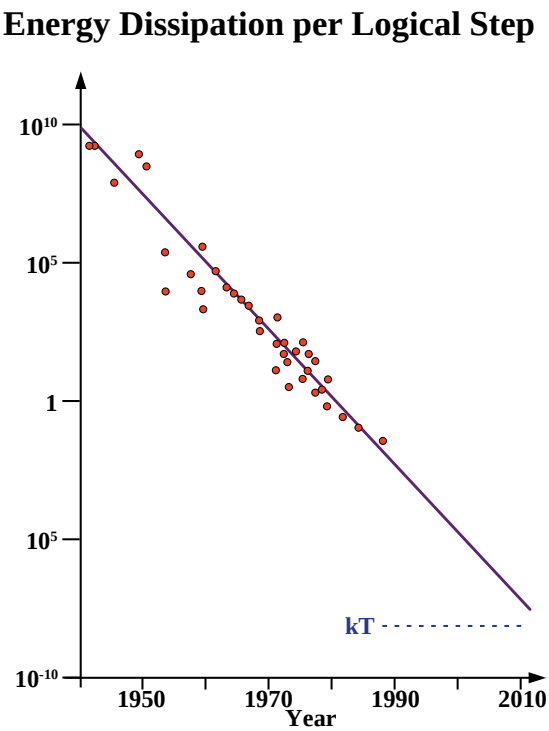
3 Physical limits to computation

Physics imposes a number of limitations to the power of computation. These are different from mathematical limitations (e.g. complexity classes) mentioned above and deal with the construction of a computational device. These limitations arise on all levels and relate to the performance of all computational steps, such as the storage of information, logical operations, or transfer of information between different parts of the computer.
While some of these issues (such as the speed of light as a limit for information transfer) are quite obvious, others have only recently been discovered. A number of limitations derive from the actual hardware used to perform the computation, such as the CMOS technology used in today's computers. In this section, however, we would like to explore limitations that are more fundamental, independent of the current implementation.
Computations are always done with devices (or brains) that obey physical laws. Correspondingly, physical laws must limit the power of these devices. Some of the issues that come up in this context are highly nontrivial, and several apparently obvious limits have later been proved to be no real limits. As an example, it was believed that, since a logic gate changes a degree of freedom of a physical system, it must therefore dissipate at least the energy kT. As the figure shows, the energy dissipated per logical step has decreases by more than ten orders of



magnitude over the last fifty years. Within 10-15 years, this quantity will reach the thermal energy kT. This implies that some schemes for performing calculations will no longer work. It is now established, however, that information can be processed with techniques that dissipate less than kT per logical step. There is no lower limit for the energy required for a logical operation, as long as the time is not critical.

3.1 Information and Thermodynamics

As discussed in chapter 1, the flow of information corresponds to a transfer of entropy. Information processing is therefore closely tied to thermodynamics. As an introduction to these issues consider the problem the Maxwell demon: As Maxwell discussed, in his "Theory of heat" in 1871, "If we conceive a being whose faculties are so sharpened that he can follow every molecule in its course, such a being, whose attributes are still essentially finite as our own, would be able to do what is at present impossible to us. For we have seen that the molecules in a vessel full of air at uniform temperature are moving with velocities by no means uniform... Now let us suppose that such a vessel is divided into two portions, A and B, by a division in which there is a small hole, and that a being, who can see the individual molecules, opens and closes this hole, so as to allow only the swifter

molecules to pass from A to B, and only the slower one to pass from B to A. He will thus, without expenditure of work, raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics." Clearly such a device is not in contradiction with the first law of thermodynamics, but with the second. A number of people discussed this issue, adding even simpler versions of this paradox. A good example is that the demon does not have to measure the speed of the molecules; it is sufficient if he measured its direction: He only opens the door if a molecule comes towards the door from the left (e.g.), but not if it comes from the right. As a result, pressure will increase in the right hand part of the container. This will not create a heat-difference, but rather a pressure difference, which could also be used as a source of energy. Still, this device does not violate conservation of energy, since the energy of the molecules is not changed.

The first hint at a resolution of this paradox came in 1929 from Leo Szilard [15], who realized that the measurement, which must be taken on the molecules, does not come for free: the information required for the decision, whether or not to open the gate, compensates the entropy decrease in the gas. It is thus exactly the information processing, which prevents the violation of the second law.

While Szilards analysis of the situation was correct, he only assumed that this had to be the case, he did not give a proof for this assumption. It was Rolf Landauer of IBM [16] who made a more careful analysis, explicitly discussing the generation of entropy in various computational processes. Other researchers, including Charles Bennett, Edward Fredkin, and Tommaso Toffoli showed that it is actually the process of erasing the information gained during the measurement (which is required as a step for initialising the system for the next measurement) creates the entropy, while the measurement itself could be made without entropy creation. Erasing information is closely related to dissipation: a reversible system does not destroy information, as expressed by the second law of thermodynamics. Obviously most current computers dissipate information. As an example consider the calculation $3 + 5 = 8$: It is not possible to reverse this computation, since different inputs produce this output. The process is quite analogous to the removal of a wall between two containers, which are filled with different pressures of the same gas.

The creation of entropy during erasure of information is always associated with disspation of energy. Typically, the erasure of 1 bit of information must dissipate at least an energy of kT. This can be illustrated in a simple picture: We assume that the information is stored in a quantum mechanical two-level system, the two states being labeled $|0>$ and $|1>$. Erasing the information contained in this bit can be achieved by placing it in state $|0>$, e.g., independent of its previous state. This is obviously impossible by unitary operations, i.e. by (energy-conserving) evolution under a Hamiltonian. A possible procedure uses spontaneous emission (which is non-unitary) from a third state $|2>$, which must have an energy higher than that of state $|0>$. We use an electromagnetic pulse that is resonant with the transition $|1> - |2>$. If the system is initially in state $|1>$, the pulse puts in in state $|2>$. If it is initially in state $|0>$, the pulse does nothing. From state $|2>$, the system will undergo spontaneous emission By a proper choice of state, it can be arranged that the system is most likely to end up in state $|0>$. If this probability is not high enough, the procedure must be repeated. The minimum energy expenditure for this procedure is defined by the photon energy for bringing the system into the excited state. This energy must be larger than kT, since the system could otherwise spontaneously undergo this transition, driven by the thermal energy. Similar requirements hold in classical systems, where dissipation is typically due to friction.

3.2 Reversible Logic

As discussed before, conventional computers use Boolean logic, which includes the operations AND and OR. Both these operations, which have two input bits and one output bit, discard information, i.e. they reduce the phase space. This requires the generation of entropy at some other place. The entropy generated by ereasing a bit of information is $\Delta S = kT \ln 2$. Computers based on Boolean logic are therefore inherently dissipative devices. This generation of heat during the computational process represents an obvious limitation, since no physical device can withstand arbitrary amounts of heat generation.

It turns out, however, that computers do not have to rely on Boolean logic. They can use, instead, reversible logic, which preserves the information, generating no entropy during the processing. Examples of reversible computation include quantum mechanical processes, where the computation can be written as a sequence of unitary operations, which are always reversible. For this particular case, it is easy to prove that the energy dissipated during the computation is zero:

$$< E > (t) = tr(H\rho(t)) = tr(H \exp(-iHt)\rho(0) \exp(iHt)) = tr(\exp(iHt)H \exp(-iHt)\rho(0)) =< E > (0)$$

A general reversible computer can be represented as a system of states, corresponding to the information stored in the computer, and a sequence of logical operations, transforming one such state into the next. Since no information is discarded, it is possible to reverse to complete computation and bring the system back to its

initial state, simply be reversing each of the logical operations. No minimum amount of energy is required to perform reversible logical operations.

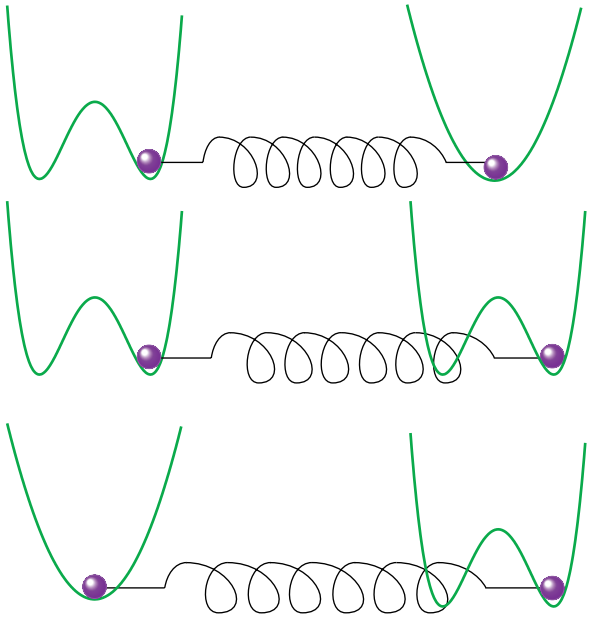
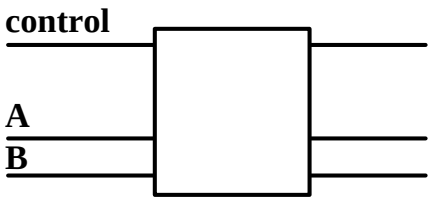


Figure 3.1: Reversible copy operation in time-modulated potential

The potential can be modulated between a monostable and a bistable state in such a way that no energy is expended. The modulation must be sufficiently slow that the system can follow it adiabatically. The spring, which is a passive device, assures that the bead in the second well falls into the left or right subwell, depending on the position of the other bit.
The first proof that reversible logic gates can form the basis of a universal computer is due to Fredkin. He proposed a three-bit gate that is now known as the Fredkin gate. The three lines are labeled as control, A and B. If the control bit is 1, the Fredkin gate interchanges A with B, otherwise it leaves them untouched. The Fredkin gate can be used to implement a reversible AND gate by identifying the inputs of the AND gate with the control and A input of the Fredkin gate and setting the B input line of the Fredkin gate to 0. As can be seen from the truth table of the Fredkin gate, the output B then contains the output of the AND gate, while the control and A lines contain bits of information which are not used by the boolean logic, but would be required to reverse the computation.
A possible mechanical realisation of the Fredkin gate is shown in this figure.
Another reversible computational architecture is a reversible Turing machine. A Turing machine consists of an infinitely long tape storing bits of information, a read/write head that can be in a number of different states, and a set of rules what the machine is to do depending on the value of the bit at the current position of the head and the state of the head. A reversible set of rules would be

head state	bit read	change bit to	change state to	move to
A	1	0	A	left
A	0	1	B	right
B	1	1	A	left
B	0	0	B	right

The information processing corresponds to a motion of the head. The motion is driven by thermal fluctuations and a smal force defining the direction. The amount of energy dissipated in this computer decreases without limit as this external force is reduced, but at the same time the processing speed decreases. Overall the best picture to describe the operation of a revesible computer is that is is driven along a computational path. The same path may be retraced backward by chaning some external parameter, thereby completely reversing the effect of the computation.



Input			Output		
control	A	B	control	A	B
1	1	1	1	1	1
1	1	0	1	0	1
1	0	1	1	1	0
1	0	0	1	0	0
0	1	1	0	1	1
0	1	0	0	1	0
0	0	1	0	0	1
0	0	0	0	0	0

Figure 3.2: Fredkin gate

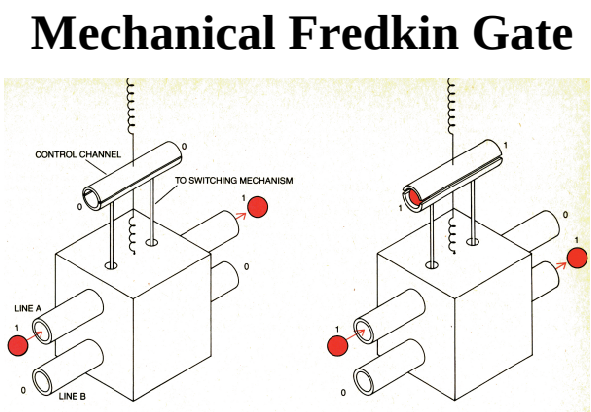


Figure 3.3: Fredkin gate

3.3 The ultimate Laptop

Some limits to the performance of computers have been summarised by Seth Lloyd [17] in a very popular style: he discusses the "ultimate laptop", i.e. the maximum performance that a computer of 1 kg mass and 1 l of volume may ultimately achieve.
One limit can be derived from the uncertainty principle: It can be shown [18] that it takes at least a time

$$\Delta t = \frac{\pi\hbar}{2E}$$

for a quantum mechanical state to evolve into an orthogonal state, if E is the energy of the system. This condition is a requirement for two states to be distinguishable, which is one condition to qualify as a computational step. If we take the mass of the computer equal to its energy $E = mc^2 = 8.9874 \times 10^{16} J$, one obtains an upper limit of 5.4258×10^{50} operations per second.
While this limit seems very remote, it is interesting that quantum computers work close to this limit, if the energy is not equated with the rest mass, but, for the example of an NMR quantum computer, with the Zeeman energy of the spins. This system also permits a verification of the condition: Setting the energy of the ground

state $|\alpha>$ to zero, the excited state $|\beta>$ has an energy $\hbar\omega_L$. An initial state

$$\Psi(0) = \frac{1}{2}(|\alpha\rangle + |\beta\rangle)$$

then evolves into

$$\Psi(t) = \frac{1}{2}(|\alpha\rangle + e^{-i\omega_L t}|\beta\rangle)$$

Apparently the two states are orthogonal for $\omega_L t = \pi$, i.e. after $t = \pi/\omega_L$. Since the (constant) energy of this state is $E = \hbar\omega_L/2$, we recover the condition given above.
An interesting aspect of this limit is that it does not depend on the architecture of the computer. While we generally expect computers containing many processors working in parallel to be faster than purely serial computers, this is no longer the case for a computer working at the limit just discussed: if the number of processors increases, the available energy per processor decreases and correspondingly its speed. The total number of logical operations per unit time remains constant.
A limit on the amount of data stored in the computer can be derived from thermodynamics. According to statistical mechanics, the entropy of a system is

$$S = k_B \ln W,$$

where W is the number of accessible states. To store N bits of information, we need N two-level systems, which have 2^N states. Accordingly, a system that stores N bits has an entropy

$$S = N k_B \ln 2,$$

It should be realized here, that the entropy that we calculate is the entropy of an ensemble at a given energy, while the actual system doing the computation is in a well-defined (pure) state, thus having zero entropy.
An additional limit derives from the necessity to include error correction. Detecting an error can in principle be achieved without energy dissipation. However, correcting it implies eliminating information (to the environment), thus generating dissipation. We will assume here that energy dissipation is limited by blackbody radiation. At a temperature of $T = 5.87 \times 10^8 K$, with a surface area of $0.01m^2$, the blackbody radiation amounts to $4 \times 10^{26} W$. This energy throughput (which is required for error correction, not for operation) corresponds to a mass throughput of

$$dm/dt = P/c^2 = 1kg/ns,$$

which must be fully converted to energy. If this is possible, the number of error bits that can be rejected per second is 7.195×10^{42} bits per second. With a total of 10^{50} logical operations per second, this implies that its error rate must be less than about 10^{-8} to achieve reliable operation.
A limit that may be easier to approach is given by the number of atoms in the system. For a mass of 1 kg, this would be of the order of 10^{25}. In NMR and ion trap quantum computers, each atom stores one bit of information. At this density, it would thus be possible to store 10^{25} qubits of information in a computer. If optical transitions of these atoms are used for logical operations, gate times of the order of 10^{-15}s would be feasible, allowing a total of 10^{40} logical operations for the whole computer.
At such data rates, the different parts of the computer would not be able to communicate with each other at the same rate as the individual logical operations. The computer would therefore need a highly parallel architecture. If serial operation is preferred (which may be dictated by the algorithm), the computer needs to be compressed. Fully serial operation becomes possible only when the dimensions become equal to the Schwarzschild radius ($= 1.485 \times 10^{-27} m$ for $m = 1kg$), i.e. when the computer forms a miniature black hole.
While all these limits appear very remote, it would only take of the order of 100-200 years of progress at the current rate (as summarised by Moore's law) to actually reach these limits. It is therefore very likely that a deviation from Moore's law will be observed within this time frame.

3.4 Literature

C.H. Bennett, "Demons, eingines and the second law", Scientific American 257, 88-96 (1987).
S. Lloyd, "Ultimate physical limits to computation", Nature 406, 1047-1054 (2000).
L. Szilard, "Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen", Z. Physik 53, 840-856 (1929).
R. Landauer, "Irreversibility and heat generation in the computing process", IBM Journal Res Dev 5, 183-191 (1961).
R. Landauer, "Information is physical", Physics Today May 1991, 23-29 (1991).

C.H. Bennett and R. Landauer, “The fundamental physical limits of computation”, Scientific American July 1985, 38-46 (1985).
Feynman, R. P., 1996, “The Feynman Lectures on Computation”, edited by A. J. G. Hey and R. W. Allen (Reading, MA: Addison-Wesley).
N. Margolus and L.B. Levitin, in Proceedings of the Fourth Workshop on Physics and ComputationÑ PhysComp96 (eds Toffoli, T., Biafore, M. and Le?o, J.) (New England Complex Systems Institute, Boston, MA, 1996).