Chapter 9: Physics of Computational Systems

Sections:

Thermodynamic View of Computation - - - Energetics of Bistable Devices - - - Thermal Limit - - - Quantum Limits - - - Granularity of Charge - - - Voltage Limit - - - An Example - - - Energy Management - - - Discreteness in Quantum Mechanical Systems - - - Conclusion

Computation is in the end a *physical process.* Data elements must be represented by some physical quantity in a physical structure, for example, as the charge on a capacitor or the magnetic flux in a superconducting ring. These physical quantities must be stored, sensed, and logically combined by the elementary devices of any technology out of which we build computing machinery. At any given point in the evolution of a technology, the smallest logic devices have a definite *physical extent*, require certain minimum *time* to perform their function, and dissipate a *switching energy* when switching from one logical state to another. From the system viewpoint, these quantities are the units of cost of computation. They set the scale factor on the size, speed, and power requirements of a computing system. Some of the relationships between these elementary quantities are discussed in this chapter, and an example is given of application to technology comparison.

Thermodynamic View of Computation

{ in preparation }

Energetics of Bistable Devices

Any physical structure we use to represent information must be *reliable.* We must be able to stably store all bits of information in our computing machine over the period of any computation. Binary information implies elementary memory elements of a bistable nature; one state denoting a logical zero, the other a logical one. A mechanical system which behaves in this way is the inverted pendulum shown in figure 1a. The force of gravity holds the pendulum stably in either the rightmost or the leftmost position. Switching from one state to the other can be accomplished by pushing the weight up to its maximum position and letting it fall onto the opposite stop.

Physicists view bistable systems of this sort in terms of a diagram such as that shown in figure 1b. What is plotted here is the potential energy of the physical system as a function of its spatial or

electrical coordinate. If the pendulum is left in one of its stable states, given by the minima in the potential diagram, it will stay there indefinitely until enough external energy is provided to surmount the potential maximum and allow the system to re-equilibrate in the other potential minimum. Note that the energy provided by the external switching source is lost in the impact when the pendulum falls to its stop (and perhaps bounces a bit until the energy is dissipated). Others have considered particles in potential wells of this shape to derive minimum switching energies for computation[5].

The slope of the energy curve, i.e. the derivative of the energy with respect to the angle of the pendulum, has the units of a torque. This torque is being supplied by gravity, and pushes the pendulum towards one of its stable positions. Note that gravity acts as the "power supply" for this mechanical logic device.

The energy required to switch from one state to the other can be supplied deliberately, or by some random occurrence. Suppose our pendulum were mounted on a railroad car. While the train is stationary, we expect the device to remain in its initial state. However when the train passes over a very rough stretch of track, the pendulum may bounce into the other state. The potential maximum must be high enough to prevent such random events.

An electronic circuit with the same logical behavior as the pendulum is the ordinary flip-flop shown in figure 2. The detailed behavior of the flip-flop is, however, somewhat different from that of the inverted pendulum. Let us attempt to change the state of the device by supplying a current into that side of the flip-flop which is at the the lower potential. If the current we supply is large enough, we will raise the potential on that side of the flip-flop, turn on the transistor on the opposite side, and change the state of the flip-flop. We can, however, supply a lower current (and therefore power) for an indefinite period of time without changing the state of the device.

In the pendulum we could support the weight part way up the potential curve for a long time and not change the pendulum's state. However, while supporting the weight in a fixed position, we would not be supplying power to the pendulum. We supply power to the pendulum only when we are increasing the elevation of its mass in the gravitational field. It is thus clear that while in general the behavior of the flip-flop and the inverted pendulum are similar, the detailed energetics are quite different. In particular, we can supply a large quantity of energy to the flip-flop without changing its state, provided we supply the energy slowly enough. This is not true of a system like the inverted pendulum.

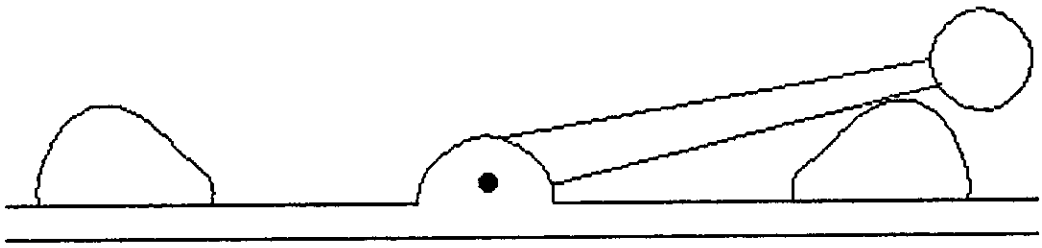

Fig. 1a Inverted Pendulum

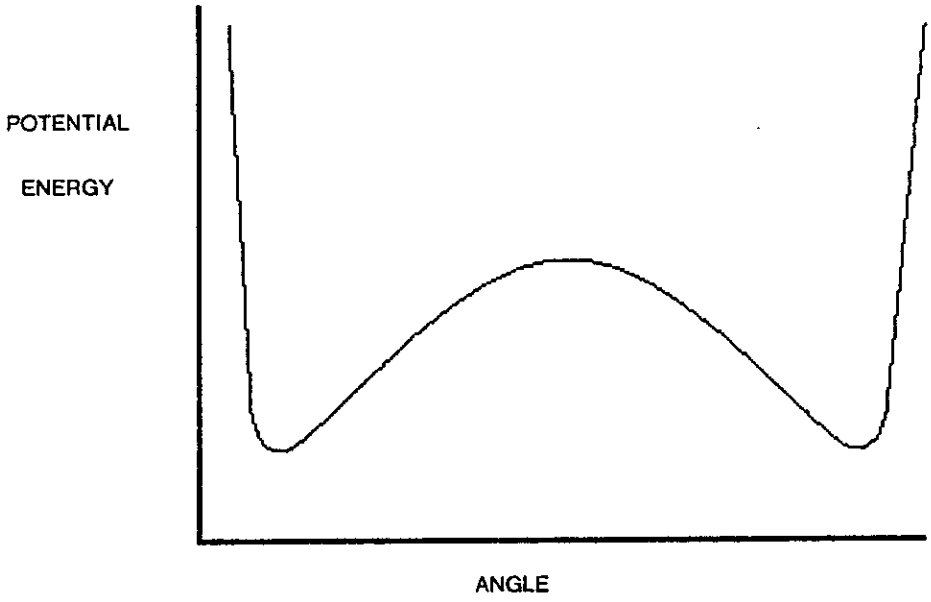

Fig. 1b Potential energy of Pendulum

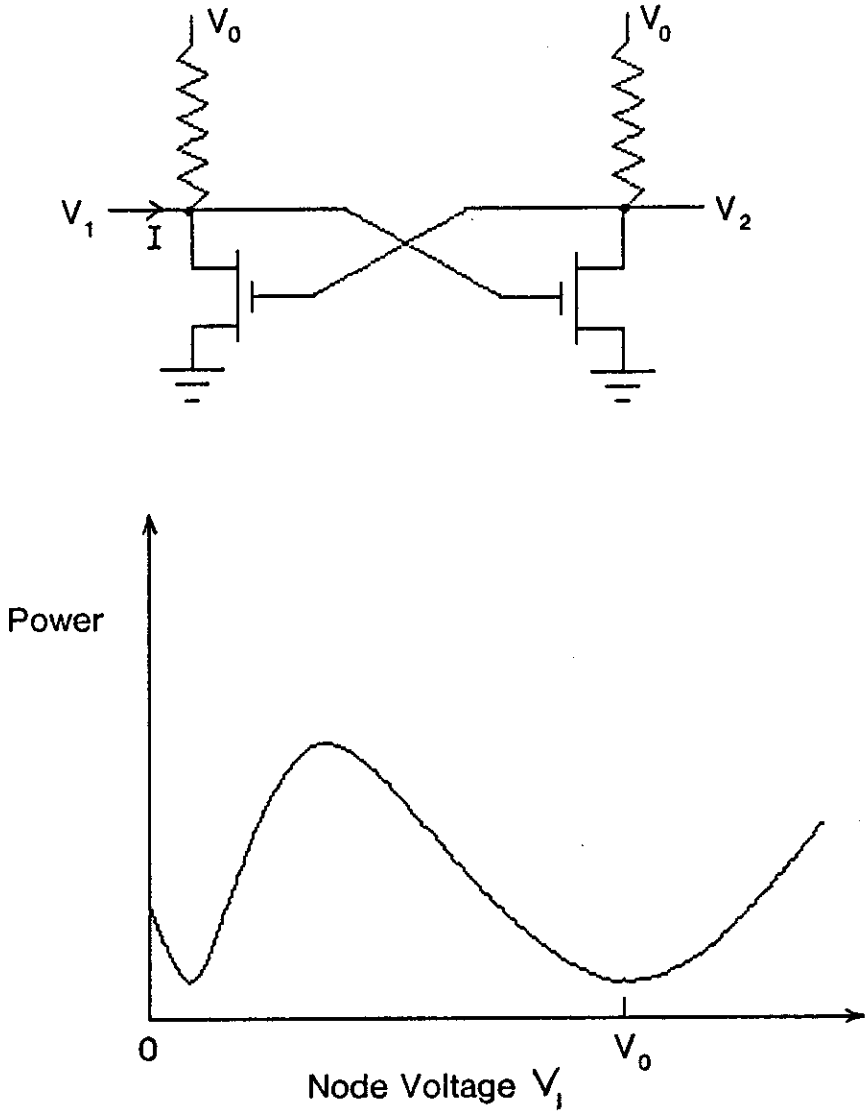

Fig. 2 The Energetics of a Flip-Flop

Principle of Least Power

The basic physical law governing the behavior of any electrical circuit involving resistors is called the Principle of Least Power[1]. Any electrical network composed of resistors comes to equilibrium by adjusting the potentials in such a way that the power dissipated in the network is a minimum. This principle holds true even if the resistors which form the network are not ideal linear resistors. Any network composed of dissipative electrical elements, such as the MOS or Bipolar transistors, will behave in this way.

Energetics of the Flip-Flop

The power dissipated by our ordinary MOS flip-flop when we forcibly hold one node (V_1 for example) at an arbitrary voltage, is plotted in figure 2. Notice that the curve has the same general shape as the energy curve for our inverted pendulum. The two minima correspond to the two stable states of the flip-flop. The maximum corresponds to the point at which no external power need be supplied to hold the flip-flop in its intermediate state, i.e. it is the metastably balanced condition for the flip-flop.

The derivative of the total power with respect to the voltage V_1 has the dimensions of a current, and is equal to twice the current we must force into the node to hold it at a particular voltage.

Although the principles we derive for circuits of this sort are quite general in nature, it is instructive to work a simple, idealized example. Let us represent the pullup transistors of our flip-flop as ordinary resistors with resistance R, and the pulldown transistors as current sources whose magnitude is some mutual conductance G_m multiplied by the voltage above threshold of the transistor gate.

This idealized equivalent circuit is shown in figure 3a. The transfer characteristic of each individual inverter in the flip-flop is shown in figure 3b. The output voltage of the inverter is constant at voltage zero until the input voltage exceeds the threshold voltage V_{th} of its pull down transistor. V_{out} then varies linearly with a slope $-G_mR$. This slope is the gain of the inverter. After the output voltage reaches 0, the inverter saturates and its output remains 0 for further increases in the input voltage.

The power dissipated by this circuit for any given value of the voltage V_1 can be computed analytically and is plotted in figure 4 for various values of the transistor transconductance G_m.

Notice that for values of $G_mR > 1$ the power curve shows a distinct maximum in the center separating the two minima corresponding to the stable states of the flip-flop. However, when the gain is one or less, the power curve shows only one minimum near the threshold voltage. This minimum corresponds to a single stable state.

A cross-coupled circuit must have a loop gain greater than unity, in order to develop two independent stable states. [anon]

This analysis based on the Principle of Least Power agrees with this standard electrical engineering model. It provides us with a very general and fundamental viewpoint from which to analyze the energetics of computer circuits.

Let us perform a conceptual experiment on the flip-flop of figure 2. In its initial state, V_1 is a logical one and V_2 is a logical zero. We delicately remove the connection from V_2 to the gate of the left transistor. At some instant of time we place a charge equivalent to a logical one on the (now floating) gate. At the identical instant we force V_2 to a logical one.

As time progresses we observe the power we must supply to keep V_2 at a logical one. For a while the device absorbs a large amount of power. However, after the signal from the gate has propagated through the two inverters, no more power will be absorbed, and we may reconnect the gate to V_2. Try as we might, we can find no path from one state to the other which can be traversed without supplying an amount of power at least as high as the maximum in the power curve. The total energy required is at least the product of this power and the inverter pair delay.

The essential difference between the "static" and "dynamic" storage devices discussed in chapter 7 is thus clear. Both forms use the same physical element to store the energy which represents information. However the "dynamic" form requires only that the requisite amount of energy be supplied to change its state. It does not matter how slowly that energy is supplied. The "static" form requires in addition that the energy be supplied within the inverter pair delay of the technology.

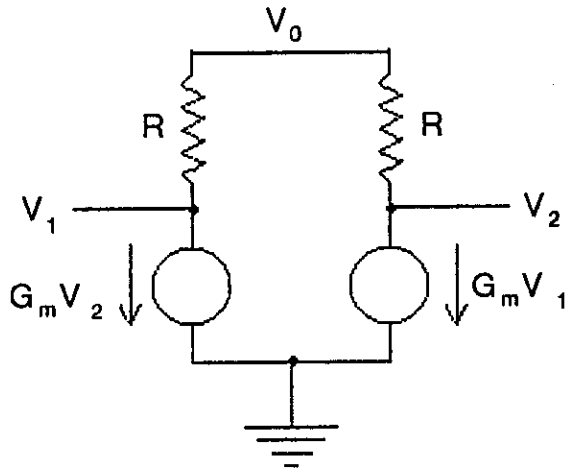

Fig. 3a Equivalent Circuit

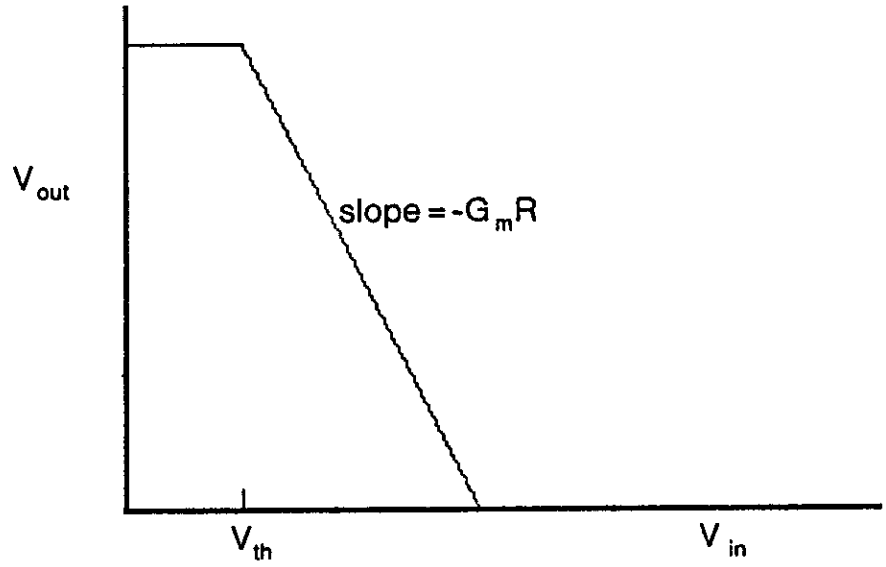

Fig. 3b Inverter Transfer Characteristic

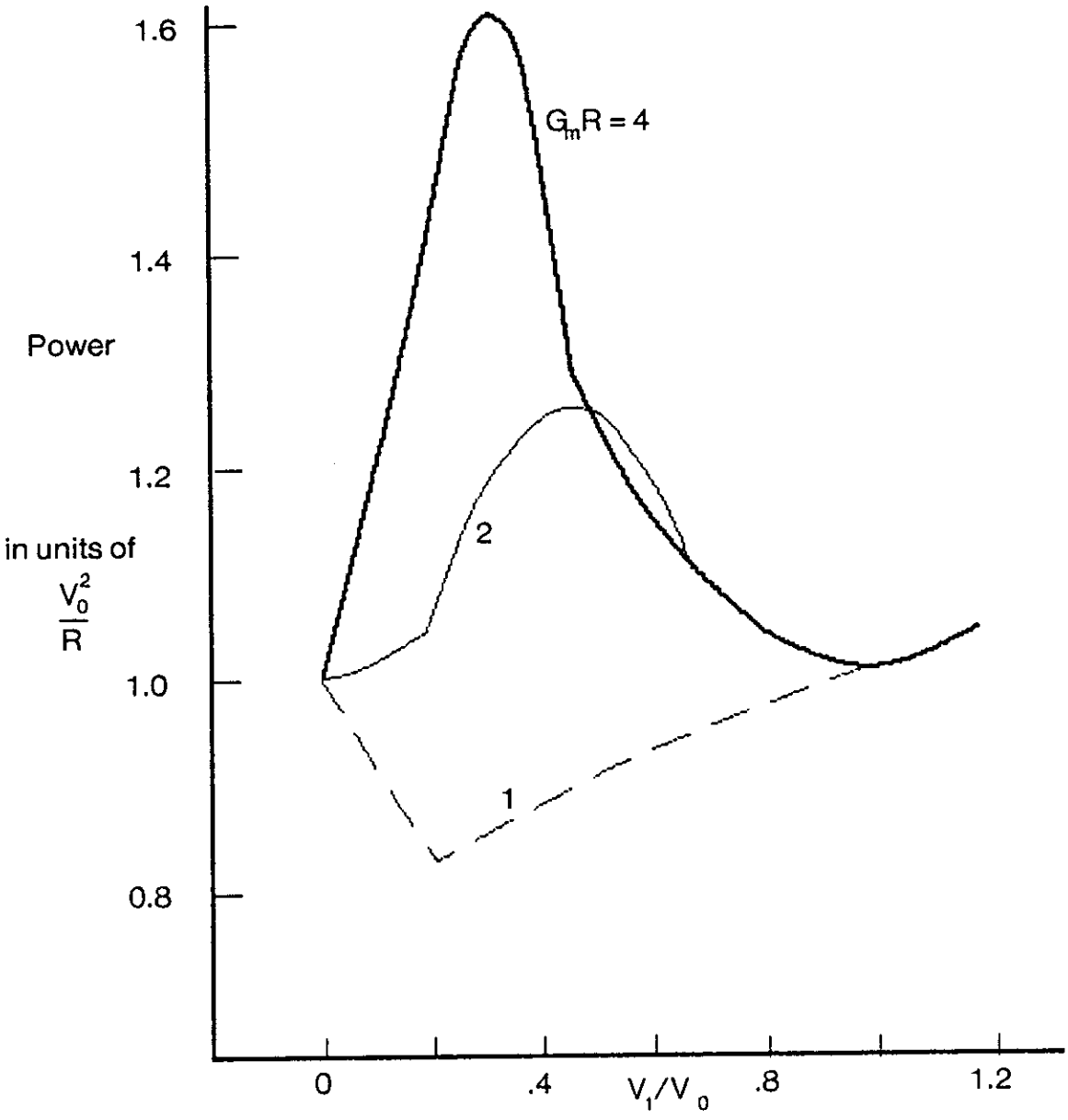

Fig. 4 Power curves for Flip-Flop of fig. 3

Thermal limit

We have just illustrated how to compute the energy required for switching the flip-flop. An external influence must supply an *additional power, P*, equal to the *difference* between the power at the minimum and maximum of the power curve in order to switch the flip-flop from one stable state to the other. That power must be supplied long enough for the information to propagate through both inverters, and back to the node where we applied the signal. This time is just the inverter pair delay τ for the technology out of which the flip-flop is built. The externally input energy required to flip the flip-flop thus becomes:

$$E_{sw} = \tau P$$

The switching energy must be sufficient to prevent random occurrances from changing the state of the device. Electrical noise is always present in any real system. It is generated by heavy electrical equipment and propagated along power mains. Radio and television transmitters of all varieties create electromagnetic radiation which can induce voltages in a circuit. Modern electronic devices allow single electronic occurrances to control relatively large currents. Atomic imperfections randomly capture and release electrons, thus creating an unsteady environment. Techniques exist for minimizing the effect of each such hazard. However, one fundamental source of irreduciable randomness remains.

Any device operating at a finite temperature is subject to the random thermal motions of the elements of which it is composed. The energy of any element, large or small, is not fixed, but fluctuates over a range of energies due to interactions with its environment. Each time we measure the energy E of an element, it will have some value which differs by some ΔE from its equilibrium value. The probability that any given *independent* measurement will yield a given ΔE is given by the Boltzmann equation: Probability $= e^{-\Delta E/kT}$.

What constitutes "independent" measurements depends on the response time τ of the system. Two measurements should be made at least τ apart to be considered independent. Similarly we have seen that in order to switch a bistable system from one state to the other, a certain amount of power P has to be supplied for the duration of the response, or switching, time τ of the system. Therefore, systems with a faster response time are more likely to be switched by thermal fluctuations, since occasions where the critical power level P is exceeded for the necessary amount of time τ occur more often. This is equivalent to the view that systems with a wider bandwidth capture more energy out of the spectrum of the thermal energy. Thus the probability per unit

time that an element will achieve some large variation ΔE is:

$$\text{Probability per unit time} = (1/\tau)e^{-\Delta E/kT}$$

The probability per unit time that random thermal noise will change the state of a bistable device is thus:

$$\text{Probability per unit time} = (1/\tau)e^{-E_{sw}/kT}$$

Typical computations may involve many millions of individual memory elements over a period of many hours. Hence we must insist that the probability of spontaneous switching be less than one part in 10^{12} or so, which requires a switching energy energy of the order of 30kT.

Viewing these energetic considerations from a system level, we have established an absolute minimum for the energy required for doing any given computation.

The energy required for a computation has a lower bound given by the minimum switching energy multiplied by the number of elementary switching events which must occur during the computation.

This estimate of minimum energy completely ignores the energy cost of communicating data from one location to another. In many systems, the total communication energy is much larger than the total switching energy.

In realizable electronic systems, the switching cost for elementary storage elements is much larger than the limit given above. In typical 1978 MOS technology, switching an elementary flip-flop requires 10^{-12} Joule, or approximately 10^{8}kT at room temperature. Even a 1/4 micron MOS transistor will require approximately 10^{4}kT; more than 100 times the energy necessary for reliable computation as given above.

Note that this view of computation makes it perfectly clear that there is no possibility of 100% reliable computing systems. There is always a finite chance that some storage element will switch spontaneously due to thermal noise. However, in today's systems, and even foreseeable VLSI systems, the probability of such a random switching event due to thermal noise is much less than that of a failure due to electrical noise, cosmic rays, or mundane device failure mechanisms. In systems with poorly designed timing constraints, synchronizer failures occur many orders of magnitude more frequently than thermal failures. This observation is the origin of the *Seitz Criterion* given in Chapter 7.

Quantum Limits

The thermal limit given above represents only one way in which an immutable law of nature places bounds on what can be physically realized. Other such limitations come from other physical laws. The lower bound on the size of an FET which will operate properly is determined, not by thermal considerations but by the *uncertainty principle* and the *discreteness of electrical charge.* From the uncertainty principle, an electron of mass m will, because of its wave nature, have an uncertainty Δx in its position x related to the uncertainty Δp in its momentum p by:

$$\Delta p \Delta x \approx \hbar$$

The energy is related to the momentum by:

$$E = p^2/2m$$

Hence an energy barrier of thickness δ and height E_b can contain an electron only if:

$$\delta \gg \Delta x \approx \hbar/(2mE_b)^{1/2}$$

For a barrier of height 1eV, Δx is about 0.001 micron. Gate oxides and junction depletion layers must be many times this thickness. In 1978, gate oxide is already less than 0.1 micron thick. We are thus within sight of a fundamental size limitation due to quantum phenomena.

Granularity of Charge

An even more severe limit results from the discreteness of impurity ion charges in the depletion layer under the FET channel. Let us attempt to reduce all voltages V and distances d by the same factor. A charge layer of q charges per unit area produces an electric field. This field across a depletion layer of thickness d results in a voltage V:

$$V \propto qd$$

The charge is due to impurity ions of density N per unit volume. Hence:

$$q \propto Nd$$

The voltage is therefore proportional to the square of the depletion layer thickness:

$$V \propto Nd^2$$

In order to scale both V and d by the same factor, N must therefore be proportional to 1/d.

The total number of charges in the channel is the number per unit volume times the volume of the region under the gate. By our scaling convention, all volumes must be proportional to d^3:

$$N_{tot} \propto Nd^3 \propto d^2$$

For randomly distributed impurities, the expected statistical variation of the total number N_{tot} is:

$$\Delta N_{tot} = (N_{tot})^{1/2} \propto d$$

Or a fractional variation of:

$$\Delta N_{tot}/N_{tot} \propto 1/d$$

This statistical variation of the number of impurity ions under the channels of different transistors results in a similar distribution of threshold voltages:

$$\Delta V_{th}/V_{th} = \Delta N_{tot}/N_{tot} \propto 1/d$$

The variation in threshold voltages thus becomes larger as devices become smaller. A detailed treatment of this effect is given in [Ref. R3 of Ch.1], which concludes that a device of ¼ micron channel length described in [Ref. 3 of Ch.1] would have an expected variation in its threshold voltage of $\approx$.08 Volts. There are techniques which can greatly reduce this statistical variation. A thin film of undoped silicon just under the gate oxide will largely isolate the threshold voltage from the granularity of charge in the substrate. Such a structure complicates an already difficult task of sub-micron fabrication. It therefore appears that, aside from the fabrication process, the first barrier we face in the sub-micron FET world is a difficulty in scaling voltages to low enough values. We consider the fundamental limit on supply voltage in the next section.

Voltage Limit

We have seen that a storage device must exhibit a maximum in its power curve in order to retain information. There are two independent ways in which this maximum may become too small. The first is that the elementary logic gates may become too small to store enough energy. We see that this limit does not constrain ordinary FET logic since FET gate lengths must be greater than 1/4 micron for other reasons. Another way is that the operating voltage may become too small to assure that the gain of an elementary circuit exceeds unity. Semiconductor technology is evolving

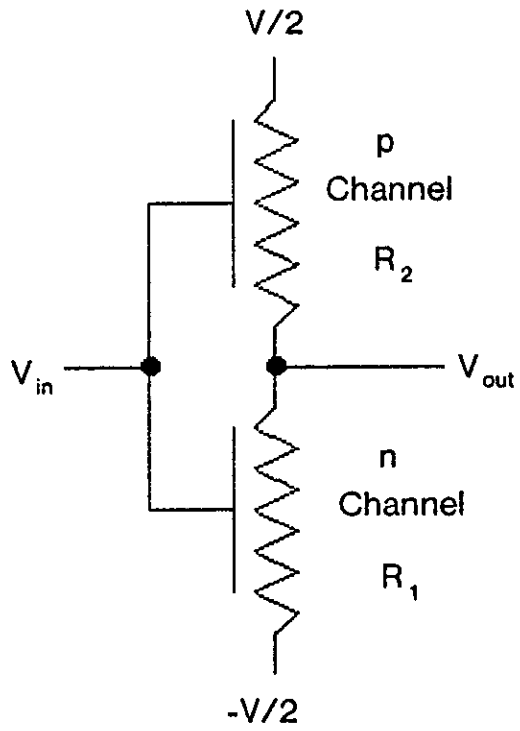

Fig. 5 Conceptual Model of a Complementary Inverter at very low Voltages

under a scaling law in which operating voltage must be decreased along with device dimensions. Hence it is important to establish a lower limit on the operating voltage of FET circuits.

As mentioned in Chapter 1, when a MOS device is operated near its threshold, the channel resistance R_{ch} is exponentially dependent upon the gate voltage V_g:

$$R_{ch} \propto e^{-qV_g/nkT}$$

The factor n is due to the substrate effect, and is approximately 1.2 for most processes.

A model of a complementary device (such as a CMOS inverter) is shown in figure 5. The resistances R1 of the lower n-channel device and R2 of the upper p-channel device are exponentially dependent on the input voltage V_{in} as follows:

$$R_1 = R\, e^{-qV_{in}/nkT}, \qquad R_2 = R\, e^{qV_{in}/nkT}$$

Therefore the output voltage is:

$$V_{out} = [V/2(R_1+R_2)]\,[R_1-R_2]\,V_{in}$$

We are interested in the gain near the switching threshold, which because of the supply voltage convention is at $V_{in}=0$. We may expand the exponentials as a power series and ignore all but the first order terms in V_{in}.

$$V_{out} \approx -(qV/nkT)V_{in}$$

The gain of the circuit is thus equal to qV/nkT. Hence realistic supply voltages for complementary circuits should be a few kT/q. At room temperature $(kT/q) \approx 25$ mV. Ratio logic families, such as nMOS, can be analyzed by the same technique. Since they have only one non-linear device rather than two, their gain is approximately half that given above. They will therefore require twice the supply voltage required by complementary devices. Routine CMOS circuits with $\approx$5 micron geometries operate with a 5V supply. Scaling in a straightforward way, we would expect ¼ micron devices to operate with a ¼ Volt supply.

While the gain of such a circuit would be adequate if all its transistors had the same threshold voltages, it is possible that the pullup transistor of an inverter could have a particularly high threshold voltage, while its companion had a particularly low threshold. If the difference in threshold voltages exceeded the supply voltage V_0, the device output would always remain in one state. The probability of such an occurrence, computed from the variation in threshold

mentioned in the last section is:

$$P = e^{-2V_0/\Delta V_{th}}$$

We might for example require that, even in a VLSI system containing 10^7 inverters, the probability of all the system's transistors being within threshold limits be greater than 0.9. Such a criterion would require a supply voltage of $\approx$0.7V. Unless special attention is directed toward reducing threshold variations, systems with ¼ micron device geometries will be forced to operate with higher supply voltages than the straightforward scaling would indicate. However the inherent nonlinearity of the FET near threshold lowers the effect of threshold variations, and system operation at a supply voltage in the 100 - 200 mV range appears feasible.

An Example

In this section we will apply physical considerations to the comparison of two very different technologies for constructing computational systems. The technologies selected for this example are based on (i) semiconductor FET devices and (ii) Josephson junction devices. The material presented in this example demonstrates the importance of considering not only device physics and device design, but also system physics and system architecture, when making such comparisons.

Several types of limits on the performance of semiconductor FET logic families have been noted in the foregoing discussion; those dealing with the temperature of operation, those arising from quantum phenomena, those associated with the granularity of charge in the semiconductor substrate, and voltage constraints arising from gain considerations. Of these, the limit due to quantum phenomena appears the least restrictive.

It would thus appear that a physical process not involving a doped semiconductor and operating at very low temperature would merit serious study. Superconducting logic families have, for this reason, attracted much attention. Information is stored as a magnetic flux trapped in a superconducting ring, and is switched by means of a Josephson (or similar) junction. Devices have been demonstrated which exhibit very fast switching times and low operating power. It is important to understand the relative merits of such a radically different technology from the point of view of overall system design. We should therefore find some way to compare it directly with semiconductor technology, and to extend the comparisons to scaling into sub-micron dimensions.

In real systems, the cost of energy, and energy conversion and distribution, often exceeds the cost

of the chips themselves. Hence any discussion of the cost of computation must include the energy cost of individual steps of the computation process. The fundamental figure of merit of a logic device is its *switching energy* discussed previously. This quantity is a measure of the power-delay product of the technology. Propagation delay can be traded off against power dissipation over a wide range in any given technology, but their product cannot be reduced below the switching energy. In a charge controlled semiconductor device such as the MOSFET, the irreducible switching energy is $E_{sw} = C_g V^2/2$, for gate capacitance C_g and supply voltage V.

In a superconducting device $E_{sw} = LI^2/2$, where L is the inductance of the superconducting loop plus the associated junction, and I is the supply current. In both technologies, parasitics will increase E_{sw} to several times the values computed for minimum devices. However, for purposes of comparison, we will consider only the minimum devices themselves.

Since all energies in both types of logic are multiples of kT, it might appear that operating a computer at very low temperatures would reduce the total power required. That this is not the case is easily demonstrated. Suppose that to perform a computation a machine dissipates energy $E_L = nkT_L$ as heat at some low temperature T_L. To maintain the low temperature, this heat energy must be transported to and released at room temperature, T_H, by some refrigerator. The total energy to run the system is equal to E_L plus the work required to run the refrigerator. Thermodynamics shows us that a refrigerator operating on the Carnot cycle requires the least amount of work input per unit of heat transported from the low temperature environment to the high temperature environment.[3] On input of work W, a Carnot refrigerator can transport, from the T_L to T_H environments, a quantity of heat energy Q given by:

$$Q/W = T_L/(T_H - T_L)$$

Thus the work W required to transport E_L from T_L to T_H is in general:

$$W \geq E_L(T_H - T_L)/T_L$$

The total energy , E_{tot}, required for the computation is therefore:

$$E_{tot} \geq nkT_L + nkT_L[(T_H - T_L)/T_L] = nkT_H$$

As T_L is lowered, the switching energy is lowered, but the work input to the refrigerator must be increased by at least an equal amount. The total energy cost, including that necessary to run the refrigerator, is thus independent of the temperature of the computer's switches. This energy cost

is, at minimum, identically equal to nkT at the temperature of the ultimate heat sink. In some space applications a heat sink at very low temperatures is available. However, for terrestrial computers, refrigerating electronic devices in order to reduce the energy of computation is logically equivalent to constructing a perpetual motion machine. For this reason, we will use kT at the heat sink temperature in system energy calculations, independent of the actual temperature at which the switching devices operate.

Now we turn to the details of the technology comparison. The switching energy of MOSFET logic is: $E_{sw} = C_g V^2/2$. The most straightforward MOSFET scaling results from reducing all dimensions by the same scaling factor. If this type of scaling is applied to the MOS family, the gate capacitance decreases linearly with the scaling factor. In order to keep the electric fields constant, the supply voltage is scaled by the same scaling factor. The switching energy is thus reduced by the *third power* of the scaling factor, as illustrated in the top curve in figure 6. The lower size limit shown is a conservative estimate set by device physics factors previously discussed.

Were it possible to build FET devices which operated with one electronic charge on their gate, their performance would not benefit from scaling to smaller dimensions. In such a device, the switching energy can be expressed in terms of q_0, the charge of the electron:

$$E_{sw} = CV^2/2 = q_0^2/2C$$

Since C decreases as the device dimensions are scaled down, the switching energy actually increases. This relationship illustrates a general principle: *A logic device working at its quantum limit requires a higher switching energy as the dimensions of the device are made smaller.*

Even at present dimensions, superconducting logic operates at or near its quantum limit. The flux in a superconducting ring must be an integral multiple of the flux quantum $\Phi_0 \simeq 2 \times 10^{-15}$ Webers. The switching energy for a device operating with one flux quantum can be written as:

$$E_{sw} = LI^2/2 = \Phi_0^2/2L$$

Note that the inductance $L = \Phi_0/I$ is directly proportional to the size of the loop. The above dependence for superconducting logic is illustrated in the bottom curve in figure 6. The lower size limit shown is set by the penetration depth λ of the superconductor[4]. Magnetic field strength decreases with distance, x, into the superconductor as $e^{-x/\lambda}$. If the thickness of the superconducting ring is less than a few λ, the ring cannot localize the flux within it. A typical

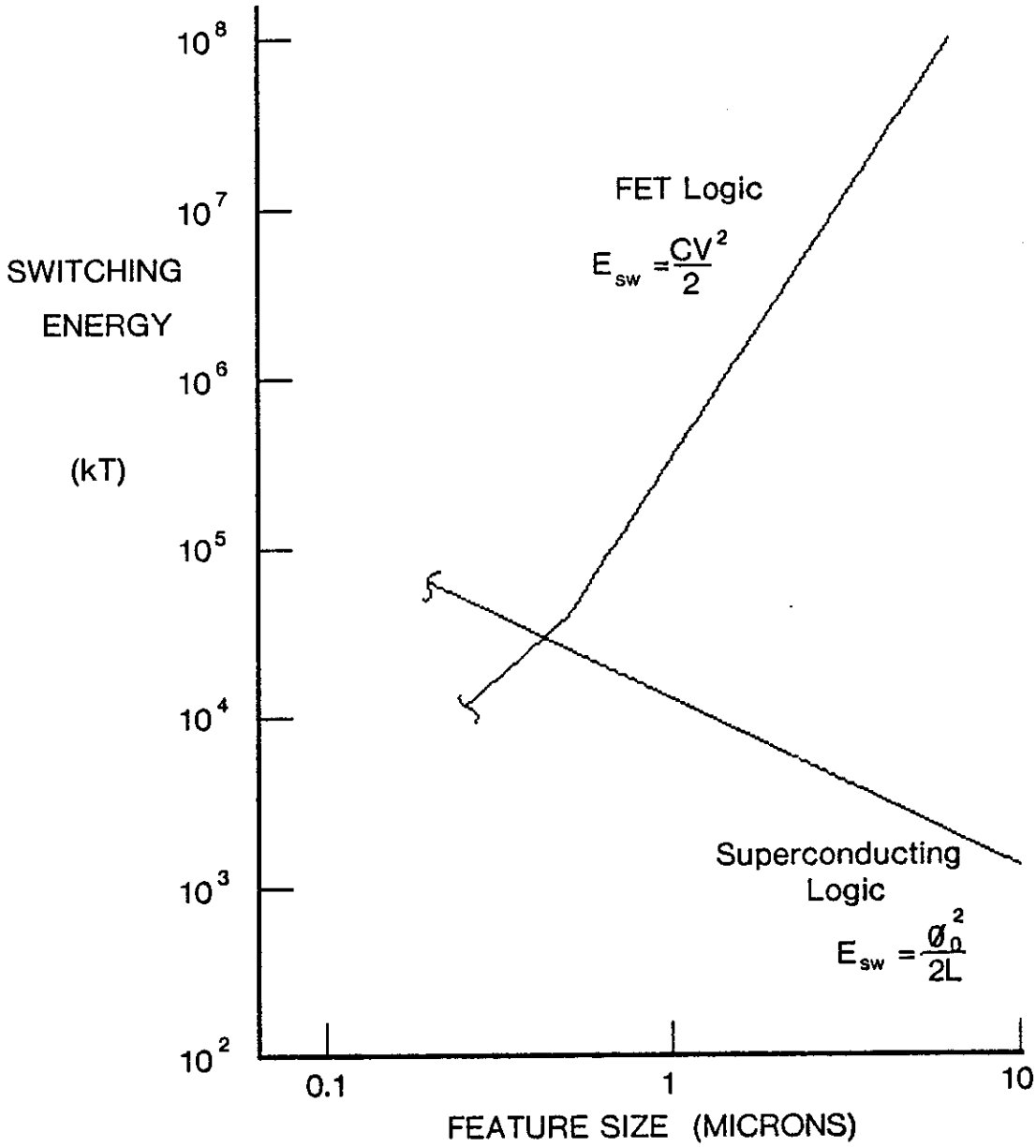

Fig. 6 Comparison of FET and Superconducting Logic

value of λ is 0.1 micron.

Comparing the upper and lower curves, it is clear that, when an accounting is made of the total energy, and when the effects of scaling to sub-micron dimensions are taken into account, room temperature FET logic is a remarkable technology. At achieveable sub-micron dimensions, it can actually outperform its superconducting counterpart. Lower switching energies in the superconductor technology can be achieved only by sacrificing density. This trade-off may be desirable under some circumstances. It seems more likely, however that maximum computation per unit cost will be achieved by jointly minimizing switching energy and maximizing circuit density.

The absolute speed attainable with the superconducting logic is, however, considerably better than that of its FET counterpart. For a critically damped Josephson junction, the time response τ is

$$\tau \simeq (LC)^{-\frac{1}{2}} \ ,$$

where L is the loop inductance used above, and C is the junction capacitance. Since the normal resistance of the Josephson junction varies exponentially with dielectric thickness, the thickness can be assumed approximately constant as the devices are scaled. Hence the delay time τ will scale down as the 3/2 power of the scaling factor. For the FET, the oxide thickness must be scaled, and the delay time varies linearly with the scaling factor.

At 1 micron feature size, for example, the switching time of a superconducting device is $\approx 2 \times 10^{-13}$ sec, while for a FET with the same feature size the transit time is $\approx 10^{-11}$ sec.

One basic problem with low temperature logic is that the lower switching energy levels result in poor noise immunity. They therefore require better shielding to reduce the effect of external electromagnetic occurrences to a level well below the switching energy.

Another problem is that the low switching energy creates a mismatch to the outside world for which a penalty in additional power consumption has to be paid, since the drivers to the outside world consume a large amount of power, and introduce extra delays. As long as information is not required to exit the low temperature environment, chip to chip communication can be done at high bandwidth. Note that in this respect, superconducting logic is superior since it is somewhat better matched to the impedance of transmission lines than is FET logic. In any event, exponentially staged drivers are required when driving from the low energy environment to the

outside world, as discussed in chapter 1. These drivers introduce a minimum delay τ_{dr}:

$$\tau_{dr} \geq \tau e \, \mathrm{Ln}(Y) ,$$

where Y is the ratio of energy required at the destination to that of the elementary logic device. If the switching energy of a logic element is a factor of 100 smaller due to operation at low temperature, a factor of at least 10 in driver delay is introduced. Furthermore, the dissipation of the last stage of the driver is determined by the energy level necessary *in the outside world*, not in the low temperature environment. The cost of this driving energy is at least 100 times higher than that for a room temperature driver of the same capability, due to the constraints imposed by the laws of thermodynamics.

It is important to recognize that the trade-off between power and delay time extends to much shorter times for Josephson devices than than it does for FET's. The speed advantage of the Josephson devices, in the scaled environment of the future, will be about a factor of fifty. Although their switching energy will be about the same as that of FETs, we would have the option of inputting fifty times more power into a system composed of Josephson devices, and then being able to switch them fifty times faster than the fastest FETs.

Architects comparing alternative technologies for building computing systems take into account many costs other than just total switching energy. The weights assigned to the various factors usually depend upon their proximity to absolute constraints imposed by physical law or by system performance and cost considerations. In certain situations, we may be perfectly willing to pay the price for large increments in energy, energy conversion equipment, mass, volume, and structural and operational complexity, in order to achieve an increment of system performance.

Suppose, for example, we now had to specify a very high performance general purpose computer for the late 80's or early 90's. Since switching speed translates directly into time performance in the classical stored program computer, we might see no other alternative for high performance than a machine based on superconducting devices. Such a decision recognizes that no present alternatives exist for trading off processing speed against concurrency in multiple processors for general purpose computation. That such alternatives must ultimately exist is of course evident by observation of the information processing capability of living organisms.

Superconducting devices meet the requirement for high speed in the classical computer, and a number of machines based on that technology will likely be built before viable high concurrency

alternatives appear. However, in the longer term, in applications where mass, volume, structural complexity, and cost are real constraints, semiconductor devices operated at heat sink temperature will generally have the advantage. Thus, the switching technology likely to dominate the terrestrial environment, used for personal computing and personal communications on a vast scale in an enormous number of different applications, is semiconductor technology. Recall that semiconductor technology itself may benefit in a variety of ways from low temperature operation [Ref.7 of Ch.1], as for example in the reduction of subthreshold current in submicron MOSFETs.

Energy Management

{ in preparation }

Discreteness in Quantum Mechanical Systems

{ in preparation }

Conclusion

We opened this book with a discussion of the physical properties of elementary switching devices. We have now closed with a discussion of fundamental physical principles which profoundly influence the higher-level properties of computing systems.

The communication of information over space and time, the storage of information by change of state at storage sites, and the transport of energy into and heat out of systems depend not only on abstract mathematical principles, but also on physical laws. The generation and synthesis of very large scale systems, whether artificial or natural, proceed under and indeed are directed by the constraints imposed by the laws of physics.

We look forward to a time when quantitative measures can be given for the true cost and complexity of any required computation. At present we are very far from this goal. The examples of this chapter do, however, serve to illustrate that concrete physical arguments can be applied to the properties of information systems. We hope others will provide insights and examples in this important area of investigation, for reporting in future editions of this text.

References

1. R. P. Feynman, R. B. Leighton, M. Sands, "The Feynman Lectures on Physics", vol. 1-3, Addison-Wesley, 1963-65. Highly readable and understandable, these are an excellent series of tutorial textbooks.

2. R. B. Leighton, "Principles of Modern Physics", McGraw-Hill, 1959. This is an excellent reference text.

3. F. W. Sears, "An Introduction to Thermodynamics, the Kinetic Theory of Gases, and Statistical Mechanics", Addison-Wesley, 1953. This is a classic text on the subject.

4. E. A. Lynton, "Superconductivity", Third edition, Chapman and Hall Ltd., 1969, and Science Paperbacks, 1971.

5. R. W. Keyes, R. Landauer, "Minimum Energy Dissipation in Logic", IBM Journal of Research and Development, vol.14, pp. 152-157, 1970.