2

Room at the Bottom,
Plenty of Tyranny
at the Top

CONTENTS

2.1 Rising to the Feynman Challenge
2.2 Tyranny at the Top
2.3 New Forms of Switching and Storage
2.4 New Architectures
2.5 How Does Nature Do It?

Karl Hess
University of Illinois

2.1 Rising to the Feynman Challenge

Richard Feynman is generally regarded as one of the fathers of nanotechnology. In giving his landmark presentation to the American Physical Society on December 29, 1959, at Caltech, his title line was, “There's Plenty of Room at the Bottom.” At that time, Feynman extended an invitation for “manipulating and controlling things on a small scale, thereby entering a new field of physics which was bottomless, like low-temperature physics.” He started with the question, can we “write the Lord’s prayer on the head of a pin,” and immediately extended the goal to the entire 24 volumes of the Encyclopaedia Britannica. By following the Gedanken Experiment, Feynman showed that there is no physical law against the realization of such goals: if you magnify the head of a pin by 25,000 diameters, its surface area is then equal to that of all the pages in the Encyclopaedia Britannica.

Feynman’s dreams of writing small have all been fulfilled and even exceeded in the past decades. Since the advent of scanning tunneling microscopy, as introduced by Binnig and Rohrer, it has been repeatedly demonstrated that single atoms can not only be conveniently represented for the human eye but manipulated as well. Thus, it is conceivable to store all the books in the world (which Feynman estimates to contain $10^{15}$ bits of information) on the area of a credit card! The encyclopedia, having around $10^9$ bits of information, can be written on about $1/100$ the surface area of the head of a pin.

One need not look to atomic writing to achieve astonishing results: current microchips contain close to 100 million transistors. A small number of such chips could not only store large amounts of information (such as the Encyclopaedia Britannica); they can process it with GHz speed as well. To find a particular word takes just a few nanoseconds. Typical disk hard drives can store much more than the semiconductor chips, with a trade-off for retrieval speeds. Feynman’s vision for storing and retrieving information on a small scale was very close to these numbers. He did not ask himself what the practical difficulties were in achieving these goals, but rather asked only what the principal limitations were. Even he could not possibly foresee the ultimate consequences of writing small and reading fast: the creation of the Internet. Sifting through large databases is, of course, what is done during Internet browsing. It is not only the
microrepresentation of information that has led to the revolution we are witnessing but also the ability to browse through this information at very high speeds.

Can one improve current chip technology beyond the achievements listed above? Certainly! Further improvements are still expected just by scaling down known silicon technology. Beyond this, if it were possible to change the technology completely and create transistors the size of molecules, then one could fit hundreds of billions of transistors on a chip. Changing technology so dramatically is not easy and less likely to happen. A molecular transistor that is as robust and efficient as the existing ones is beyond current implementation capabilities; we do not know how to achieve such densities without running into problems of excessive heat generation and other problems related to highly integrated systems. However, Feynman would not be satisfied that we have exhausted our options. He still points to the room that opens if the third dimension is used. Current silicon technology is in its essence (with respect to the transistors) a planar technology. Why not use volumes, says Feynman, and put all books of the world in the space of a small dust particle? He may be right, but before assessing the chances of this happening, I would like to take you on a tour to review some of the possibilities and limitations of current planar silicon technology.

2.2 Tyranny at the Top

Yes, we do have plenty of room at the bottom. However, just a few years after Feynman’s vision was published, J. Morton from Bell Laboratories noticed what he called the tyranny of large systems. This tyranny arises from the fact that scaling is, in general, not part of the laws of nature. For example, we know that one cannot hold larger and larger weights with a rope by making the rope thicker and thicker. At some point the weight of the rope itself comes into play, and things may get out of hand. As a corollary, why should one, without such difficulty, be able to make transistors smaller and smaller and, at the same time, integrate more of them on a chip? This is a crucial point that deserves some elaboration.

It is often said that all we need is to invent a new type of transistor that scales to atomic size. The question then arises: did the transistor, as invented in 1947, scale to the current microsize? The answer is no! The point-contact transistor, as it was invented by Bardeen and Brattain, was much smaller than a vacuum tube. However, its design was not suitable for aggressive scaling. The field-effect transistor, based on planar silicon technology and the hetero-junction interface of silicon and silicon dioxide with a metal on top (MOS technology), did much better in this respect. Nevertheless, it took the introduction of many new concepts (beginning with that of an inversion layer) to scale transistors to the current size. This scalability alone would still not have been sufficient to build large integrated systems on a chip. Each transistor develops heat when operated, and a large number of them may be better used as a soldering iron than for computing. The saving idea was to use both electron and hole-inversion layers to form the CMOS technology. The transistors of this technology create heat essentially only during switching operation, and heat generation during steady state is very small. A large system also requires interconnection of all transistors using metallic “wires.” This becomes increasingly problematic when large numbers of transistors are involved, and many predictions have been made that it could not be done beyond a certain critical density of transistors. It turned out that there never was such a critical density for interconnection, and we will discuss the very interesting reason for this below. Remember that Feynman never talked about the tyranny at the top. He only was interested in fundamental limitations. The exponential growth of silicon technology with respect to the numbers of transistors on a chip seems to prove Feynman right, at least up to now. How can this be if the original transistors were not scalable? How could one always find a modification that permitted further scaling?

One of the reasons for continued miniaturization of silicon technology is that its basic idea is very flexible: use solids instead of vacuum tubes. The high density of solids permits us to create very small structures without hitting the atomic limit. Gas molecules or electrons in tubes have a much lower density than electrons or atoms in solids typically have. One has about \(10^{18}\) atoms in a cm\(^3\) of gas but \(10^{23}\) in a cm\(^3\) of a solid. Can one therefore go to sizes that would contain only a few hundred atoms with current silicon technology? I believe not. The reason is that current technology is based on the doping of silicon

© 2003 by CRC Press LLC
with donors and acceptors to create electron- and hole-inversion layers. The doping densities are much lower than the densities of atoms in a solid, usually below $10^{20}$ per cm$^3$. Therefore, to go to the ultimate limits of atomic size, a new type of transistor, without doping, is needed. We will discuss such possibilities below. But even if we have such transistors, can they be interconnected? Interestingly enough, interconnection problems have always been overcome in the past. The reason was that use of the third dimension has been made for interconnects. Chip designers have used the third dimension — not to overcome the limitations that two dimensions place on the number of transistors, but to overcome the limitations that two dimensions present for interconnecting the transistors. There is an increasing number of stacks of metal interconnect layers on chips — 2, 5, 8. How many can we have? (One can also still improve the conductivity of the metals in use by using, for example, copper technology.)

Pattern generation is, of course, key for producing the chips of silicon technology and represents another example of the tyranny of large systems. Chips are produced by using lithographic techniques. Masks that contain the desired pattern are placed above the chip material, which is coated with photosensitive layers that are exposed to light to engrave the pattern. As the feature sizes become smaller and smaller, the wavelength of the light needs to be reduced. The current work is performed in the extreme ultraviolet, and future scaling must overcome considerable obstacles. Why can one not use the atomic resolution of scanning tunneling microscopes? The reason is, of course, that the scanning process takes time; and this would make efficient chip production extremely difficult. One does need a process that works “in parallel” like photography. In principle there are many possibilities to achieve this, ranging from the use of X-rays to electron and ion beams and even self-organization of patterns in materials as known in chemistry and biology. One cannot see principal limitations here that would impede further scaling. However, efficiency and expense of production do represent considerable tyranny and make it difficult to predict what course the future will take. If use is made of the third dimension, however, optical lithography will go a long way.

Feynman suggested that there will be plenty of room at the bottom only when the third dimension is used. Can we also use it to improve the packing density of transistors? This is not going to be so easy. The current technology is based on a silicon surface that contains patterns of doping atoms and is topped by silicon dioxide. To use the third dimension, a generalization of the technology is needed. One would need another layer of silicon on top of the silicon dioxide, and so forth. Actually, such technology does already exist: silicon-on-insulator (SOI) technology. Interestingly enough, some devices that are currently heralded by major chip producers as devices of the future are SOI transistors. These may be scalable further than current devices and may open the horizon to the use of the third dimension. Will they open the way to unlimited growth of chip capacity? Well, there is still heat generation and other tyrannies that may prevent the basically unlimited possibilities that Feynman predicted. However, billions of dollars of business income have overcome most practical limitations (the tyranny) and may still do so for a long time to come. Asked how he accumulated his wealth, Arnold Beckman responded: “We built a pH-meter and sold it for three hundred dollars. Using this income, we built two and sold them for $600 … and then 4, 8, … .” This is, of course, the well-known story of the fast growth of a geometric series as known since ages for the rice corns on the chess board. Moore’s law for the growth of silicon technology is probably just another such example and therefore a law of business rather than of science and engineering. No doubt, it is the business income that will determine the limitations of scaling to a large extent. But then, there are also new ideas.

### 2.3 New Forms of Switching and Storage

Many new types of transistors or switching devices have been investigated and even mass fabricated in the past decades. Discussions have focused on GaAs and III-V compound materials because of their special properties with respect to electron speed and the possibility of creating lattice-matched interfaces and layered patterns of atomic thickness. Silicon and silicon-dioxide have very different lattice constants (spacing between their atoms). It is therefore difficult to imagine that the interface between them can be electronically perfect. GaAs and AlAs on the other side have almost equal lattice spacing, and two crystals
can be perfectly placed on top of each other. The formation of superlattices of such layers of semiconductors has, in fact, been one of the bigger achievements of recent semiconductor technology and was made possible by new techniques of crystal growth (molecular beam epitaxy, metal organic chemical vapor deposition, and the like). Quantum wells, wires, and dots have been the subject of extremely interesting research and have enriched quantum physics for example, by the discovery of the Quantum Hall Effect and the Fractional Quantum Hall effect. Use of such layers has also brought significant progress to semiconductor electronics. The concept of modulation doping (selective doping of layers, particularly involving pseudomorphic InGaAs) has led to modulation-doped transistors that hold the current speed records and are used for microwave applications. The removal of the doping to neighboring layers has permitted the creation of the highest possible electron mobilities and velocities. The effect of resonant tunneling has also been shown to lead to ultrafast devices and applications that reach to infrared frequencies, encompassing in this way both optics and electronics applications. When it comes to large-scale integration, however, the tyranny from the top has favored silicon technology. Silicon dioxide, as an insulator, is superior to all possible III-V compound materials; and its interface with silicon can be made electronically perfect enough, at least when treated with hydrogen or deuterium.

When it comes to optical applications, however, silicon is inefficient because it is an indirect semiconductor and therefore cannot emit light efficiently. Light generation may be possible by using silicon. However, this is limited by the laws of physics and materials science. It is my guess that silicon will have only limited applications for optics, much as III-V compounds have for large-scale integrated electronics. III-V compounds and quantum well layers have been successfully used to create efficient light-emitting devices including light-emitting and semiconductor laser diodes. These are ubiquitous in every household, e.g., in CD players and in the back-lights of cars. New forms of laser diodes, such as the so-called vertical cavity surface emitting laser diodes (VCSELs), are even suitable to relatively large integration. One can put thousands and even millions of them on a chip. Optical pattern generation has made great advances by use of selective superlattice intermixing (compositionally disordered III-V compounds and superlattices have a different index of refraction) and by other methods. This is an area in great flux and with many possibilities for miniaturization. Layered semiconductors and quantum well structures have also led to new forms of lasers such as the quantum cascade laser. Feynman mentioned in his paper the use of layered materials. What would he predict for the limits of optical integration and the use of quantum effects due to size quantization in optoelectronics?

A number of ideas are in discussion for new forms of ultrasmall electronic switching and storage devices. Using the simple fact that it takes a finite energy to bring a single electron from one capacitor plate to the other (and using tunneling for doing so), single-electron transistors have been proposed and built. The energy for this single-electron switching process is inversely proportional to the area of the capacitor. To achieve energies that are larger than the thermal energy at room temperature (necessary for robust operation), extremely small capacitors are needed. The required feature sizes are of the order of one nanometer. There are also staggering requirements for material purity and perfection since singly charged defects will perturb operation. Nevertheless, Feynman may have liked this device because the limitations for its use are not due to physical principles. It also has been shown that memory cells storing only a few electrons do have some very attractive features. For example, if many electrons are stored in a larger volume, a single material defect can lead to unwanted discharge of the whole volume. If, on the other hand, all these electrons are stored in a larger number of quantum dots (each carrying few electrons), a single defect can discharge only a single dot, and the remainder of the stored charge stays intact.

Two electrons stored on a square-shaped “quantum dot” have been proposed as a switching element by researchers at Notre Dame. The electrons start residing in a pair of opposite corners of the square and are switched to the other opposite corner. This switching can be effected by the electrons residing in a neighboring rectangular dot. Domino-type effects can thus be achieved. It has been shown that architectures of cellular neural networks (CNNs) can be created that way as discussed briefly below.

A new field referred to as spintronics is developing around the spin properties of particles. Spin properties have not been explored in conventional electronics and enter only indirectly, through the Pauli principle, into the equations for transistors. Of particular interest in this new area are particle pairs that
exhibit quantum entanglement. Consider a pair of particles in a singlet spin-state sent out to detectors or spin analyzers in opposite directions. Such a pair has the following remarkable properties: measurements of the spin on each side separately give random values of the spin (up/down). However, the spin of one side is always correlated to the spin on the other side. If one is up, the other is down. If the spin analyzers are rotated relative to each other, then the result for the spin pair correlation shows rotational symmetry. A theorem of Bell proclaims such results incompatible with Einstein’s relativity and suggests the necessity of instantaneous influences at a distance. Such influences do not exist in classical information theory and are therefore considered a quantum addition to classical information. This quantum addition provides part of the novelty that is claimed for possible future quantum computers. Spintronics and entanglement are therefore thought to open new horizons for computing.

Still other new device types use the wave-like nature of electrons and the possibility to guide these waves by externally controllable potential profiles. All of these devices are sensitive to temperature and defects, and it is not clear whether they will be practical. However, new forms of architectures may open new possibilities that circumvent the difficulties.

2.4 New Architectures

Transistors of the current technology have been developed and adjusted to accommodate the tyranny from the top, in particular the demands set forth by the von Neuman architecture of conventional computers. It is therefore not surprising that new devices are always looked at with suspicion by design engineers and are always found wanting with respect to some tyrannical requirement. Many regard it extremely unlikely that a completely new device will be used for silicon chip technology. Therefore, architectures that deviate from von Neuman’s principles have received increasing attention. These architectures invariably involve some form of parallelism. Switching and storage is not localized to a single transistor or small circuit. The devices are connected to each other, and their collective interactions are the basis for computation. It has been shown that such collective interactions can perform some tasks in ways much superior to von Neuman’s sequential processing.

One example for such new principles is the cellular neural network (CNN) type of architectures. Each cell is connected by a certain coupling constant to its nearest neighbors, and after interaction with each other, a large number of cells settle on a solution that hopefully is the desired solution of a problem that cannot easily be done with conventional sequential computation. This is, of course, very similar to the advantages of parallel computation (computation by use of more than one processor) with the difference that it is not processors that interact and compute in parallel but the constituent devices themselves. CNNs have advantageously been used for image processing and other specialized applications and can be implemented in silicon technology. It appears that CNNs formed by using new devices, such as the coupled square quantum dots discussed above, could (at least in principle) be embedded into a conventional chip environment to perform a certain desired task; and new devices could be used that way in connection with conventional technology. There are at least three big obstacles that need to be overcome if this goal should be achieved. The biggest problem is posed by the desire to operate at room temperature. As discussed above, this frequently is equivalent to the requirement that the single elements of the CNN need to be extremely small, on the order of one nanometer. This presents the second problem — to create such feature sizes by a lithographic process. Third, each element of the CNN needs to be virtually perfect and free of defects that would impede its operation. Can one create such a CNN by the organizing and self-organizing principles of chemistry on semiconductor surfaces? As Dirac once said (in connection with difficult problems), “one must try.” Of course, it will be tried only if an important problem exists that defies conventional solution. An example would be the cryptographically important problem of factorizing large numbers. It has been shown that this problem may find a solution through quantum computation.

The idea of quantum computation has, up to now, mainly received the attention of theoreticians who have shown the superior power of certain algorithms that are based on a few quantum principles. One such principle is the unitarity of certain operators in quantum mechanics that forms a solid basis for the
possibility of quantum computing. Beyond this, it is claimed that the number of elements of the set of parameters that constitutes quantum information is much larger than the comparable set used in all of classical information. This means there are additional quantum bits (qubits) of information that are not covered by the known classical bits. In simpler words, there are instantaneous action at a distance and connected phenomena, such as quantum teleportation, that have not been used classically but can be used in future quantum information processing and computation. These claims are invariably based on the theorem of Bell and are therefore subject to some criticism. It is well known that the Bell theorem has certain loopholes that can be closed only if certain time dependencies of the involved parameters are excluded. This means that even if the Bell theorem were general otherwise, it does not cover the full classical parameter space. How can one then draw conclusions about the number of elements in parameter sets for classical and quantum information? In addition, recent work has shown that the Bell theorem excludes practically all time-related parameters — not only those discussed in the well-known loopholes. What I want to say here is that the very advanced topic of quantum information complexity will need further discussion even about its foundations. Beyond this, obstacles exist for implementation of qubits due to the tyranny from the top. It is necessary to have a reasonably large number of qubits in order to implement the quantum computing algorithms and make them applicable to large problems. All of these qubits need to be connected in a quantum mechanical coherent way. Up to now, this coherence has always necessitated the use of extremely low temperatures, at least when electronics (as opposed to optics) is the basis for implementation. With all these difficulties, however, it is clear that there are great opportunities for solving problems of new magnitude by harnessing the quantum world.

2.5 How Does Nature Do It?

Feynman noticed that nature has already made use of nanostructures in biological systems with greatest success. Why do we not copy nature? Take, for example, biological ion channels. These are tiny pores formed by protein structures. Their opening can be as small as a few one-tenths of a nanometer. Ion currents are controlled by these pores that have opening and closing gates much as transistors have. The on/off current ratio of ion channels is practically infinite, which is a very desirable property for large systems. Remember that we do not want energy dissipation when the system is off. Transistors do not come close to an infinite on/off ratio, which represents a big design problem. How do the ion channels do it? The various gating mechanisms are not exactly understood, but they probably involve changes in the aperture of the pore by electrochemical mechanisms. Ion channels do not only switch currents perfectly. They also can choose the type of ions they let through and the type they do not. Channels perform in this way a multitude of functions. They regulate our heart rate, kill bacteria and cancer cells, and discharge and recharge biological neural networks, thus forming elements of logic and computation. The multitude of functions may be a great cure for some of the tyranny from the top as Jack Morton has pointed out in his essay “From Physics to Function.” No doubt, we can learn in this respect by copying nature. Of course, proteins are not entirely ideal materials when it comes to building a computer within the limits of a preconceived technology. However, nature does have an inexpensive way of pattern formation and replication — a self-organizing way. This again may be something to copy. If we cannot produce chip patterns down to nanometer size by inexpensive photographic means, why not produce them by methods of self-organization? Can one make ion channels out of materials other than proteins that compare more closely to the solid-state materials of chip technology? Perhaps carbon nanotubes can be used. Material science has certainly shown great inventiveness in the past decades.

Nature also has no problems in using all three dimensions of space for applying its nanostructures. Self-organization is not limited to a plane as photography is. Feynman’s ultimate frontier of using three dimensions for information storage is automatically included in some biological systems such as, for example, neural networks. The large capacity and intricate capability of the human brain derives, of course, from this fact.

The multitude of nanostructure functionalities in nature is made possible because nature is not limited by disciplinary boundaries. It uses everything, whether physics or chemistry, mechanics or electronics.
— and yes, nature also uses optics, e.g., to harvest energy from the sun. I have not covered nanometer-size mechanical functionality because I have no research record in this area. However, great advances are currently made in the area of nanoelectromechanical systems (NEMS). It is no problem any more to pick up and drop atoms, or even to rotate molecules. Feynman’s challenge has been far surpassed in the mechanical area, and even his wildest dreams have long since become reality. Medical applications, such as the insertion of small machinery to repair arteries, are commonplace. As we understand nature better, we will not only be able to find new medical applications but may even improve nature by use of special smart materials for our bodies. Optics, electronics, and mechanics, physics, chemistry, and biology need to merge to form generations of nanostructure technologies for a multitude of applications.

However, an area exists in which man-made chips excel and are superior to natural systems (if man-made is not counted as natural). This area relates to processing speed. The mere speed of a number-crunching machine is unthinkable for the workings of a biological neural network. To be sure, nature has developed fast processing; visual evaluations of dangerous situations and recognition of vital patterns are performed with lightening speed by some parallel processing of biological neural networks. However, when it comes to the raw speed of converting numbers, which can also be used for alphabetical ordering and for a multitude of algorithms, man-made chips are unequaled. Algorithmic speed and variability is a very desirable property, as we know from browsing the Internet, and represents a great achievement in chip technology.

Can we have both —the algorithmic speed and variability of semiconductor-based processors and, at the same time, three-dimensional implementations and the multitude of functionality as nature features it in her nanostructure designs? I would not dare to guess an answer to this question. The difficulties are staggering! Processing speed seems invariably connected to heat generation. Cooling becomes increasingly difficult when three-dimensional systems are involved and the heat generation intensifies.

But then, there are always new ideas, new materials, new devices, new architectures, and altogether new horizons. Feynman’s question as to whether one can put the Encyclopaedia Britannica on the head of a pin has been answered in the affirmative. We have proceeded to the ability to sift through the material and process the material of the encyclopedia with lightning speed. We now address the question of whether we can process the information of three-dimensional images within the shortest of times, whether we can store all the knowledge of the world in the smallest of volumes and browse through gigabits of it in a second. We also proceed to the question of whether mechanical and optical functionality can be achieved on such a small scale and with the highest speed. Nature has shown that the smallest spatial scales are possible. We have to search for the greatest variety in functionality and for the highest possible speed in our quest to proceed in science from what is possible in principle to a function that is desirable for humanity.