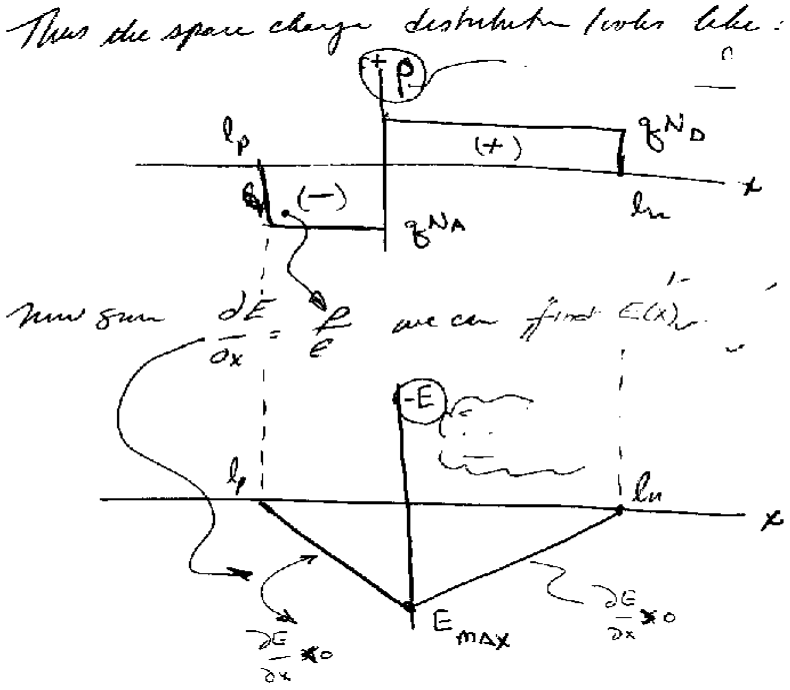


Figure 11.11: Charge density distribution across the p-n junction

We can use this charge distribution in Poisson's equation to find the electric field and, since the electric field is related to the gradient of the electrostatic potential, we can then find the potential distribution as well.

Since this problem only involves an electrostatic electric field, we can write $E = -\frac{\partial \psi}{\partial x}$ which then allows us to write Poisson's equation as $\frac{\partial^2}{\partial x^2}\psi = -\frac{\rho}{\xi_s}$

Thus for x < l_p (where $\phi = 0$) ; $\rho = 0$ at $x \to -\infty$ and therefore the potential obeys the equation

$$\frac{\partial^2 \phi}{\partial x^2} = 0$$

in this region.

For l_n < x < 0 the potential obeys the equation

$$\frac{\partial^2 \phi}{\partial x^2} = -\frac{-qN_A}{\varepsilon_s} = const = \frac{qN_A}{\xi_s} > 0$$

and for 0 < x < l_n the potential obeys the equation

$$\frac{\partial^2 \phi}{\partial x^2} = -\frac{qN_D}{\varepsilon} = const = -\frac{qN_D}{\xi_s} < 0$$

Finally, for x > l_n the potential obeys the equation

$$\frac{\partial^2 \phi}{\partial x^2} = 0$$

.

The boundary and matching conditions apply at x = l_p, x = l_n, and x=0

$$\phi'(l_p) = 0$$
$$\phi'(-\delta x) = \phi'(+\delta x)$$
$$\phi''(-\delta x) = \phi''(+\delta x)$$
$$\phi'(-l_n) = 0$$

The potential distribution that satisfies these requirements will have the following shape.

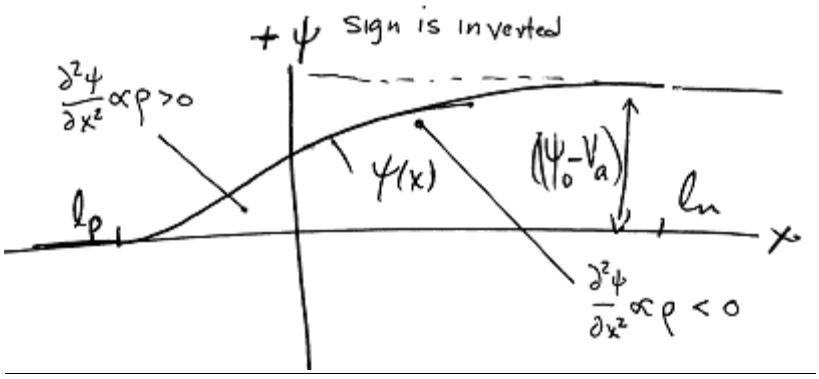


Figure 11.12: Electrostatic potential distribution across the p-n junction.

The maximum electric field will be given as

$$E_{\max} = -\left[\frac{2q}{\varepsilon}(\psi_o - V_A)\Big/\left(\frac{1}{N_A}+\frac{1}{N_D}\right)\right]^{1/2}$$

and the width, W, of the charged region is given as

$$W = l_n + l_p = \left[\frac{2\varepsilon}{q}(\psi_o - V_A)\left(\frac{1}{N_A}+\frac{1}{N_D}\right)\right]^{1/2},$$

where $l_p = W\dfrac{N_D}{N_A + N_D}$ and $l_n = W\dfrac{N_A}{N_A + N_D}$ denote the length of the charged regions in the p-type and the n-type regions. Note that in these expressions we have also included the possibility that an external voltage, Va, might also be applied. The sign of this voltage was assumed to be such that the p-type region would be biased to a positive value relative to the n-type region.

Unbiased (V_a = 0) Case:

Let us now consider what happens when no external voltage (usually refered to as a "bias") is applied to the junction. In this case, for x < a and for x > b, $\rho = 0$ and $\vec{E} = 0$, $\dfrac{\partial n}{\partial x} = 0$, $\dfrac{\partial p}{\partial x} = 0$. As a result, there can be no charge motion since the electric field vanishes and there is no density gradient present to drive diffusion of charge carriers. Now, for the region a < x < b the electric field and charge distribution is as discussed above. However, in steady-state there will be no current across this region. Thus, using the charge current equation we can then write

$$J_{c,h} = 0 \Rightarrow \begin{aligned} q\mu_c nE &= -qD_C \frac{\partial n}{\partial x} \\ q\mu_h pE &= +qD_h \frac{\partial p}{\partial x} \end{aligned}$$

Now we note that *E* is large in the <u>depletion</u> region which forms at the edge of the p and n type materials. This is also when $\partial n/\partial x$ and $\partial p/\partial x$ are largest. The distribution of charge is then as shown in Figure below.

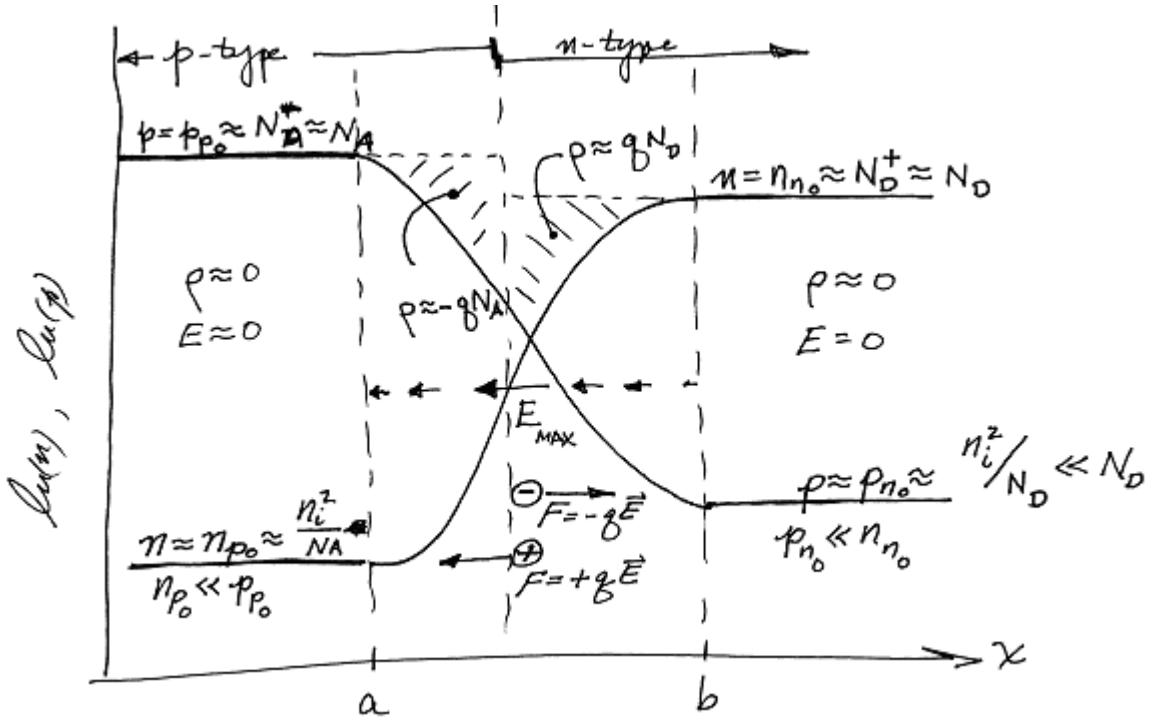


Figure 11.13: Charge distribution across an unbiased p-n junction.

Forward Biased (V_a >0) Case:

If the external Voltage Va is finite, then a current can begin to move across the junction.

The net current through this region is the difference between two large terms the

for example for holes the current density is given a:

$$J_h = qp\mu_h E - qD_h \partial p / \partial x .$$

Thus, we make a Second Approximation in the depletion region: we will assume that the relation

$$qp\mu_h E \approx qD_h \frac{dp}{dx}$$

will always hold – even when the current J is not zero. Now the Einstein relation (which is derived in kinetic theory) relates μ and D as $\frac{D}{\mu} = kT$. This expression can then be used to write the electric field in terms of the hole density gradient:

$$E \approx \frac{kT}{q}\frac{1}{p}\frac{dp}{dx}$$

Now since $E = -\partial\psi/\partial x$ we can then integrate p across the depletion region:

$$\psi_o - V_a = \frac{kT}{q}\ln p\Big|_a^b$$

$$= +\frac{kT}{q}\ln\frac{p_{p_a}}{p_{n_b}}$$

Where $p_{p_a} = p(x=a) \approx p_{p_o}$ and $p_{n_b} = p(x=b)$ with $V_a \neq 0$

After rearranging we can then write

$$p_{n_b} = p_{p_a}\exp\left[\frac{q\psi_o}{kT}\right]\exp\left[\frac{qV_a}{kT}\right]$$

i.e. the hole density (which is the minority carrier in the n-type material) at the location beginning of the charge neutral n-type region, P_{nb}, increases exponentially with V_a A similar set of expressions will hold for the electron density at the edge of the quasi-neutral p-type region.

Now, at point a, we have charge neutrality ($x \leq a$). We now use the next approximation.

Approximation 3: Charges in the minority of a much lower density than those in the majority for all values of V_a. (i.e. $p_{p_0} >> n_{p_0}$ and $n_{n_a} >> p_{n_a}$ always). In this case,

$$p_{p_a} \approx N_A + n_{p_a} \approx p_{p_0} \approx p_{n_0} \exp\left(\frac{q\psi_0}{kT}\right)$$

and thus we can write $p_{n_b}(V_A)$ and $n_{p_a}(V_a)$ in terms of p_{n_o} and V_a:

$$p_{n_b}(V_a) \approx p_{n_0} exp(qV_a/kT) \approx \frac{n_i^2}{N_D} exp(qV_a/kT)$$

and

$$n_{p_a}(V_a) \approx n_{p_0} exp(qV_a/kT) \approx \frac{n_i^2}{N_A} exp(qV_a/kT)$$

Thus, the application of a forward bias (Va>0) to the junction *increases the minority carrier density exponentially at both sides of the junction.* The external voltage is said to *inject* minority carriers into the quasi-neutral regions located on each side of the junction. The resulting charge motion is illustrated in Figure , which shows how the charges move through the junction, into the quasi-neutral regions, and then through an external circuit. The corresponding charge distribution within the junction is shown in Figure below.

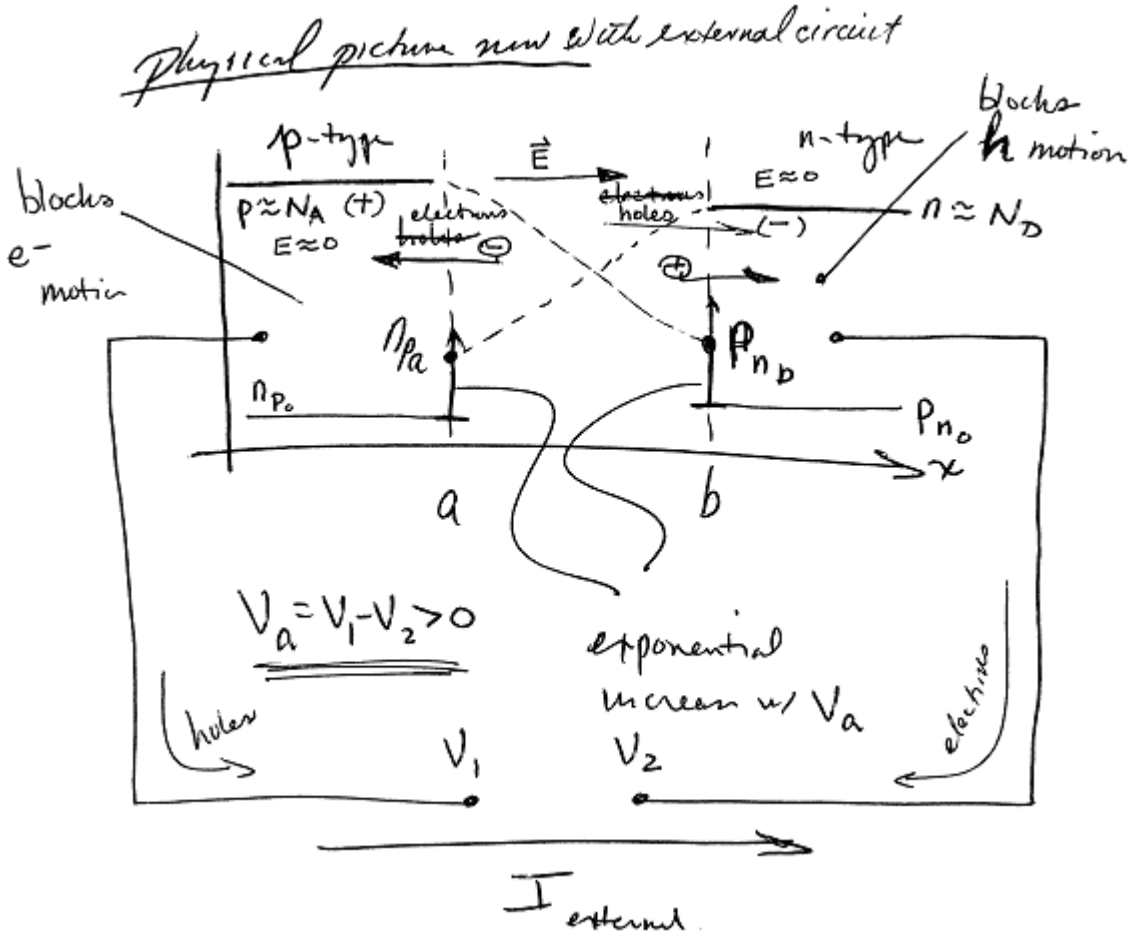


Figure 11.14: Schematic of a forward biased p-n junction connected to an external circuit.

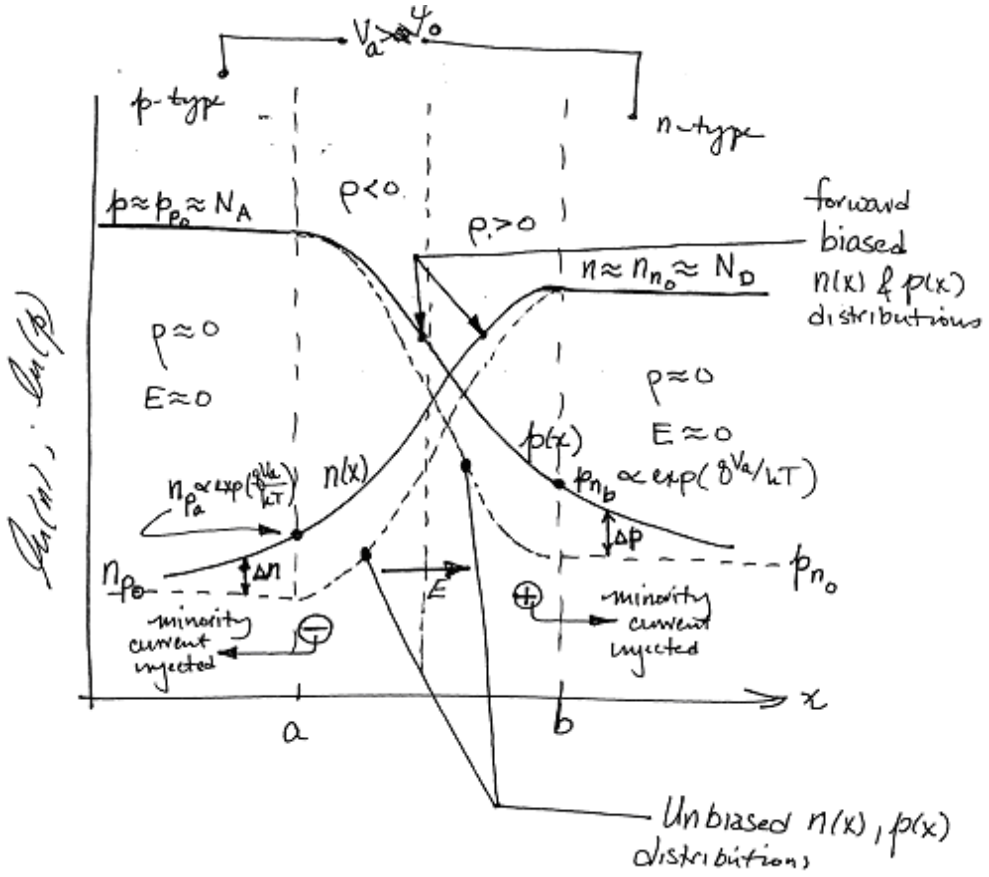


Figure 11.15: Qualitative Picture of Charges and E field in the p-n diode with forward bias

The external bias, V_a, injections electrons into the p-type region and holes into the n-type region. These charges then diffuse through these regions that have E=0 until they reach the external circuit connections located at the outer boundaries of the diode. The electrons then move through the external circuit around to the connection between the p-type side and the external circuit, where they recombine with the holes. The motion of these charges then forms the current through the p-n diode and the external circuit. Note that a reverse bias (i.e. one with Va<0) will not produce the same magnitude of current as

an equal magnitude forward bias voltage. Thus, the device acts like a one-way valve for electrical charge.

We now wish to determine the quantitative relationship between current, J, and the external bias voltage. We do this by considering the diffusion of the minority charges through the quasi-neutral regions. We examine the transport of holes through n-type material first. Thus, on the n-type side we have the hole current density given as:

$$J_h = -qD_h \frac{dn}{dx};\ x \geq b$$

Now the equation of continuity gives:

$$\frac{1}{q}\frac{dJ_n}{dx} = -(U - G)$$

Where U denotes the volumetric loss rate for holes, and G denotes the production rate per unit volume for holes. The loss term, U, can be expressed as:

$$U = \frac{p - p_{n_o}}{\tau} = \frac{\Delta p}{\tau},$$

where Δp is the local minority carrier density in excess of the V_a=0 value, p_{n_o} and τ is the *minority carrier lifetime*. Combining these equations, we thus have a diffusion equation for the minority carrier transport:

$$D_h \frac{d^2}{dx^2} p_n(x) = \frac{p_n(x) - p_{n_0}}{\tau} - G$$

Let us first consider what happens when there is no production of electrons and holes (i.e. when there is no illumination of the PV cell with light). In this case, G=0 and thus we have

$$D_h \frac{d^2}{dx^2} p_n(x) = \frac{p_n(x) - p_{n_0}}{\tau} = \frac{\Delta p_n}{\tau}()$$

$$D_h \Delta p'' = \frac{\Delta p}{\tau}$$

where ()'' denotes the second derivative and we have used the fact that $p_{n_0} = const.$. For convenience let us define the diffusion length as $L_h^2 = D\tau$. The whole diffusion equation is now given

$$\Delta p'' = \frac{\Delta p}{L_h^2}$$

This equation has a general solution given as

$$\Delta p = A \exp\left[{+x}/{L_h} \right] + B \exp\left[{-x}/{L_h} \right]$$

Because $\Delta p \to 0$ when $x \to \infty$ we can then write A=0. Using the boundary condition for x=b then gives

$$\Delta p = p_{n_0} \left(\exp(qV_a / kT) - 1 \right) \text{ for x=b}$$

Thus the solution for p(x) for x>b is then given as

$$p_n(x) = p_{n_o} + p_{n_o} \left[e^{{qV}/{kT}} - 1 \right] e^{-{(x-b)}/{Lh}} \text{ for x > b}$$

Similarly in the p-type region the electron density is given as

$$n_p(x) = n_{p_o} + n_{p_o} \left[e^{{qV}/{kT}} - 1 \right] e^{{(x-a)}/{Le}} \text{ for x < -a}$$

A schematic of these charge densities is given below.

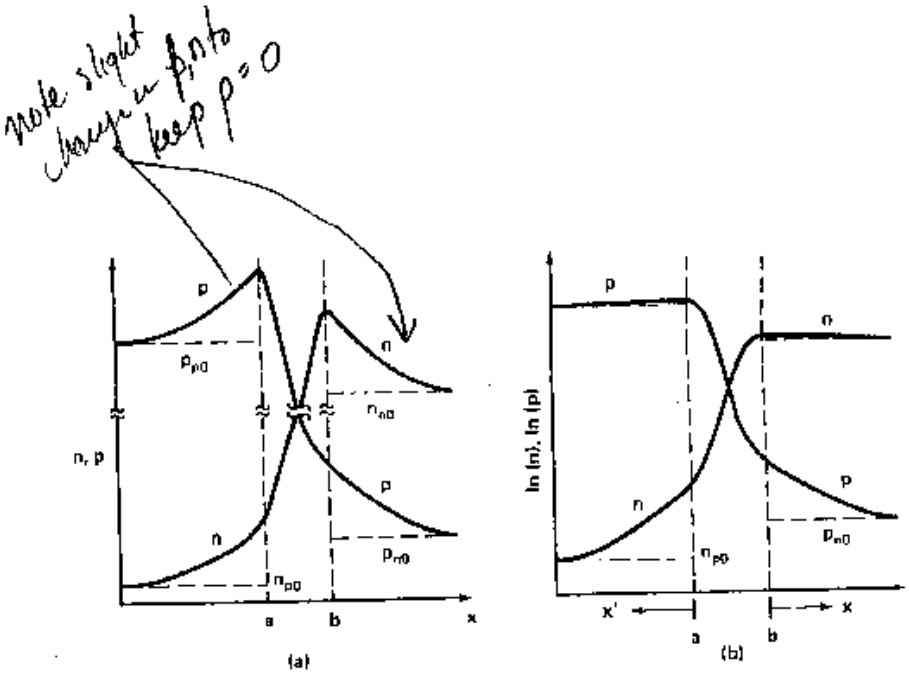


Figure 4.8. (a) Linear plot of the distributions of carriers throughout the *p-n* junction diode under forward bias. (b) Corresponding semilogarithmic plot. Note the differences with respect to majority carriers.

NEED TO RE-DO THIS FIGURE WITH OUR OWN LABELING AND DISCUSSION.

We can now use these minority carrier densities to determine the current due to the diffusion of the minority carriers in their respective regions of the diode. Recalling that the electric field is zero in the quasineutral regions, we can write these current densities as

$$J_h = -qD_h \frac{dp}{dx}$$
$$J_e = +qD_e \frac{dn}{dx}$$

Using our previous solutions for p(x) and n(x) we can then write the current densities as

$$J_h = +\frac{qD_h p_{n0}}{L_h}\left[\exp\left(qV/k_BT\right)-1\right]\exp\left(-x/L_h\right)$$

and

$$J_e(x') = +\frac{qD_e n_{p0}}{L_e}\left[\exp\left(qV/k_BT\right)-1\right]\exp\left(-x'/L_e\right)$$

where the x (x') coordinates extend to the right (left) of points b and a respectively.

These solutions are shown in Figure below.

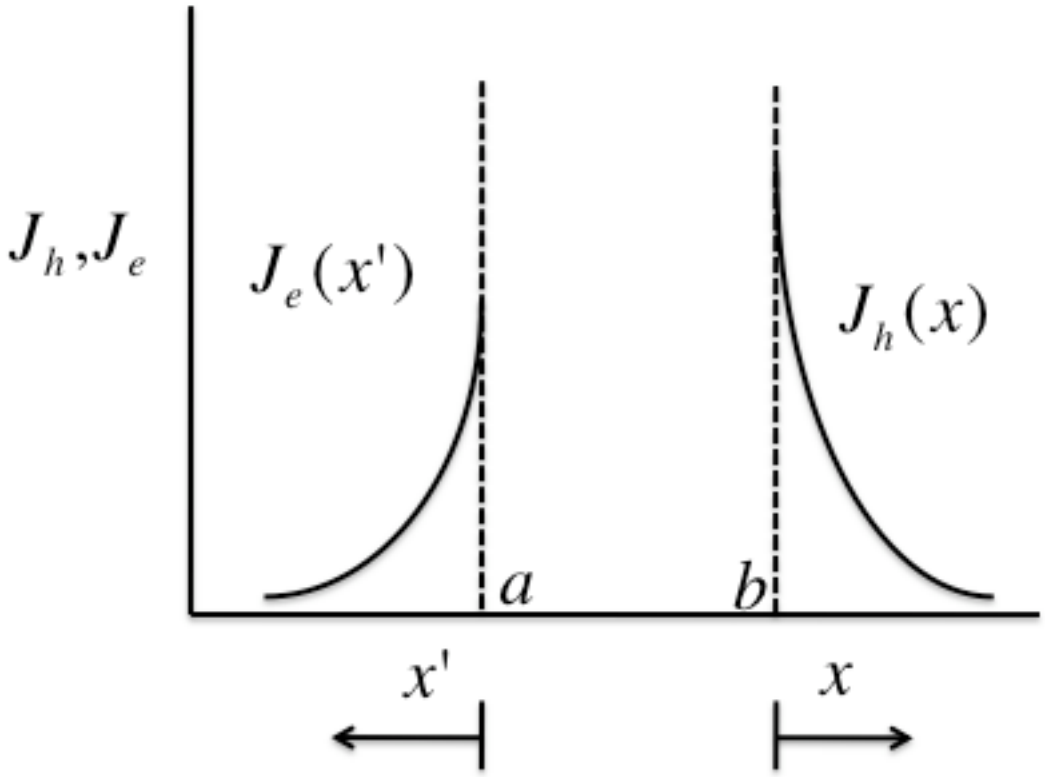


Figure 11.16: Minority carrier current densities.

Finally, consider current flow across depletion region (a < x < b), which must satisfy:

$$\frac{1}{q}\frac{dJ_e}{dx} = U - G = -\frac{1}{q}\frac{dJ_h}{dx}$$

$$dJ_e = q(U-G)dx = -dJ_h$$

$$\delta J_e\Big|_a^b = q(U-G)W = -\delta J_h\Big|_a^b$$

Now the width W = b-a, is usually small such that W << L_h, L_e, thus $|\delta J_e|_{a<x<b} \approx |\delta J_h|_{a<x<b} \approx 0$ i.e. there is *no change* in current density across depletion region. In this case, we can then determine the current everywhere in the diode by simply adding the two minority carrier currents as shown schematically in the figure below.

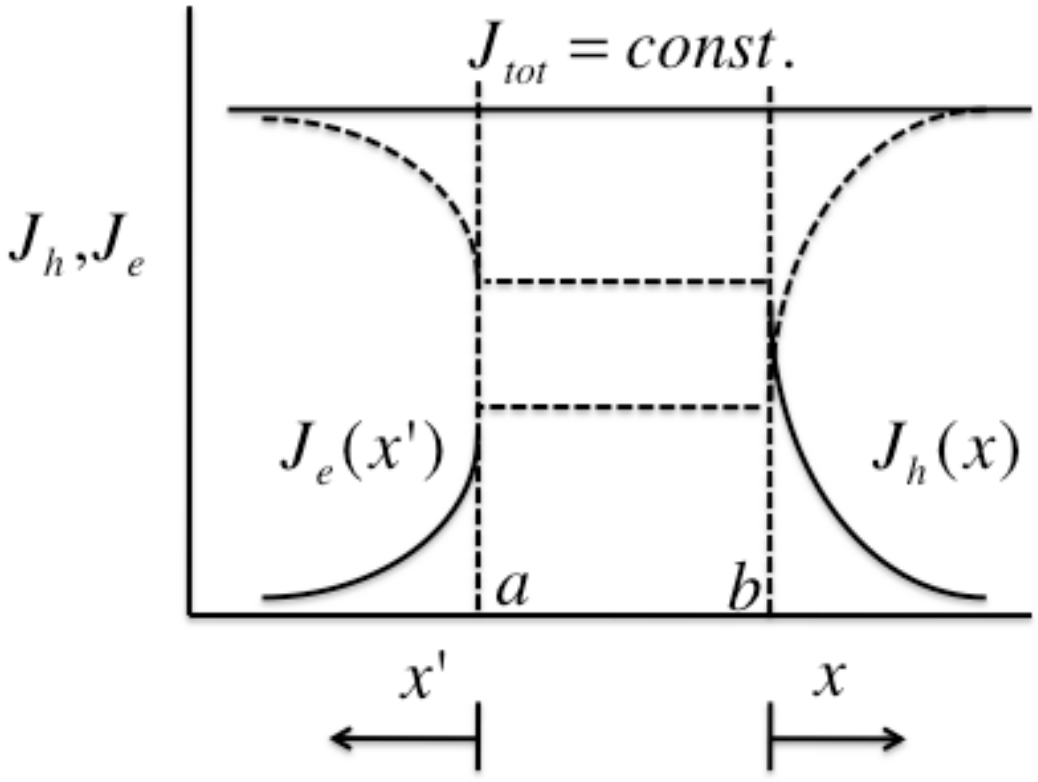


Figure 11.17: Total current density distribution across an unilluminated p-n diode.

Using this assumption that the current density does not change across the junction, we can now write the total current density J_{total}

$$J_{tot} = J_e + J_h$$
$$J_{tot} = J_e\big|_{x=a} + J_h\big|_{x=b}$$
$$J_{tot} = \left(\frac{qD_e n_{p_o}}{L_e} + \frac{qD_h p_{n_o}}{L_h} \right)\left(e^{qV/kT} - 1 \right)$$
$$J_o = \left(\frac{qD_e n_{p_o}}{L_e} + \frac{qD_h p_{n_o}}{L_h} \right)$$
$$I = J_{tot} \cdot A$$
$$I = I_o\left(e^{qV/kT} - 1 \right)$$
$$I_o \equiv J_o \cdot A$$

This solution is shown in Figure above. We have found the key result for the un-illuminated diode: the current through the diode thus increases *exponentially* with Va as long as Va>0. If Va<0 only a very small current flows through the diode. This analysis assumed that the volumetic production rate of charge carriers, G=0. Now, what happens when the diode is illuminated by a light source with photons whose energy exceeds the bandgap voltage? In this case, G>0, and we have to include the effect of a source of charges in the diffusion equation.

The Illuminated Diode Model of a Solar PV Cell

Our previous discussion has led to the description of an un-illuminated p-n diode. We found that minority carriers are injected into the quasi-neutral regions (i.e. holes with a distribution p(x) are injected into the n-type region, and electrons with a distribution n(x)

are injected into the p-type region). The resulting gradient of the minority carrier density leads to the thermal diffusion of these particles through the quasi-neutral region. We found a solution to the diffusion equation for these charges, and that in turn led to the charge distribution within the p-n diode. By differentiating this distribution we were then able to determine the current density distribution vs. applied bias. Finally, when we took note that the junction region is thin compared to the diffusion lengths, we could find the total current vs. applied voltage in the un-illuminated ideal p-n diode. In this discussion, we now turn our attention the behavior of the diode when a source of illumination is present such that the absorption of the photons from the light source can create electron-hole pairs within the device. We then wish to know how these charges move within the device and, ultimately, wish to determine the power production from such an illuminated diode.

In the presence of illumination and carrier generation, the diffusion equation for non-equilibrium minority carrier population now acquires a non-zero source term proportional to G, the generation rate of charge carrier pairs. Thus, we have

$$\frac{d^2\Delta P}{dx^2} = \frac{\Delta P}{L_h^2} - \frac{G}{D_h}$$

$$\frac{G}{D_h} \sim const$$

where the source term G denotes the volumetric production rate of e/h pairs and D_h ~ diffusion coefficient for holes in the n-type region.

The complete solution to this equation is given as

$$\Delta P = G\tau_h + Ce^{x/L_h} + De^{-x/L_h},$$

where $\tau_h = \frac{L_h^2}{D_h}$ is the hole lifetime.

We apply the same boundary conditions as in the G=0 case to find the particular solution for $p_n(x)$; we then write the solution for $n_p(x)$ by inspection.

$$P_n(x) = P_{n_o} + G\tau_h + \left[P_{n_o}\left(e^{qV/kT} - 1 \right) - G\tau_h \right] e^{-x/L_h}$$

$$N_p(x') = n_{p_o} + G\tau_e + \left[n_{p_o}\left(e^{qV/kT} - 1 \right) - G\tau_e \right] e^{-x/L_e}$$

Following the same procedure as used in the un=illuminated diode model, we use the diffusion approximation to write the minority current density with illumination since minority carrier current occurs via diffusion.

$$J_h(x) = \frac{qD_h P_{n_o}}{L_h}\left(e^{qV/kT} - 1 \right) e^{-x/L_h} - qGL_h e^{-x/L_h}$$

$$J_e(x') = \frac{qD_e n_{p_o}}{L_e}\left(e^{qV/kT} - 1 \right) e^{-x'/L_e} - qGL_e e^{-x'/L_e}$$

Neglecting recombination in the depletion region <u>but</u> including effect of G on charge carrier generation: $|\delta J_e| = |\delta J_h| = qGW$, we find that there is a "jump" in the current density across the junction. If we then add this current jump to the solutions $J_e(x'=0)$ and $J_h(x=0)$ we can then find the total current across the diode in the presence of illumination. If we write the frontal area of the cell as A and note that the total current I=Jtot*A we can then write the I-V characteristic of the illuminated diode as

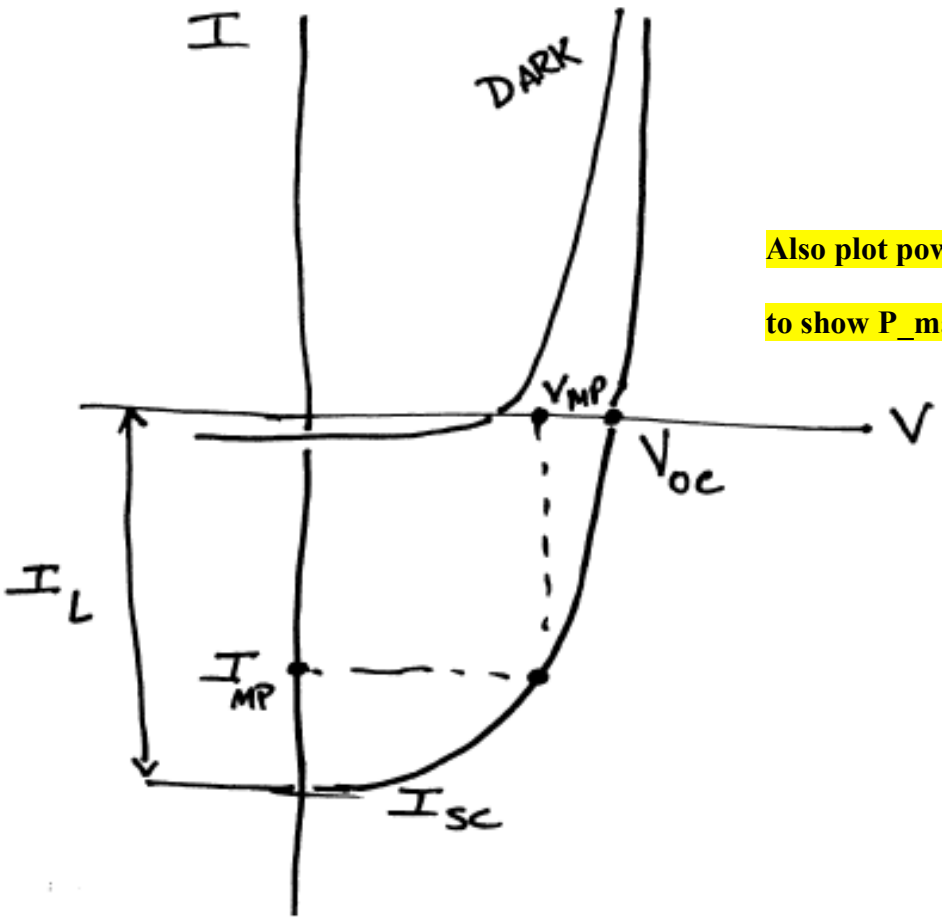
$$I = I_o\left(e^{qV/kT} - 1\right) - I_L$$
$$I_L = qAG\left(L_e + W + L_h\right)$$

The region located within L_e, L_h of the junction is usually termed the "Active Region". The key quantities in the I(V) expression are given as:

$$I_o = A\left(\frac{qD_e n_i^2}{L_e N_A} + \frac{qD_h n_i^2}{L_h N_D}\right)$$
$$L_{e,h} = \sqrt{D_{e,h}\tau_{e,h}}$$
$$\frac{1}{\tau_{net}} = \sum_{\substack{i=all \\ loss \\ mechanisms}} \frac{1}{\tau_i}$$

$\tau_{e,h}$ ~ net recombination time

Figure below illustrates the I(V) response of the diode for both the G=0 case and the finite G case. Note that there is a region where I<0 and V>0 in which net power production occurs. This is the region where we are interested in operating the solar PV cell.



Also plot power vs. V?

to show P_max

Figure 11.18: Current-voltage characteristic of an unilluminated (or "dark") p-n junction and an illuminated p-n junction.

There are several macroscopic parameters (i.e. parameters that are measured externally to the diode itself) that are used to characterize the I(V) characteristic curve. These parameters include:

(1) the Short Circuit Current I_{sc} determined by evaluating I(V) for the case V=0:

Ideally, $I_l = qAG\left(L_e + W + L_h\right)$

(2) the Open Circuit Voltage V_{oc} determined by setting I(V)=0 and solving for V:

$$V_{oc} = \frac{kT}{q} \ln\left(\frac{I_L}{I_o} + 1 \right)$$

and

(3) the maximum power operating point, denoted as MP on the figure above and located at coordinates [V_{MP}, I_{MP}]; maximum power output of cell is then given as $P_{MP} = I_{MP} V_{MP}$.

It is common practice to define the so-called Fill Factor, $FF \equiv \frac{V_{MP} I_{MP}}{V_{OC} I_{SC}}$ which in a sense measures how "square" the I-V characteristic is. Figure below shows the variation of the FF with Voc (normalized to the operating temperature of the cell. For Si PV cells, typically V_{oc}~0.7. If T=300 deg K then kT/q~0.02 or so and thus we estimate FF~0.8-0.85 or so.

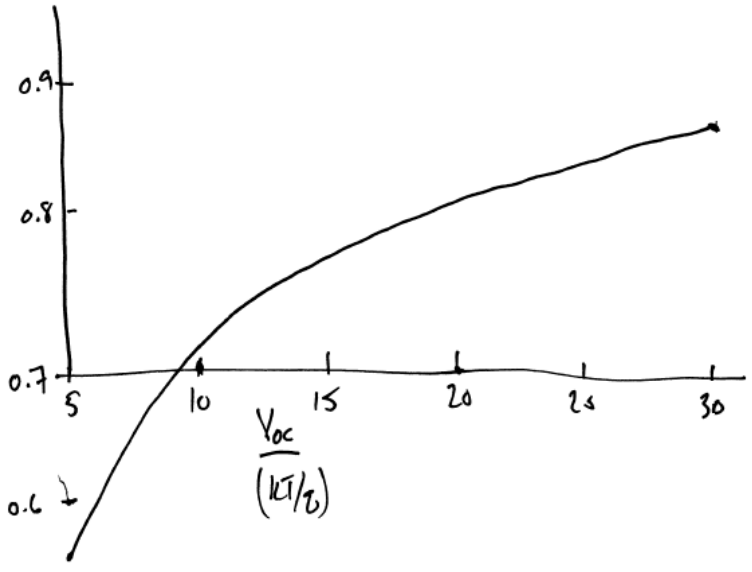


Figure 11.19: Variation of fill factor, FF, with normalized open circuit voltage.

We can now use these results to learn what microscopic processes determine the macroscopic efficiency of the PV cell. We write the maximum efficiency as

$$\eta = \frac{V_{MP} I_{MP}}{P_{in}} = \frac{V_{OC} I_{SC} FF}{P_{in}}$$

Typically, η ~ 15-18% for commercial PV cells intended for large-scale power production. The quantity FF does not change much with Voc and thus we will take it as a constant. Thus the efficiency is determined by Isc and Voc. Let us then examine the factors that influence these two key parameters.

(A) Short Circuit Current, I_{sc}

It is clear that I_{sc} is proportional to I_L which in turn is determined by $I_L = qAG\left(L_e + W + L_h\right)$. From the definitions of the diffusion lengths we have $L_{e,h}^2 \equiv D_{e,h} \cdot \tau_{e,h}$. The diffusion coefficient for minority carriers is a material property primarily and thus is determined by the choice of materials for the PV cell. However, the minority carrier lifetime is controlled in a significant part by the manufacturing processes and technologies used to make the PV cell. This lifetime is determined by multiple loss processes than can occur within the cell. Thus we can see that $I_{SC} \propto \frac{1}{E_{gap} \tau_{life}^{1/2}}$, and that the $I_{SC} \propto G \propto \frac{1}{E_{gap}}$ scaling comes from the fact that for a given photon energy spectrum (determined by the sun's blackbody radiation temperature), the e^--h^+ pair generation rate is then inversely proportional to the bandgap energy that must be overcome to create the charge carrier pair.

C ı the solar

c rface and

interface connections. Let us denote the average number of these defects per unit volume

as N_{defect}. It then stands to reason that the charge carrier lifetime varies inversely with the defect density, i.e. we can write

$$\tau_{life} \propto \frac{1}{N_{defect}}.$$

I_L also depends upon G, the number of electron-hole pairs produced per unit volume per unit time in the active region. From our earlier discussions about charge carrier production from a blackbody spectrum, we know that this parameter is determined by the flux of photons incident upon the PV cell with an energy $E>E_{gap}$ i.e. we can write

$$G \propto \phi_g = \frac{1}{h}\int_{f_{gap}}^{\infty} \frac{1}{f} I_{bb}(f)df$$

where $f_{gap}=E_{gap}/h$ and $I_{bb}(f)$ denotes the frequency distribution of the incident solar radiation which is approximated by a blackbody spectrum. Thus clearly $G \propto \frac{1}{E_{gap}}$. Thus we can see that Isc scales according to $I_{SC} \propto \frac{1}{E_{gap}N_{defect}^{1/2}}$.

(B) V_{oc}:

The open circuit voltage is given as $V_{oc} = \frac{kT}{q}\ln\left(\frac{I_L}{I_o}+1\right)$; with

$$I_o = A\left(\frac{qD_e n_i^2}{L_e N_A} + \frac{qD_h n_i^2}{L_h N_D}\right);$$

and $I_L = qAG\left(L_e + W + L_h\right)$

i.e. V_{oc} depends on:

i) Material Properties (n_i^2 , N_A, N_D, D_e, D_h)

ii) photon – Material Properties (G, L_e, L_h)

$V_{oc} \sim 700$ mV for Si

Now

$$I_o \propto n_i^2$$

$$n_i^2 = N_C N_V \exp\left[-E_{gap} \Big/ kT \right]$$

$$I_o = K \exp\left[-E_{gap} \Big/ kT \right]$$

$$K = 1.5 \cdot 10^5 (?)$$

Thus $\left.\begin{array}{l} V_{oc} \propto E_{gap} \\ I_{sc} \propto 1 \Big/ E_{gap} \end{array}\right\} \Rightarrow$ and we thus conclude that for a given blackbody spectrum, there must be an optimum E_{gap} that maximizes the efficiency η_{max}.

Major Contributions to η being significantly lower than intrinsic or theoretical efficiency:

(1) Reflection at surface of cell

(2) Bulk and surface recombination

This discussion implies that the solar PV cell efficiency can be increased by:

a) Reduce E_{gap} or use multi-junction design

b) Reduce impurity density

c) Reduce effective surface areas (increase crystal size) which can trigger charge carrier destruction.

Typically these changes impose a cost penalty e.g. due to the increase in manufacturing complexity and/or cost, and increase in material costs. The question will then arise: is the increase in efficiency outweighed by the increase in cell. This can only be answered by a detailed analysis of the particular device in question.

Add discussion on the tradeoff between photons absorption vs. exciton diffusion?

Optimum thickness of the PV cell? For polycrystalline Si PV & thin film amorphous Si PV.

Chapter 12: Solar Thermal Electricity Production

Introduction

Solar thermal electricity generation (SEG) technology is a reasonably well-developed technique for producing electricity from solar radiation, and has been implemented in a number of locations where the solar resource is high and which are also reasonably close to electricity demand locations and/or long-distance transmission systems. At this writing, the costs of electricity produced with this technology are approaching those of fossil fueled power plants in regions with adequate solar insulation, and thus the technology is beginning to be deployed in these regions. In this chapter, we briefly summarize the essential elements of this technology.

Basic Schematic of Solar Thermal Electricity Production

A schematic view of a SEG power plant is shown in Figure below. The plant is composed of a collector system (usually composed of mirrors), which act to collect direct (i.e. non-diffuse) solar radiation and then focus it onto a heat absorbing target through which a high temperature coolant is circulated. Often, this coolant is a molten salt or other substance which is maintained in a liquid state at moderately high temperatures (500 deg C or higher) at close to atmospheric pressure. Ideally, this coolant would also have a high specific heat. This high temperature coolant received heat input from the collected solar radiation, and is circulated through a primary loop as shown in the schematic; it can also be stored at high temperature if necessary.

A secondary coolant loop runs through a heat exchanger which transfers heat from the high temperature primary coolant to a secondary coolant. This coolant is then run through a conventional heat engine which then converts the heat in the secondary loop to mechanical work and rejects waste heat. For example, this loop could be a Rankine cycle turbine in which high pressure water is heated into high pressure steam; this steam would then be expanded through a Rankine cycle turbine producing mechanical work. The low grade steam would then be condensed back into water via a conventional heat rejection system before then being run through the pump, thereby beginning the cycle again. Alternatively, the heat engine could be a closed Brayton cycle turbine in which a working gas is compressed, run through the heat exchanger where it received high grade heat. The resulting high pressure, high temperature gas is then expanded through the turbine which produces mechanical work and lower temperature, lower pressure gas. This gas would then be run through a conventional heat rejection system before being re-introduced into the closed Brayton cycle again. Note that in all cases the SEG system requires a means to reject the waste heat. Since these systems are usually located in regions of high solar incident radiation, and these regions tend to be geographically isolated from large quantities of water, providing for adequate heat rejection must then be carefully considered in the design of a SEG system.

SEG plant designs differ in the details of the collection system. In the first common approach, a one-dimensional trough-like parabolic mirror focuses the solar radiation onto a radiation onto a tubular pipe through which the primary coolant is circulated (see Figure a below). In the second approach, also finding significant use, a

series of flat mirrors are arranged on a large flat parcel of land and are used to reflect incident solar radiation onto a single thermal collector which is usually located at the top of a tower (see Figure b below). In the third approach, which is more suited to smaller scaled SEG systems, a parabolic dish is used to focus solar radiation onto a single collector through which a coolant is circulated. In the latter two approaches, the collecting mirrors must be articulated on a joint which provides two degrees of freedom in order to maintain operations; the parabolic trough arrangement only requires one degree of freedom (or in some cases, a stationary arrangement can be used). The first two approaches tend to scale favorably to large systems (i.e. many 10's MW and larger), while the dish approach is suited to smaller scaled (10's kWs) systems.

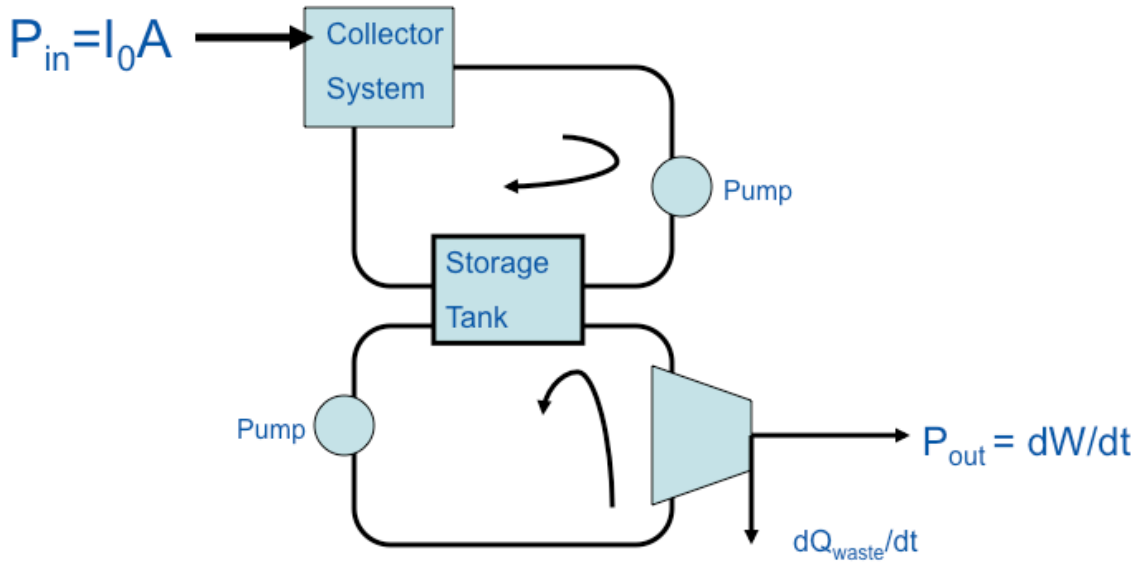


Figure 12.1: Schematic of a SEG power plant

Figure 12.2: Photographs of (a) linear parabolic trough array, (b) parabolic dish and (c) power tower SEG power plant configurations.

Analysis of Solar Thermal Power Plant – Steady State Conditions

To proceed with an analysis of a SEG system, let us consider the conceptual system shown in Figure below. The solar radiation power input $P_{in} = I_0 A$ where I_0 is the incident solar radiation intensity and A is the collector area normal to the incident radiation. In principle, an auxiliary heat source, P_{aux}, from e.g. combustion of natural gas can also be used to heat the primary coolant. This coolant loop, including both the loop and pump itself as well as an storage tank or plenum, has a volume V and is filled with a coolant with density ρ and specific heat C_p. The secondary coolant loop removes heat at a rate dQ_{out}/dt from the primary loop, and will have a temperature that must be at or below the temperature of the primary coolant (which in turn sets the thermal efficiency of the

system). This secondary coolant then is converted to work at a rate dW/dt at an efficiency determined in part by the operating temperature and in part by the thermal cycle chosen for the system. The waste heat is then rejected at a rate dQ_{waste}/dt.

A simple heat balance applied to the primary loop gives the time rate of change of the primary coolant temperature, T, as

$$\rho C_p V \frac{\partial T}{\partial t} = P_{in} + P_{aux} - \frac{dQ_{out}}{dt}$$

Application of the first law, combined with the definition of thermal efficiency gives

$$\frac{dQ_{out}}{dt} = \frac{dW}{dt} + \frac{dQ_{waste}}{dt}; \quad W = \eta_{th} Q_{out}$$

thus

$$\rho C_p V \frac{\partial T}{\partial t} = P_{in} + P_{aux} - \frac{1}{\eta_{th}} \frac{dW}{dt}$$

Obviously then in steady state, with no auxiliary power input we simply have

$$\rho C_p V \frac{\partial T}{\partial t} = P_{in} + \cancel{P_{aux}} - \frac{1}{\eta_{th}} \frac{dW}{dt} \Rightarrow P_{in} = \frac{1}{\eta_{th}} \frac{dW}{dt} \text{ in 2.2.}$$

i.e. the power output is simply given by the input power multiplied by the thermal efficiency of the heat engine. The remainder of the power input must be dissipated to the environment.

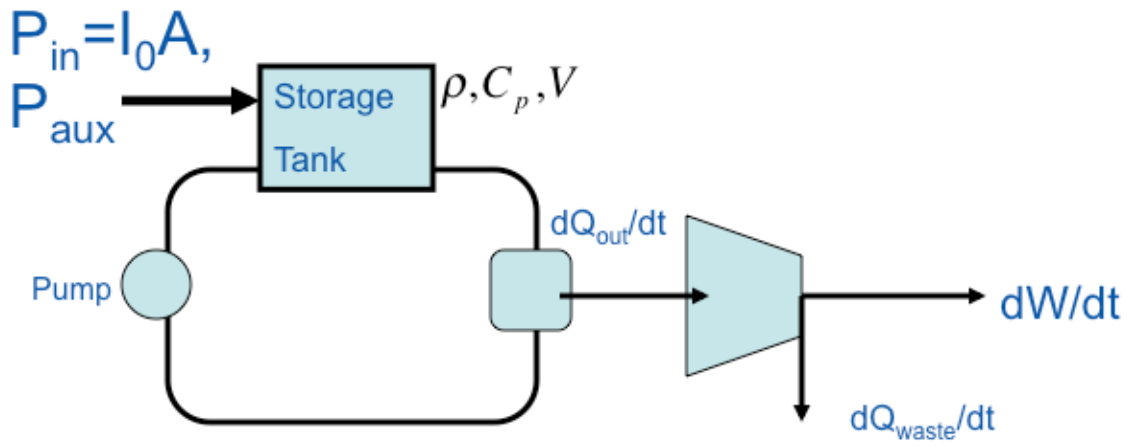


Figure 12.3: Conceptual arrangement of a SEG power plant

Analysis of Solar Thermal Power Plant – Transient Conditions

Let us now consider one unique capability of a SEG system – namely the ability to store energy for later recovery and conversion. In particular, we consider the case where the SEG system is to be operated at night with no auxiliary heat input such that $P_{aux} = 0$ & $P_{in} = 0$. In this case the primary coolant loop heat balance equation is given as

$$\rho C_p V \frac{\partial T}{\partial t} = -\frac{1}{\eta_{th}} \frac{dW}{dt}$$

For simplicity we assume that the power output, dW/dt, and the conversion efficiency are constant. In this case, the primary coolant temperature then decreases on a timescale τ given by

$$\rho C_p V \left(-\frac{T}{\tau} \right) = -\frac{1}{\eta_{th}} \frac{dW}{dt} \Rightarrow \quad \tau = \frac{\rho C_p V T \eta_{th}}{dW / dt}$$

Thus, a given power output, dW/dt, can be sustained for roughly a period τ after the removal of solar power input. This period can be increased by increasing the volume of the primary coolant storage, increasing the heat capacity of the primary fluid, or decreasing the power extraction rate from the system. Clearly, if an auxiliary heat input is provided e.g. from the combustion of natural gas, such a system could be operated continuously, cycling between solar power input and fossil fuel power input.

Chapter 13: Energy from Nuclear Fission

Introduction and elementary considerations

An understanding of the key elements of nuclear fission as a power source requires us to review a few elementary concepts and ideas from physics. In particular, we recall that atoms are composed of electrons, which have a negative electrical charge and which are (usually) bound to a positively charged atomic nucleus that is composed of protons and neutrons which are, in turn, bound together by nuclear forces. The typical spatial scale of the electron orbitals is Angstroms, while the scale of the atomic nucleus is about six orders of magnitude smaller. The proton has an electrical charge that is equal in magnitude and opposite in sign to that of the electrons, while the neutron is electrically neutral. Since these oppositely charged particles have an attractive force, the electrons are then bound to the atomic nucleus. One can consider this force to be transmitted by mass-less particles called photons which mediate the electromagnetic force between the nucleus and the bound electrons. The positive charge of the protons within the nucleus exerts a repulsive force on other protons within the nucleus, but the nuclear particles are also bound together via another type of force, the so-called strong nuclear force. This force is mediated by another type of particle known as mesons which only persist across very short distances on the scale of the atomic nucleus. The resulting net force is then sufficiently strong to hold the nucleus together. Table below provides a summary of these key particles, their charges, rest mass (i.e. their mass in the absence of any motion)

and the so-called rest energy, which is the energy equivalent of the rest mass, computed from the relation where m refers to the particle rest mass.

Electron:

Rest mass ~ 9.

Charge ~ e = -1.602 x 10^{-19} C

$E = mc^2$ = 511 keV (?)

Proton:

Rest mass ~ m_p = 1.67 x 10^{-27} kg

Charge ~ e = +1.06x 10^{-19} C

E = 931 MeV (?)

Neutron:

Rest mass ~ m_n = 1.674 x 10^{-27} kg

Charge ~ e = 0

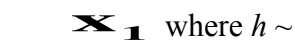
Positron (positive electron):

Rest mass ~ 9.1 x 10^{-31} kg

e = +1.602 x 10^{-19} C

Photon:

Mass-less particle of electromagnetic energy with energy and momentum given by:

 where *h* ~ Planck's constant and ν ~ frequency

And momentum 

Neutrino:

Nearly mass-less, neutral particle

Table 2: Summary of elementary particles of interest for nuclear energy

The chemical composition of an atom is determined by the number of protons in the atomic nucleus together with the bound electrons whose arrangements and allowed energy levels determine the types of chemical interactions (i.e. those interactions that are mediated by the electrons of the atom). For a given number of protons in a nucleus, there are atoms with varying numbers of neutrons within the nucleus. Atoms with the same number of protons but differing numbers of neutrons are referred to as isotopes. Thus, for example, oxygen always has 8 protons, but has multiple isotopes, including one with 7 neutrons and one with 8 neutrons. The chemical symbol for these two isotopes are usually written as $\mathbf{x}_2$and $W = \int_{x_1}^{x_2} F(x) \cdot dx$. In general for a chemical element, X, the atomic number (i.e. the number of protons in the nucleus) is written as Z and the atomic weight (given by the number of protons and neutrons in the nucleus) is given as A, and the element is then referred to by the symbol $W = \int_{x_1}^{x_2} F(x) \cdot dx$.

In our earlier discussions, we have already encountered the idea that bound electrons have a variety of allowed energies, and that electrons can move between these so-called allowed states by either absorbing energy (if the transition involved moving from a lower energy state to a higher energy state) or by giving off energy (if the transition involved moving from a higher energy state to a lower energy state). In either case, the energy exchange with the external environment is mediated by the absorption or emission of a photon, which represents a quantum of electromagnetic energy. For transitions involving weakly bound valance-band electrons the photon energy typically ranges from 1-10 eV; transitions involving more tightly bound inner orbital electrons have energies that can range up to several keV.

Experiments show that the atomic nucleus likewise can exhibit similar transitions. In this case, the transitions involve the excitation or decay of the nuclear particles (i.e. the protons and neutrons). For example, one can think of a transition into an excited state as moving the nucleus from a condition, or state, of low energy (think of a weakly vibrating liquid droplet as a reasonable analog) to a state of higher energy (think of a more strongly

vibrating liquid droplet as a reasonable analog). Just as in the case of the electronic transition, this excitation process requires the nucleus to absorb enough energy from the outside environment to drive the excitation. Likewise, a decay from a higher energy state to a lower energy state requires the nucleus to shed the corresponding amount of energy. Thus, we see an important principle at work here: transitions in the energy state of a nucleus can and do occur in nature, and require an exchange of energy with the surrounding environment. These transitions are again mediated by the emission of quanta of electromagnetic energy, i.e. the emission of photons. However, the energy transitions associated with nuclear transitions are much higher than those encountered in electronic transitions; typically nuclear transitions involve energy changes that range from many keV up to MeV or more of energy – i.e. energy transitions that are many orders of magnitude larger than those found in electronic transitions.

Excited States in Atoms:

- Electron Excitation
- Nuclear Excitation

Decay process emits a photon of energy

few eV → photon (visible light / IR / mm-wave)

100 eV > E > few eV → UV / SXR

E ~ keV → X-rays

E ~ MeV → γ-rays

Radioactive Decay

The number of protons and neutrons in the nuclei found in nature have been measured and, when plotted against each other as shown in Figure below, it is found that nuclear composition generally is concentrated on a band as seen in the figure. Nuclei located close to the center of the band tend to be stable – that is they do not undergo spontaneous

changes in the number of protons or neutrons within the nucleus. Nuclei whose composition places them above the middle region of the band are termed "proton rich" nuclei, while those that lie below the middle region of the band are termed "neutron rich" nuclei. It is found that such nuclei can sometimes be unstable, and can occasionally undergo spontaneous changes in the numbers of protons and/or neutrons within the nucleus. This process is referred to as radioactive decay and is an essential concept for the operation and evaluation of nuclear energy systems.

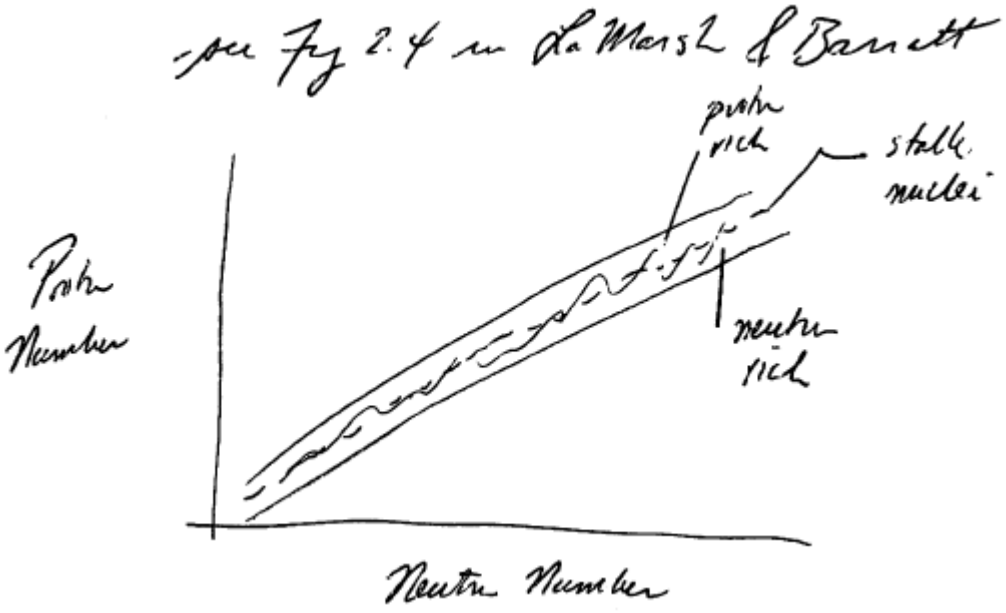


Figure 13.1: Plot of atomic nuclei proton number vs. number of neutrons.

Perhaps the easiest way to introduce the concept of radioactive decay is to consider an explicit example that can (and has been) observed in nature and in the laboratory. These experiments show that E_{src} is the stable form of oxygen nucleus while E_o, which is missing one neutron is observed to undergo a spontaneous transition or decay by emitting a positron (for our purposes think of this particle as a positively

charged electron). Emission of an electron or positron from the nucleus is referred to as beta decay. Again, experiments show that when this happens, the oxygen nucleus transforms into a nitrogen nucleus containing 8 neutrons and 7 protons. Thus, apparently a proton in the original nucleus has been transformed into a neutron; the overall electrical charge has been conserved by the emission of the positron. Some energy is also carried off by this positron (which has finite kinetic energy) as well as by the emission of a particle known as a neutrino (which is a very low mass particle which interacts only very weakly with nuclei and thus is not of particular interest for nuclear energy applications). This reaction is written schematically as

$$E_o$$

where E_{re} is a neutrino and E_{re} is the "daughter nucleus. Note that, just like in chemical reactions, the total number of interacting particles (i.e. in this case the total number of protons and neutrons) and the electrical charge are both conserved as can be seen by summing the subscript and superscripts of the reaction. If one were to measure the momentum of the particles involved, we would also find that the momentum is also always conserved. Thus, experiments show that during such nuclear decay events this is always the case, and thus we take away an important point which impacts our understanding of nuclear energy systems: during radioactive decay processes the total number of neutrons and protons together, the electrical charge and the momentum is always conserved.

Often the daughter Nucleus can be unstable and can subsequently decay. For example, another isotope of oxygen can undergo the following beta decay process, which leads to the formation of an intermediate unstable fluorine nucleus, which then undergoes a second beta decay and transforms into a neon nucleus which is then stable.

$$^{20}_{8}O \xrightarrow{\beta^-} {}^{20}_{9}F \xrightarrow{\beta^-} {}^{20}_{10}Ne(stable)$$

These reactions are forms of beta decay, in which an electron or positron is emitted.

There are other types of decay that are also important for nuclear energy systems. In heavier nuclei (i.e. those whose nuclear mass exceeds that of lead nuclei) radioactive decays sometimes occurs by the emission of an alpha particle (which is really just a bare helium nucleus, and is written as E_{re}) and the corresponding change in composition (commonly referred to as nuclear transmutation). For example, an isotope of uranium can undergo the spontaneous alpha decay reaction given as

$$^{238}_{92}U \xrightarrow{\alpha-decay} {}^{234}_{90}Th + {}^{4}_{2}He$$

Just as in the beta decay examples considered above, if the daughter nucleus is unstable, then that nucleus can then subsequently decay via an alpha or beta decay process. This can continue multiple times and will stop only when the nucleus is transmuted into a stable isotope (i.e. one with a very long or nearly infinite lifetime).

There is one additional nuclear transmutation process that occasionally can be important in nuclear energy systems. In this process, a large (i.e. atomic number A>>1) proton rich nucleus can capture a K-shell electron (i.e. one of the electrons in the most tightly bound inner electronic orbital). This capture event transforms a proton in the nucleus into a neutron, and is observed to be accompanied by the emission of one or more gamma ray photons that carry off energy from the atom. A schematic of the transition is shown in Figure below.

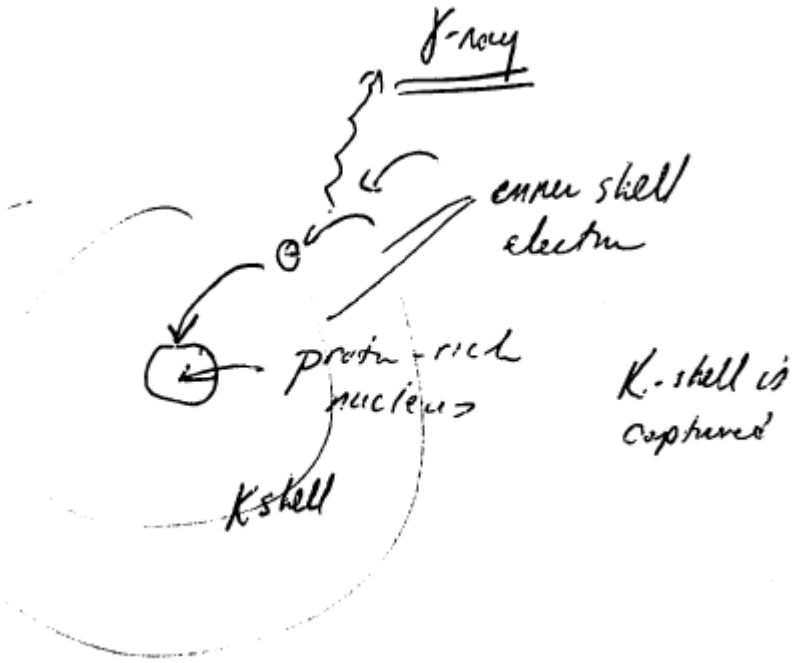


Figure 13.2: Schematic of a K-shell capture transition.

The discussion above summarizes several types of nuclear decay, including β^-, β^+, and α decay, that are important for nuclear energy systems, and was focused on considering the behavior of a single atomic nucleus. We now wish to determine how to quantify the decay of a collection of unstable nuclei. To make progress on this issue, it is necessary to make a key observation from experiments, which show that the probability that a given nucleus will decay per unit time is a constant characteristic for that particular nucleus and is independent of the number of nuclei present in the experiment. Let us refer to this constant as λ which has units of 1/time. Now, suppose we have a sample of radioactive material containing n(t) nuclei at time t. The question is then: how many will decay in during observations made over a small time interval (t, t + dt)? Following the experimental observations discussed above, it is clear that the answer will be the product λ n(t) dt. It is common to sometimes denote the Activity α as $\alpha = \lambda n(t)$ with units

[a]: { Curie (Cu) = $37\,10^{10}$ /sec
Becquerel, Bq = 1/sec
1 Bq = $2.703\,10^{-11}$ Cu

It follows that the decrease in n(t) is then $E_{net} = E_o - E_{re}$. This can then be recast as a differential equation which has the solution:

$$E_{net} = E_o - E_{re}$$

or equivalently

$$n(t)\, E_R \equiv \frac{E_o}{E_{re}}$$

And thus using the definition of activity given above we then see that it decays as

$$E_R \equiv \frac{E_o}{E_{re}}$$

The so-called half-life is defined as the time interval required for the activity of a sample to decrease by a factor of two, i.e. for $\alpha \rightarrow \alpha / 2$. Using the results above, we can then write the half-life in terms of the decay rate as

$$\alpha\left(T_{1/2}\right) = \alpha_o / 2$$

$$T_{1/2} = \frac{\ln 2}{\lambda} \simeq \frac{0.69}{\lambda}$$

And the activity in terms of the half-life as

The half life is commonly tabulated for many isotopes and can range from nanoseconds or less all the way up to billions of years depending upon the isotope being considered.

One can show that the average life expectancy of any nucleus or mean-life, τ is then given by

$$E_R = \frac{E_o}{E_{re}} = \frac{E_{re} + E_{net}}{E_{re}} = G + 1$$

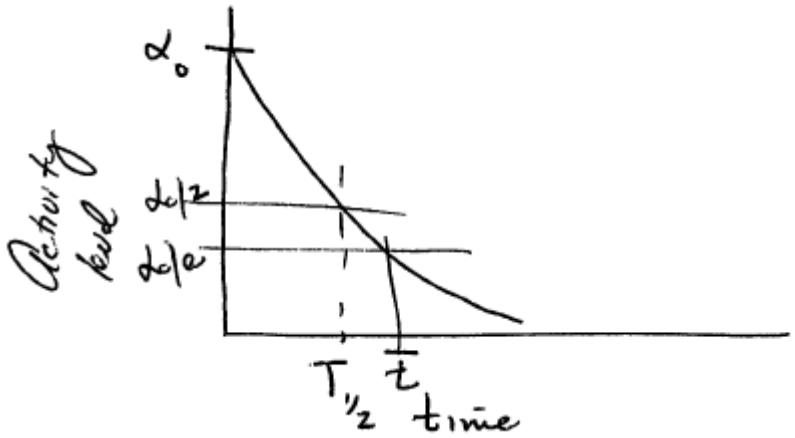


Figure 13.3: Exponential decay of activity level of a sample composed of unstable nuclei.

We can use these simple results to gain insight into the behavior of an evolving population of unstable nuclei. For example, suppose we wish to create an unstable isotope in a reactor or accelerator e.g. for medical purposes. Typically, this would be done by bombarding a collection of stable nuclei with a beam of neutrons, protons, or other heavier nuclei within the reactor or accelerator. The resulting interactions between the beam and the target would then transmute the target nuclei into a new unstable nucleus. Later on, we will determine how to calculate this transmutation rate, but for now let us suppose that these unstable nuclei are produced at a rate R (which has units of inverse seconds). The unstable nuclei have a decay rate given λ, and we wish to

determine the time evolution of the number of unstable nuclei, n(t), within the reactor or accelerator.

We begin by writing the time rate of charge of the unstable nucleus population n(t) as

(rate of production) – (rate of decay)

$$\frac{E_{net}}{E_{\mathrm{Re}}} = E_R - 1$$

where R and λ are two constants. This is a simple inhomogenous linear first order ordinary differential equation which has the solution for n(t) given as

$$\frac{E_{net}}{E_{\mathrm{Re}}} = E_R - 1$$

Thus, for example, if we begin this process at t=0 with n(t=0)=0, i.e. if at the beginning of the beam irradiation of the target there are no unstable nuclei present, then the unstable nuclei population builds up as shown in below,

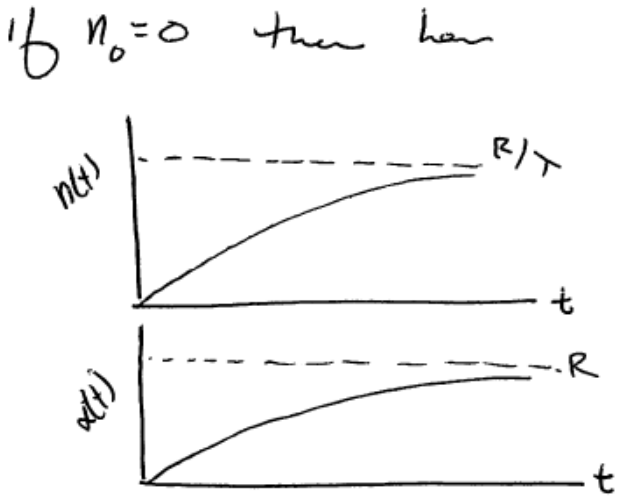


Figure 13.4: Build-up of unstable nuclei within a reactor or accelerator due to the combined effects of a fixed production rate, R, and a fixed decay rate, l.

We can gain further insight by now considering the following. Suppose we continue exposing the target to the particle beam and wait a sufficiently long time such that $t >> 1/\lambda$. In this case, the total number of unstable nuclei population will (very nearly) come to an equilibrium given by the ratio R/λ. Then, after coming to equilibrium we then shut off the production of the isotope (i.e. we suddenly set the production rate $R \rightarrow 0$). Based upon the results above, we know that the population would then decay as shown in below.

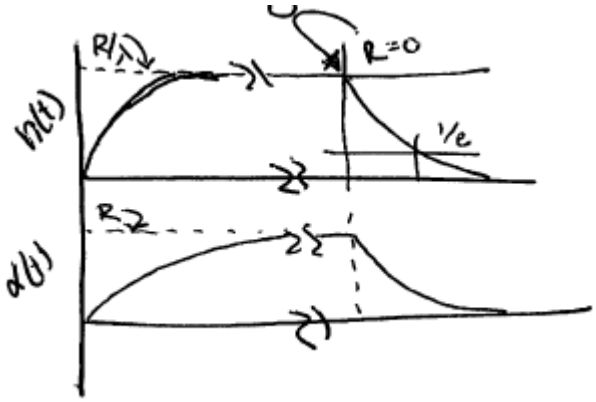


Figure 13.5: Time evolution of the population of unstable nuclei with a decay rate l in the case where first the production rate of the nuclei is fixed at a rate R and then, at some late time, the production rate is quickly set to zero.

To solidify understanding of radioactive decay processes, let us consider the behavior of a so-called decay chain, in which an unstable species A decays into a daughter product B which in turn is unstable. This species then undergoes a subsequent decay into species C and so forth. We can write this symbolically as $A \rightarrow B \rightarrow C \rightarrow \ldots$.and now we need to calculate the activity α, for a product in the decay chain. This process describes the decay of nuclear waste products and thus is of fundamental importance in considering nuclear energy technologies for large scale use.

If at t=0 we begin this process with a population n_{A0} of unstable isotopes, we can then write particle balance equations for species A and B as

$$\frac{dn_B}{dt} = -\lambda_B n_B + \lambda_A n_A$$

$$n_A(t) = n_{A_\circ} e^{-\lambda_A t}$$

$$\frac{dn_A}{dt} = -\lambda_B n_B(T) + \lambda_A n_{A_\circ} e^{-\lambda_A t}$$

$$n_B(t) = n_{B_\circ} e^{-\lambda_B t} + \frac{n_{A_\circ}\lambda_A}{\lambda_B - \lambda_A}\left(e^{-\lambda_A t} - e^{-\lambda_B t}\right)$$

FIX THESE EQUATIONS FROM NOTES

Plotting the evolution of the populations of species A, B and C as shown in Figure , we see that at first, the populations of the decay products B and C are zero at t=0, while the initial population of species A at t=0 is known from the initial value specified. Then, as species A decays its population dies off exponentially; however the decay of A leads to the birth of species B and thus that population begins to build up as time proceeds. As species B builds up, some of those nuclei begin to decay into species C, leading to the build-up of that species. When sufficient time passes, the population of species B will reach a maximum and will then begin to decay away as the total decay rate for species A declines as most of those nuclei decay away. Eventually the decay chain stops when the nuclear decay leads to the production of a stable isotope that has an indefinite lifetime.

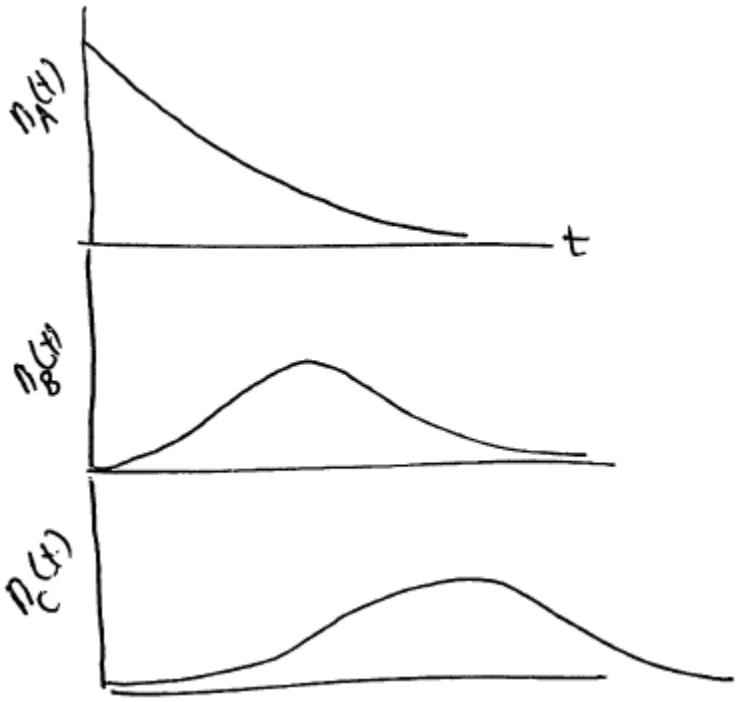


Figure 13.6: Evolution of the populations of species A, B and C in a hypothetical decay chain described by the reactions A→B→C.

Up until this point, we have only considered the behavior of one or more nuclei due to spontaneous nuclear decay. However, in nuclear energy systems, many times we are interested to know what happens when two or more nuclear particles (i.e. neutrons, protons, and atomic nuclei) interaction. These so-called nuclear reactions form an essential part of the operation of existing and proposed nuclear energy systems and thus are essential to understand. For our purposes, a nuclear reaction is said to occur when 2 nuclear particles (e.g. 2 nuclei, or nucleus and proton or neutron) interact to produce 2 or more nuclear particles. If *coincident* particles are *a* and *b* and *exiting* particles are *c* and *d* then such a reaction can be denoted via the notation

$$a + b \rightarrow c + d$$

Alternatively the reaction is denoted via the notation

a(b,c)d

Experiments with a variety of types of nuclear reactions (we will have more to say about this topic in subsequent sections of this chapter) show that four fundamental conservation laws are followed by such reactions. In particular, the number of nucleons, total electrical charge, momentum and total energy (we will define what we mean by total energy in the next section) are all conserved. These conservation laws provide powerful constraints on what types of reactions can and cannot occur, and must always be kept in mind when examining nuclear reactions within nuclear energy systems.

Nuclear Binding Energy

Suppose we were to measure the mass of an atomic nucleus, X, with a very precise instrument and we designate this mass as $E_{\mathrm{Re}}$and furthermore suppose that this nucleus is composed of Z protons and A-Z neutrons, where A is the atomic number of the nucleus (i.e. using the notational scheme we introduced earlier, we could write the nucleus as $W = \int_{x_1}^{x_2} F(x) \cdot dx$ where here Z is the atomic number and A the atomic weight of the nucleus). Such measurements have indeed been carried out for many nuclei. Invariably these measurements show that

$$E_{\mathrm{Re}}$$

Where $E_p = E_{\mathrm{Re}} + E_n$ and $E_n = E_{\mathrm{Re}} + E_n$ denote the rest mass of the proton and neutron respectively and $\frac{E_{src}}{E_{net}} = \frac{E_R}{E_R - 1} + \frac{1}{G} = \frac{E_R + 1}{E_R - 1}$ is the number of neutrons within the nucleus. In other words, the mass of the nucleus is *less* than the sum of the masses of each constituent nuclear particle. It is as if, were we to assemble an atomic nucleus from the constituent nuclear particles, some of the mass of these components has disappeared. We can define the missing mass, usually described as the mass defect, $\frac{E_{src}}{E_{net}} = \frac{E_R}{E_R - 1} + \frac{1}{G} = \frac{E_R + 1}{E_R - 1}$ by the expression

$$\frac{E_{src}}{E_{net}} = \frac{E_R + 1}{E_R - 1}$$

We can get a better idea of what has happened to this missing mass by considering a complementary (and hypothetical) experiment. Suppose we were to take this atomic nucleus and, using a hypothetical tool, disassemble the nucleus piece by piece by removing each proton and neutron one by one. In order to do this, we would find that we must do work on the constituent particles, that is to say, we would have to exert a force on each particle to remove it from the nucleus and remove it to a distance that is located a long distance away. Apparently, these nuclear particles are held within the nucleus by some effective force or, viewed from the standpoint of energetics, these particles sit within a potential well and must be lifted up out of that well in order to disassemble the nucleus. Let us denote the amount of work needed to disassemble nucleus in this manner as the binding energy, which we will denote via the symbol E_B. It turns out that this work is related to the mass defect by the relation

$$\frac{E_{src}}{E_{net}} = \frac{E_R + 1}{E_R - 1}$$

We can then define the binding energy per nuclear particle via the ratio of this energy to the total number of nuclear particles within the nucleus, i.e.

NEED TO FINISH THIS DISCUSSION AND ADD PLOT OF "CURVE OF BINDING ENERGY"

The above plot shows us how tightly bound a nuclear particle is to a given nucleus, and shows that nuclei with middle ranged atomic numbers (i.e. with A ~200 and Z~80-100) have the highest binding energy, while lighter and heavier nuclei have lower binding energies per unit nuclear particle.

Let us now consider the implications of this finding. We have already discussed that the total energy (defined as the sum of the rest mass energy and the kinetic energy) of the constituent particles within a nuclear reaction must be conserved. Now, suppose that we arrange for a nuclear reaction to occur in such a way that the incident particles

have a lower binding energy per unit nuclear particle than the reaction products. To put it another way, the nuclear particles of the reaction products sit in a deeper potential well than did the nuclear particles of the incident nuclei. In such a case, the conservation of total nuclear energy then forces only one conclusion: the products leaving this reaction must have a higher kinetic energy than did the incident nuclear particles. Such reactions are endothermic, in that they release stored nuclear energy in the form of high kinetic energy of the reaction byproducts, in an analogous manner to the formation of energetic reaction products during endothermic chemical reactions.

Such endothermic reactions are possible in two cases: a) when a heavy nucleus breaks up into two or more smaller nuclei, and b) when two light nuclei combine to form a more tightly bound heavier nucleus. The first reaction is an endothermic fission event, while the second reaction is an endothermic fusion reaction. These two fundamental types of reactions form the basis for nuclear fission and nuclear fusion energy technologies. Fission reactors already exist and are used today around the world; fusion reactors are still in the research stage and may, in the coming decades, provide a second fundamentally distinct type of nuclear energy.

Particle Energy Distributions

Statistical thermodynamics shows that the particles in a gas (e.g. a neutral gas in the room) have a variety of kinetic energies. When such a collection is in thermal equilibrium, designated by a temp *T*, the kinetic energy distribution N(E) is given as:

$$E_{src} / E_{net}$$

Where

N ~ nuclear density (number per unit volume) of the particles

k ~ Boltzmann's Constant

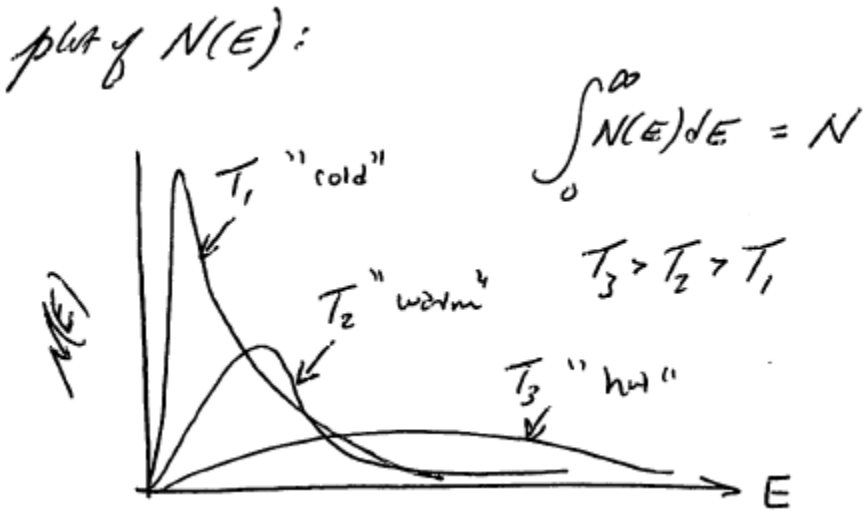
here N is related to the mass density, ρ [mass per unit volume] by the relation:

$$E_{src} / E_{net}$$

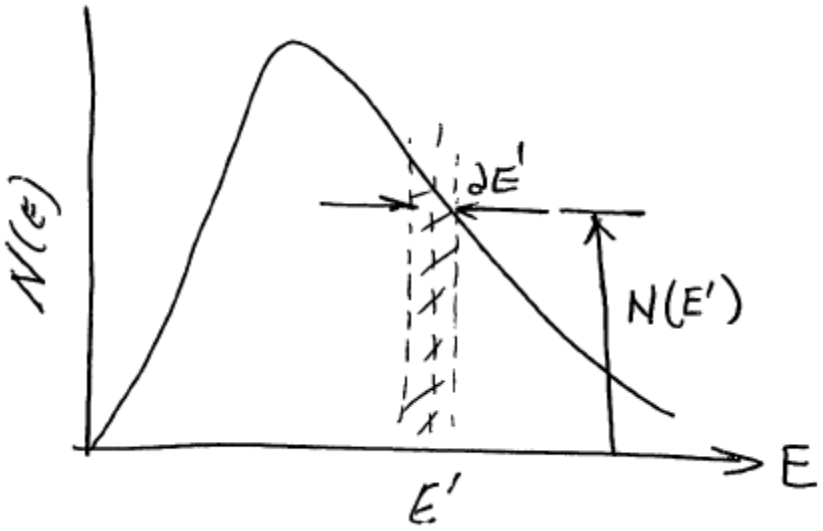
Where

E_R ~ Avogadro's number

M ~ mass of one mole of atoms or particles



To understand physical meaning of N(E) let us look in more detail:



Consider the shaded region above. It has an area given approximately by $E_{src} \approx E_{net}$

If we considered a variety of values for E' ranging from E=0 to E→infinity, and add them all up etc (?) we find:

$$E_R \to 1$$

But the integral has a value of 1, thus $E_R \to 1$.

i.e. the total area under the curve is equal to the nuclear density N of the particles. This then suggests the physical interpretation of N(E'):

N(E')dE' is equal to the number of particles per unit volume with energy in range (E', E' + dE')

We could also define a related function:

$$E_{src} >> E_{net}$$

Question: what would p(E)dE then correspond to?

This distribution function can be used to describe a number of useful quantities:

A) Most probable energy, E_p, given by the peak in N(E) or p(E)

Can show that $E_{src} \gg E_{net}$

B) Average Energy, :

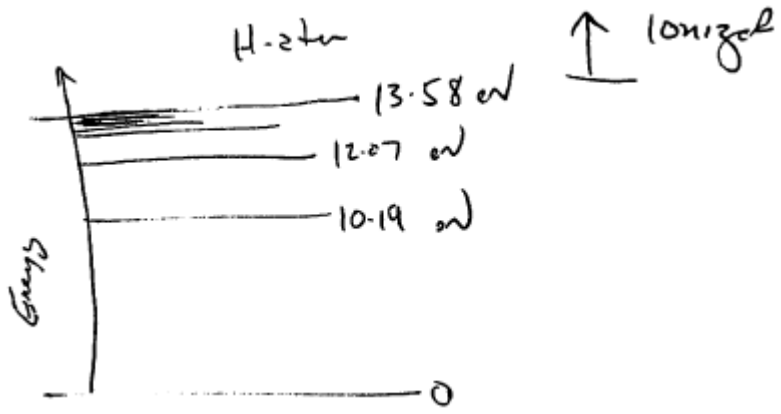
$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right)$$

Why is this so important?

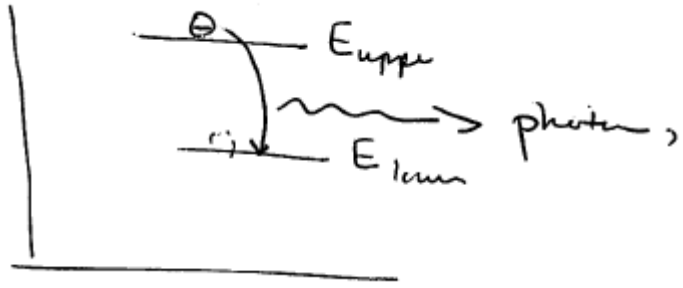
- The nuclear reaction rates of interest are functions of particle energy, and will need to be evaluated appropriately.
- Nuclear reaction products typically have very high kinetic energies, E>>>kT. These fast particles can collide with background particles and transfer some or all of their energy to them
- This "slowing down" process, and what can happen to the particles while they slow down can have a big impact on the system behavior (e.g. thermal stability)

Excited States and Radiation:

Bound electrons have electronic excited states



When a bound electron is given some energy, it can move into one of the "excited states." However, they usually don't stay there long—usually such states quickly decay down to a lower energy state or even a ground state. The energy difference between the upper and lower states is taken up by the emission of a photon.



This photon will have an energy E given as $E = E_{up} - E_{low}$

The frequency and "wavelength" of the photon is related to E as follows:

$$P(t) = \frac{KP_0}{P_0 + (K - P_0)e^{-rt}}$$

Likewise, the atomic nucleus can have excited states which undergo decay procedures. However, the energies of these states are typically on the order of MeV and thus result in the emission of more energetic photons.

Ex 2.5 pg 16

In order to derive a basic understanding of the operation of a fission reactor, there are a number of interactions that can occur between neutrons and atomic nuclei which need to be summarized. These include elastic scattering, inelastic scattering, neutron capture, charged particle reactions, and neutron producing reactions, including nuclear fission. In the following, we summarize key aspects of these types of interactions.

Elastic scattering involves interactions in which an incident neutron with known kinetic energy approaches a nucleus. When the neutron gets close enough to the nucleus the two particles then interact via the strong nuclear force. In this type of interaction, the neutron then exits the near-nucleus region with a different kinetic energy and momentum; the nucleus then recoils with some other kinetic energy and momentum. The key aspect that then distinguishes elastic interactions from other types of interactions is that the total kinetic energy and momentum of the neutron and nucleus are conserved, while in the other interactions, distinct and/or additional particles (e.g. photons, other neutrons, alpha particles or beta particles) are emitted and can carry off some of the energy and momentum of the incident neutron and nucleus. Figure below provides a schematic of the elastic neutron-nuclear interaction.

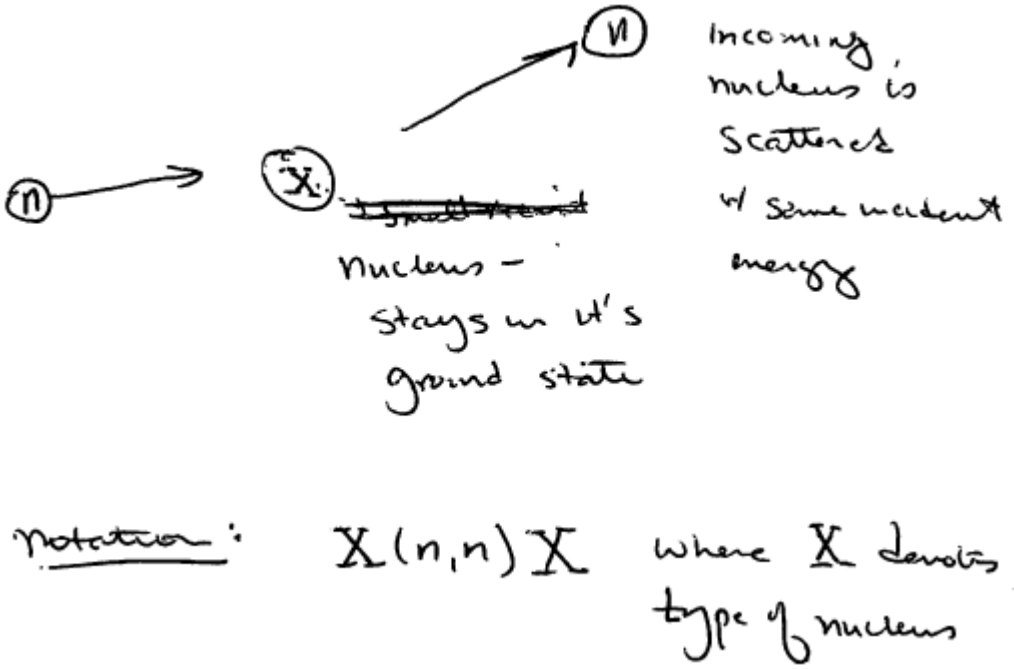


Figure 13.7: Schematic of an elastic interaction between an incident neutron and a target nucleus.

Inelastic scattering is a second type of important neutron-nuclear interaction to be considered in the operation of a fission reactor. Figure below provides a schematic of the interaction. In this case, an incident neutron approaches a nucleus and begins to interact with it. The neutron has sufficient incident energy that it can then put the nucleus into an excited state, in which the protons and neutrons within the nucleus can be thought of as having higher energies, much like the electrons in an atom can be put into a higher energy state by the absorption of a photon of sufficient energy. The neutron then exits the nuclear region with a reduced kinetic energy; usually the lifetime of the excited nucleus is rather short (with a decay timescale typically measured in microseconds or less) and the

nucleus then spontaneously decays to a lower energy state or to the ground state, and emits an energetic photon (i.e. an x-ray or a gamma ray).

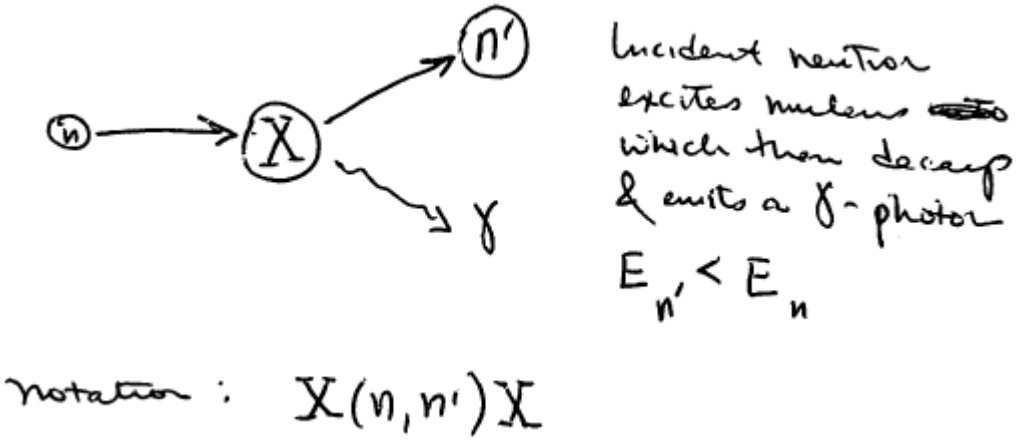


Figure 13.8: Schematic of an ineslastic neutron-nuclear scattering event

A third type of interaction is termed the radiative capture interaction, and is shown schematically in Figure below. In this interaction, an incoming neutron is absorbed by nucleus, which then emits a gamma ray photon. The neutron is not subsequently ejected; instead the nucleus retains it and now has one additional neutron. This absorption usually leaves the nucleus in an excited state; the subsequent decay of the nucleus is then accompanied by the emission of an energetic photon as shown in the figure. The incident energy and momentum is then carried out of the reaction by the energy and momentum of the combined nucleus and photon.

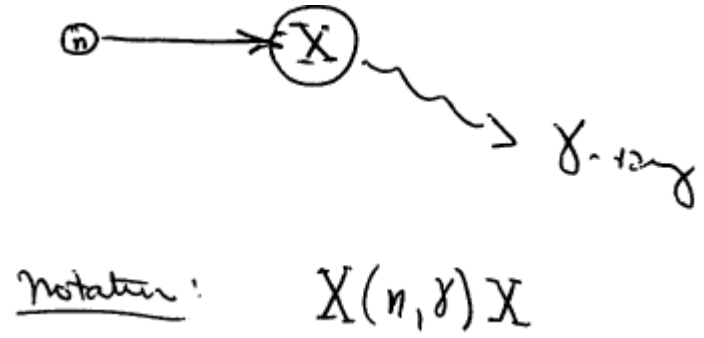


Figure 13.9: Schematic of radiative capture neutron-nucleus interaction

The so-called charged particle reactions form a slightly different type of neutron-nucleus interaction. Figure below provides a schematic of this type of interaction. In this case, an incident neutron is absorbed by the nucleus. The resulting modified nucleus is unstable, and decays by emitting a charged particle (usually either proton or an alpha particle) and one or more energetic photons. These exiting particles together carry off the momentum and energy of the incident particles.

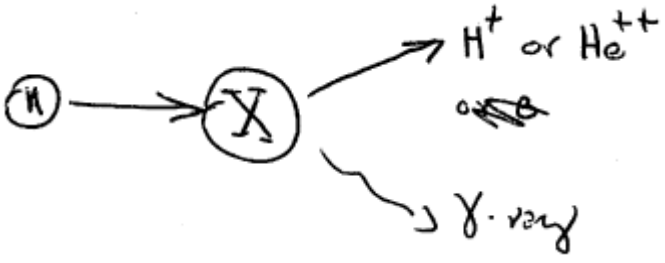


Figure 13.10: Schematic of a neutron-nucleus interaction that leads to the subsequent emission of charged particles.

There are two classes of neutron-nucleus interactions which emit more than one neutron from the interaction. The first type, illustrated in Figure below, has an incident

neutron absorbed by the nucleus which is then transformed into an unstable isotope. This isotope then undergoes a decay event and emits 2 or 3 neutrons while retaining all of the nuclear protons. This interaction is usually describes as an n-2n or n-3n reactions; the choice is then determined by the number of neutrons that are emitted during the nuclear decay.

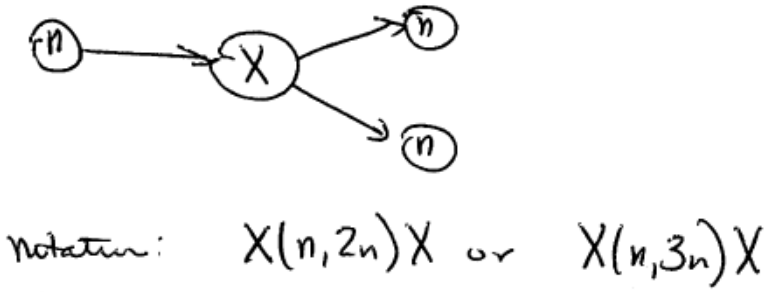


Figure 13.11: Schematic of 2n and 3n neutron-nucleus interaction

The final neutron-nucleus interaction which we are interested in is the fission reaction, shown schematically in Figure below. In this reaction, the incident neutron is absorbed by the nucleus. The resulting isotope is unstable, and subsequently decays by splitting into two lighter nuclei, and emits 2 or 3 neutrons from the interaction. This reaction is of particularly importance in nuclear reactors, and will be discussed in more detail below, after we introduce several other critical ideas.

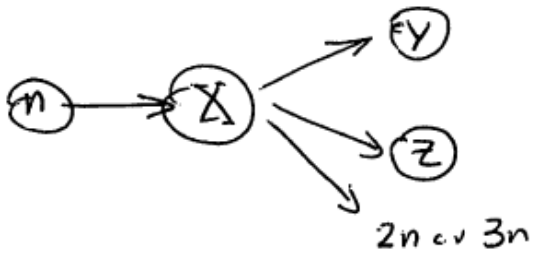


Figure 13.12: Schematic of a nuclear fission reaction

Having introduced these important neutron-nucleus interactions, we need to now determine a scheme to measure the likelihood, or probability for such interactions. This then leads to the concept of the nuclear reaction cross section, which can then be used to characterize each class of these neutron-nucleus interactions. To proceed, let us consider a beam of mono-energetic, mono-directional neutrons impinging uniformly on a target of cross sectional area A as shown in Figure below. The neutrons have a number density *n* neutrons/unit volume and all have the same speed *l*. We can define the quantity I=nv which denotes the intensity of the beam which has units of particles per unit area per unit time. Now, in a time interval δt, the neutrons will travel a distance ν = υδt. Thus, neutrons in the volume *lA= nυA* will impinge on the target during this time interval. Thus, we can conclude that *I* corresponds to the number of neutrons striking the target per unit area per unit time.

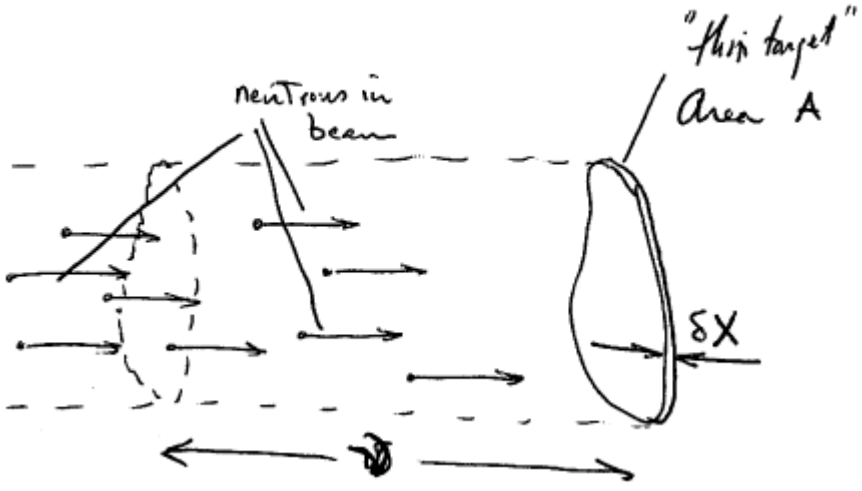


Figure 13.13: Schematic of a beam of monoenergetic, mono-directional neutrons incident on a thin target.

Next, we wish to determine how many of these neutrons interact with the nuclei in the target (remember that the neutrons have no electric charge and thus do not interact with the electrons nor with the Coulomb field of the nucleus, but instead interact with the nuclei in the target via the strong nuclear force). Let us denote the number of neutron-nuclei interactions occurring in the target per unit volume and unit time as . It stands to reason that: , where *N* is the nuclear density of nuclei in the target, N is the number of nuclei per unit volume in the target, and denotes the target thickness.

If we examine the units of each term in the inequality we that

:

$$[I] \sim \frac{\#}{area \cdot time}$$
$$[N] \sim \frac{\#}{volume}$$
$$[A] \sim area$$
$$[\delta X] \sim length$$
$$\therefore [INA\delta X] \sim \frac{\#}{area \cdot time}$$
$$[\zeta] \sim \frac{\#}{time}$$

Thus introducing a constant of proportionality σ (with units of Area) then allows us to write

.

Here σ is called the microscopic cross section and has the units of an area. The value of this quantity is determined by the neutron energy, the composition of the target nucleus, and the type of interaction of interest. We can obtain insight into the meaning of this quantity by noting that the factor $NA\delta x$ is simply the total number of nuclei in the target. Thus, it follows that σI denotes the number of interactions with a single nucleus per unit time. Thus, one interpretation of the microscopic cross-section σ is that it denotes the number of interactions/unit-time/per unit beam intensity.

There is also a second way to view the concept of nuclear cross-section. In order to gain insight into this interpretation of the microscopic cross-section, consider Figure below which shows the thin target to be composed of distinct atoms with their associated atomic nuclei.

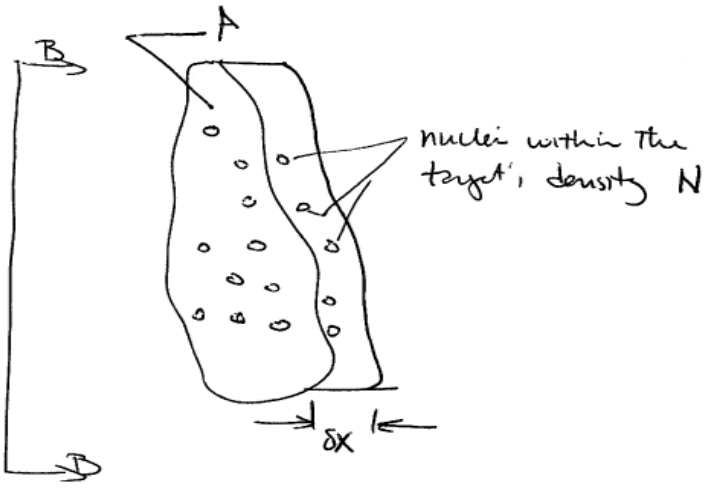


Figure 13.14: Schematic of a thin target showing it composed of a number of distinct atoms with their associated nuclei.

Now, consider the view of the thin target from perspective B-B shown in Figure above, i.e. consider a face-on view of the target as shown in Figure below. The total number of nuclei in the target is given as NAδx. Now, if each nucleus has an effective cross-sectional area δA, then the total area density of the nuclei is given by the product (NAδx)δA. It then follows that the fraction *f* of the frontal area that is taken up by the nuclei in the target is:

As shown in Figure below, let us now define the effective cross-sectional area δA of the nucleus is defined by noting that if a neutron's incident trajectory causes it to intersect this area then it is to be understood that there an interaction will then occur

between the neutron and the nucleus, while if the trajectory always falls outside of this area then no interaction takes place.

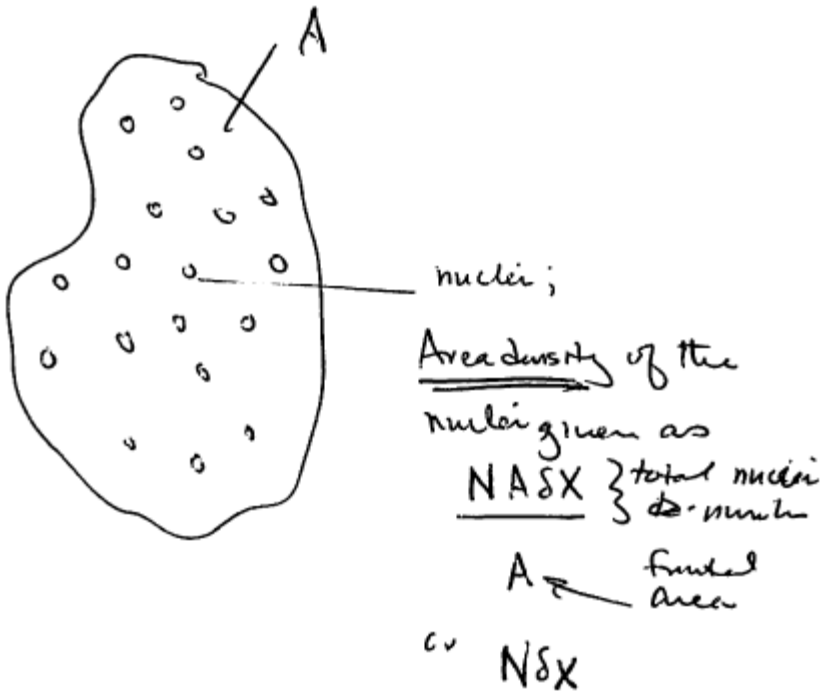


Figure 13.16: Face-on view of the thin target, showing it composed

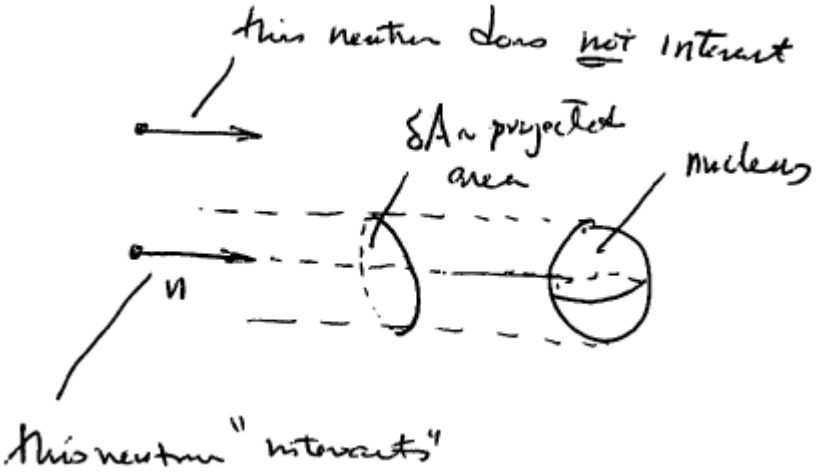


Figure 13.17: Schematic of the microscopic cross section of an individual nucleus defined in terms of whether or not the neutron "interacts" with the nucleus.

With these elementary concepts, we can now determine an alternate interpretation of the cross-section σ. Suppose that we have *I* neutrons per unit area per unit time incident on the target. Then, by the foregoing definition for the area δA, a fraction *f* of these particles will "interact" with the target while the fraction 1-f will not interact with the target. Now, examining the earlier results in which we first introduced the concept of the microscopic cross section, we see that σ is simply given by the interaction area, δA, i.e. $\sigma = \delta A$. Thus σ denotes the effective area of a target nucleus for an interaction to take place between the neutron and the target nucleus.

Each of the different types of nuclear reactions can be characterized by a corresponding cross-section, i.e. σ_e denotes the elastic scattering cross-section, σ_i denotes the inelastic scattering cross-section, σ_γ denotes the neutron capture cross-section, σ_f denotes the fission cross-section, σ_p denotes the proton emission cross-section, σ_α denotes

the α-particle emission cross-section and so forth. It is sometimes convenient to define the total cross-section as

$\sigma_t = \sigma_e + \sigma_i + \sigma_\gamma + \sigma_f + \ldots,$

the absorption cross-section as

$\sigma_a = \sigma_\gamma \sim \sigma_f \sim \sigma_p \sim \sigma_\alpha + \ldots$

and the scattering cross-section as

$\sigma_s = \sigma_e + \sigma_i$

$\sigma_\varepsilon = \sigma_s + \sigma_a$

Since $A\delta x$ is the volume of the target, we can conclude that the total number of interactions per unit volume per unit time is:

$$F = I\, N\, \sigma_t$$

A similar interaction rate per unit volume and unit time can be defined for specific types of interactions, e.g. the total number of fission events per unit volume and unit time can be defined similarly.

It is also common to sometimes see the so-called macroscopic cross section, Σ, written in as $\Sigma = N\,\sigma$. Note here that Σ has units of 1/length and thus the terminology "cross section" may be a bit misleading; nonetheless this is standard usage in the nuclear engineering literature and we will maintain the convention here.

Our previous considerations focused on a "thin" target. Let us now consider what happens in a "thick" target, i.e. one with a thickness that is large compared to the typical distance a neutron travels between collisions with atomic nuclei. Referring to Figure below, consider a finite thickness target that scatters some of the incident neutrons away from their original incident direction. A detector is placed behind the target at a distance large enough that its detection area is negligible as seen from the target. Thus, the detector only measures unscattered neutrons.

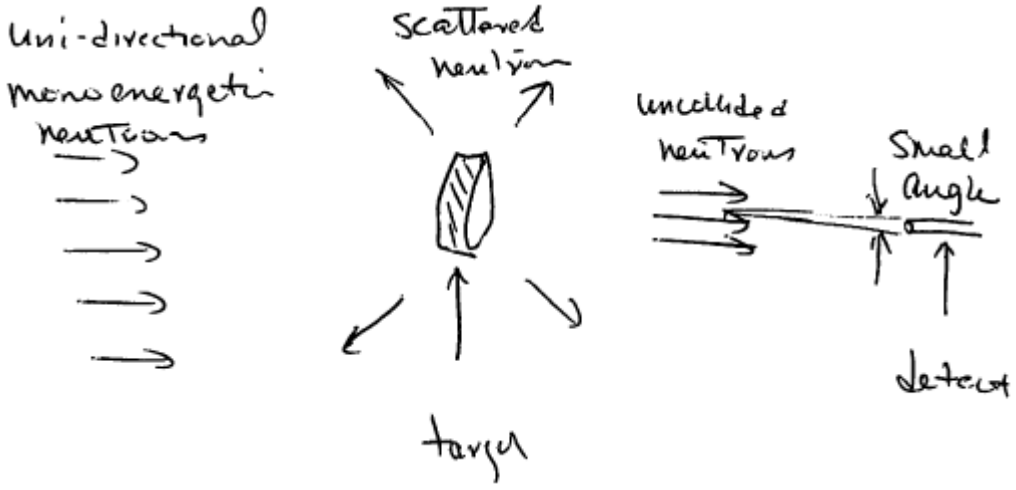


Figure 13.18: Schematic of a unidirectional monoenergetic neutron beam interacting with a finite thickness target.

Now let us concentrate on the thick target, and let I(x) denote the intensity of the beams that have NOT interacted with the target at position x, as shown in Figure . By convention we define x=0 to be the front face of the target where the beam particles enter

the target volume. The change in the intensity, dI(x), due to neutron-nuclei interactions in the region (x, x+dx) and the subsequent solution for I(x) can then be written as

Thus, the macroscopic cross-section is simply the inverse of the e-folding decay length of the uncollided beam intensity.

We can also see a slightly different interpretation of Σ. From above we can write

$$\Delta t$$

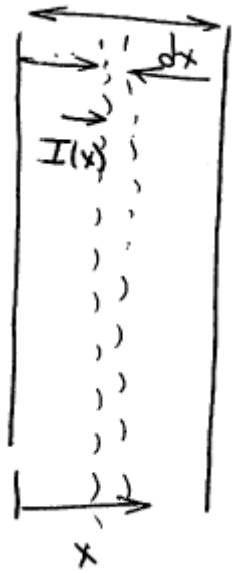


Figure 13.19: Schematic of a thick target denoting the un-interacted beam intensity I(x) at position x within the target.

Now dI(x) gives the number of neutrons out of a total of I(x) that collide in (x, x+dx), and thus  gives the probability that a neutron first survives uncollided up to position x, and then suffers a collision within (x, x+dx) and Σdx also gives this same probability. Thus Σ can be interpreted as the probability of a neutron-nucleus interaction per unit path length in the target medium.

Finally, we note that these concepts can be extended to mixtures of materials. For example, suppose we have 2 types of atoms, X and Y , in target. Each has a different number density, N_x, N_y and microscopic cross-sections σ_x andσ_y. It is common to then define an effective macroscopic cross-section is then

$$\Delta P = P_0 \left[(1 + r_0)^{t+\Delta t} - (1 + r_0)^t \right]$$

This can be extended to any number of target nuclei, and can be applied to any or all of the neutron-nuclei interactions to define e.g. effective scattering cross sections, absorption cross sections, and so forth.

We can now extend these concepts to examine neutron-target interactions for neutrons that are traveling in more than one direction. Suppose we have several beams incident on an incremental volume as shown below and suppose that each beam has an intensity I_A, I_B, I_C and each comes at the target from a different direction. Our previous result then allows us to write the total volumetric interaction rate as

$$F = \Sigma_t (I_A + I_B + I_C + \ldots)$$

If we write the beam intensity I as $I_i = n_i \upsilon_i$ and assume all of the beams have the same speed υ_i, then we have

$$F = \Sigma_t (n_A + n_B + n_C + \ldots)\upsilon$$

Now suppose that we continue this process until there are such a large number of beams incident on the target that the target sees an isotropic distribution of neutrons incident on the target. In this case, we can generalize the above results and write an isotropic neutron flux as

$$F = \Sigma_t n \sigma$$

where now υ is understood to be some part of characteristic speed for the neutrons and n denotes the number of neutrons per unit volume. The term $n\upsilon$ occurs so often that it is usually written as $\varphi = n\upsilon$ and is denoted as the *neutron flux* which has units of neutrons per unit area unit time. Note that this quantity is a scalar quantity, and that the interaction rate per unit volume is then given as $F = \Sigma_t \varphi$.

In reality, neutrons within a reactor have a variety of speeds and kinetic energies and thus the above concepts are generalized to this more complex situation. Before taking up the essential aspects of this fact, let us first consider the kinematics of neutron-nuclei collisions, and in particular examine how neutrons slow down from their very high birth energy from fission (at ~2MeV)

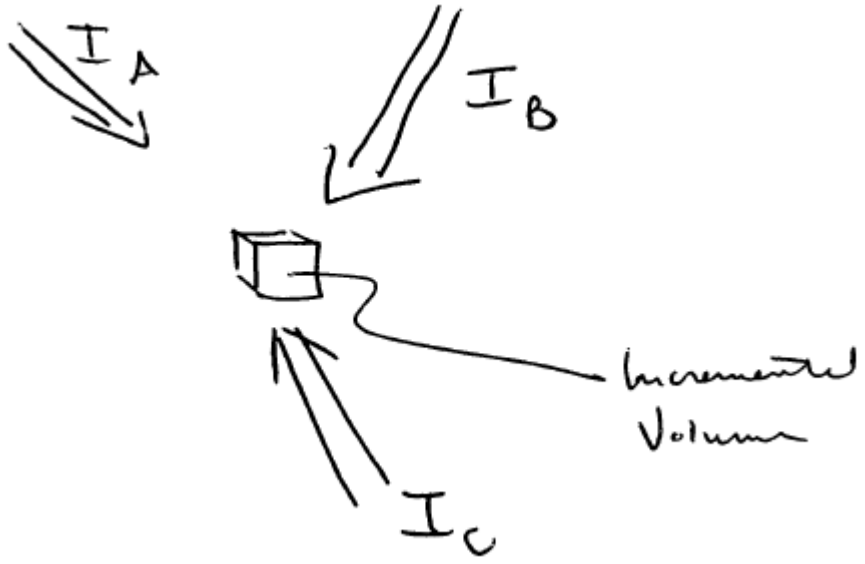


Figure 13.20: Multiple single speed beams incident on an incremental volume.

Energy Loss in Elastic Scattering

In the core of a nuclear reactor, elastic scattering between a neutron and a core atomic nucleus is one of the most common types of interactions and thus is an important process to consider here. Of particular importance is the energy of the neutron and nucleus before and after the collision. We present the key results here and refer the interested reader to the references for more detail. The particle momentum vectors are shown in Figure below.

The incident and scattered neutron energy and momentum are denoted as (E, p) and (E', p') respectively, and (E_A, P) denote the recoiling nucleus energy and momentum.

The nucleus is usually assumed to be at rest prior to the collision. The conservation of energy and momentum are then written as

$E = E' + E_A$

And

$P = p' + P$

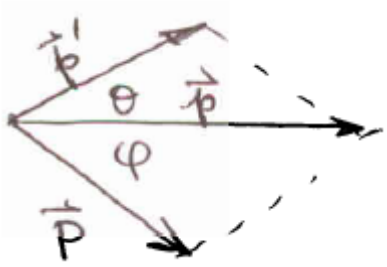


Figure 13.21: Incident (p) and scattered (p') neutron momentum vectors, and scattered nucleus momentum vector, P.

Using these conservation laws, it can be shown that the scattered energy is related to the incident energy and the scattering angle via the relation

$$r_2 = f^2 r_0$$

where A denotes the atomic mass number of the nucleus and q denotes the scattering angle of the scattered neutron with respect to the incident neutron momentum vector.

We can identify the essential physics by considering several special cases of scattering. First, consider a very small angle scattering event in which the incident neutron scattering angle is vanishingly small, i.e. $\theta \rightarrow 0$. In this case, it is then clear that E=E' and $E_A = 0$, i.e the neutron does not lose any energy and the nucleus does not

receive any transferred kinetic energy. This is the trivial case of a vanishly weak elastic scattering event.

Next, consider a head-on collision such that the neutron is directly scattered backwards, i.e. the case where $\theta \rightarrow \pi$. In this case, it can be shown than E' is minimized and E_A is maximized, and that the neutron momenta and energy satisfy the expressions

$$r_2 = f^2 r_0$$

The leading factor in this last expression is an important quantity and is often referred to as the collision parameter α defined as

$$r_i = f^i r_0$$

A plot of α vs. A is given in Figure below. We note that it has the following important limits: if A = 1 (i.e. if the neutron scatters from a hydrogen nucleus (such as found in water), then we have

$\alpha = 0$

$E' = 0$

$E_A = E$

$\Theta_{max} = \pi/2$

i.e. the neutron can give up all of it's energy in a *single* head-on collision. In the opposite limit, when A>>1, then α=>1 and thus E'~E and thus the recoil energy E_A<<E. Thus in this limit it takes many elastic scattering events for the neutron to lose a significant fraction of its initial energy.

Note that we can write for any value of A: $(E')_{min} = \alpha E$

Average energy of a scattered neutron is E'_{avg} =

Average energy loss is $\Delta E_{avg} = E - E'_{avg} = 0.5(1-\alpha)E$

Average fractional energy loss $r_i = f^i r_0$

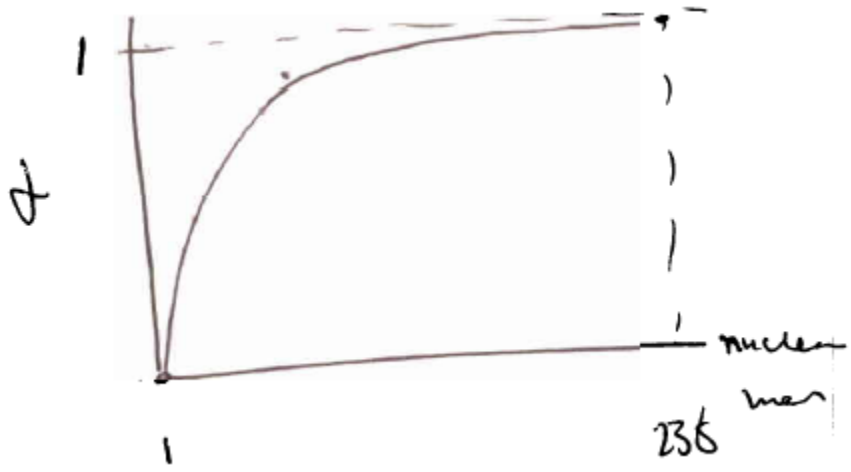


Figure 13.22: Plot of the elastic collision parameter vs. scattering nucleus atomic number.

This analysis, which is based on very elementary considerations, gives a key result which has a significant impact on the design of a nuclear reactor. If a reactor core contains a lot of low-mass nuclei and relatively few heavy nuclei, the neutrons from fission will lose their energy quickly (i.e. in few collisions) and will have an energy distribution that is (nearly) in thermal equilibrium with the reactor core. As shown schematically in Figure below, the net result is that such a reactor will have an energy spectrum in which the low energy neutron population very nearly follows a Maxwellian thermal distribution. Such a reactor is usually referred to as having a thermal neutron

spectrum, and sometimes is referred to as a thermal reactor. Conversely, if a core is composed of mostly heavier nuclei (carbon, sodium, …), then the fission neutrons will need many more collisions to lose their energy. Since fission reactions result in the birth of very high energy neutrons (as we saw earlier typically their birth energy is ~2MeV or higher), then the high energy (i.e. E>>kT where k is the Boltzmann constant and T is the reactor temperature) neutron population will significantly exceed the population that would occur at that energy if the reactor had a Maxwellian neutron energy distribution, i.e. if $N(E)=N_{Max}(E)$. Such a reactor is often referred to as a fast reactor. The presence of this elevated population of energetic neutrons can have an important impact on the reactor operation and behavior, and is particularly important for those reactions whose cross-section becomes large at high neutron energy. In particular, such reactions play a key role in the design of so-called breeder reactors, which transform normally inert nuclei such as U^{238} into fissile nuclei, and thus create, or breed, fissile material.

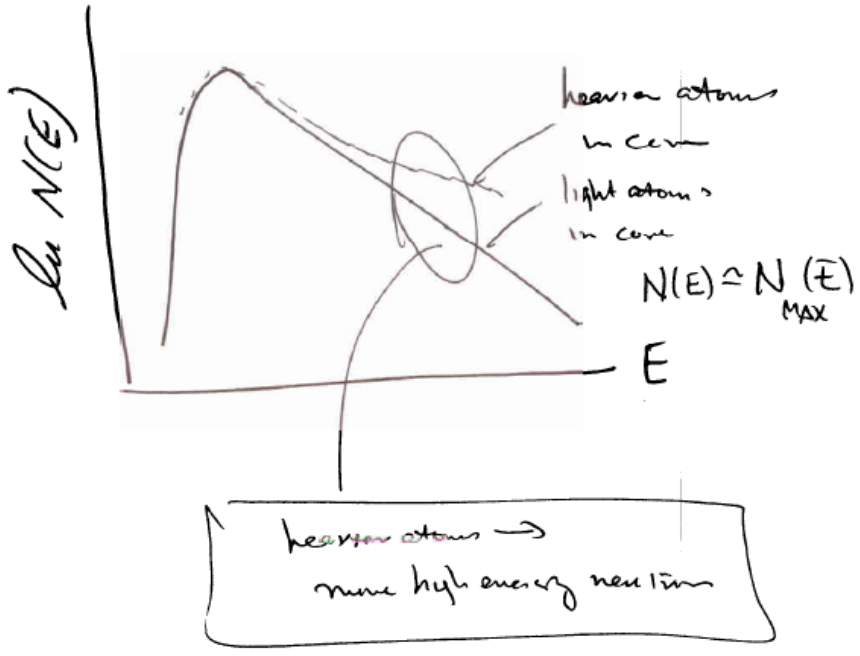


Figure 13.23: Schematic of the neutron energy distribution function, N(E), for a thermal neutron population and an energetic, or "fast" neutron spectrum.

The flux, $\varphi_o = nv_o$ is referred to as the 2200 m/sec flux. Thus

$F_a = \Sigma_a(E_o)P\varphi_o$ is the thermal absorption rate

There are a few nuclei that do not have Σ_a (E) [proportional] to $1/v$(E). Departure from $1/v$ behavior is characterized by a "non-$1/v$ factor" g_a(T):

$$\Delta P_0 \equiv \Delta P|_{t=0.\Delta t=1} = r_0 P_0$$

See table 3.2 in La Marsh (?)

For most nuclei, $\Delta P_0 \equiv \Delta P|_{t=0.\Delta t=1} = r_0 P_0$ for low energies.

Thus we can write $P_i = P_0 + \Delta P_0\left(1 + f + f^2 + \ldots + f^{i-1}\right)$ where E_o is an arbitrary reference energy and

$P_i = P_0 + \Delta P_0\left(1 + f + f^2 + \ldots + f^{i-1}\right)$ is the corresponding speed.

We can then write F_a as

$$F_a = \Sigma_a(E_o)v_o\int n(E)dE$$
$$= \Sigma_a(E_o)v_o n$$

$$P_i = P_0 + \Delta P_0\left(1 + f + f^2 + \ldots + f^{i-1}\right)$$
$$= P_0 + r_0 P_0\left(1 + f + f^2 + \ldots + f^{i-1}\right)$$
$$= P_0\left(1 + r_0\sum_{j=0..i} f^j\right)$$

Thus it is common to refer to v_o=2200 m/s, E_o=0.0253 eV

$i \longrightarrow \infty$ ~ "thermal cross-section"

The total interaction density is then obtained by integrating over all $i \longrightarrow \infty$

$$\sum_{\substack{j=0,i \\ \lim i\to\infty}} f^{\,j} = \frac{1}{1-f}; \quad f<1$$

If we wish to calculate *F* for only a specific type of interaction we just make the substitution for Σ_t, e.g $F_{fission}$ would be given as

$$\sum_{\substack{j=0,i \\ \lim i\to\infty}} f^{\,j} = \frac{1}{1-f}; \quad f<1$$ and so forth

Absorption in Thermal Reactors

When $n(E) \sim n_{max}(E)$ (i.e. neutrons are in thermal equilibrium w/ reactor core) the reactor is referred to as a "Thermal Reactor."

The previous discussion assumed that the neutrons have a simple energy. In any real system the neurons will have a distribution of energy. Thus, the question arises: how to calculate interaction ratios for such a collection.

Let n(E)dE denote the number of neutrons per unit volume with energy in the range (E, E + dE). This collection of neutrons is incident on a thin target much like in the previous mono-energetic problem.

The intensity dI(E) = n(E)dE *v*(E)

This beam will then interact w/ target at a volumetric rate

$$f = 1 - r_0\left(\frac{P_\infty}{P_0} - 1\right)^{-1}$$

Fission

Let us recall the so-called curve of binding energy which was introduced earlier in this chapter. We found that for energy release to occur during a nuclear reaction, we need to split heavier nuclei into smaller light nuclei, or fuse light nuclei into heavier nuclei. The question then arises: how can an incident neutron split the nucleus into two less massive components and thus release energy? To answer this query, it important to note that there are 2 important energies to keep in mind: First, there is a critical neutron energy need to split the nucleus, E_{crit}, and then second, there is the binding energy of the last neutron in a nucleus, $f = 1 - r_0\left(\frac{P_n}{P_0} - 1\right)^{-1}$

Consider first the energy needed to split a nucleus, E_{crit}: A crude model of the nucleus pictures it much like a liquid droplet which can exhibit a variety of oscillatory behavior. Droplet vibration requires an energy input, while splitting the nucleus (or droplet in this crude representation) requires a minimum, or critical, energy input as shown schematically in Figure below. This process typically requires ~ 4-6 MeV of energy for most Th, U, Pu isotopes and thus fission in this case requires very high energy neutrons. If such nuclei are exposed to a flux of thermal neutrons within a reactor core, they will not fission and thus will not contribute to sustaining the reactor at its operating condition.

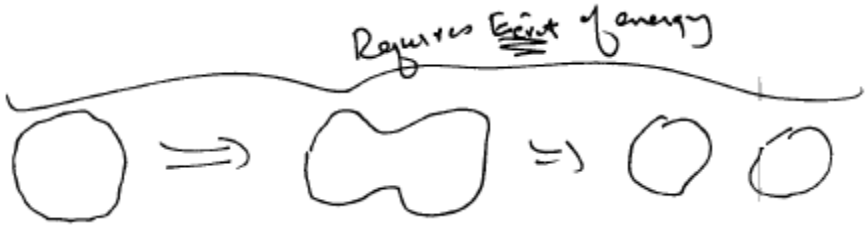


Figure 13.24: Representation of fission event as a large amplitude oscillation of a liquid droplet.

However, for certain special isotopes (^{233}Th, ^{235}U, ^{233}U, ^{239}Pu, ^{241}Pu) the binding energy of the last neutron of the next heavier isotope, $P_\infty = 10^{10}$ (e.g. ^{236}U for ^{235}U) exceeds E_{crit}. This fact then has a crucial implication, which we consider next. To make the problem specific, let us consider the case of ^{235}U which absorbs a low energy thermal neutron via the reaction

.

Note that this captured neutron can be considered to have fallen down a potential well that has a depth given by the neutron's binding energy, which is given simply as $P_\infty = 10^{10}$. Since this neutron is captured by the nucleus, the neutron itself does not acquire this amount of kinetic energy (otherwise it would not have been captured by the nucleus). Instead, this amount of energy is then transferred into the vibrational and rotational states of the newly formed nucleus. The nucleus is then said to be in an excited state, which is often denoted by representing the nucleus as .

To understand what happens next, let us recall that it takes an amount of energy to split the nucleus into two pieces. For a few select atomic nuclei, it turns out that . Thus, when these types of nuclei absorb a low energy neutron, they immediately split apart and thus are said to be *fissile*. Known fissile isotopes include ^{235}U, ^{233}U, ^{239}Pu and ^{241}Pu.

The fission of an atomic nucleus results in the production of various fission products. The typical probability distribution of the mass of fission products produced by fast neutrons and from thermal neutrons is shown in Figure below. These fission products are important for two primary reasons: first, they usually become trapped within the volume of the reactor and often have significant neutron absorption cross sections. Thus, they can act as a volumetric sink for neutrons in the reactor core, making it more difficult to maintain steady-state of reactor operation (in the parlance of nuclear engineering they can "poison" the reactor core), and second, these fission are often

radioactive and thus must be isolated from the environment until their activity has decayed to acceptably low levels. These waste products can have a wide range of half-lives ranging from seconds to minutes all the way to many millennia, and thus can present a challenge to manage safely.

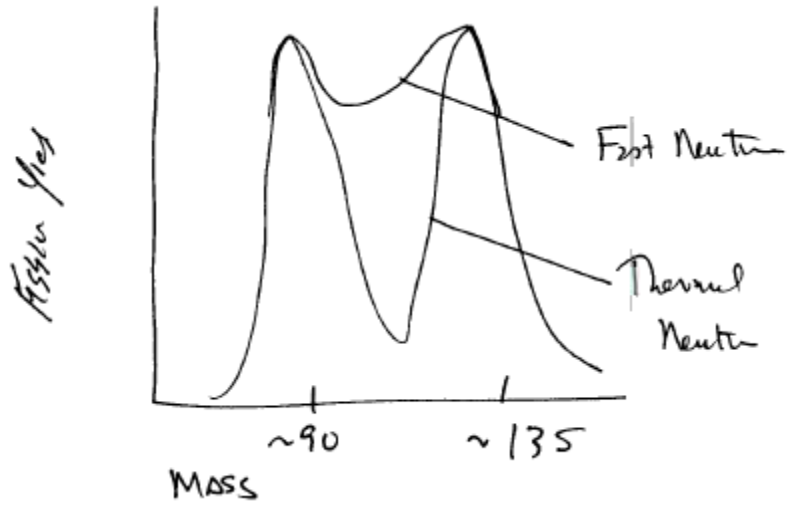


Figure 13.25: Probability distribution of fission product masses for both thermal and energetic incident neutrons.

In addition to emitting fission by-products, the fission event also emits two or occasionally three neutrons from the reaction. If these neutrons can then be used to cause subsequent fission events, then the possibility of a self-sustaining (or even growing) reaction can be realized. As we will see in later section, there are two classes of neutrons that are emitted by a fission event. The large majority (>99%) of the fission neutrons are emitted nearly instantaneously from the fission event. These neutrons are termed the "prompt" neutrons and play the dominant role in maintaining a reactor in operating condition. However, there is a small fraction (<1%) of the neutrons which are emitted with a measurable time delay after the fission event (actually there is a range of time delays, ranging from approximately seconds up to minutes, but for our purposes we shall

only consider a single class of such neutrons). This population is referred to as the "delayed" neutrons, and plays a critical role in the stability of a reactor to perturbations in the operating conditions.

It is convenient to define the number of neutrons released per fission event and the number of neutrons released in fission per neutron absorbed in fissile material as

$$\text{Average \# fission neutrons} \Big/ \text{fission} \equiv \nu$$

$$\frac{\text{\# neutrons released in fission}}{\text{\# neutrons absorbed in fission}} \equiv \eta$$

$$\eta = \nu \frac{\sigma_f}{\sigma_a} = \nu \frac{\sigma_f}{\sigma_\gamma + \sigma_f} = \frac{\nu}{1+}$$

$$\alpha \equiv \frac{\sigma_\gamma}{\sigma_f}$$

$$\eta = \frac{1}{\Sigma_a} \sum_i \nu_t(i) \Sigma_f(i)$$

Because the fission process produces more than two particles leaving the reaction, and these particles leave with a variety of exit angles, the prompt neutrons (i.e. those released at the moment of fission) have a distribution of energies. Experiments show that for fission by thermal neutrons, the energy distribution looks as shown in below. The average prompt fission neutron energy , while the maximum energy can lie in the range of 5-6 MeV.

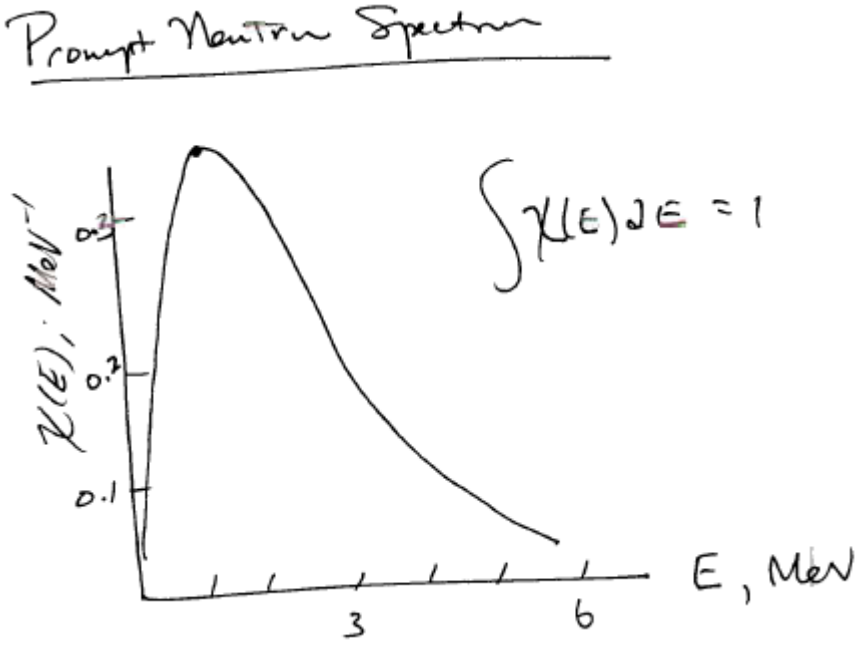


Figure 13.26: Prompt neutron energy spectrum for fission by thermal neutrons.

We saw that reaction rate per unit volume and unit time, F, is given as

for a single energy population of neutrons interacting with a target. This was also generalized to the case with a distribution of neutron energizes by writing

Now suppose that , i.e. the flux varies with position. This might occur if the neutron density, n, varies with position as shown schematically in below.

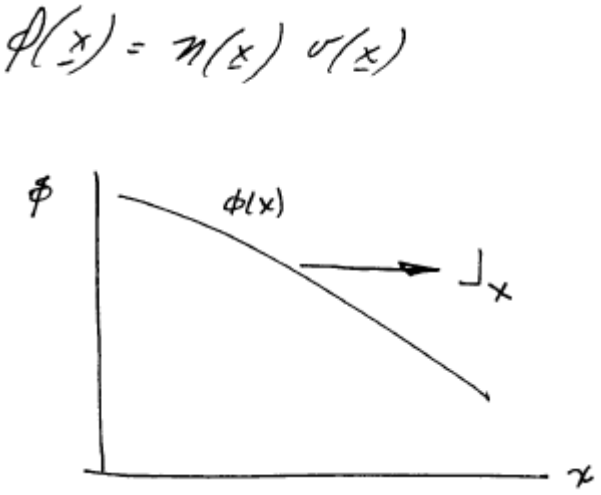


Figure 13.27: Schematic of spatially varying flux due e.g. to a spatially varying neutron density, n(x).

To a good approximation, neutrons in the core behave like a concentrate in solution in a chemical system. Thus, if n(x) is higher at some place and lower in another, then there will be a net flow of neutrons from , and we can say that a current J(x) of neutrons will flow between the two regions as shown in Figure above. Usually it is acceptable to assume that the current density is proportional to the gradient of the flux, i.e. for a one-dimensional problem we can write the current density as:

Examining the units we see that

More generally the flux is a function of the vector position, $\underline{x}$, i.e. and then the current density is given as

where is the neutron current density vector if is a unit vector pointing in an arbitrary direction, then .

In many cases, the diffusion coefficient is given approximately by the expression

Example

Suppose (flux at distances r from a point source in an infinite medium)

(a) Find an expression for J_r =

(b) Find the total # neutrons leaving a sphere of radius r?

$$\nabla_r = \frac{d}{dr}\vec{\imath} \text{ in spherical coordinates, then}$$

$$J_r = -D\nabla_r\phi = -D\frac{d}{dr}\left\{\delta_\circ \frac{e^{-r/L}}{4\pi Dr}\right\}\vec{\imath}$$

$$J_r = \vec{\imath}\frac{S}{4\pi}\left\{\frac{1}{r^2}+\frac{1}{rL}\right\}e^{-r/L}$$

b.

We note that this diffusion approximation is not exact and can, in some conditions, break down. In particular, it breaks down when there is a strong source or sink of neutrons, when the boundary of the reactor is approached, or when neutron scattering is strongly anisotropic. However, for the purposes of this text it will suffice and we shall make use of it from this point forward in our simplified analysis of the essential elements of operation of a fission reactor.

We can now use this diffusion approximation to write a neutron continuity equation, similar to that encountered in fluid mechanics. Considering the possibility that we can release and absorb neutrons from nuclear reactions, we can write a balance equation for the number of neutrons within a volume V. Just like the phenomenological derivation of the fluid continuity equation, this equation will need to consider the time rate of change of neutrons within an incremental volume element, the rate of neutron

production and absorption within the elements, and the rate of loss of neutrons from the boundary of the element. Let us consider each term in order.

1) total # neutrons in

   rate of change is then given

   if V=const, then can bring inside the integral to write (1) =

2) For now, let's denote production rate per unit volume of neutrons (we will later incorporate expressions for S to denote e.g. fission processes). Then the second term is given by

3) As has already been discussed earlier, we can write term 3 for the volumetric absorption rate of neutrons as (3) =

Finally, let us consider term 4, which denotes the loss of neutrons through a boundary. To make progress, consider the schematic representation of the volume V as shown in Figure below.

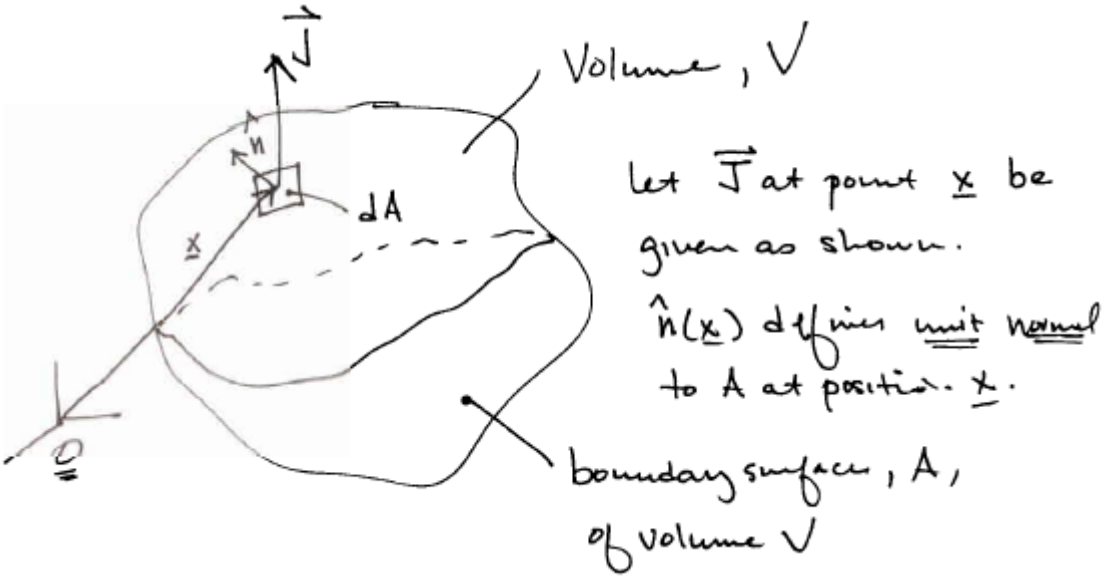


Figure 13.28: Schematic representation of a volume V which has neutrons escaping through the boundary surface A.

At some point x on the surface of the volume V, we have a neutron current J with projection to the local surface normal vector n, . The total # of particles through the incremental surface area dA=dAn is then given by

We can then find the total leakage rate R through the whole surface A by simply integrating over the entire surface, and thus write

R= .

Now recall the divergence theorem which states that for any vector field, *B*. Thus we can write

R= .

We can now write our *neutron continuity equation* as:

and since V is arbitrary can re-write this as a differential equation

.

This differential form of the neutron continuity equation will play a critical role in our subsequent analysis of neutron diffusion within a reactor volume as well as our analysis of reactor criticality and stability.

In order to gain further insight, let us consider first the case when = 0. The steady-state neutron continuity equation then becomes

.

In this text, we shall assume that the diffusion approximation is valid, i.e.

.We then substitute this result into the above the continuity equation to find:

This is the neutron diffusion equation, and can be used to solve for if S, D and are known and suitable boundary conditions are specified. If the diffusion coefficient is spatially uniform so that D = const. then the equation becomes:

it is common to re-write it in terms of :

and in steady state we have:

where . Note that in writing this expression, we have implicitly assumed that the neutrons have a single energy and speed.

Boundary Conditions

The diffusion equation is a 2nd order partial differential equation and thus requires boundary conditions in order to find a solution. There are several which can be applied

here. First, because the flux is associated with the motion of neutrons within a physical system, we require

Second, as was noted earlier, the diffusion approximation breaks down near the surfaces. Referring to

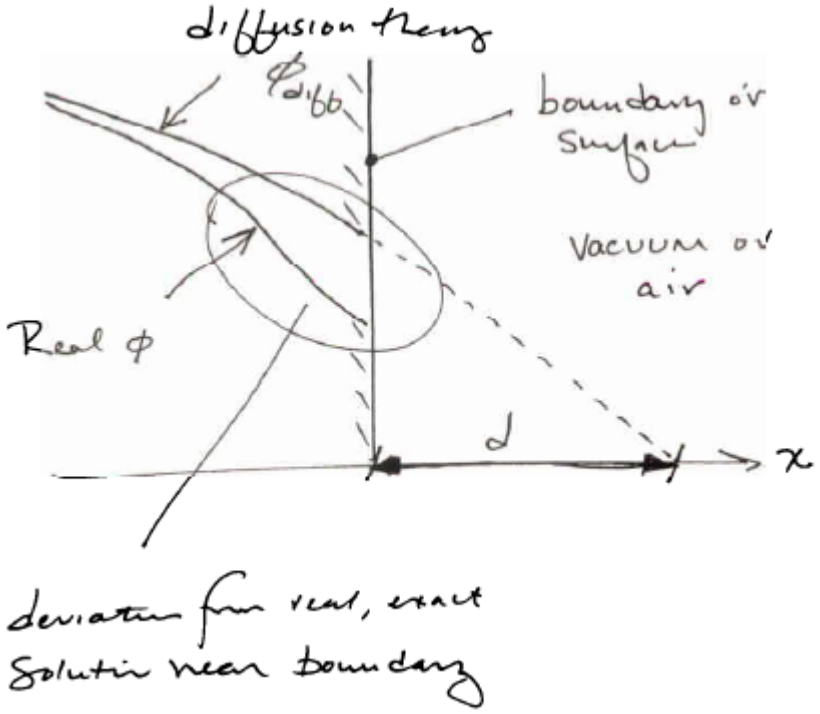


Figure 13.29: Actual flux, f , and flux computed from the solution to the diffusion equation, fdiff, near the boundary of a reactor volume. Near the boundary, the actual and diffusion equation solutions deviate. The diffusion solution vanishes at a distance d outside of the actual system boundary.

Can consider at an extrapolated boundary which lies a distance *d* beyond the actual boundary. Comparison with more complex exact solutions shows that approximately

This distance d for most solid materials is less than or approximate to 1cm and thus as long as $d << L_{sys}$ where L_{sys} is the size of the reactor core then *d* is negligible and is it common to then take at the system boundary and only incur a small error by doing so.

The third type of boundary condition applies if we have an interface between 2 types of materials, A and B. In this case, then and J at the interface must be continuous, i.e.

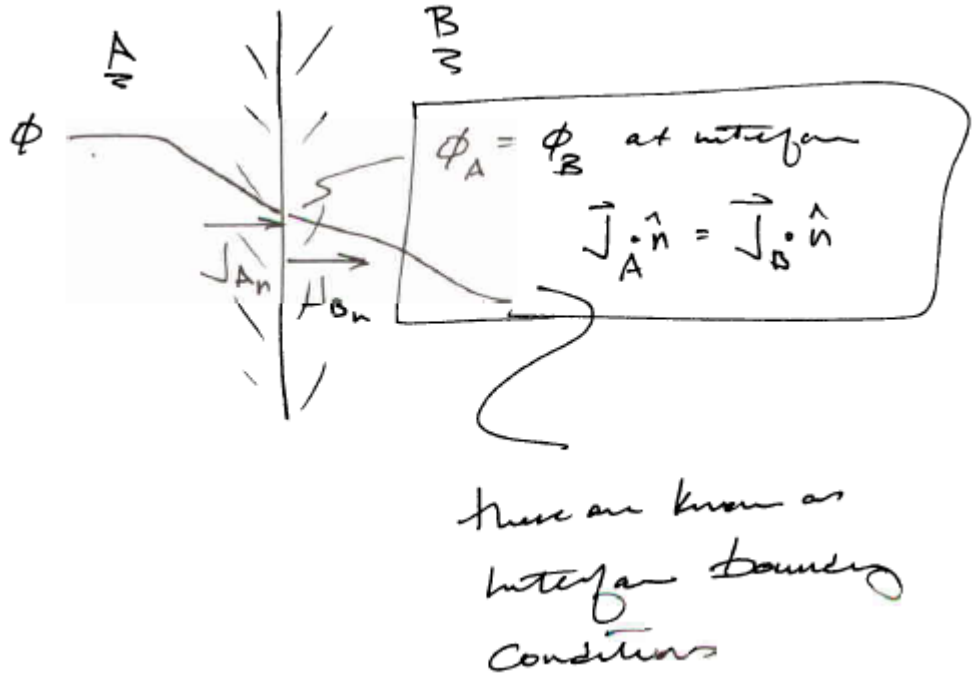


Figure 13.30: Boundary conditions at an interface between two distinct regions, A and B. The flux and the current density must be continuous across the interface.

We can gain further insight into the neutron diffusion equation by noting an analogy with another physical problem that students might be more familiar with – namely the equations that govern heat conduction within a solid material. The table below shows the analogous terms

Heat Conduction	Neutron Diffusion
Temperature T	Flux

heat flux, neutron current

$$\vec{q} = -K\nabla T$$

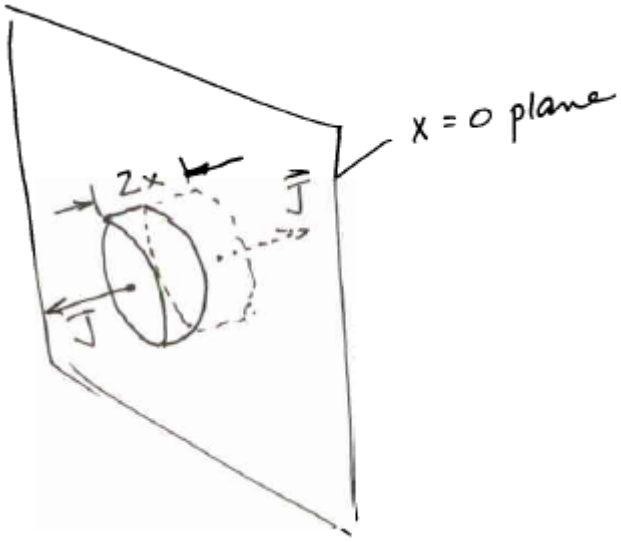
To gain insight into the behavior of this neutron diffusion equation, let us consider an infinite planar source that emits S neutrons per second at position x=0 which then diffuse into an infinite medium, and we wish to find the neutron flux everywhere. To begin, we note that this is a 1D planar problem that is symmetric about x=0 where source is located.

At this point, let us consider the x>0 region. For this simple geometry, the diffusion equation is given as:

with the general solution and boundary conditions given as

$$\phi = Ae^{-x/L} + Be^{+x/L}$$
$$B.C.'s:$$
$$\phi \to 0 \text{ as } x \to \infty \to B = 0$$
$$\therefore \phi = Ae^{-x/L}$$

To determine the constant A, consider a Gausian pillbox centered on the source:



The net flux through the surface of the pill box is 2J(x). Letting the thickness of the box vanish, i.e. taking x→0 gives . From the diffusion approximation for the current density J, we can write J in terms of the flux as:

Thus we find:

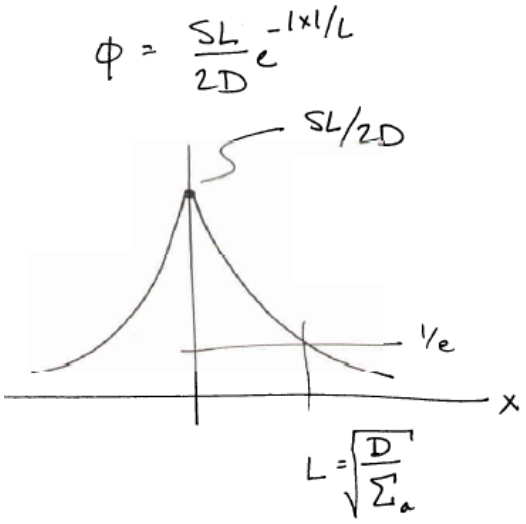


Figure 13.31: Solution to neutron flux emitted by an infinite planar source S located at x=0, diffusing through an infinite medium.

We consider a second simple problem, this time of a point source located at r=0, emitting S=constant neutrons/second which then diffuse into an infinite medium. Because of the spherical symmetry we use spherical coordinates. The diffusion equation is then given as:

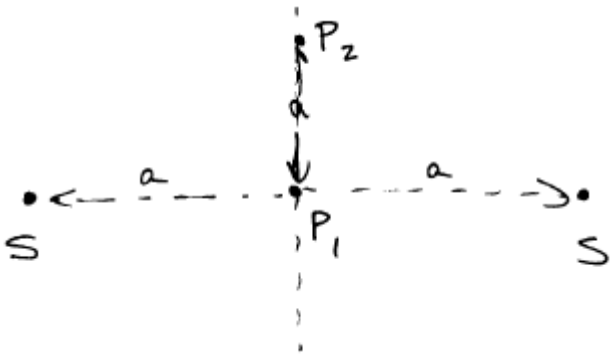
for r>0. Considering a small sphere of radius r surrounding the source, we can write the current density at this position as .

Let us define . The differential equation then becomes:

Using the boundary conditions at the source we then find:

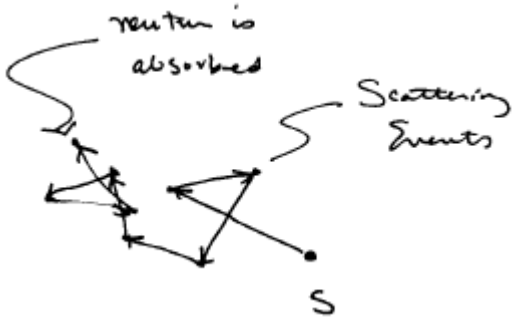
Note the fall off is quite different than in a vacuum (which has)

Note that the diffusion equation is linear and therefore we can superimpose solutions. For example, consider the flux from two point sources located in an infinite medium as shown below.



We wish to find the flux at P_1. Noting that we can superimpose solutions, we can then note that at P_1 is simply the summation for each of the two point sources and thus we find

Since we have now introduced the neutron diffusion equation, and have examined a few elementary solutions to it, let us consider the physical interpretation of the diffusion length, L, which is defined as . The figure below shows schematically the actual path followed of one particular neutron emitted from a point source into an infinite medium which then acts to scatter the neutron.



We know that the number of neutrons absorbed per second at a distance (r, r+dr) from the source is given as

and thus

Now, since S neutrons per second are emitted, and *du* are absorbed between (r, r+dr), it follows that the probability that a source neutrons emitted at r=0 is then absorbed within the volume (r, r+dr) is given as:

i.e. is the probability that a neutron emitted at r=0 is absorbed in within a shell bounded by (r, r+dr).

Can calculate mean quantities, e.g.

This gives the physical interpretation for L:

The discussion above applies to neutrons that have a single speed. However, in any realistic system, neutrons will have a distribution of speeds as has already been discussed. For a system where the neutrons can be described accurately by the Maxwellian thermal distribution, we can generalize the above results and arrive at a very

similar equation. Let us consider a population of thermal neutrons with a temperature $T_{neutron}=T$. Then we know that we have a Maxwellian energy distribution given as

,

where *n* denotes the number of neutrons per unit volume and n(E)de gives the number of neutrons per unit volume w/ energy (E, E + dE).

We can now average each of the terms in the single speed neutron diffusion equation over n(E) to find the corresponding thermally averaged quantities:

if we define , we then have . Similarly we can write the averaged absorption cross-section as

In the above, the integral equals , where is the "2200 m/s flux" seen earlier. denotes the absorption cross

section at a reference energy that is usually taken to correspond to room temperature. We can then write the thermal absorption as

Taking these results together, we can now write the so-called one-group steady state diffusion equation for thermal neutrons:

Which can be re-arranged to give

We note that the equation has the same form as single speed neutron diffusion equation, but all quantities have now been averaged over n(E) and thus can be applied to a thermal reactor.

For most of the analysis in this text, we shall use the thermal diffusion equation described above. However, in some problems in nuclear fission, it is necessary to keep track of the varying neutron energies in the system. In such cases, the diffusion equation in then used, but is modified to keep track of the various ranges of neutron kinetic energy. Neutron energy is then broken out into different groups, with the first group corresponding to the highest energy neutrons in the system (i.e. those that still have their original high energy from nuclear reactions), the second, third and higher groups then corresponding to lower energy bands the correspond to neutrons in the so-called slowing-

down part of the energy spectrum and then finally the highest group, N, which corresponds to thermal energies in the system. A schematic of this labeling scheme is shown in Figure 13.32 below.

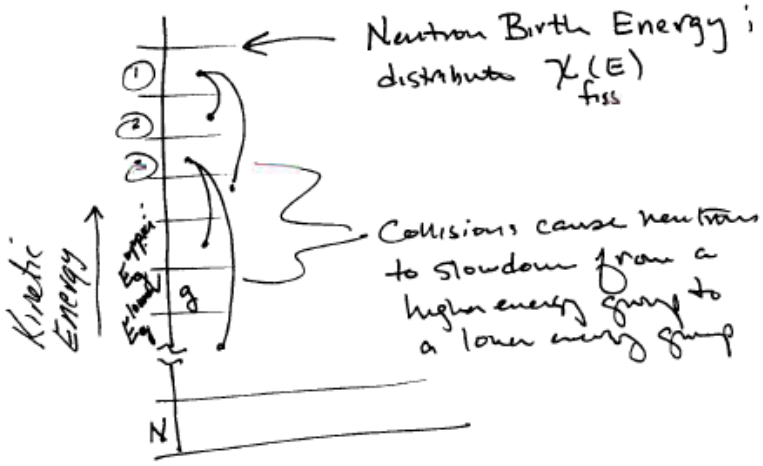


Figure 13.32: Schematic of neutron energy groups used in a multi-group description of neutron diffusion.

The multi-group diffusion equation can then be found from the single speed diffusion equation by integrating each term in the equation over a particular energy range. Thus for example, the neutron flux in a particular energy group, g, is given as

The absorption Rate for g^{th} group =

Defining the absorption cross section as

then gives the absorption rate as =

The rate at which neutrons scatter from the *g-th* group into the *h* group due to collisions is then given by:

[here we are assuming E>>kT so that neutrons only lose energy and do not gain energy in a collision].

The total rate of loss of neutrons from the *g-th* group into all lower energy bands is given as:

Likewise, neutrons are transferred via collisions from higher energies into group *g*. The expression for this rate is given as .

Combining all of these terms gives the multi-group diffusion equation for neutrons

This somewhat more complex treatment of neutron diffusion is sometimes useful because many cross-sections can vary strongly (i.e. by many orders of magnitude) with energy, and thus we need to keep track of neutron energy to some amount. In this text we only introduce the simplest possible version of such a model – the so-called two-group model for neutron diffusion.

Simplest approach: 2-Group Model

Group 1: E > 5kT "Fast Neutrons"

Group 2 (Thermal): E < 5kT "Thermal Neutrons"

Equations:

Fast_____

Thermal_____

Consider a power source of strength S_1 for Group 1; $S_2 = 0$. Source is at r = 0 in infinite medium.

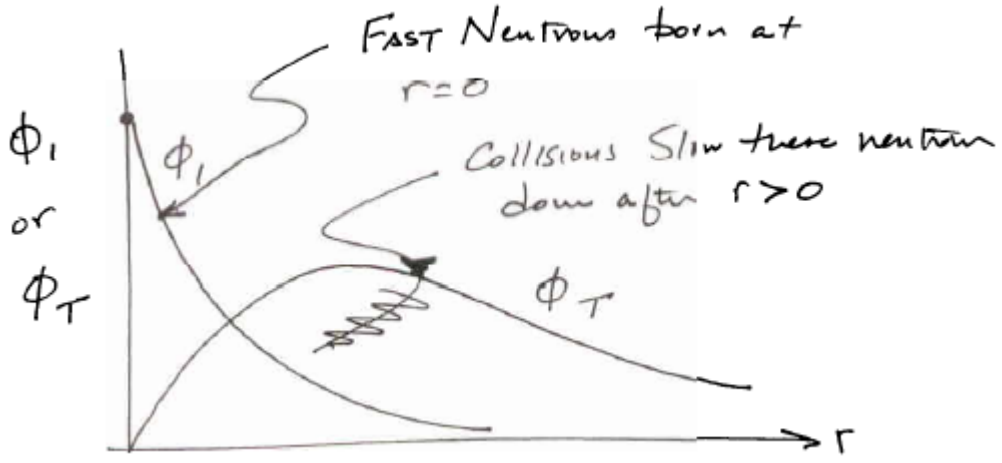
Then have

for r > 0

where

Solution:

Use this in equation for → linear inhomogeneous ODE. Solution is:



Our diffusion equation model for the neutron transport contains the source term, S, which drives the ordinary differential equation. Mathematically this source turns the diffusion equation into an inhomogenous ordinary differential equation, and requires a solution which satisfies both this equation and the appropriate boundary conditions.

For the analysis of a nuclear fission reactor, we are interested in source terms which are driven by fission processes. If we denote the average number of neutrons emitted by a fission event as , then we can write the fission source term as

for a single-speed population of neutrons (we will comment briefly on multiple speed neutron effects later; a detailed discussion of this important topic lies beyond the scope of this text). Here denotes the macroscopic fission cross-section and is the neutron flux at position x. With this form for S, the one-speed diffusion equation is then given as

.

This equation becomes the key equation which will then guide our further analysis of fission reactor behavior.

Examining this equation, we note that it expresses a balance between the neutron source term from fission events against the absorption of neutrons within the volume of the reactor and diffusion of neutrons out of the volume in question. Any imbalance

between these terms then leads to a time-dependent neutron flux which, since we are only considering a single speed of neutrons, corresponds to a change in neutron density at a particular location.

In the analysis of such systems, it is common to introduce an adjustable factor, k, into the steady-state neutron diffusion equation which is then written as

.

This term can be viewed in the following sense. For a given reactor configuration there will be values for the diffusion coefficient and for the absorption and fission cross-sections. The system geometry will then set the boundary conditions, flux and its gradients. Together, these terms will in general give a growing, decaying or steady-state neutron flux. When used in the above equation, the value for k will then be determined. If , this implies that in the time-dependent diffusion equation the fission source term is larger (smaller) than the other terms, and thus we will have a temporally growing (decaying) neutron flux. If k=1 then the system will be in steady-state.

It is also common to define the so-called "geometric buckling" factor, B as

which then allows us to write the steady-state neutron diffusion equation as

.

This is simply an eigenvalue problem, where B^2 denotes the eigenvalue and denotes the eigenmode. Using this result in the one-speed steady-state neutron diffusion equation then allows us to write

we can then solve this equation for the factor k:

.

The key point to keep in mind is the fact that if k=1, then we have a steady-state solution for the neutron diffusion equation. Examining this expression for k, we see that it is determined by the eigenvalue, B which is in turn determined by the reactor geometry, and by constants that are properties of the reactor composition alone. In particular, for a given macroscopic fission and absorption cross sections, there must be a minimum eigenvalue, B, which gives k=1. Since B is determined by the solution to a second order homogenous linear differential equation, there must then be a minimum system size which will then give k=1. Alternatively, for a given system size there must be a combination of fission and absorption cross-sections that gives k=1. The problem facing us is to then determine quantitatively how to provide for a steady-state solution to the neutron diffusion equation, which will then allow a time-invarient neutron flux to be maintained in the reactor.

It is common to see the source term re-written in terms of the fuel absorption cross-section, , i.e.

.

Here the absorption cross-section is computed for a homogenous mixture of fuel, moderator and coolant with atomic number density N and microscopic absorption cross section , i.e.

and the factor f denotes the fraction given by the ratio of neutrons absorbed in fuel to neutrons absorbed in the fuel, moderator and coolant.

We can gain further insight by considering the (idealized) case of an infinitely large uniformly homogenous reactor core in which the neutron flux becomes spatially uniform. In this case, and therefore and thus we can write for this special case

where the subscript denotes the case of the infinite reactor. Thus, we find simply for this case that . If the composition of this hypothetical infinite reactor core is arranged such that we have , then the source and absorption are balanced in the

infinite reactor, and thus a steady-state condition with finite flux, , everywhere can be maintained.

Let us now apply this concept back to a finite geometry reactor. First, we write the neutron fission source as

and write the steady-state neutron diffusion equation as

which can then be solved for $B^2$to give the so-called bare reactor critical equation:

where we have defined the characteristic length .

There are two essential points to keep in mind at this point. First, the eigenvalue B, and thus the left side of the bare reactor critical equation is determined entirely by the geometry of the system since it satisfies a second order eigenvalue problem. Second, the right hand side of the bare reactor critical equation is determined entirely by the composition the reactor core, which in this simplified analysis is assumed to be composed of a homogenous (or uniform) mixture of fuel, moderator and coolant. If B>0 as it must be for any finite geometry reactor, and L>0, then maintaining a steady-state in the system will then require that . As the reactor system becomes very large in size, then B->0 and then steady state then requires that .

To make these concepts more concrete, let us consider the simplest possible finite geometry reactor: the so-called slab reactor. Here, we assume that the system is infinite in the y and z directions, but is finite in the x direction. Below is a schematic of the system geometry.

INSERT FIGURE HERE

Figure 13.33: Schematic of a one-dimensional slab reactor geometry

For this system, the eigenvalue equation is then given as

with boundary conditions that and . Also we can argue that the symmetry of the problem requires $P(t) = \frac{dQ(t)}{dt}$ and $P(t) = \frac{dQ(t)}{dt}$, i.e. the flux is even.

The general solution is easily seen to be given as

$$Q(t) = \int_{t'=0}^{t} P(t')dt'$$

from the symmetry we can see immediately that C=0. Applying the boundary conditions at –a/2 and +a/2 then forces the eigenvalue to satisfy $Q(t) = \int_{t'=0}^{t} P(t')dt'$ where n=1,3,5… The lowest order, or fundamental, solution is then given as

and

$$Q_{res} = Q_0 - Q(t) = Q_0 - \int_{t'=0}^{t} P(t')dt'.$$

Note at this point the constant A is still undetermined and, without additional information, is arbitrary (note that since the diffusion equation is linear with respect to the neutron flux this doesn't impact the solution). To find the magnitude of the flux and thus the constant A we must introduce a new piece of information – the power density of the reactor.

Suppose that each fission event releases an energy E_R. Then for this slab reactor, the power per unit cross-sectional area of the slab is given as

$$\frac{Q(t)}{Q_0} \approx 0.$$

Using the above result for $\frac{Q(t)}{Q_0} \approx 0$ then allows us to solve for the constant A which is given as

$$q(t) = \frac{Q(t)}{Q_0}.$$

A similar analysis can be performed for other simple shapes such as infinite cylinder and a sphere. Using a separation of variables, it can also be applied to more

complex (and realistic) shapes such as a finite height cylinder, cubic shapes and so forth. Again these are left as exercises.

Let us summarize our findings up to this point. We found that the one-speed diffusion equation could be reduced to the expression

which when k=1 gives a steady-state non-trivial solution for the neutron flux. This can then be solved for the critical buckling, B_C, needed to provide this condition:

$$q(t) = \frac{Q(t)}{Q_o}$$

This can also be re-written as

$$\frac{dq}{dt} = rq(t)\left(1 - q(t)\right)$$

or

$$\frac{dq}{dt} = rq(t)\left(1 - q(t)\right).$$

For this steady-state critical condition to be maintained in a finite geometry reactor, then this critical buckling must be equal to the lowest order eigenvalue to the diffusion equation, $\frac{1}{1+\exp(-r(t-t_0))}$ with $q(t_0)=0.5$, i.e. we must have $\frac{1}{1+\exp(-r(t-t_0))}$ with $q(t_0)=0.5$. Since $P(t)=\frac{dQ(t)}{dt}=Q_o\frac{dq(t)}{dt} - rQ_o\frac{\exp(-r(t-t_0))}{\left[1+\exp(-r(t-t_0))\right]^2}$ for a finite geometry reactor, we then require for critical steady-state operation that $P(t)=\frac{dQ(t)}{dt}=Q_o\frac{dq(t)}{dt} - rQ_o\frac{\exp(-r(t-t_0))}{\left[1+\exp(-r(t-t_0))\right]^2}$. From the

definition, we then can write finally that finite reactors require for steady-state operation. We can therefore conclude that for a given reactor core geometry and macroscopic absorption cross-section, there must exist a minimum fissile fuel nuclei density that will allow critical core operation. The corresponding mass of fissile material is then known as the critical mass.

Reactor Stability

The foregoing discussion and analysis was focused on determining if a reactor system has achieved a critical state – i.e. one in which the nuclear reactions are self-sustaining at a desired power level. However, the analysis says nothing about the stability of such a system – i.e. if the system is disturbed away from the critical state will the system then move back towards a stable operating point, or will the reactions either grow or die away in time. This issue of reactor stability is important for both safety and operational considerations, and is the subject of our discussion here. The fundamental question which we seek to address is: how does a fission reactor respond to a disturbance away from an operational equilibrium?

We first wish to define the time-scales that we are interested in. Let us define a timescale such that is the time needed for the neutron flux to change in a significant amount (the precise amount of the change is to still be determined). There are other system timescales which can also be defined, e.g. one can define the fuel burn-up timescale given as

which describes the time needed to burn-up a significant quantity of fissile material and modify the neutron absorption by the creation of fission byproducts within the reactor core. Another important timescale is the thermal transient timescale, defined as

which describes the time needed for the reactor core temperature, T, to change in response to a change in operational power level.

With these timescales in mind, we now consider long timescale transient effects, which are those in which the inequalities $I(\omega)$ hold. For these types of problems, one can then view the system as being at an equilibrium at any point in time, and then simply apply the time-independent analysis which was developed above to the particular conditions of the core that are of interest. This analysis is not fundamentally different than the previous discussion.

There are a second class of transient phenomena that are also of interest – namely fast transients in which the ordering of the inequality is reversed, i.e. in which $I(\omega)$. Furthermore, we shall further assume that the flux timescale is such that the shape of the neutron flux within the reactor core doesn't change. In this case, the core can be viewed as a mathematical point, and the analysis then carried out in only one dimension – time. This approach is the subject of reactor kinetics which we take up here.

Prompt and Delayed Neutrons

When a fissile nucleus absorbs a neutron, most of the time the resulting compound nucleus undergoes an immediate fission event, and emits two or more prompt neutrons. On rare occasions, the compound nucleus does not immediately undergo a fission event, but instead stays in an excited state for a period ranging from several milliseconds up to many tens of seconds of time. Eventually, the nucleus then undergoes a fission event and releases two or more delayed neutrons.

Let us first take a the case of a reactor which is maintained in a critical state using only prompt neutrons, and determine the stability of such a reactor to disturbances in the rate of fission events. For simplicity, we consider only an infinitely large reactor in which there are no diffusive losses. Let us denote the average time between the birth or emission of a prompt neutron and the eventual absorption of that neutron within the reactor as the prompt neutron lifetime, . This timescale is determined by two distinct processes that occur within the reactor: neutron slowing down from fission birth energies (typically in the MeV energy range) down to thermal energies (less than 1eV energy). Let us denote this timescale as , and neutron diffusion time, ,which is the time taken for neutrons to diffuse out of the reactor core. In a finite size core, typically we have ; this inequality certainly holds for our simplified analysis since we have assumed that the core is so large that diffusion is negligibly small compared to

absorption. Thus, the neutron lifetime is dominated by the slowing down time, i.e. we can write approximately that $l_p \approx t_{slow}$.

We have already seen that on average a neutron with energy E travels a distance $\lambda_a(E) = 1/\Sigma_a(E)$ before being absorbed, and the corresponding time interval t(E) is given as $t_a(E) = \lambda_a(E)/v(E)$. For most absorbing materials that are used in reactor cores, the absorption cross section satisfies the expression

$$\Sigma_a(E) = \Sigma_a(E_0)\frac{v(E_0)}{v(E)}$$

where E_0 is room temperature which corresponds to a neutron thermal speed of ~2200 m/sec, we can then find that $t_a(E) = 1/\Sigma_a(E_0)v(E_0) \approx constant$, independent of the neutron energy. Averaging this expression over the thermal neutron energy distribution then gives the average thermal neutron lifetime, t_d, given as $t_d \frac{\sqrt{\pi}}{2\bar{\Sigma}_a v_{th}}$ where $\bar{\Sigma}_a$ and v_{th} denote the neutron absorption cross-section averaged over the thermal neutron distribution and thermal neutron speed respectively. In reactors with light mass moderators (i.e. water, Be, etc…) typically t_d~0.1 – few milliseconds or so. For heavier moderators (e.g. graphite, Na, etc…) the value is typically 10's milliseconds.

In the case of an infinite reactor core, we can neglect the $D\nabla^2\phi$ term in the neutron diffusion equation. If such a system is operating at or near the critical condition, i.e. if $k_\infty \approx 1$, then by definition neutron production and absorption are very carefully balanced in the core. Thus, it stands to reason that the prompt neutron lifetime, l_p, must

be very nearly equal to the prompt neutron generation time, Λ, which is given by the time between the neutron birth from the n-th generation of fission events, and the subsequent neutron absorption in a fissile nucleus which leads to the emission of neutrons in n+1-th generation of fission events (another way to look at this generation time is that it is equal to the time needed for a neutron to be emitted from fission, slow down in the reactor, diffuse around in the core, and then be absorbed either in the moderator or fuel nuclei). Writing the number of fission events per unit volume and unit time at time t as $N_F(t)$, we can then write

$$N_F(t+\Lambda) \approx N_F(t+l_p) = k_\infty N_F(t)$$

If we Taylor expand $N_F(t+l_p)$ it is easy to show that

$$\frac{dN_F(t)}{dt} = \frac{k_\infty - 1}{l_p} N_F(t).$$

This expression can be re-arranged and then integrated over the range (0,t). The final result is then the solution

$$N_F(t) = N_F(0) exp(t/T)$$
$$where$$
$$T = \frac{l_p}{k_\infty - 1}$$

is known as the reactor period in the absence of delayed neutrons. Now, typically $l_p \sim 1-10 msec$. Thus, if such a reactor had a disturbance such that $k_\infty = 1 \Rightarrow k_\infty = 1.01$ then the corresponding reactor period would lie in the range T~0.1-1 second. Thus, the reactor flux (and power release(!)) would grow exponentially on this timescale. Clearly,

such a system would be very difficult to control and would in fact not even be licensed for operation anywhere in the world. Thus, we find a key result: a reactor operating on prompt neutrons alone is very difficult to control and is unstable.

Effect of Delayed Neutrons

Most neutron emission within the core of a fission reactor occurs by the prompt emission of neutrons from fission events. However, a small fraction of neutron emission is delayed with respect to the formation of an unstable compound nucleus. Neutrons emitted in this manner are referred to as delayed neutrons, and they can have a significant effect on the dynamic response of a reactor core to operating transients. Thus, we provide an analysis of this important effect here.

Again, we consider an infinite homogenous reactor core which allows us to neglect changes in the spatial distribution of neutron flux and thus we can treat the reactor as a point. In this case, the thermal neutron conservation equation becomes

$$S_T - \overline{\Sigma}_a \phi_T = \frac{\sqrt{\pi}}{2v_T} \frac{d\phi_T}{dt}$$

where we note that we have dropped the diffusion term due to our assumption of an infinitely sized reactor. This can be re-written in terms of the neutron lifetime by dividing through by the absorption cross-section

$$\frac{S_T}{\overline{\Sigma}_a} - \phi_T = l_p \frac{d\phi_T}{dt}.$$

Let us now carefully consider the neutron source term. We allow that a fraction β of the neutrons being emitted per unit volume and unit time are delayed neutrons. This fraction is bounded by $0 < \beta < 1$. Then the prompt neutron source is then given by

$$S_T\big|_{prompt} = (1-\beta)k_\infty \overline{\Sigma}_a \phi_T.$$

However, we must also allow for the fact that there is a source of delayed neutrons. These neutrons are released by the radioactive decay of a precursor nucleus following the processes introduced early in this chapter. Let us denote the precursor density as C and the decay rate of these precursors as λ. Thus, the delayed neutron source rate is given as

$$S_T\big|_{delayed} = \lambda C(t).$$

The total source rate is then given by the sum of these two source terms. We can thus write the time-dependent infinite reactor diffusion equation including delayed neutron effects as

$$(1-\beta)k_\infty \overline{\Sigma}_a \phi_T + \lambda C - \overline{\Sigma}_a \phi_T = l_p \frac{d\phi_T}{dt}.$$

Of course, to solve this equation we need to find C(t) and thus we need a simple model for the production of delayed neutron precursors.

We note that in equilibrium the rate of production of delayed neutrons can be equated to the rate of production of the precursor nuclei C(t), and that in equilibrium this quantity must be equal to the decay rate of C(t). Thus, we can write a balance equation for the delayed neutron precursor nuclei as

$$\frac{dC}{dt} = \beta k_\infty \bar{\Sigma}_a \phi_T - \lambda C.$$

These two equations form a closed system and can thus be solved.

The solution can be found as follows. We take

$$C(t) = C_0 exp(\omega t)$$

and

$$\phi_T(t) = A exp(\omega t).$$

It is then straightforward to show that

$$C_0 = \frac{\beta k_\infty \bar{\Sigma}_a A}{\omega + \lambda}.$$

We then use this solution and the assumed form for the flux to write

$$(1-\beta) k_\infty \bar{\Sigma}_a A exp(\omega t) + \lambda C_0 exp(\omega t) - \bar{\Sigma}_a A exp(\omega t) = l_p \omega A exp(\omega t).$$

Re-arranging gives

$$k_\infty (1-\beta) - \frac{k_\infty \lambda \beta}{\omega + \lambda} = l_p \omega$$

or finally

$$\frac{k_\infty - 1}{k_\infty} = \frac{l_p \omega}{1 + l_p \omega} + \frac{\omega}{1 + l_p \omega} \frac{\beta}{\omega + \lambda}.$$

The left hand side of this equation is often defined as the reactivity, $\rho \equiv \frac{k_\infty - 1}{k_\infty}$, which determines how far a particular disturbance moves the system away from equilibrium. The equation is second order in ω and thus in general two solutions will exist such that

$$\phi(t) = A_1 exp(\omega_1 t) + A_2 exp(\omega_2 t).$$

It is possible to show that one root, denoted here as the ω_2 root, always has $\omega_2<0$ and thus gives a rapidly decaying disturbance, while the second root ω_1 can give either a growing or decaying solution depending on the sign of the reactivity. Furthermore, these roots can be shown to be bounded by the limits $-\frac{1}{l_p} < \omega_2 < -\lambda$ for any value of ρ and $-\lambda < \omega_1 < 0$ for $\rho<0$, while for $\rho>0$ it is possible to show that $\omega_1 > 0$.

For small values of reactivity such that $\omega l_p << 1; \omega << \lambda$ we can then write that $\rho \approx \omega\left(l_p + \frac{\beta}{\lambda}\right)$. Solving for ωin this limit gives

$$\omega \approx \frac{\rho}{\left(l_p + \frac{\beta}{\lambda}\right)} = \frac{k_\infty - 1}{k_\infty\left(l_p + \frac{\beta}{\lambda}\right)}.$$

Note that if we have no delayed neutron population then $\beta \to 0$ and we recover the earlier prompt neutron response found above. A positive reactivity will then give a solution that grows in time, while a negative reactivity will give a solution that decays in time.

If delayed neutrons dominate the dynamics of the system, then $\frac{\beta}{\lambda} >> l_p$ and we will then have

$$\omega \approx \frac{\lambda}{\beta}\frac{k_\infty - 1}{k_\infty} << \frac{1}{l_p}\frac{k_\infty - 1}{k_\infty}$$

i.e. a reactor whose rate of change is much slower than a reactor that is dominated by prompt neutron analysis. Typically, for a reactor operating on fissile uranium, $\beta \sim 10^{-3}$ and $\frac{1}{\lambda} \sim 10's - 100's$ seconds while $l_p \sim 1-10$ milliseconds. Thus, such a system would respond to changes in reactivity on time scales of minutes or longer.

Changes in the reactivity are usually made by introducing (or removing) materials from the reactor core which absorb neutrons. In our homogenous model of the reactor core, this would correspond to a change in the density of the absorbing material, which then changes the macroscopic absorption cross-section which, in turn, induces a change in k_∞. These results have very important implications for the operation of fission reactors. To see the essential physics behind this statement, let us consider a scenario in which a change is made in the reactivity on a time scale that is rapid compared to the time-scale for changes in the delayed neutron precursor density, C, i.e. we assume that the reactivity change occurs on a timescale $\varepsilon << \frac{1}{\lambda}$. In this case, C will be constant across the reactivity change and,for $k_\infty = 1$ before the reactivity change, will be given by $C = \frac{\beta \Sigma_a \phi_{T_0}}{\lambda}$ where here ϕ_{T_0} denotes the neutron flux in the system before the change in reactivity. Over the short timescale then, this value for C can be used in the time-dependent infinite reactor diffusion equation to write

$$l_p \frac{d\phi}{dt} = \left[(1-\beta)k_\infty^+ - 1\right]\phi + \beta\phi_0$$

where here k_∞^+ denotes the condition just after the rapid change in reactivity. The solution is then given as

$$\phi(t) = \phi_0\, exp(t/T) + \frac{\beta\phi_0}{1-(1-\beta)k_\infty^+}\left[1 - exp(t/T)\right]$$

where

$$T = \frac{l_p}{(1-\beta)k_\infty^+ - 1} \approx \frac{l_p}{k_\infty^+ - 1}.$$

As long as $(1-\beta)k_\infty^+ < 1$ (i.e. as long as we do not have a prompt critical change in reactivity such that the reactor would become critical on prompt neutrons alone, leading to a very short reactor period), then T<0 and the perturbation would then decay away on the timescale T.

NEED TO CHECK ABOVE AND THEN ADD TWO CASES:

A) SMALL POSITIVE REACTIVITY GIVES SLOWLY GROWING SOLUTION

B) LARGE NEGATIVE REACTIVITIES. GIVES TWO TIMESCALES: RAPID DECAY AND THEN A SLOW DECAY ➔ IMPORTANT IMPLICATION FOR SHUTDOWN OF REACTOR!

THEN NEED TO ADD DISCUSSION ABOUT THERMAL STABILITY OF REACTOR CORE, RESONANCE ESCAPE PROBABILITY, ETC…

ALSO FEW COMMENTS ABOUT REACTOR DESIGNS, FUEL CYCLES AND WASTE DISPOSAL ISSUES.

Chapter 14: Transitioning to a Carbon-free Global Energy Economy

Assuming that one or more of the climate neutral energy technologies discussed above becomes technologically and economically feasible, several questions naturally emerge. These include:

1) Can we estimate how quickly these technologies will move into the market place?
2) If so, then when might the technology become a significant contributor to world energy demand?
3) Given the anticipated rate of adoption of new climate neutral energy technologies, what is the likely impact upon carbon emission trajectories?
4) Are additional approaches needed to then adequately address the C emission problem?

In this chapter, we take up these questions by first introducing an existing model of technology substitution and of learning, and then applying it to the carbon-free primary energy sources discussed in this book. This analysis addresses the first two questions posed above. When coupled with the climate and carbon balance models which have previously been summarized, we can then obtain some insight into the last two questions summarized above.

The Fischer-Pry Substitution Model of Technological Change

The penetration of a new product or technology into the market place was studied by Fischer & Pry [Fischer & Pry, Technological Forecasting and Social Change 3, pp. 75-88, (1971)]. In this work, the displacement of an established solution to a human need with a newer solution which, in some sense of the word is "better" than the older solution, was studied and a mathematical model of the substitution process was introduced. We summarize their model here, and then consider applying it to primary energy sources. Our discussion is based closely upon that introduced in the Fischer & Pry paper, and is motivated by the work of Marcetti and his colleagues at the IIASA [REFERENCE].

In their work on the subject of technological change in a market place, Fischer & Pry (1971) make three key assumptions:

1) Many technological advances can be considered as a competitive substitution of one technique or approach which satisfies a human need which up until that point had been met by some other approach or technique.
2) If the new technique or approach begins to acquire a few percent market fractions, then it will proceed until its substitution is "complete".
3) The fractional rate of fractional substitution of new for old is proportional to the remaining amount of the old left to be substituted.

If one examines the history of the adoption of a new technique that replaces a pre-existing technique, one finds that initially the market fraction that is met by the new approach exhibits exponential growth. Then, as the market penetration increases, the growth rate of the new then slows and eventually saturates. The process can then be reversed as the novel approach becomes widely accepted, and is then eventually overtaken by yet another new solution. In this manner, human needs are satisfied by the progress of different solutions in a competitive marketplace where choices among competitors can be made.

To make this discussion quantitative, let us first define the quantity f to denote the fraction of the available market for which the new has supplanted or replaced the old. Assumption 3 above can then be quantitatively written as a differential equation:

$$\frac{1}{f}\frac{df}{dt} = r_0(1-f)$$

where r_0 denotes the annual growth rate of f in the early portion of the new innovation's implementation in the marketplace. The effective growth rate at anytime t is then given by the left hand side of the equation and tends to decrease as f approaches unity. Thus, this model equation captures the essential features expected from the three postulates given above.

This equation has a solution for f(t) given by

$$f(t) = \left[1+\exp\left(-r_0\left(t-t_0\right)\right)\right]^{-1}$$

where t_0 denotes the time when f=0.5. Obviously, this model has the flaw that it assumes that f(t)>0 for infinite time in the past; however, application of the model to real data for many different types of markets and technologies shows that, once f becomes appreciable (say f>0.01 or so) then the model works reasonably well for finite times. As we will discuss later in this chapter, there are clearly hurdles (sometimes quite high) that must be overcome for f to even reach a value f=1%. But for the present time, let us assume that a new product reaches the necessary critical stage of development that this model then holds. **Error! Reference source not found.** illustrates this solution.

We can gain insight by considering a few characteristics of the model. It is convenient to define a "take over time", Δt, defined as the time required for f to go from f=0.1 to f=0.9. The corresponding data points are shown as + symbols in the figures above. This time is obviously inversely proportional to the early growth rate, r_0, and has an approximate value given by

$$\Delta t \equiv t_{f=0.9} - t_{f=0.1} \; ; \; \frac{4.4}{r_0}.$$

which can be seen by direct substitution into the solution f(t) given above. We thus draw an important conclusion: the long-term takeover time is determined by the early growth rate of the technology's market fraction. The precise minimum value of f needed for the model to accurately describe a particular market evolution (which then defines the starting time t=0 of the model) must be obtained from real data for the market in question.

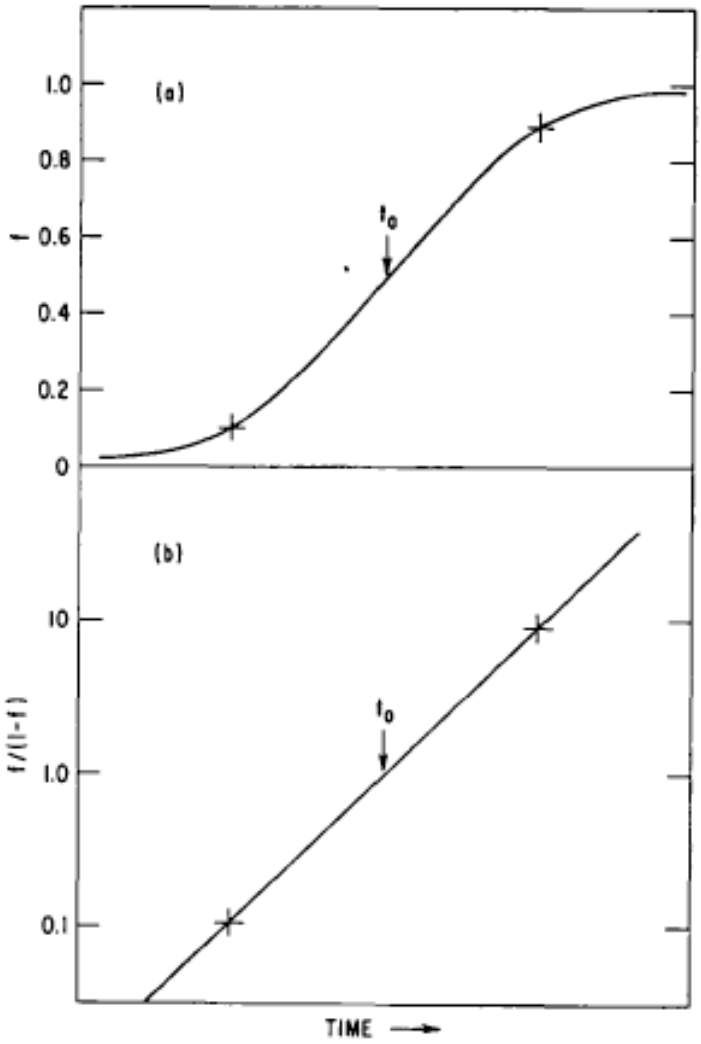


Fig. 1. General form of the substitution model function.

Figure 14.1: General form of the substitution model function. (a) f(t) and (b) f(t/(1-f(t)). Figure taken from Fischer & Pry (1971).

The substitution fraction can also be seen to satisfy the equation

$$\frac{f}{(1-f)} = \exp(r_0 t)$$

which implies that if one were to plot $\ln\frac{f}{(1-f)} = r_0 t$ for the substitution for any given innovation, one would find a linear dependence of the logarithm vs. time, with the slope giving the early growth rate for the innovation. The takeover time, T, would then be given by the time taken to go from $\frac{f}{(1-f)} = 0.11$ to $\frac{f}{(1-f)} = 9$ and the mid-point time, t_0, corresponding to f=0.5 is found where $\frac{f}{(1-f)} = 1$.

This simple model, if shown to be valid for a particular set of technological substitutions in a given market area, can then be used to estimate the onset of market saturation, and estimate T, which gives how long it will take to reach market saturation, if enough early data (say for f lying in the range of $0.01 < f < 0.1$ *are available.*

Fischer & Pry applied this analysis to a large number of products and substitutions that occurred in the twentieth century, and gave a semilog plot of f/(1-f) vs. time in normalized units $(t-t_0)/\Delta t$. The result is reproduced in **Error! Reference source not found.** below and shows that indeed the model captures the dynamics of many different types of market substitutions. The same result, shown in a linear scale is also plotted in **Error! Reference source not found.**.

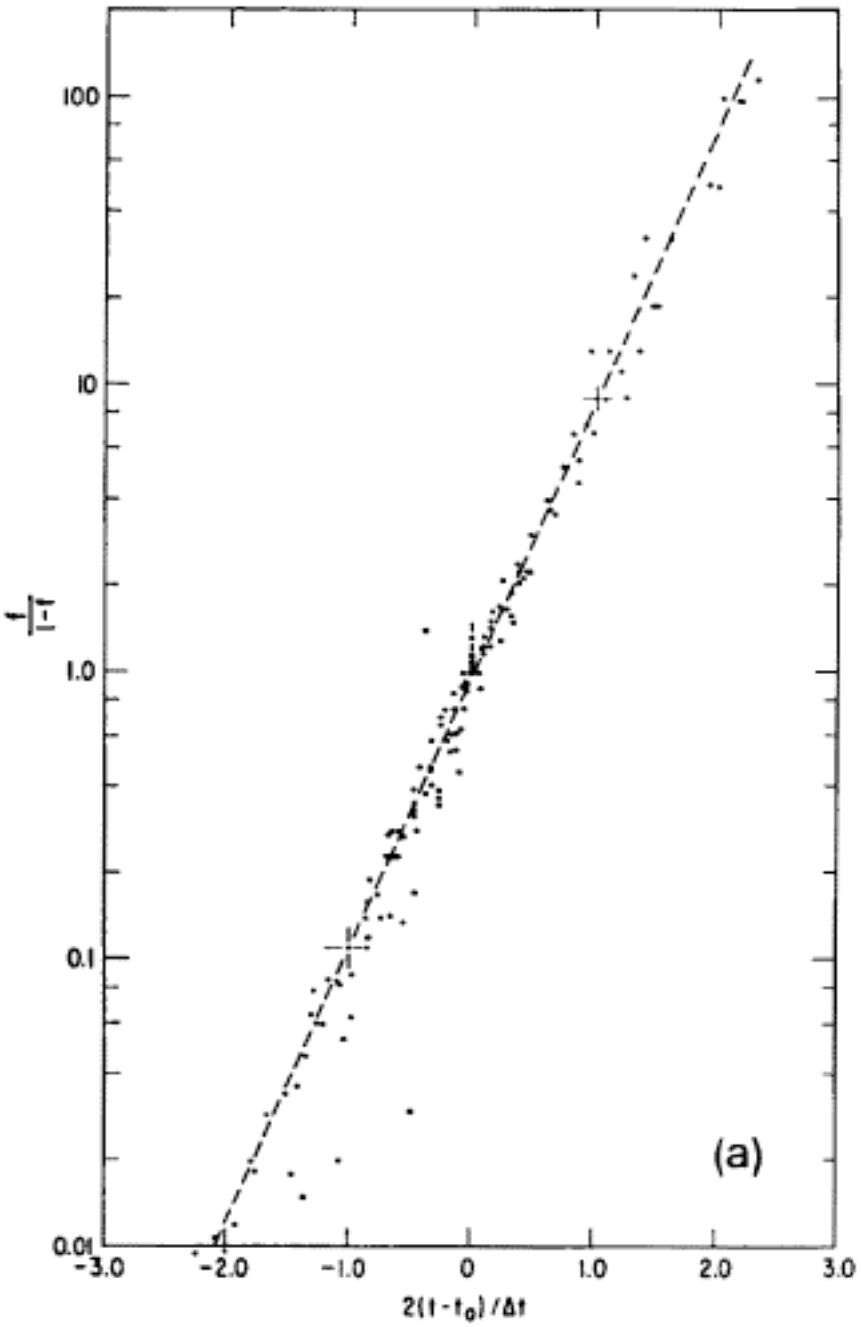


Figure 14.2: Semilog plot of f/(1-f) verses normalized time, where t0 denotes the f=0.5 point and Δt denotes the take over time. Figure taken from Fischer & Pry (1971).

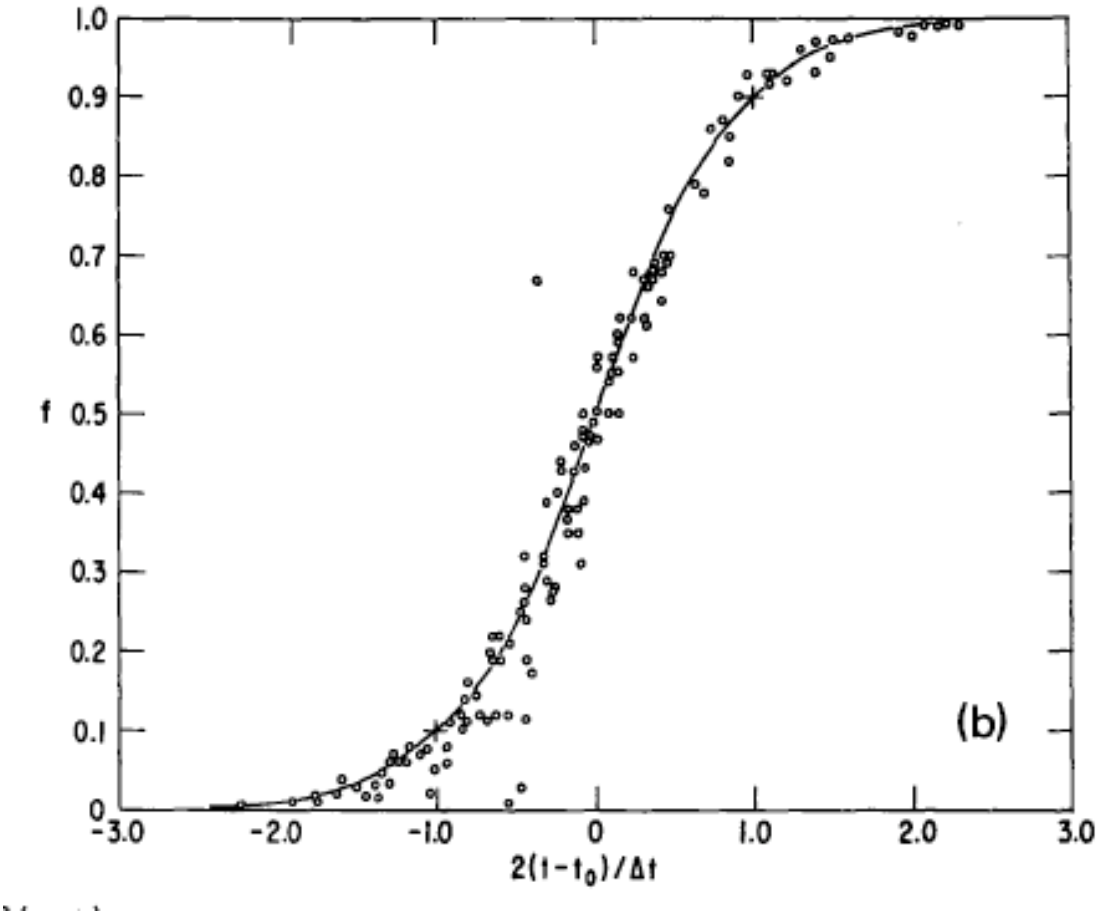


Figure 14.3: Plot of f(t) verses normalized time, where t0 denotes the f=0.5 point and Δt denotes the take over time for multiple market substitutions. Figure taken from Fischer & Pry (1971).

Thus, this simple logistics model (which interestingly is identical to the population dynamics model introduced in the beginning of this book) seems to hold for many of the market substitutions that occurred in the 20th century.

Application of the Substitution Model to Primary Energy Source Evolution

In a set of key papers, Marchetti and others applied this substitution model to the evolution of primary energy sources over the last 150 years [see Marchetti, Technological Forecasting and Social Change, 10, pp. 345-356 (1977) for the first major reviewed paper on the subject]. However, because the original Fischer-Pry model assumed just two competing techniques, while primary energy is provided by several competing sources, Marchetti had to introduce an additional assumption: namely that the "first in" technique would be the "first out" technique – i.e. older primary energy sources would be supplanted by newer ones, and the older sources would then be first to die out. With this assumption, then if the i-th primary energy source provides a fraction $f_i(t)$ of the primary energy requirement at time t, and there are a total of N primary energy sources available, then the requirement that $\sum_{i=1}^{N} f_i(t) = 1$ is ensured by forcing the oldest source, denoted here by the index J, to satisfy the requirement that $f_J = 1 - \sum_{\substack{i=1 \\ i \neq J}}^{N} f_i(t)$, i.e. the oldest, least effective primary energy source provides only the margin of energy needed to satisfy the requirement $\sum_{i=1}^{N} f_i(t) = 1$. The first part of Marchetti's analysis used U.S. historical data for primary energy for the period from 1850 until 1970. The absolute values of primary energy provided from various sources in the U.S. over this time period are shown in **Error! Reference source not found.** below. Marchetti and collaborators took these absolute values for primary U.S. energy sources, calculated the total U.S. energy demand vs. time, found f for each energy source, and then analyzed the resulting values using the logistics curve analysis introduced by Fischer & Pry. The results of this exercise are shown in **Error! Reference source not found.** and **Error! Reference source not found.** below. They also displayed the result of this analysis in linear plot format, which makes it clear how primary U.S. energy sources have evolved.

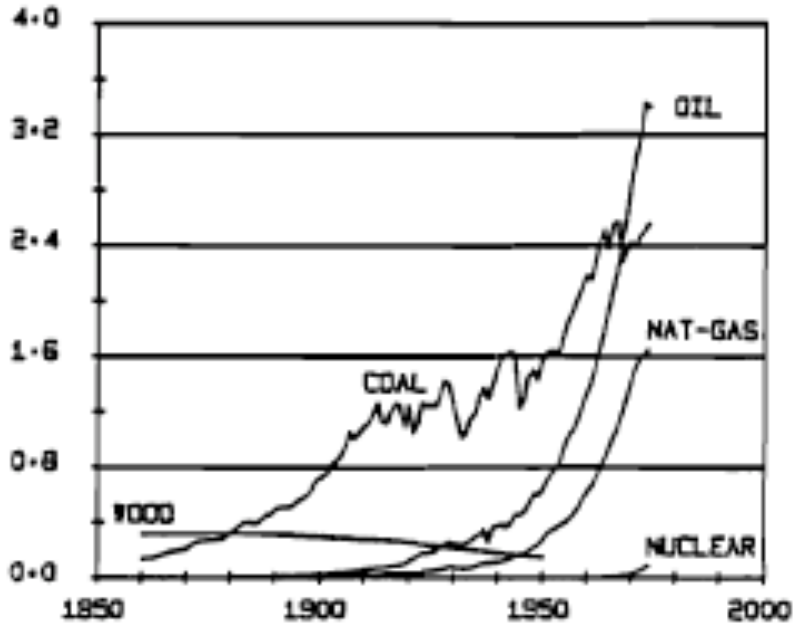


Figure 14.4: Absolute value of primary energy supplied in the US from various sources. Y-axis is in units of tonnes of coal equivalent (1 Tonne coal = 7 Gcalories). Figure taken from Marchetti & Nacinovik, IIASA RR-79-013.

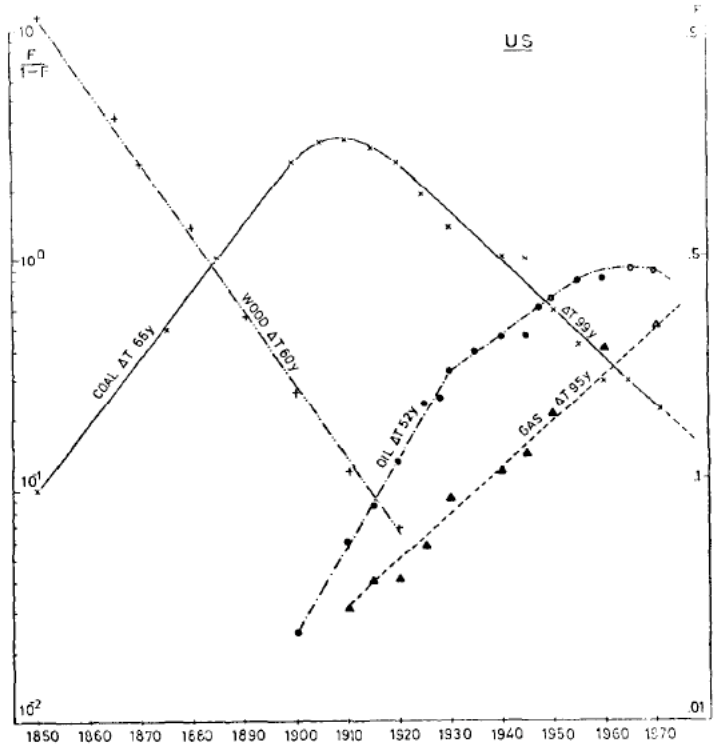
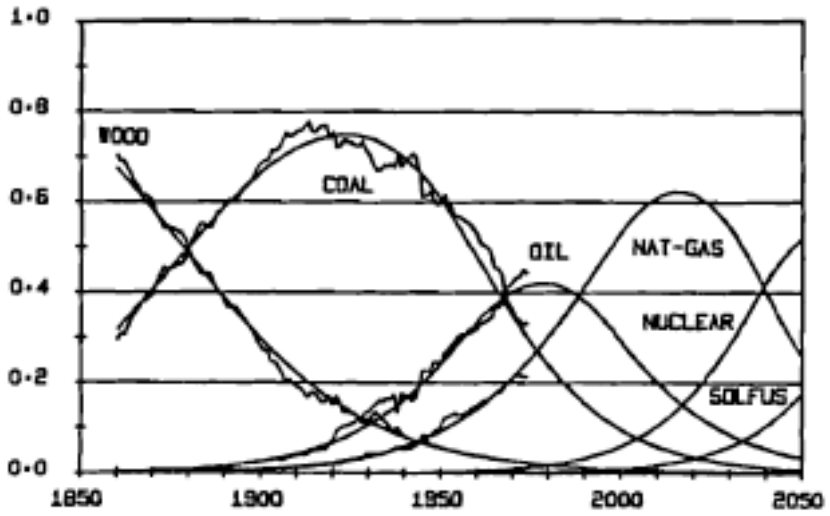


Fig. 5. Fitting of the statistical data on primary energy consumption in the U.S. Straight lines are represented by equations of type (2). Rates of penetration are indicated by the time to go from 1% to 50% of the market (ΔT years). The knee in the oil curve and the saturation regions can be calculated by the rule "first in–first out".

Figure 14.5: Evolution of market fraction of primary energy in the U.S. Figure from Marcetti (1977).



Here the contributions of the various primary sources are shown as fractions of the total market. The smooth curves are two-parameter logistics assembled in a system of equations as described in the text. *The fitting appears perfect for historical data.*

Figure 14.6: Evolution of market fraction of primary energy sources in the U.S. along with fits to substitution model. Figure taken from Marchetti & Nakicenovic, IIASA RR-79-013 (1979).

These results showed that, for a 120 year period, the evolution of US primary energy sources follows the Fischer-Pry substitution model reasonably well. In particular, the analysis shows that the typical substitution time scales are of the order of 50 years for wood, coal, oil, and natural gas. The analysis also shows that in the period from 1975-1980, the historical evidence suggested that natural gas would become the dominant primary energy source sometime around 1990 or so, and that its use in the primary energy market would peak sometime around 2015-2020.

Next, Marchetti played an interesting game. He took the fraction of primary energy provided by oil in the U.S. for the period from 1930-1940, used it to extract the

market fraction model fit, and then projected that model forward into the period from 1940-1970. He then used plotted those projections against the actual fraction of primary energy in the U.S. provided by petroleum. The result is reproduced in **Error! Reference source not found.** below and shows that the model (open squares) and the actual data (filled triangles) are in excellent agreement. This result suggests that this approach to energy substitution may provide a reasonable means to forecast the evolution of climate neutral energy sources.

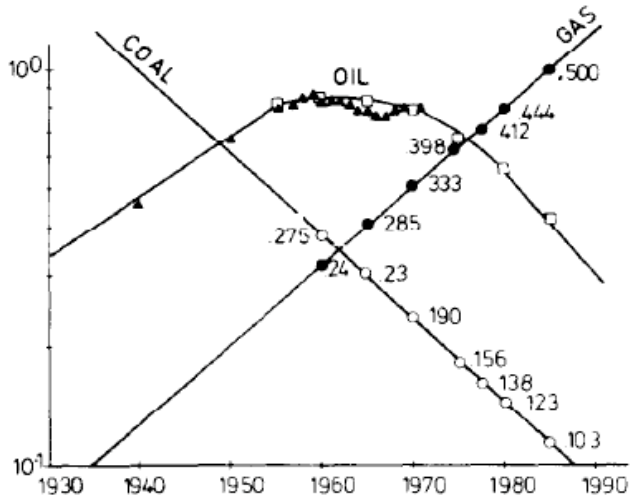


Fig. 6. Forecasting U.S. oil comsumption as a fraction of total energy consumption from 1930–1940 trends. ▫ calculated values, △ statistical data. Other symbols and figures represent intermediate steps in the calculation, the graph having been drawn from my notebook.

Figure 14.7: Forecasting of U.S. oil consumption as a fraction of total energy consumption for the period up to 1985 based upon substitution model. Figure taken from Marchetti (1977).

It is important to keep in mind that these studies refer to only the relative fraction of the market supplied by a given primary energy source. If the overall market for

primary energy is growing (which it has been), then the absolute amount of energy provided by a given energy source will be the product of the relative fraction, f, of that source, and the total primary energy demand. The total primary energy demand, as we have already seen, is linked to the total human population and the per-capita energy consumption, with per-capita energy consumption in turn being linked to human quality of life and specifically to the human population growth rates which likely a proxy for societal measures such as urbanization, agricultural mechanization, increased literacy, access to clean water and a minimal level of health care, and so forth.

Nonetheless, we can draw a crucial conclusion from these studies: *The substitution of a new primary energy source for a declining source proceeds on a time scale of approximately 50 years (i.e. it takes about 50 years for the source to growth from a ~10% contribution to a ~50% contribution).*

Application of Logistic Substitution Model to Climate-neutral Energy Technologies

Marchetti commented in passing on the significant of this finding of a long substitution time on the climate change problem. About 10 years later, Laurmann picked up the analysis and focused exclusively on looking at the substitution of climate neutral primary energy sources for existing fossil fuel sources that emit CO_2 into the atmosphere [see Laurmann, Energy 10, pp. 761-775 (1985)]. Laurmann first looked at the penetration of nuclear fission as a primary energy source; in the earlier work by Marchetti, fission had a

small presence in the primary energy market and thus Marchetti framed his projections for fission with very large uncertainties. However, with a larger market base and another ~10 years of data, Laurmann was able to track the projection for the role of fission in the energy economy. The actual data along with several logistic curve analyses of the data are shown in **Error! Reference source not found.** below. The red dot also shows actual data from the IEA 2004 World Energy Outlook for the year 2002.

Figure 14.8: Past and projected world nuclear energy production as a function of total energy use. Logistic growth curves assume a 1% market share by 1975. Tp is the time for an increase from 1% to 50% of the market share. Figure from Laurmann (1985).

The results show that fission appears to be following a logistic replacement curve with a displacement time of about 60 years. If these trends continue, one could expect fission to provide ~15-30%% of world-wide primary energy demand sometime around the year 2030-2040 according to this analysis. In absolute units, this model would predict that fission would produce ~3-6TW of power in the 2030-2040 timeframe, about a factor

5-10 higher than present 2009 values. A careful examination of this result also shows that the early growth rate of fission in the 1970-1980 period in the world's energy market was quite rapid with a replacement time as short as ~30-40 years or so. However, this growth obviously slowed in the more recent period since the actual fraction of power from fission (based on 2004 IEA data) lies on a curve corresponding to a ~60 year replacement time. The difference is perhaps related to the significant slowing in the growth rate of nuclear fission power in the 1980s and 1990s in the face of the Three Mile Island and Chernobyl nuclear accidents, combined with the large up-front capital costs of fission power and the relatively low capital costs of fossil fuel power plants.

Laurmann then did an interesting exercise (recall that this paper was written in 1985). He assumed that the primary energy sources could be lumped into two categories: climate neutral and CO_2 emitting, and then applied the Fischer-Pry analysis to this hypothetical energy system. He combined these results with then-available carbon balance models and an assumed a rate of growth in global primary energy demand that was consistent with the recent past, and then estimated the future C-inventory in the atmosphere that would occur assuming various dates relative to the year 1975 when the hypothetical C-free energy technology had a 1% share of the world's primary energy demand. The results of this exercise are shown in **Error! Reference source not found.** below, taken from the Laurmann paper. Given that nuclear fission occupied about 2% of the total world primary energy demand in 1975, it would seem that the t_0=0 case of this figure would be about right. A review of the IEA 2004 World Energy Outlook reveals

that global energy demand growth has been in the neighborhood of 2% for the period from 2002-2010, and the IEA projections anticipate a growth rate of ~1.7% for the next 20 years or so. The figure above shows that the displacement time for fission is about 60 years, which lies somewhat between the two sets of curves which Laurmann provides. Thus, based upon Laurmann's modeling, one would expect that the atmospheric CO_2 concentration might reach a saturated value of approximately 2x that of pre-industrial times in the late part of the 21st century. However, this conclusion is quite sensitive to the assumed carbon-neutral energy technology displacement time. If this value is increased to 75 years, then the CO_2 concentration would continue to grow throughout the 21st century.

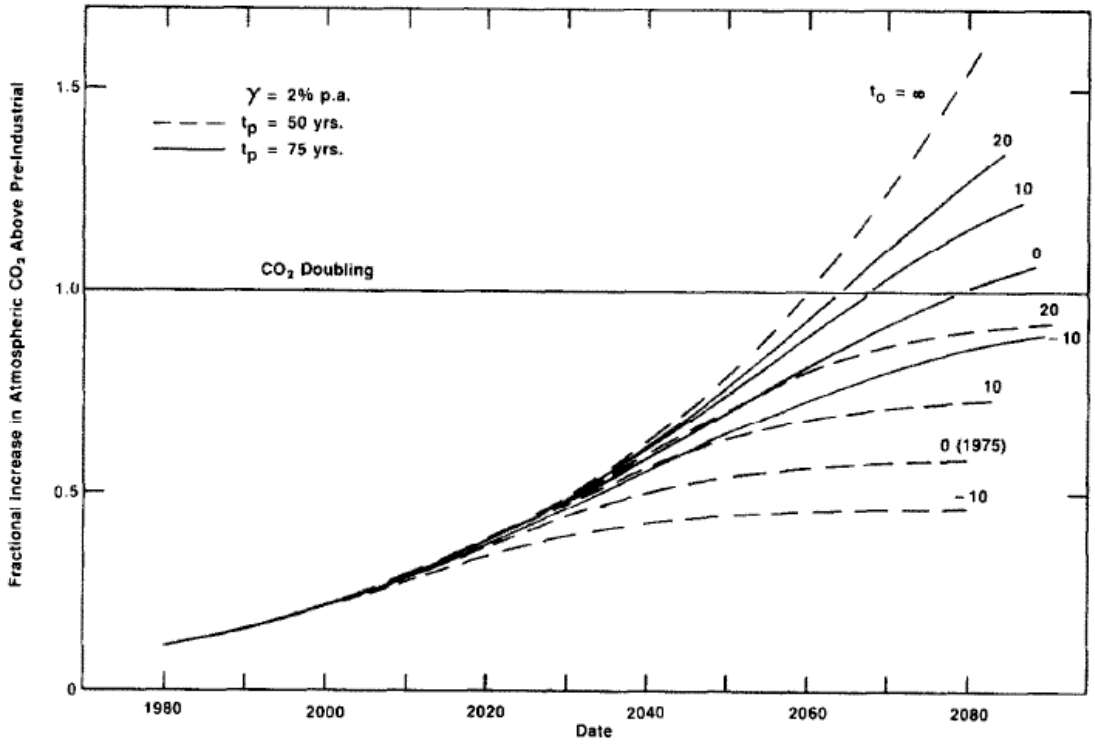


Fig. 5. Atmospheric CO_2 increase for a total exponential energy growth rate of 2% per annum.

(

Figure 14.9: Fractional increase in atmospheric CO_2 concentration for a total exponential energy growth rate of 2% annum. Figure taken from Laurmann, Energy (1985).

Application of Displacement Model to Renewable Energy Sources

At this writing the contribution of wind, solar PV, and solar thermal energy technologies to the world's energy supply are growing rapidly and at least one (wind) has reached a market penetration approaching 1% of world electricity demand. Thus, in this section we apply the model to these emerging primary energy systems and try to get a rough idea of how they can be expected to enter the market over the coming decades. Since these carbon-neutral technologies presumably displace fossil fuel sources, we can then estimate

the resulting impact upon carbon emissions and compare those estimates against emission trajectories required for specific CO_2 concentration targets.

The methodology is as follows. We first will examine the market data for installed capacity for wind, solar PV, and solar thermal technologies. From the worldwide demand for electricity we can then find the market fraction evolution, f(t), for each of these technologies. Assuming that the logistics model holds for these new technologies, we can then estimate the early growth rate, r_0, for each technology and then project f(t) into the future since the early growth rate determines the replacement time, Δt. The plot of ln(f/(1-f)) also allows us to estimate the midpoint time, t_0, at which point f=0.5 for each technology.

With these parameters known, then f(t) is then determined. We then translate this result back into an estimate for the absolute energy production as follows. First, we can estimate global energy demand in the future if the growth rate follows historical patterns (of course it may or may not!) to find future electrical demand $P_{demand}(t)$. Second, since all of these energy sources are intermittent, and we know that with current electrical power distribution grid technologies, the maximum fraction of power coming from e.g. wind or solar is limited to a f_{max} of this maximum power. Thus, in any given year the maximum power from any one of these sources will be given by the product $f_{\max} P_{demand}(t)$. The absolute power then delivered by each of these sources will then be given as $f(t) f_{\max} P_{demand}(t)$ where f(t) is the market fraction determined by the logistics model. We can then estimate the amount of C emissions avoided by assuming a C-

intensity for a fuel and a thermal conversion efficiency (we will use 50 MJ/kg which is comparable to the value for CH4 and an efficiency of 45% comparable to a good gas turbine system).

Let us first examine the wind energy results. **Error! Reference source not found.** below shows the growth in global cumulative installed capacity for wind energy generation.

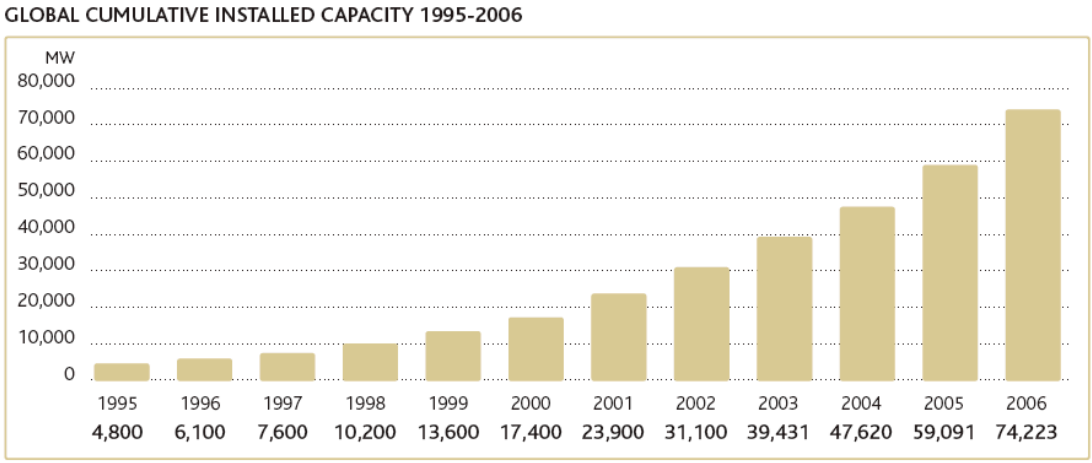


Figure 14.10: Global cumulative installed capacity of wind power for the 1995-2006 period. Source: GWEC, Global Wind 2006 Report

Using this data along with global electricity demand over the same time period, we can find f(t). The results are shown in **Error! Reference source not found.** below as the blue diamond data points. Using these historical values for f(t), we can then plot ln(f/(1-f)) (see **Error! Reference source not found.** below) and then estimate the growth rate, r0. The results show r_0~20%/year which then gives a takeover time of ~22 years.

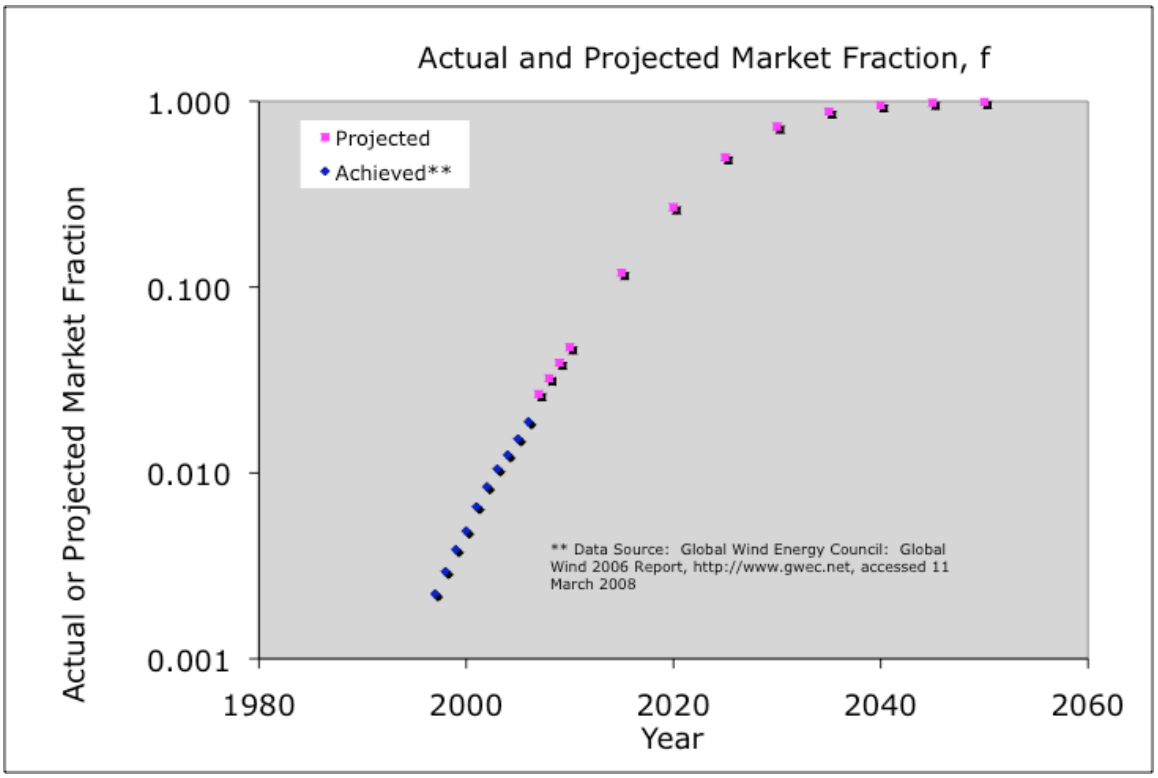


Figure 14.11: Historical (blue diamonds) and projected (pink squares) market fraction for wind energy.

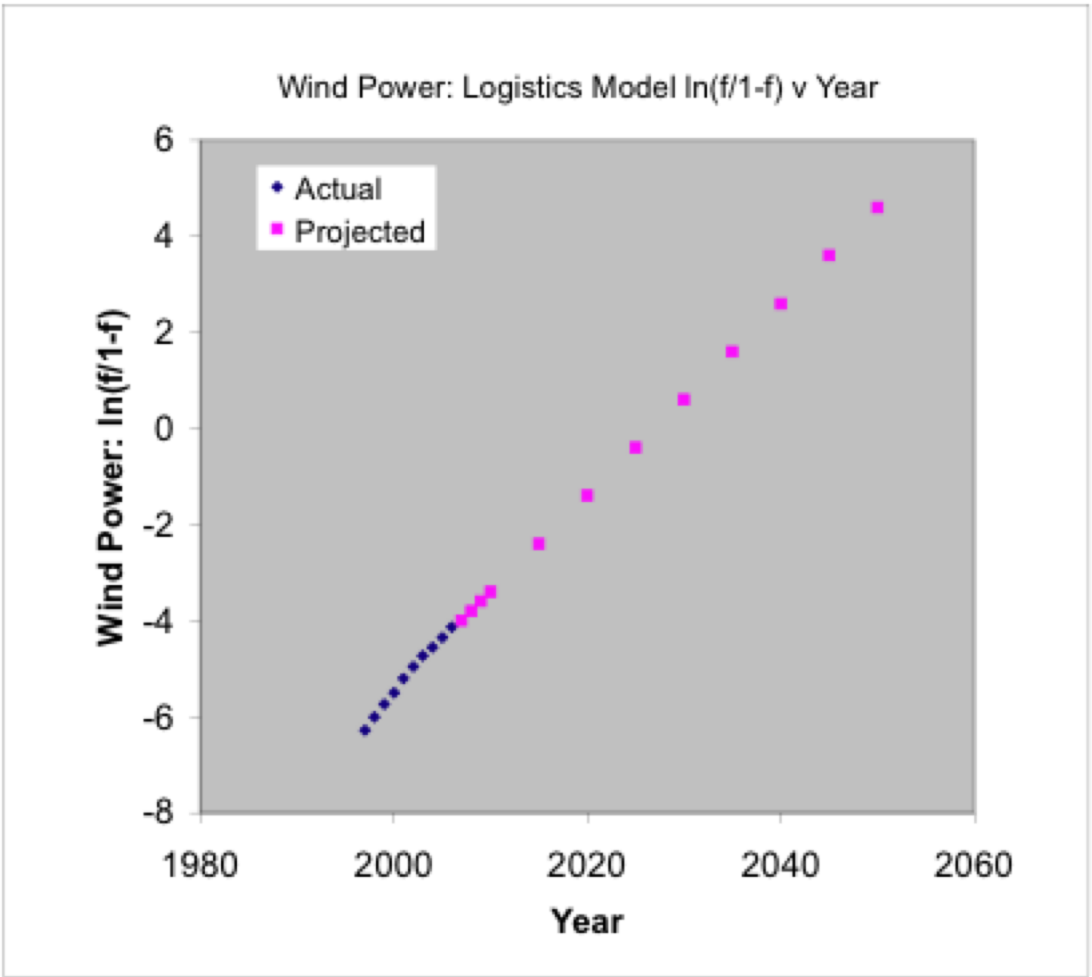


Figure 14.12: Historical (blue diamonds) and projected (pink squares) ln(f/(1-f)) evolution for wind energy. The historical data up until 2007 suggest an annual growth rate of ~20% for cumulative global installed capacity, which then yields a takeover time of ~22 years.

From the plot above we see that the midpoint time, t_0, would be expected to occur in ~2025, and f~0.9 by about 2035 or so. If due to intermittency and grid stability considerations, wind energy is then limited to 25% of the total power on the grid, then we estimate that globally wind energy will contribute about 1TW in ~2025 and about 3-5TW in 2035 or so. Note that due to the turbulent boundary physics, we have already discussed

these figures correspond to installing wind turbine arrays over very large (~10^6 km^2 in 2025, 3-5x10^6 km^2 in 2035) land areas, which correspond to 1-10% percent of the total land area of the Earth. The exact land area required depends upon a more careful evaluation of wind speed distribution over all continents.

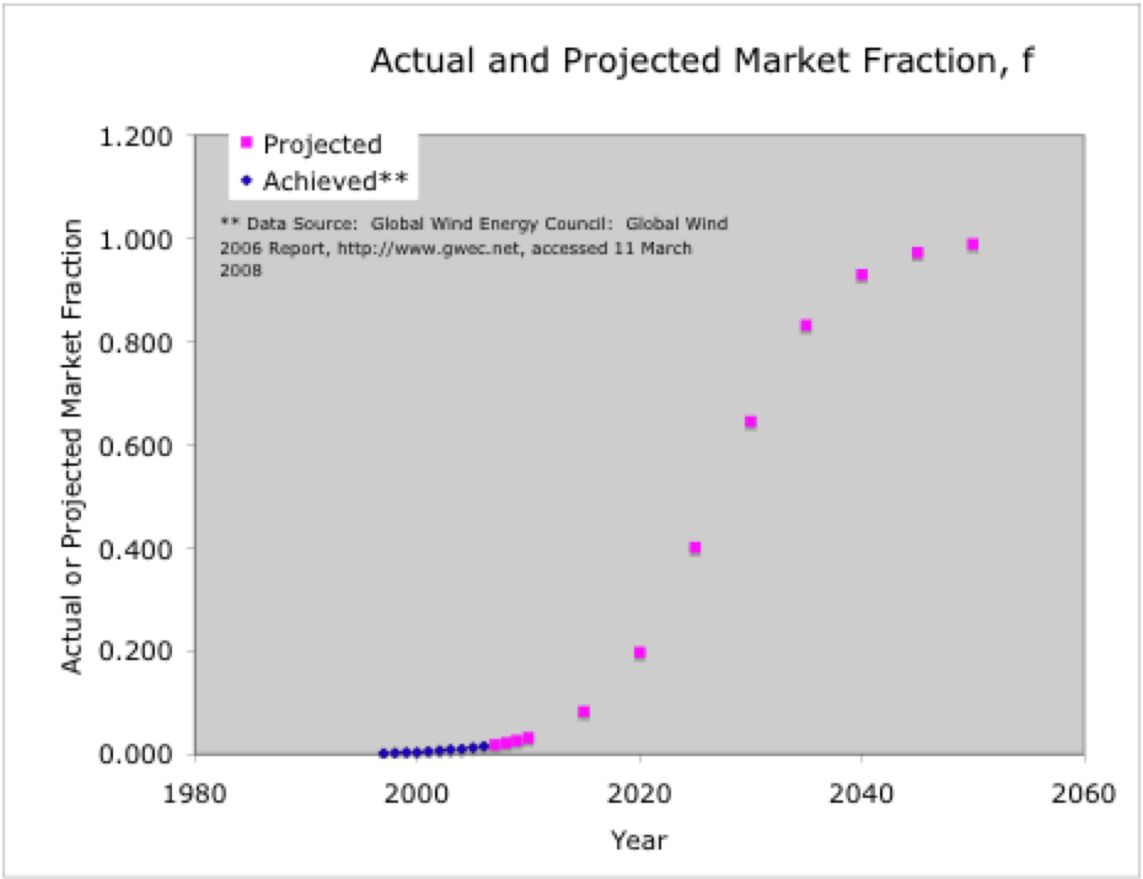


Figure 14.13: Actual and projected market fraction for wind power based upon installed capacity evolution and the Fischer-Pry replacement model.

The carbon displaced would then be as shown in **Error! Reference source not found.** below. In this case, we assume that wind energy can safely provide 25% of the global demand. For comparison, a single "wedge" from Socolow is shown as well. This

rather optimistic growth scenario for wind energy would displace several GTonnes of C emission/year by ~2040. The net C displaced is sensitive to the maximum fraction of wind power that can be accommodated by a grid – which in turn is dependent upon both the availability of large energy storage as well as on a 'smart grid" in which the demand can be controlled by some network mechanism in response to available sources. We illustrate this point by showing the case where the maximum permissible wind energy fraction is 10%. In this case, the maximum annual C emission displacement is only ~1GTonne in 2050, and the midpoint time, t0 is now 2021 instead of 2025.

Figure 14.14: Projected displaced carbon emissions due to market penetration and saturation of wind power.

A similar analysis for solar PV shows that, based upon recent growth rates, r0~30% for this technology, implying a takeover time of less than 20 years. One would then expect f=0.1 in ~2020, and f=0.5 in 2025 or so. If solar PV can contribute 10% of the power demand due to intermittency impacts on grid stability, then we could expect the absolute PV power generation would follow the projections shown in **Error! Reference source not found.** below. With the same carbon intensity and thermal conversion efficiency, then this power corresponds to the C emission displacement shown in **Error! Reference source not found.** below.

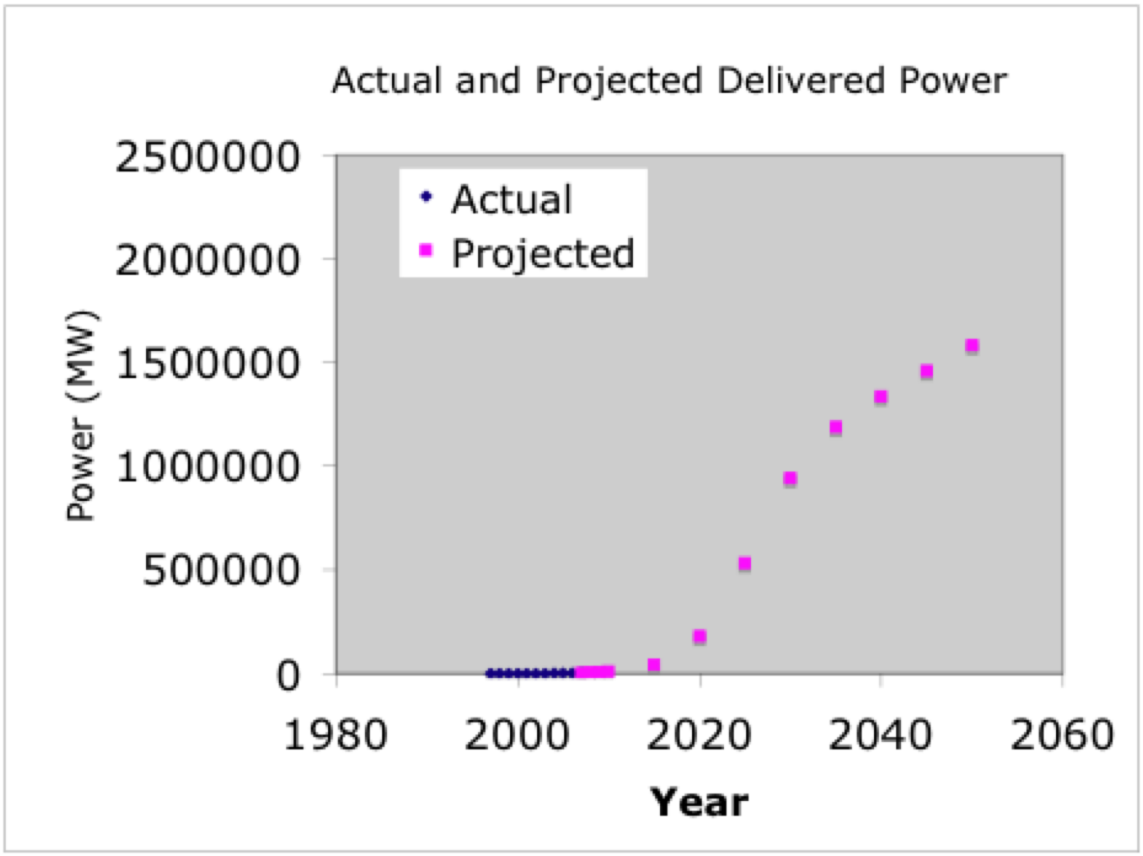


Figure 14.15: Actual and projected delivered power from solar PV based upon growth in installed capacity up to 2006.

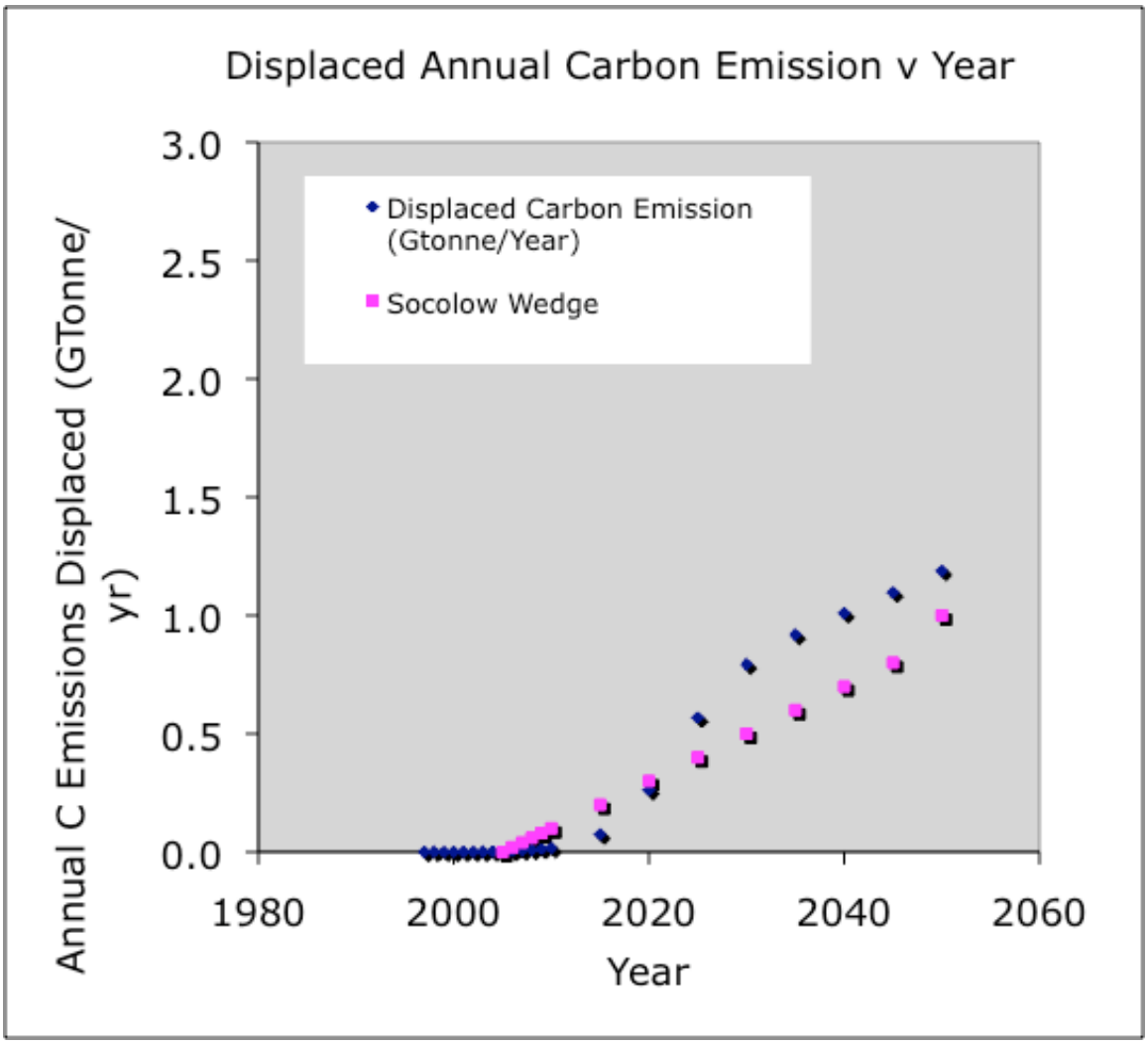


Figure 14.16: Annual carbon emission rate displaced by projected solar PV market penetration.

If both wind and solar PV are limited to generating 10% each of the electrical power demand, then together they will result in the C emission trajectory shown in **Error! Reference source not found.** below. Increasing the maximum contribution for both wind and PV up to 25% each (i.e. 50% of electrical power is met by renewables in 2050) they will together displace about 4-6 GTonnes of C emission, leaving net C

emissions at about 9-11 GTonnes/year – i.e. still higher than current values. Thus, even this optimistic scenario does not allow humanity to begin to decrease global C emissions.

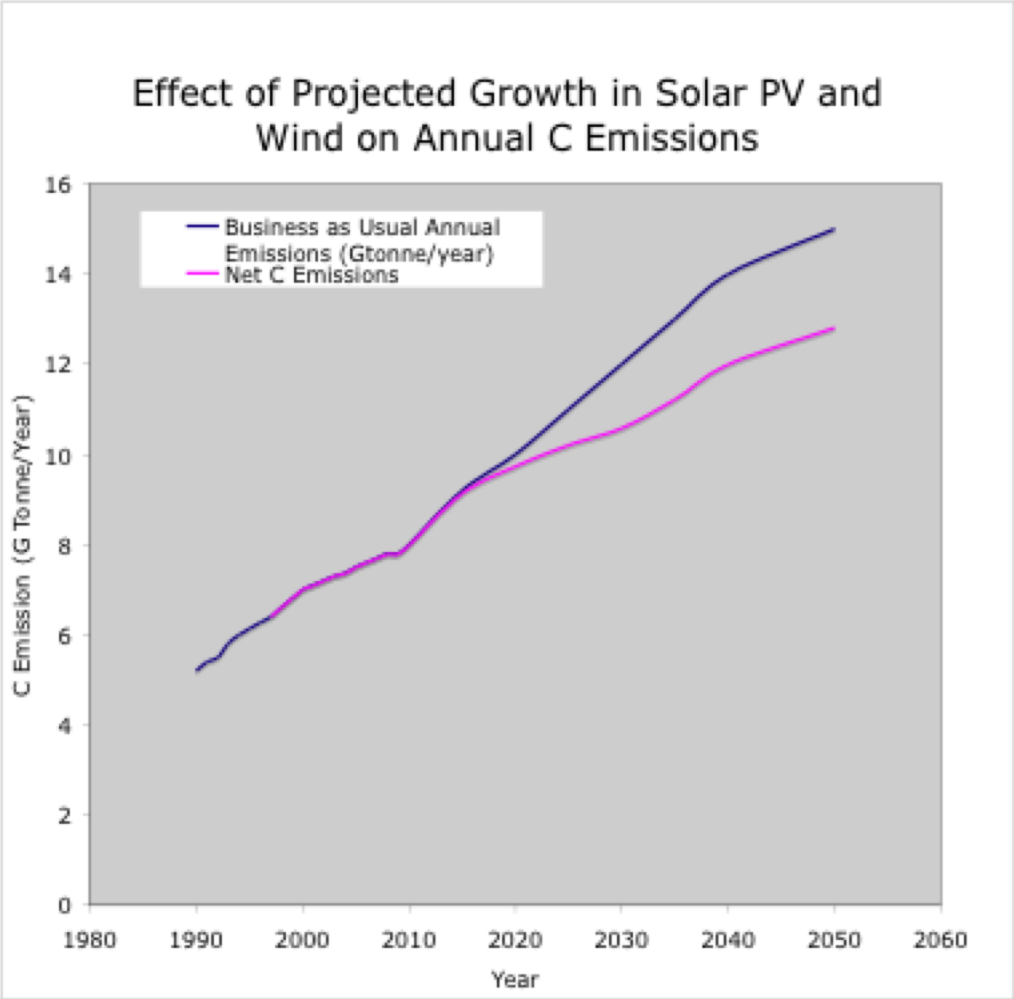


Figure 14.17: Impact of market penetration of wind and solar PV on net global carbon emissions, assuming that wind and solar PV each saturate at 10% of world energy demand.

Chapter 15: Emerging Primary Energy Sources

Liquid Fuels from Biological and Synthetic Sources - Pending

Nuclear Fusion - Pending

Chapter 16: EROEI and Its Effect on Total Energy Demand

Epilogue: Where to Now?

Near term outlook: Continuing fossil fuels and continued development and deployment of renewables

Intermediate term: Rising fossil fuel prices & falling prices for renewables and nuclear; lead to peak in demand for fossil fuels and continued transition to C-free sources; slowing (and eventual stop) of population growth; Reduced drivers for contuing growth in absolute size of economy. Impacts on long term demand for energy?

Long term: Climate change effects are serious and could have major impact. Transition to a static or even declining global human population? Implications?