

THE PHYSICAL LIMITS OF COMPUTING

Many of the fundamental limits on information processing from thermodynamics, relativity, and quantum mechanics are only a few decades away. Novel physically motivated computing paradigms such as reversible computing and quantum computing may help in certain ways, but even they remain subject to some basic limits.

Computational science and engineering and computer science and engineering have a natural and long-standing relationship. Scientific and engineering problems often require extreme computational power, thereby driving the development of new bit-device technologies and circuit architectures. In return, research using computational methods fosters more efficient computing systems.

Impelled by this positive feedback loop between increasing demand and improving technology, computational efficiency has improved steadily and dramatically since computing's inception. When looking back at the last 40 years, and forward to the next 10 or 20, this empirical trend is often characterized with reference to the famous Moore's Law,^{1,2} which describes the increasing density of microlithographed transistors in integrated semiconductor circuits.³ Naturally, we wonder how far we can reasonably hope this fortunate trend will take us. What are the ultimate limits? *Are* there limits? When semiconductor technology reaches its technology-specific limits, can we hope to maintain the

curve by jumping to some alternative technology and then to yet another, ad infinitum? Or, is there a foreseeable and technology-independent endpoint in sight?

Obviously, forecasting future technological developments is always a difficult and risky proposition. However, 20th-century physics has given forecasters a remarkable gift in the form of the sophisticated modern understanding of fundamental physics embodied in the Standard Model of particle physics. Although, of course, many interesting unsolved problems remain in physics at higher levels,⁴ all available evidence tells us that the Standard Model, together with general relativity, explains the foundations of physics so successfully that no currently experimentally accessible phenomenon fails to be encompassed by it. We expect the fundamental principles of modern physics have "legs"—that is, they will last us many decades as we try to project what will and will not be possible in the coming evolution of computing. By taking our best theories seriously, and exploring the limits of what we can engineer with them, we test the bounds of what we think we can do. If our present understanding of these limits eventually turns out to be seriously wrong, then the act of pushing against those limits will most likely lead us to that very discovery.⁵

Forecasting future limits, even far in advance,

is a useful research activity. It gives us a roadmap suggesting where we can expect to go with future technologies and helps us know where to look for advances. Fortunately, by considering fundamental physical principles and reasoning in an abstract and technology-independent way, we can arrive at several firm conclusions regarding upper bounds on the limits of computing. In this article, I review fundamental, technology-independent limits because it would take too much space to survey the technology-specific limits of all present and proposed future computing technologies.

Physical information and entropy

Before we can talk sensibly about information technology in physical terms, we must define information in physical terms. For the purposes of discussing the limits of information technology, the relevant definition relates closely to the physical quantity known as *entropy*. Entropy is really just one variety of a more general sort of entity, which we call *physical information*, or *information* for short. This abbreviation is justified because all information that we can manipulate is ultimately physical in nature.⁶

Rudolph Clausius introduced the entropy concept in thermodynamics in 1850, before it was understood to be an informational quantity. Ludwig Boltzmann later identified the maximum entropy S of any physical system with the logarithm of its total number of possible, mutually distinguishable states. (This discovery is carved on his tombstone.) I call this same quantity the total physical information in the system, for reasons soon to become clear.

In Boltzmann's day, presuming that the number of states for typical systems was finite and allowed a logarithm was a bold conjecture. Today, we know that operationally distinguishable states correspond to orthogonal quantum state vectors, and that the number of these for a given system is well-defined in quantum mechanics and finite for finite systems.

Any logarithm, by itself, is a pure number, but the logarithm base that we choose in Boltzmann's relation determines the appropriate unit of information. Using base 2 gives us the information unit of 1 bit (b), whereas the natural logarithm (base e) gives a unit that I call the *nat*, which is simply $(\log_2 e) \approx 1.44$ b. In situations where the information in question happens to be entropy, the nat is more widely known as Boltzmann's constant k_B or the ideal

gas constant R , depending on the context.

We can associate any of these units of information with physical units of energy divided by temperature because temperature itself can be defined as a measure of energy required per increment in the log state count, $T = \partial E / \partial S$, holding volume constant. For example, we can define the temperature unit 1 Kelvin as a requirement of 1.38×10^{-23} Joules (or $86.2 \mu\text{eV}$) of energy input to increase the log state count by 1 nat—that is, to multiply the number of states by e . A bit, meanwhile, is associated with the requirement of 9.57×10^{-24} Joules ($59.7 \mu\text{eV}$) energy per Kelvin to double the system's total state count.

This is information, but what distinguishes entropy from other kinds of information? The distinction is fundamentally observer-dependent, but in a way that is well defined and that coincides for most observers in simple cases.

Let *known information* be the physical information in that part of a system whose state is *known* by a particular observer, and entropy be the information in the part that is *unknown*. We can clarify the meaning of "known" by saying that system A (the observer) knows the state of system B (the observed system) to the extent that some part of A's state (some record or memory) correlates with B's state and that the observer can access and interpret this record's implications regarding B's state.

To quantify things, the maximum known information or maximum entropy of any system is the log of its possible number of distinguishable states. If we know nothing about the state, all the system's physical information is entropy from our viewpoint. But, as a result of preparing or interacting with a system, we might learn something more about its actual state besides it being one of the N states originally considered "possible" (see Figure 1).

Suppose we learn that the system is in a particular subset of $M < N$ states; only the states in that set are then possible, given our knowledge. Thus, the system's entropy from our new viewpoint is $\log M$, whereas to someone without this knowledge, it is $\log N$. For us, there is $(\log N) - (\log M) = \log(N/M)$ less entropy in the system. We say we now know $\log(N/M)$ more information about the system because $\log(N/M)$ of the physical information that it contains is known information from our viewpoint. The remaining $\log M$ amount of information—the physical information still unknown in the system—we call *entropy*.

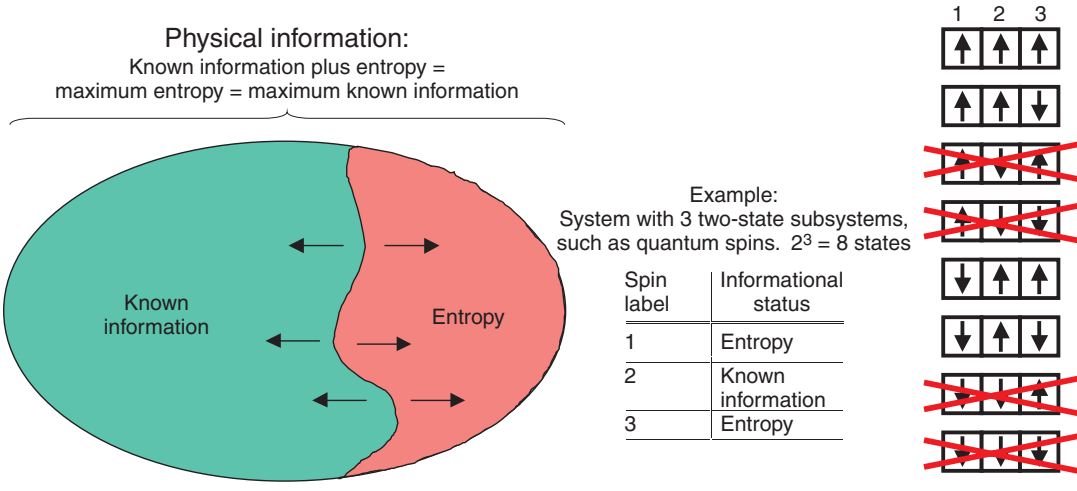


Figure 1. Physical information, entropy, and known information. Quantum mechanics helps determine the exact number N of states. We define the total physical information in a system as the logarithm of this number of states; we can express it equally well in units of bits or nats. On the right, a system of 3 two-state quantum spins shows $2^3 = 8$ distinguishable states. It therefore contains a total of 3 bits = $2.08 k_B$ of physical information. Relative to some knowledge about the system's actual state, the physical information has a part that is determined by that additional knowledge (*known information*) and a part that is not (*entropy*). In the example, suppose we learn that the system is not in any of the four crossed-out states. In this case, the 1 bit ($0.69 k_B$) of physical information associated with spin number 2 is then known information, whereas the other 2 bits ($1.39 k_B$) of physical information in the system are entropy.

Claude Shannon showed how to appropriately generalize the definition of entropy to situations where our knowledge about the state x is expressed not as a subset of states, but as a probability distribution p_x over states. In this case, the entropy is

$$H = -\sum_x p_x \log p_x .$$

The known information is then $(\log N) - H$. The Boltzmann definition of entropy is thus the special case of Shannon entropy where p_x happens to be a uniform distribution over all N states (see Figure 2), also called the maximum entropy distribution.⁷

Known information and entropy are two forms of the same fundamental conserved quantity, somewhat analogous to kinetic versus potential energy. We can convert a system's entropy to known information by measurement, and known information into entropy by forgetting or erasing it. However, the sum of the two in a given system is always a constant, unless the maximum number of possible states in the system is itself changing, which could happen if the system's size changes, or if energy is added or removed. For example, in an expanding universe,

the number of states (and thus the total physical information) is increasing, but in a small, local system with constant energy and volume, it stays constant.

Saying that we can convert entropy to known information through observation seems like a contradiction of the second law of thermodynamics; entropy always increases in closed systems. However, if we measure a system, then it isn't completely closed from an informational viewpoint. The measurement requires an interaction that manipulates the measurement apparatus' state in a way that depends on the system's state. From a global viewpoint, the entropy, even if extracted from the original system through measurement, still exists and is still entropy (see Figure 3).

Boltzmann developed his definition of entropy in the context of classical mechanics by making the ad hoc assumption that even the seemingly continuous states of classical mechanics were somehow discretized into a finite number that admitted a logarithm. Max Planck and the entire subsequent development of quantum mechanics vindicated this notion, showing that the world was discretized in the relevant respects. The entire classical understanding of the rela-

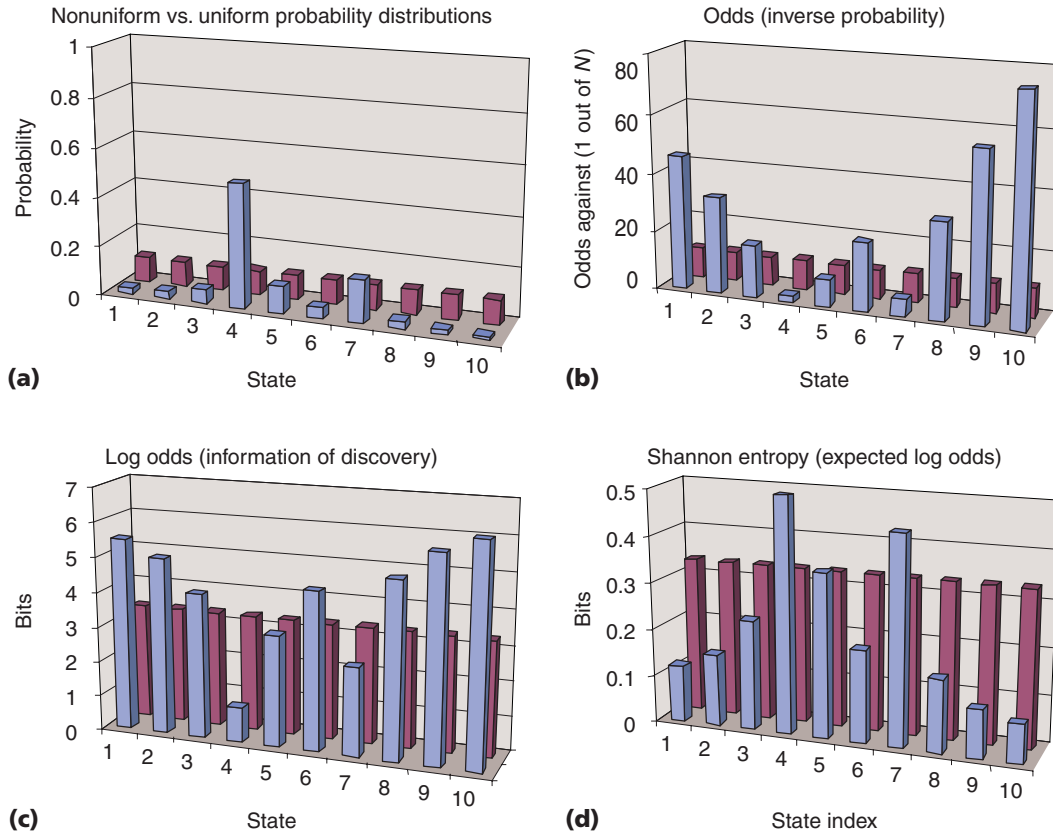


Figure 2. An example of Shannon’s generalization of Boltzmann entropy for a system with 10 distinguishable states. The blue bars correspond to a specific nonuniform probability distribution over states; the purple bars show the case with a uniform Boltzmann distribution. (a) The two probability distributions. (b) An inversion of the probability to get the odds against the state; state 4 is found in one case out of two, whereas state 10 appears in one case out of 70. (c) The logarithm of this number of cases is the information gain if we actually encounter this state—in state 4, we gain 1 bit; in case 10, more than 6 bits ($2^6 = 64$). (d) Weighting the information gain by the state probability gives the expected information gain. Because the logarithm function is concave-down, a uniform distribution minimizes the expected log-probability, maximizes its negative (the expected log-odds, or entropy), and minimizes the information (the expected log-probability, minus that of the uniform distribution).

tions between entropy, energy, temperature, and so on remained essentially valid, forming the whole field of quantum statistical mechanics—a cornerstone of modern physics. Only the definition of entropy had to be further generalized.

Partially known states in quantum mechanics are described by a generalization of a probability distribution called a *mixed state* or *density operator*, which can be represented with a *density matrix*. However, entropy is defined for these more

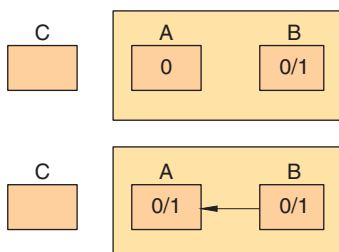


Figure 3. Entropy and measurement. Suppose system B, which contains 1 bit of entropy (two possible states, labeled 0 and 1), is measured by system A (arrow). System A is thus correlated with B, and B’s entropy is now information from A’s viewpoint. However, from outside observer C’s viewpoint, the combined system still has 1 bit of entropy because it could be either in state (A = 0, B = 0) or state (A = 1, B = 1). The number of the whole system’s possible states can’t decrease from the viewpoint of an outsider who isn’t measuring the state. However, it can increase, for example, if C loses track of the interactions occurring between A and B, in which case all four joint states of the AB system become possibilities from C’s viewpoint.

Figure 4. A density matrix representation of probabilistic mixtures of quantum states. The rows and columns of ρ are indexed by a maximal set of mutually distinguishable states (a basis). Each diagonal element ρ_{ii} gives the probability of basis state i . The off-diagonal elements ρ_{ij} , $i \neq j$ are complex numbers that specify quantum coherences between the basis states. Any density matrix ρ has a unique basis such that when ρ is reexpressed in that basis, the resulting matrix ρ' is diagonal and represents a classical mixture of $\leq n$ basis states. The basis-independent von Neumann entropy of a mixed state is given by $H = -\text{Tr } \rho \ln \rho$. This quantity is exactly the same as the Shannon entropy of the probability distribution specified along the diagonal of the diagonalized density matrix ρ' . The von Neumann entropy of a density matrix ρ is always less than or equal to the Shannon entropy of ρ 's own diagonal, which in turn always less than or equal to the Boltzmann entropy, $\ln n$.

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{bmatrix} \Rightarrow \rho' = \begin{bmatrix} \rho'_{11} & 0 & \cdots & 0 \\ 0 & \rho'_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho'_{nn} \end{bmatrix}$$

von Neumann entropy:

$$\begin{aligned} H(\rho) &= -\text{Tr } \rho \ln \rho \\ &= -\text{Tr } \rho' \ln \rho' \\ &= -\sum_i \rho'_{ii} \ln \rho'_{ii} \\ &\leq -\sum_i \rho_{ii} \ln \rho_{ii} \\ &\leq \ln n \end{aligned}$$

complex objects in a way that remains perfectly consistent with the more restricted cases Boltzmann and Shannon addressed (see Figure 4).

Information storage limits

Now that we know what information physically *is* more or less, let's talk about some of the limits that we can place on it, based on known physics.

As an abstract mathematical entity, an arbitrary quantum state or wavefunction could require infinite information to describe precisely. In principle, there is a continuous, uncountable set of possible wavefunctions, but there are only countably many finite descriptions,

or computable wavefunctions. Ever since Boltzmann, the key definition for physical information is not the number of states that might mathematically exist, but the number of operationally distinguishable ones. Quantum mechanics gives distinguishability a precise meaning: two states are 100 percent distinguishable if and only if, considered as complex vectors, they are orthogonal.

A textbook result of quantum statistical mechanics is that the total number of orthogonal states for a system consisting of a constant number of noninteracting particles is roughly given by the numerical volume of the particles' joint configuration space, or *phase space* (whatever its shape).⁸ (The volume of phase space must be expressed in units where Planck's unreduced constant \hbar is equal to 1.) So, as long as the number of particles is finite, and the volume of space occupied by the particles and their total energy is

bounded, then even though the number of point particle states and possible quantum wavefunctions is uncountably infinite, the amount of information in the system is finite.

This model of a constant number of noninteracting particles is a bit unrealistic because in quantum field theory (the relativistic version of quantum mechanics), particle number is not constant; particles can split (radiation) and merge (absorption). To refine the model, we must think about possible field states with varying numbers of particles. However, this still does not fundamentally change the conclusion of finite information content for any system of bounded size and energy. In independent papers, Warren Smith⁹ and Seth Lloyd¹⁰ have given an excellent description of the quantitative relationships involved.

In his paper, Smith argues for an upper bound to entropy S per unit volume V of

$$\frac{S}{V} = \left(\frac{q}{2}\right)^{1/4} \frac{16\sqrt{\pi}}{3 \cdot 60^{1/4}} \left(\frac{c}{\hbar} \cdot \frac{M}{V}\right)^{3/4} \text{ nat},$$

where q is the number of distinct particle types (including different quantum states of a given particle type), c is the speed of light, \hbar is Planck's constant, and M is the system's total (gravitating) mass energy. As a numerical example, using only photons with two polarization states (argued to be the dominant entropy carriers at high temperatures), a 1-m³ box containing 1,000 kg of light could contain at most 6×10^{34} bits, or 60 kb per cubic Ångstrom (1Å = 10⁻¹⁰ m; 1Å³ is roughly a hydrogen-atom-sized volume). However, achieving this limit for stable storage is probably unrealistic because light with this mass density—that of water—would have a temperature of nearly a billion degrees and exert a pressure on the order of 10¹⁶ pounds per square inch.

Lloyd presents a bound nearly identical to Smith's, derived from similar arguments. It differs from Smith's only in that it is tighter by the small constant factor of $2\sqrt{2}$. Lloyd presents the

example of a 1-kg, 1-liter “ultimate laptop”—again, at the density of water—for which, using the same two-state photon assumption as Smith, the maximum entropy would be 2.13×10^{31} bits (basically, the same entropy density as in Smith’s example, less the factor of $2\sqrt{2}$).

These field-theory-based limits do not account for the effects of gravity and general relativity. Based on general grounds, Jacob Bekenstein proved a much looser entropy limit for a system of given size and energy that holds even when accounting for general relativity:¹¹

$$S < 2\pi ER / \hbar c,$$

where E is total energy, and R is the system’s radius. The only systems known to actually attain this entropy bound are black holes. (A black hole’s “radius” has a standard, meaningful definition even in the severely warped spacetime in and around the hole.) Interestingly, a black hole’s entropy is proportional to its surface area, not to its volume, as if all the information about the hole’s state were stuck at its surface (event horizon). A black hole has exactly 1/4 nat of entropy per square Planck length of surface area (a Planck length is a fundamental unit of length equal to 1.6×10^{-35} m). In other words, the absolute minimum physical size of 1 nat’s worth of information is a square exactly 2 Planck lengths on a side.

The Bekenstein bound is truly enormous. A hypothetical 1-m radius, mainframe-sized machine that achieved this bound would have an average entropy density throughout its volume (calculated assuming a spherical shape and ignoring spacetime curvature) of 10^{39} bits per cubic Ångström, much higher than the limit for the water-density machines described earlier. However, this “machine” would also be a black hole with roughly the mass of Saturn. Needless to say, this is not very practical.

Of course, both the field-theory and Bekenstein bounds on entropy density are technology-independent upper bounds. Whether we can come anywhere close to reaching them in any realistic computing technology is another question entirely. Both of these bounds require considering all the possible states of quantum fields. However, it seems impossible to constrain or control a field’s state in definite ways without a stable surrounding or supporting structure. Arbitrary field states in general are not stable structures. For stability, we must use long-lived, bound particle states, such as we find in molecules, atoms, and nuclei.

How many bits can we store in an atom?

Nuclei have an overall spin orientation that is encoded using a 2D state-vector space, so the spin only holds 1 bit of information. Apart from its spin variability, at normal temperatures, a typical nucleus is frozen into its quantum ground state; it can only contain additional information if it is excited to higher energy levels. However, excited nuclei are not stable—they are radioactive and decay rapidly, emitting high-energy, damaging particles. Bad news for the computer user’s safety!

Electron configuration is another possibility. Outer-shell electrons have spin variability and excited states that, although still unstable, at least do not present a radiation hazard. Furthermore, a given atom’s ionization states could be reasonably stable in a sufficiently well-isolated environment. This gives us another few potential bits.

The choice of nuclear species in the atom in question presents another opportunity for variability. However, there are only a few hundred reasonably stable isotopes, so at best, even if we have a storage location that can hold any isotope, we gain at most an additional 8 bits or so.

An atom in a solid is in a potential energy well, relative to its neighbors, and generally has six restricted degrees of freedom, three of position and three of momentum. At normal temperatures, each contributes $k_B/2$ to its heat capacity, which in turn contributes an equivalent amount of entropy for each factor of e increase in temperature beyond the regime where the excited states become accessible. This gives us a few more bits per atom encoded in these vibrational states, but *phonons* (the quantum “particles” of mechanical vibration) can easily dissipate out into any mechanical supporting structure, so they do not represent stable storage.

Of course, an arbitrarily large number of bits could be encoded in an atom’s position and momentum along unrestricted degrees of freedom—for example, in infinitely large open spaces. However, given bounded spaces and energies, the limit of log phase-space-volume mentioned earlier still limits the entropy. Because entropy per atom grows only with log volume, entropy density per volume actually shrinks with increasing volume. So, spreading atoms out, although it increases entropy per atom by some small number of bits, does not increase entropy density. If we want to maximize information density using atoms, we should stick with dense, solid materials, which also have the advantage of stability.

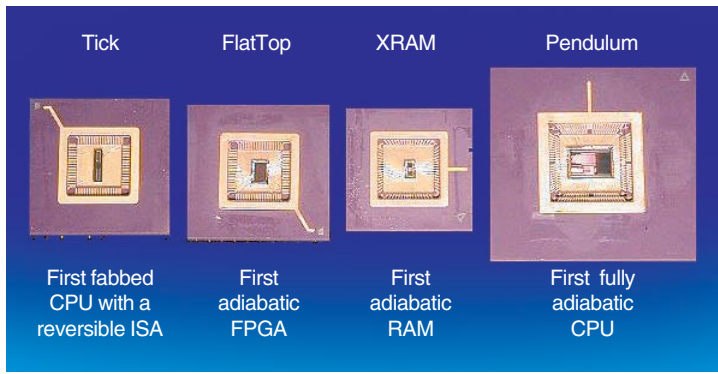


Figure 5. Reversible chips designed at the Massachusetts Institute of Technology between 1996 and 1999. As graduate students, my coworkers (Josie Ammer, Nicole Love, Scott Rixner, and Carlin Vieri) and I designed, outsource-fabricated, and tested these four proof-of-concept reversible chips using the Split-level Charge Recovery Logic (SCRL) adiabatic CMOS logic family.¹⁶ Tick was a benchmark for comparison purposes, an 8-bit, nonadiabatic implementation of a reversible instruction set architecture. Pendulum was a 12-bit fully adiabatic implementation with a similar ISA designed to achieve much lower power.¹⁷ Before Pendulum, we built the much simpler FlatTop, a fully adiabatic programmable array of 400 simple 1-bit processing elements; arrays of these chips could in principle be programmed to simulate arbitrary reversible circuits in a scalable way.¹⁸ Xram was a small fully adiabatic static RAM chip.

For example, a rough estimate I derived for the entropy density in pure copper suggests that at atmospheric pressures, the actual entropy density falls between 0.5 to 1.5 bits per cubic Ångstrom over a wide range of temperatures, from room temperature up to right below the metal's boiling point. Entropy densities in a variety of other pure elemental materials are also near this level, although copper had the highest entropy density of the materials I studied. We would expect the entropy density to be somewhat greater for mixtures of elements, but not by much.

We can try to further increase entropy densities by applying high pressures. At the moment, the ultimate limits to pressures achievable in stable structures are unclear. The only clear limit I know is the pressure at a neutron star's core, just below the critical mass (approximately 3.2 suns) for black hole collapse, or roughly 10^{30} atmospheres. Of course, stellar-scale engineering is, at best, a very long-term prospect.

Based on these observations, I would be surprised if we could achieve an information density greater than, say, 10 bits per cubic Ångstrom for stable, retrievable storage of digital information any time within the next 100 years. Even at an information density of only $1 \text{ bit}/\text{Å}^3$, a convenient 1-cm^3 lump of material could theoreti-

cally hold 10^{24} bits of information. This quantity is known, in obscure jargon, as 1 *yottabit* or 1 Yb. In more familiar units, it is roughly 100 billion terabytes, much greater than the total digital storage in the entire world today.

Minimum energy for information storage

One of the most important raw resources involved in computing, besides time, space, and manufacturing cost, is energy. When we talk about using up some amount of energy, we usually mean converting *free* energy into *degraded* (low-temperature) heat energy. Due to space limitations, I will not get into detailed definitions of free energy, heat, and so on. Essentially, what this conversion process amounts to is increasing the total entropy content.

Because entropy cannot be destroyed and the total information content of space- and energy-bounded systems is finite, entropy's sustained generation within any bounded system requires its eventual disposal in some external thermal reservoir. This acts as a garbage dump for unwanted entropy. If the thermal reservoir used for entropy disposal has temperature T , then disposing entropy S requires committing an amount of energy ST to the reservoir (by the definition of temperature). This energy is effectively spent at the moment entropy S is generated. Minimizing a system's energy usage thus boils down to minimizing the system's total entropy generation.

Suppose we (some entity, human or computer) have obtained some information of interest by whatever means and we wish to store a permanent record of it in some system's state. How much entropy must be generated in this process?

First, consider that the system in question already contains physical information—which is either known information or entropy, from the entity's perspective. This existing information, whether it is information or entropy, can't just be destroyed. This is because, at the lowest level, physics is reversible, meaning that in a closed system, it transforms one state to another over time in a mathematically invertible way. Reversibility does not require time-reversal symmetry. Particle physics indicates that we must negate all electrical charges and replace all spatial configurations with their mirror images to obtain exactly identical time-reversed laws. (This is called charge-parity-time or CPT symmetry.) Regardless of the precise symmetry, all currently tenable microphysical theories are unchanged in overall form, apart from various sign changes,

with respect to time reversals, so they remain reversible—that is, reverse-deterministic (at the level of wavefunction evolution).

Given this constraint of reversibility, how can we deal with unwanted information? One answer is to just move it to some other system. If an amount of information I is lost track of in this process, this information has become entropy, and so we have increased entropy by an amount $\Delta S = I$, implying a free energy loss of $T\Delta S$, for a (constant-volume) entropy dump at temperature T , by the definition of temperature. For example, an increase of $\Delta S = 1 \text{ bit} = k_B \ln 2$ in the entropy of a system implies investing at least $k_B T \ln 2$ energy in the system as heat. Rolf Landauer first detailed this argument, connecting a loss of known information with a loss of free energy;¹² John von Neumann discussed a similar limit but did not prove it in a 1949 lecture, published posthumously, after Landauer's work.¹³

In present-day commercial computer technology, every act of information storage, meaning every bit operation performed by each of the tens of millions of logic gates in every modern CPU every nanosecond, uses this method of disposal of old information. The storage location's previous contents are treated as unknown, thus generating new entropy, with many orders of magnitude worth of added energy inefficiencies.

However, there is an alternative: The space and energy occupied by old, unwanted (but known) information can be *recycled* by using the knowledge about the old state to transform the old state of the storage element into the new one in a thermodynamically reversible way—by a process that generates no entropy. Charles Bennett first showed the theoretical possibility (consistent with thermodynamics) of reversibly reusing storage for multiple computations that produce useful results,¹⁴ although earlier work by Landauer¹² and Yves Lecerf¹⁵ had danced around the edges of this discovery.

Based on this insight, there is a small but slowly growing research field of *reversible computing*, which is concerned with investigating this alternative of engineering systems that approach the theoretical possibility of zero dissipation as closely as possible. Realizable technologies can indeed approach these predictions, as is suggested, for example, by the adiabatic CMOS circuits that have been a popular topic of investigation and experimentation (for myself and coworkers, among others) in recent years.^{16–19} *Adiabatic* processes are those that asymptotically approach thermodynamic reversibility at low

speeds, although no highly structured system can be perfectly adiabatic because it will always be subject to some background rate of decay toward a less-structured equilibrium ensemble.

In the late 1990s, our group at the Massachusetts Institute of Technology designed and built several processors that were almost fully adiabatic (see Figure 5), to the limit set by transistor off-state leakage currents. This demonstrates that there is nothing inherently impossible or even especially difficult about building real computer architectures based on reversible logic. These techniques could even soon lead to cost-efficiency benefits in real electronics applications that demand extremely low power consumption.

However, some interesting fundamental research problems remain to be solved before we can firmly establish the practicality of these kinds of approaches for breaching sub- $k_B T$ energy levels. To save space, I won't discuss the open problems here (see www.cise.ufl.edu/research/revcomp/physlim/plpaper.html for a fuller description).

Current technology is relatively close to approaching the fundamental limits on energy dissipation for irreversible storage. Current trends have us reaching the limit of $k_B T \ln 2$ in only about 35 years (see Figure 6). At that time, the performance per unit power of ordinary irreversible computing, which does an irreversible (entropy-producing) storage operation with every logic-gate operation, will start to level off to a maximum level of 3.5×10^{22} irreversible bit operations per second in a 100-W computer that

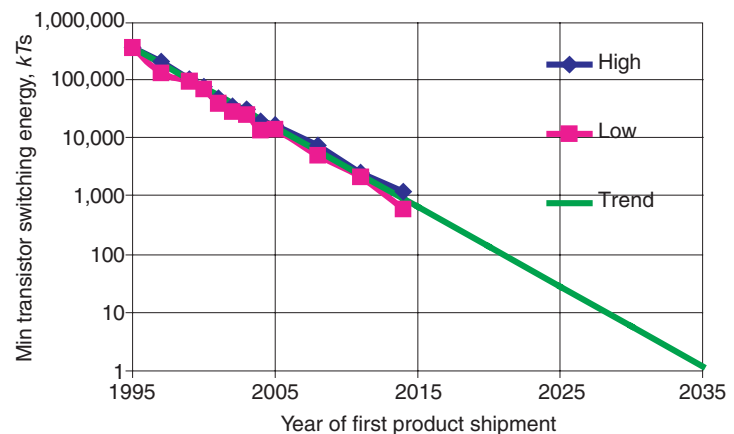


Figure 6. A trendline of minimum $\frac{1}{2}CV^2$ transistor switching energy.³ Energy is expressed as a multiple of room-temperature kT , which also is the number of nats of information associated with that energy. If the trend is followed, thermal noise will begin to become significant in the 2030s, when transistor energies approach small multiples of kT .

disposes displaced entropy into a room-temperature thermal reservoir. This rate is about a million times higher than the maximum rate of bit operations in a 30-million-gate, 1-GHz processor of today. Any possible further improvements in performance per power beyond this point require reversible computing.

Communication limits

Communication is important in computing because it constrains the performance of many parallel algorithms. In his well-known work spawning the field of information theory, Shannon derived the maximum information-carrying capacity of a single wave-based communications channel in the presence of noise.²⁰ The coding schemes used in state-of-the-art wave-based communications today closely approach Shannon's limits.

However, when considering the ultimate physical limits relevant to computation, we must go a bit beyond the scope of Shannon's paradigm. We want to know not only the capacity of a single channel, but also the maximum bandwidth for communication using any possible number of channels, given only area and power constraints.

Interestingly, the limits from the previous section on information storage density and energy directly apply to this. The difference between information storage and information communication is fundamentally only a difference in one's inertial frame of reference. Communication from point A to point B is ultimately bit transportation—a form of “storage” but in a state of relative motion. Likewise, storage is just “communication” across zero distance but through time.

If we have a limit on information density ρ and information propagation velocity v , we immediately get a limit of ρv on information flux density (flux for short)—that is, bits per unit time per unit area in communications.

Of course, we always have a limit on propagation velocity—namely, the speed of light c —so each of the information density limits mentioned earlier directly implies a limit on flux density, given suitable relativistic corrections. We can thus derive a maximum information bandwidth per unit area (in other words, information flux) as a function of per-area power density (energy flux).

For example, Smith⁹ shows that the maximum entropy flux F_S using photons, given energy flux F_E , is

$$F_S \leq \frac{4}{3} \sigma_{SB}^{1/4} F_E^{3/4},$$

where σ_{SB} is the Stefan-Boltzmann constant $\pi^2 k_B^4 / 60c^2 \hbar^3$. As a numerical example, a 10-cm square wireless tablet transmitting electromagnetically at a 1-W power level could never communicate at a bit rate of more than 6.8×10^{20} bits per second, no matter what distribution of frequencies or coding scheme we use, even in the complete absence of noise.

This limit sounds high at first, but consider that the corresponding bit rate per square nanometer is only 68 kbps. For communication among neighboring devices over a cross-section of a computer having densely packed nanoscale components, we want a much higher bandwidth density, perhaps 10^{11} bps/nm², to keep up with the 100 GHz expected rate of bit operations in a nanometer-size electronic component that is 1/100th the size of today's 0.1 μm transistors. This 10^6 times higher information flux would require a $(10^6)^{4/3} = 10^8$ times higher power density (from Smith's law)—that is, on the order of 1 MW/cm². The equivalent temperature is about 14,000 K. This seems too high to be practical—the computer would melt—so we can rule out light as a practical medium for dense interconnects at the nanoscale, at least until we find some way to build stable structures at such temperatures.

In contrast, notice that if we encoded a bit in more compact particles (atomic or electronic states), given a plausible information density of 1 bit per cubic nanometer, we could achieve our desired bit rate of 10^{11} bps/nm² by using a quite reasonable velocity of atoms or electrons of only 100 m/s.

Another interesting consideration is the minimum energy dissipation (as opposed to energy transfer) required for communications. As we saw earlier, we can look at a communication channel as being the same thing as a storage element but from a different relativistic angle. If the channel's input bit is in a known state, then swapping it with the desired information takes no energy.²¹ The channel does its thing—ideally, ballistically transporting the information and energy—and the information is then swapped out at the other end, although the receiver needs an empty place to store it. However, if the receiver's storage location is already occupied with a bit that's in the way of our new bit and that it can't uncompute, then we have to pay the energetic price to dispose of the old bit.

Computation rate limits

So far, I have focused only on limits on information storage and communication. What about computation itself? What minimum price, in terms of raw physical resources, must we pay for computational operations?

Earlier, we discussed the thermodynamic limit on computational performance of irreversible computations as a function of their power dissipation, due to the need for removal of unwanted garbage information. However, this limit might not apply to reversible computations. Are there other performance limits that apply to any type of computation, even reversible ones?

Basically, yes: We can use quantum theory to derive a maximum rate at which transitions (such as bit flips) between distinguishable states can occur.^{10,22} One form of this upper bound depends only on the system's total energy E and is given by $4E/h$, where $h = 2\pi\hbar$ is the unreduced Planck's constant.

At first, this seems like an absurdly high bound because the total energy presumably includes the system's rest mass-energy, which, if the system contains massive particles, is substantial. For example, Lloyd's 1-kg ultimate laptop has a mass-energy of 9×10^{16} Joules, so its maximum rate of operation comes out to be 5×10^{50} state changes per second.

If the system's whole mass-energy is not actively involved in the computation, presumably only the active portion of the mass-energy is relevant in this bound. This gives a much more reasonable level. For example, a hypothetical single-electron device technology in which electrons operate at 1 eV above their ground state could perform state transitions at a maximum rate of 1 PHz (10^{15} Hz) per device. As with the speed limit due to energy dissipation, this is only about a factor of a million beyond where we are today.

By changing the utilized set of distinguishable states over time, a coherent quantum computer can take drastic shortcuts through state space to quickly solve certain problems.²³ However, the rate of orthogonal transitions and the number of distinct states always still obey the limits discussed here.

All computer users, including computational scientists and engineers, naturally hope that the trend of increasing affordability of computing power will take us as far as possible. Where the ultimate limits of computing lie is obviously an important question; indeed, some have suggested it could even have a bearing on the long-term fate of life

itself.^{24–26} However, our best available knowledge of physics strongly indicates that some ultimate limits do exist, and gives us loose upper bounds on what we can achieve.

One of the most imminent of the fundamental limits appears to be the limit on the energy dissipation of irreversible computation, which might possibly be circumvented through the use of reversible computing techniques. Although reversible computing has made impressive progress, whether this “fix” can ultimately work in a scalable and cost-efficient way remains an open question, one that is the subject of active research by myself and others.

I hope that this article inspires researchers in many fields to devote increased attention to finding ways to meet the incredible challenges facing the future of computing as it approaches the many limits found at the atomic scale. These limits are now close enough to fall within the career horizons of people starting out today. For example, given present rates of improvement, computing will encounter the $k_B T$ thermodynamic barrier before today's 30-year-old PhD graduates retire. Although computing appears to be nearing various hard physical limits, the race to get as far as possible within those limits promises exciting research opportunities in many areas of the physical and computer sciences as we develop these new machines.

But, even if someday we figure out how to optimally harness all the raw computational power of physics itself, we can be sure that our ultimate power users—computational scientists and engineers—will enthusiastically tackle new problems so challenging that computers will still seem too slow. ☹

References

1. G.E. Moore, “Cramming More Components onto Integrated Circuits,” *Electronics*, 19 Apr. 1965, pp. 114–117.
2. G.E. Moore, “An Update on Moore's Law,” Intel Developer Forum Keynote Speech, San Francisco, 30 Sept. 1997, <http://developer.intel.com/pressroom/archive/speeches/gem93097.htm>.
3. Semiconductor Industry Association, *Int'l Technology Roadmap for Semiconductors: 2001 Edition*; <http://public.itrs.net>.
4. D.K.K. Lee and A.J. Schofield, “Metals Without Electrons: The Physics of Exotic Quantum Fluids,” *Visions of the Future: Physics and Electronics*, J.M.T. Thompson, ed., Cambridge Univ. Press, Cambridge, England, 2001, pp. 17–37.
5. D. Deutsch, *The Fabric of Reality: The Science of Parallel Universes—And Its Implications*, Penguin Books, New York, 1997.
6. R. Landauer, “Information is Physical,” *Physics Today*, vol. 44, May 1991, pp. 23–29.
7. E.T. Jaynes, “Information Theory and Statistical Mechanics,” *Physical Rev.*, vol. 106, no. 4, 15 May 1957, pp. 620–630; <http://bayes.wustl.edu/etj/articles/theory.1.pdf>.

Member Societies

American Physical Society
Optical Society of America
Acoustical Society of America
The Society of Rheology
American Association of Physics Teachers
American Crystallographic Association
American Astronomical Society
American Association of Physicists in Medicine
AVS
American Geophysical Union
Other Member Organizations
Sigma Pi Sigma, Physics Honor Society
Society of Physics Students
Corporate Associates

The American Institute of Physics is a not-for-profit membership corporation chartered in New York State in 1931 for the purpose of promoting the advancement and diffusion of the knowledge of physics and its application to human welfare. Leading societies in the fields of physics, astronomy, and related sciences are its members.

The Institute publishes its own scientific journals as well as those of its Member Societies; provides abstracting and indexing services; provides online database services; disseminates reliable information on physics to the public; collects and analyzes statistics on the profession and on physics education; encourages and assists in the documentation and study of the history and philosophy of physics; cooperates with other organizations on educational projects at all levels; and collects and analyzes information on Federal programs and budgets.

The scientists represented by the Institute through its Member Societies number approximately 120,000. In addition, approximately 5,400 students in over 600 colleges and universities are members of the Institute's Society of Physics Students, which includes the honor society Sigma Pi Sigma. Industry is represented through 47 Corporate Associates members.

Governing Board*

John A. Armstrong, (Chair), *Marc H. Brodsky* (Executive Director), Benjamin B. Snively (Secretary), Martin Blume (APS), William F. I. Brinkman (APS), Judy R. Franz (APS), Donald R. Hamann (APS), Myriam P. Sarachik (APS), *Thomas J. McIlrath* (APS), George H. Trilling (APS), Michael D. Duncan (OSA), Ivan P. Kaminow (OSA), Anthony M. Johnson (OSA), *Elizabeth A. Rogan* (OSA), Anthony A. Atchley (ASA), *Lawrence A. Crum* (ASA), Charles E. Schmid (ASA), *Arthur B. Metzner* (SOR), Christopher J. Chiaverina (AAPT), Charles H. Holbrow (AAPT), John Hubisz (AAPT), *Bernard V. Khoury* (AAPT), Charlotte Lowe-Ma (ACA), *S. Narasinga Rao* (ACA), *Leonard V. Kubi* (AAS), Arlo U. Landolt (AAS), Robert W. Milkey (AAS), *James B. Smathers* (AAPM), Christopher H. Marshall (AAPM), *Rudolf Ludeke* (AVS), N. Rey Whetten (AVS), Dawn A. Bonnell (AVS), James L. Burch (AGU), Robert E. Dickinson (AGU), Jeffrey J. Park (AGU), Judy C. Holoviak (AGU), *Louis J. Lanzarotti* (AGU), Fred Spilhaus (AGU), Brian Clark (2002) MAL, Frank L. Huband (MAL)

*Executive Committee members are printed in italics.

Management Committee

Marc H. Brodsky, Executive Director and CEO; Richard Baccante, Treasurer and CFO; Theresa C. Braun, Vice President, Human Resources; James H. Stith, Vice President, Physics Resources; Darlene A. Walters, Senior Vice President, Publishing; Benjamin B. Snively, Secretary

Subscriber Services

AIP subscriptions, renewals, address changes, and single-copy orders should be addressed to Circulation and Fulfillment Division, American Institute of Physics, 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502. Tel. (800) 344-6902; e-mail subs@aip.org. Allow at least six weeks' advance notice. For address changes please send both old and new addresses, and, if possible, include an address label from the mailing wrapper of a recent issue.

8. K. Stowe, *Introduction to Statistical Mechanics and Thermodynamics*, John Wiley & Sons, New York, 1984.
9. W.D. Smith, *Fundamental Physical Limits on Computation*, tech. report, NEC Research Inst., Princeton, N.J., 1995; <http://external.nj.nec.com/homepages/wds/fundphys.ps>.
10. S. Lloyd, "Ultimate Physical Limits to Computation," *Nature*, vol. 406, no. 8, Aug. 2000, pp. 1047-1054.
11. J.D. Bekenstein, "Universal Upper Bound on the Entropy-to-Energy Ratio for Bounded Systems," *Physical Rev. D*, vol. 23, no. 2, 15 Jan. 1981, pp. 287-298.
12. R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM J. Research and Development*, vol. 5, no. 3, 1961, pp. 183-191.
13. J. von Neumann, *Theory of Self-Reproducing Automata*, Univ. of Illinois Press, Champaign, Ill., 1966.
14. C.H. Bennett, "Logical Reversibility of Computation," *IBM J. Research and Development*, vol. 17, no. 6, 1973, pp. 525-532.
15. Y. Lecerf, "Reversible Turing Machines: Recursive Insolubility," *Weekly Proc. Academy Science*, vol. 257, Oct. 1963, pp. 2597-2600; www.cise.ufl.edu/~mpf/rc/Lecerf/lecerf.html.
16. S.G. Younis and T.F. Knight, Jr., "Asymptotically Zero Energy Split-Level Charge Recovery Logic," *Int'l Workshop Low Power Design*, 1994, pp. 177-182; www.ai.mit.edu/people/tk/lowpower/low94.ps.
17. C.J. Vieri, *Reversible Computer Engineering and Architecture*, doctoral dissertation, Mass. Inst. of Technology, EECS Dept., Cambridge, Mass., 1999.
18. M.P. Frank et al., "A Scalable Reversible Computer in Silicon," *Unconventional Models of Computation*, Springer-Verlag, Berlin, 1998, pp. 183-200.
19. M.P. Frank, *Reversibility for Efficient Computing*, doctoral dissertation, Mass. Inst. of Technology, EECS Dept., Cambridge, Mass., 1999; www.cise.ufl.edu/~mpf/rc/thesis/phdthesis.html.
20. C.E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical J.*, vol. 27, July and Oct. 1948, pp. 379-423; 623-656.
21. R. Landauer, "Minimal Energy Requirements in Communication," *Science*, vol. 272, no. 5270, 28 June 1996, pp. 1914-1918.
22. N. Margolus and L.B. Levitin, "The Maximum Speed of Dynamical Evolution," *Physica D*, vol. 120, 1998, pp. 188-195.
23. M.A. Nielsen and I.L. Chuang, *Quantum Computation and Quantum Information*, Cambridge Univ. Press, Cambridge, UK, 2000.
24. F.J. Dyson, "Time Without End: Physics and Biology in an Open Universe," *Rev. Modern Physics*, vol. 51, no. 3, July 1979, pp. 447-460.
25. L.M. Krauss and G.D. Starkman, "Life, the Universe, and Nothing: Life and Death in an Ever-Expanding Universe," *Astrophysical J.*, vol. 531, no. 1, 2000, pp. 22-30.
26. F.J. Dyson, "Is Life Analog or Digital?" *Edge*, vol. 82, Mar. 13, 2001; www.edge.org/documents/archive/edge82.html.

Michael Frank is an assistant professor in the Computer and Information Science and Engineering Department at the University of Florida, where he received the local ACM chapter's Teacher of the Year award while teaching his course on Physical Limits of Computing. He received a BS in symbolic systems from Stanford University, an MS in electrical engineering and computer science, and a PhD in reversible computing research from MIT. He is a member of the IEEE and ACM. Contact him at P.O. Box 116120, Gainesville, FL 32611, mpf@cise.ufl.edu.

For more information on this or any other computing topic, please visit our Digital Library at <http://computer.org/publications/dlib>.