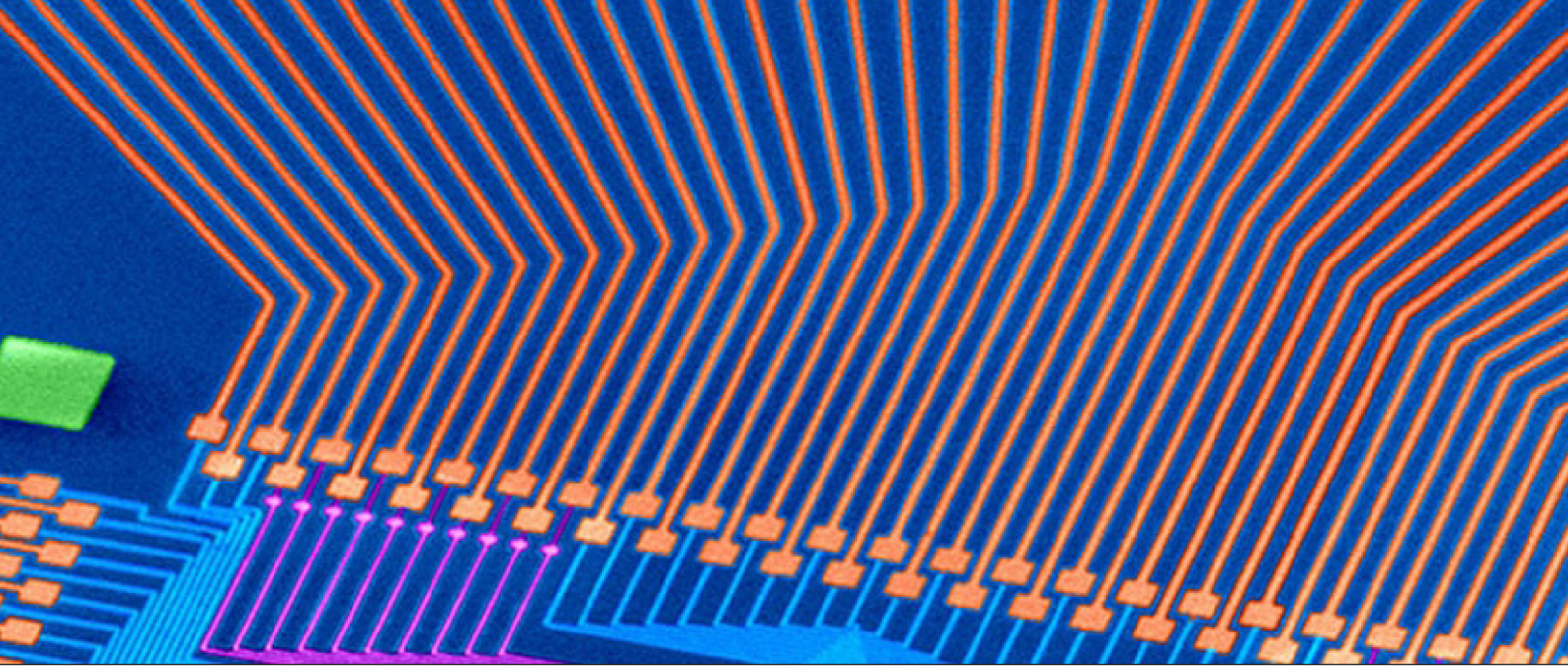
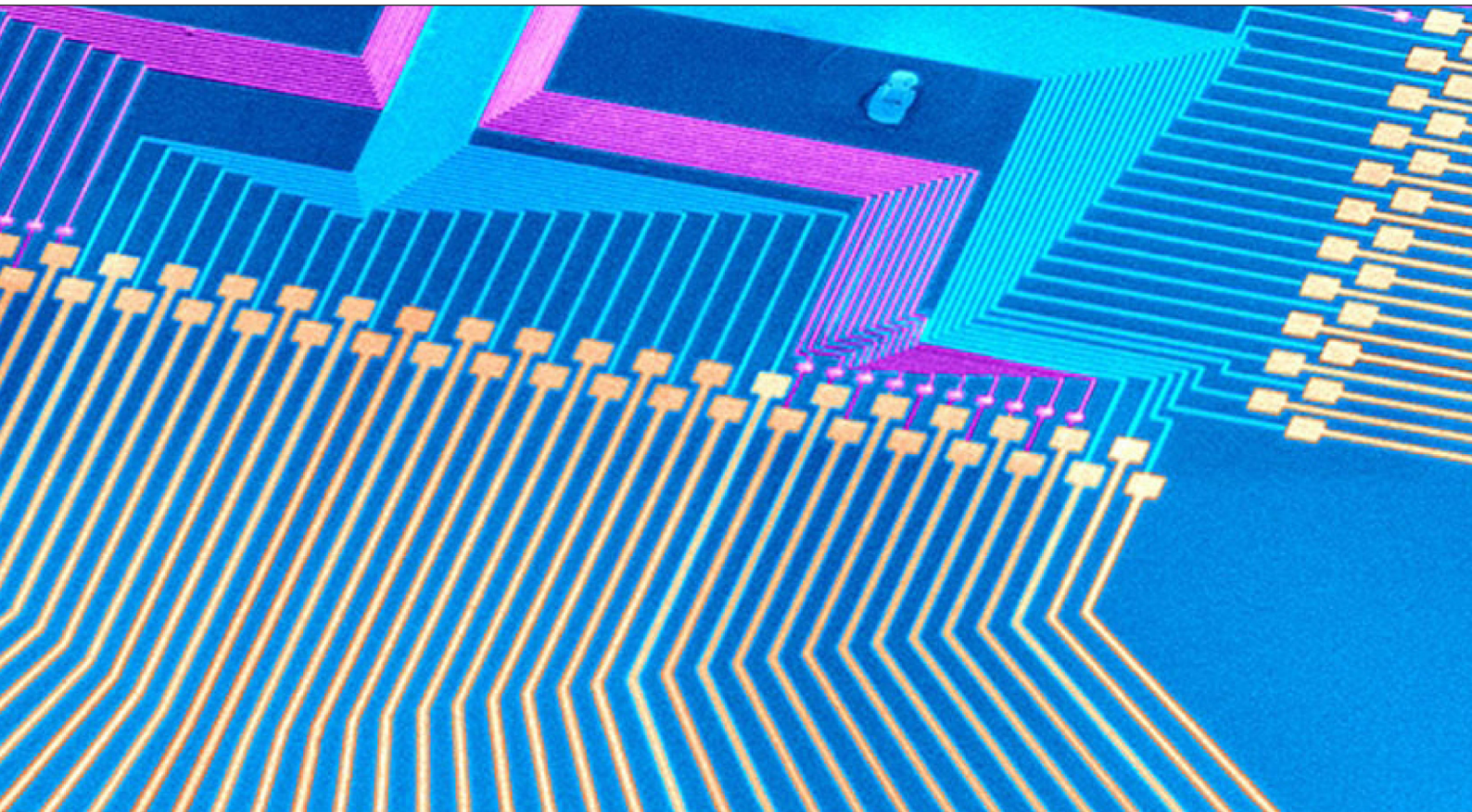


Designing a Nanocomputer

Daniel Strickland



DESIGNING A NANOCOMPUTER

DESIGNING A NANOCOMPUTER

Daniel Strickland

BIBLIOTEX
Digital Library

Designing a Nanocomputer
by Daniel Strickland

www.bibliotex.com

Ebook ISBN: 9781984664167

Published by:
Bibliotex
Canada
Website: www.bibliotex.com

Contents

1

Introduction to Nano Computer

A nanocomputer is a computer whose physical dimensions are microscopic. The field of nanocomputing is part of the emerging field of nanotechnology. Several types of nanocomputers have been suggested or proposed by researchers and futurists.

Electronic nanocomputers would operate in a manner similar to the way present-day microcomputers work. The main difference is one of physical scale. More and more transistor s are squeezed into silicon chips with each passing year; witness the evolution of integrated circuits (IC s) capable of ever-increasing storage capacity and processing power. The ultimate limit to the number of transistors per unit volume is imposed by the atomic structure of matter. Most engineers agree that technology has not yet come close to pushing this limit. In the electronic sense, the term nanocomputer is relative. By 1970s standards, today's

ordinary microprocessors might be called nanodevices. Chemical and biochemical nanocomputers would store and process information in terms of chemical structures and interactions. Biochemical nanocomputers already exist in nature; they are manifest in all living things. But these systems are largely uncontrollable by humans. We cannot, for example, programme a tree to calculate the digits of pi, or programme an antibody to fight a particular disease (although medical science has come close to this ideal in the formulation of vaccines, antibiotics, and antiviral medications). The development of a true chemical nanocomputer will likely proceed along lines similar to genetic engineering. Engineers must figure out how to get individual atoms and molecules to perform controllable calculations and data storage tasks.

Mechanical nanocomputers would use tiny moving components called nanogears to encode information. Such a machine is reminiscent of Charles Babbage 's analytical engines of the 19th century. For this reason, mechanical nanocomputer technology has sparked controversy; some researchers consider it unworkable.

All the problems inherent in Babbage's apparatus, according to the naysayers, are magnified a millionfold in a mechanical nanocomputer. Nevertheless, some futurists are optimistic about the technology, and have even proposed the evolution of nanorobots that could operate, or be controlled by, mechanical nanocomputers.

A nanocomputer is a type of computer that is commonly employed in the field of nanotechnology. Essentially, a nanocomputer is a microscopic computer device. While the

concept of the nanocomputer has been around for several decades, the perception of the technology continues to evolve as it becomes possible to store and utilize data and functions on systems that are increasingly compact in design.

In years past, the concept of a nanocomputer was thought of in terms of being a small computer that would accomplish all the tasks associated with the large electronic brains of the 1940's and 1950's. Over time, as technology made it possible for smaller mainframes to surpass the capabilities of those earlier electronic brains, the idea of the nanocomputer as a device that was not only small but possibly even microscopic in nature began to emerge.

Today, the general idea of a nanocomputer would involve the insertion of the small device within the body to help support natural organ functions, treat illnesses, and in general perform specific tasks that the body is no longer able to manage without some sort of assistance. The potential for a microscopic computer appears to be endless. Along with use in the treatment of many physical and emotional ailments, the nanocomputer is sometimes envisioned to allow for the ultimate in a portable device that can be used to access the Internet, prepare documents, research various topics, and handle mundane tasks such as e-mail. In short, all the functions that are currently achieved with desktop computers, laptops, and hand held devices would be possible with a nanocomputer that is inserted into the body and directly interacts with the brain.

At the present time, a nanocomputer that is workable for general use in home or business applications is still more fiction than fact. However, technology continues to find ways

to store increasing amounts of data on a single computer chip. As the mechanical aspects of computer technology continue to result in smaller outward devices, the concept of a nanocomputer that makes use of a biochemical interface becomes more feasible. There are proponents of the nanocomputer concept that believe these microscopic implants will be available for the general population at some point over the next couple of decades.

A quantum nanocomputer would work by storing data in the form of atomic quantum states or spin. Technology of this kind is already under development in the form of single-electron memory (SEM) and quantum dots. The energy state of an electron within anatom, represented by the electron energy level or shell, can theoretically represent one, two, four, eight, or even 16 bits of data. The main problem with this technology is instability. Instantaneous electron energy states are difficult to predict and even more difficult to control. An electron can easily fall to a lower energy state, emitting a photon ; conversely, a photon striking an atom can cause one of its electrons to jump to a higher energy state.

A computer with circuitry so small that it can only be seen through a microscope. Nanocomputers can be electronic (where nanolithography is used to create microscopic circuits), biochemical or organic (such as DNA computers), or quantum (such as quantum computers). Nanocomputers deal with materials at a molecular level and hold the promise of creating increasingly smaller and faster computers, an important concept in the realm of pervasive computing. In the future, nanocomputers will become a reality. As computer technology advances, computers became more powerful,

while their size decreases. This is because the basic unit of the computer, the transistor, has decreased in size.

According to Moore's law, however, using currently available technology, computer size should reach a lower limit by about 2006. This is due to the laws of quantum mechanics along with the limitations of fabrication techniques for transistors. Once the transistors have shrunk to less than 0.1 micrometres they will probably not function properly. Many transistors are made from silicon oxide. As the transistor width decreases, the oxide becomes porous and doesn't act as an insulator.

For functionality, transistors require the ability to be turned off and on. When the oxide is unable to act as an insulator, there is not enough resistance for the transistor to turn off. This being the case, transistors have a natural lower limit in size of about 0.1 micrometres, or possibly as low as 50 nanometres.

Any transistors that is smaller than that will not function and as such will not be of any use in a computer. The implication of this low limit size for transistors is that decreasing the size of computers further will require venturing into a new technological field.

Most researchers are turning to nanotechnology to decrease the size of modern computers. Potentially, a transistor could be made with a 1 nanometre minimum size feature. One nanometre is one billionth of a metre, and has a linear distance spanning approximately ten atomic diametres. In the event that a transistor could be made with a 1 nanometre minimum feature size, then over 10,000 nanodevices would fit into the same area as a present day

transistor. Ultimately, this means that a nanocomputer could be orders of magnitudes more powerful than modern computers.

The purpose of this paper is to introduce potential devices in the nanotechnological field that could be used to build and design a nanocomputer. Fabrication of these devices will discussed, along with an analyses of the way in which the nanocomputer will be built in terms of its architecture. Finally, the work that still needs to be done will be addressed. Currently, there is no market to drive the production of nanocomputers. Most of the work being done is occurring in the laboratories of Universities.

NANOTECHNOLOGY

Discussions of the possibility of nanometre-scale devices started in 1959. Physicist Richard Feynman lectured that manipulating atoms one at a time could theoretically lead to the assembly of large numbers of completely identical devices. The concept of building structures one device at a time didn't catch on until the 1960's or 1970's.

During this time, advances occurred in the decrease in size of transistors. Transistor size was decreased by increasing the density of the transistors. In the 1960's Moore observed that the feature size for devices on a semiconductor chip was decreasing by a factor of 2 every 18 months. As mentioned, this concept became known as "Moore's Law," which is what gives us the prediction that transistors will reach a minimum size by 2006.

Concurrently, advances in biology and chemistry allowed for the manipulation of matter atom by atom and molecule by molecule, rather than in the bulk. Geneticists began

exploring the use of natural biological processes to manipulate proteins and other molecules. Splicing of DNA and RNA into longer sequences had also been accomplished.

Eventually, scientists began to envision electronic circuit elements made from single molecules. The ability to position single atoms has become a reality in the 1990's. Also, nanometre-scale quantum effect devices such as artificial atoms, or quantum dots are candidates for building blocks for molecular electronic devices.

POSSIBLE TYPES OF NANOCOMPUTERS

Several types of nanocomputers have been proposed to build the best nanocomputer. These include mechanical, chemical, quantum and electronic nanocomputers. Briefly, mechanical nanocomputers would use molecular-scale rods of rotating molecular-scale wheels. These wheels would spin on shafts and bearings. In this way, the mechanical nanocomputer could be used to calculate. Also, it may be possible to assemble mechanical nanocomputers using the mechanical positioning of atoms or other molecular building blocks one atom/molecule at a time. This process is known as "mechanosynthesis."

A chemical computer will process information by making and breaking chemical bonds. The resultant chemical will store the logic states or other information. A variation of this type of computer is a biochemical computer. Multicellular nervous systems demonstrate that this type of computer would be possible as this is the way in which animal nervous systems work. Every single action in an animal is mediated by biochemical reactions and pathways. This evinces that biochemical reactions and pathways could be used to store

information. However, most biochemical pathways and mechanisms are not currently well understood.

A quantum computer would consist of many parallel computations performed by "natural" mechanism of interference among the quantum waves. These quantum waves are associated with the nanometre-scale elements of a multi-component, coherent quantum state. Each piece of information would be represented by a quantum state of some component in the computer. Finally, there may be developments in electronic nanocomputers. This type of nanocomputer is the most likely because this is the way in which computers are currently developed, specifically through the use of electronic digital computers. As such, we should be able to expand on this knowledge and development in order to create a nanocomputer.

One noteworthy benefit is that this type of approach would not require a fundamental change in operating principles of the transistor. As mentioned, a change in the technology will need to happen to further decrease the size of computers. Fundamentally, an electronic nanocomputer will store and move electrons in order to represent information. This is the same way in which transistors work in modern computers. From here, we will discuss some theoretical technologies that will potential serve to build a nanocomputer.

POTENTIAL TECHNOLOGIES FOR USE IN NANOCOMPUTER

As discussed, transistors will need to be replaced by some form of nanotechnology in order to decrease the size of computers. Several replacement technologies have been suggested.

The technologies to be discussed include the resonant tunneling transistor, single electron transistor, and the quantum dot cell. Each of these devices is based upon the principles of quantum mechanics.

Resonant Tunneling Transistor

The resonant tunneling device (RTD) may be used to create a nanometre-scale two-state device. Often, RTDs are constructed from semiconductor heterostructures. The resonant tunneling diode consists of two insulating barriers in a semiconductor. This creates an island between the two insulators.

This centre island should be around 10 nanometres in width. In the event that electrons are confined between two closely spaced barriers, the quantum wave properties of the electrons restrict their energies to certain discrete levels that are closely spaced in energy. As a result of this energy restriction, the potential well contains only a finite, integral number of discrete "quantized" energy levels. Energy quantization is the way in which the resonant tunneling diode operates.

Electrons are able to pass through the device by tunneling through the two barriers. The probability that an electron will tunnel through the two barriers depends on the energy of the incoming electrons as compared to the device's internal energy level. Ultimately, current will only flow when the energy of the incoming electrons is the same as the energy levels that are allowed inside the potential well. When this is the case, current is allowed to flow through the device because the energy level of the electrons outside the well is in

"resonance" with the energy level of the electrons inside the well. However, no current is able to flow when the energy of the electrons differs from the allowable levels inside the potential well. Again, the average energy of the incoming electrons must align with the internal energy level of the potential well in order for current to flow.

A resonant tunneling transistor can be made by incorporating an RTD into the emitter of a conventional bipolar junction transistors (BJTs), which is a microelectronic device. A BJT is a three terminal device that is similar to a MOSFET. The difference is that a BJT is a current-controlled amplifier, rather than being a voltage-controlled amplifier. Also, the gate, source, and drain of the MOSFET are analogous to the three terminals of the BJT, referred to as the base, emitter, and collector. In a BJT, the current that flows into the base determines the amount of current that flows between the emitter and collector. Small changes in the base current result in a large change in the collector current, allowing the BJT to be used as an amplifier.

The combination of an RTD and BJT is a resonant tunneling transistor (RTT) and this particular arrangement can allow the RTT to be a two-state device, which can be in an "on" or "off" mode. The way that this works is that the RTD is able to serve as a filter. The RTD will allow current to flow to the emitter of the BJT at certain base-emitter voltages. The voltages allowed correspond to the RTD's internal energy levels.

The "off" mode results when there is a low base-emitter voltage, which does not allow current to flow through the base-emitter RTD.. The transistor is "on" when the base-

emitter voltage is increased so as to coincide with the first internal energy level of the RTD, thereby allowing the base current to pass through the RTD. Of note, the transistors can have multiple "on" and "off" states. What this allows for is multi-state transistors. These multi-state transistors enable for a reduction in the number of devices necessary to implement logic functions. Ultimately, the density of logic in integrated circuits will be increased. A problem with this RTT is that we have a nanometre device integrated into a microelectronic device.

This combination means that the size will decrease only so much, due to the presence of the microelectronic device. Hopefully, some combination of two nanodevices will be possible in the future in order for the size of computers to decrease.

Single Electron Transistor

A single electron transistor (SET) is another device that has the potential to be used in developing a nanocomputer. The SET does not experience quantum effects such as the interference among a few electrons. An SET operates by moving single electrons that have electrostatic interactions with a large number of surrounding electrons.

An SET consists of a source, an island, and a drain. The source and drains are wires. Electrons are able to enter the island only one at a time. The electrons tunnel onto the island from the source and then leave the island via the drain. This flow of electrons produces a flow of current. Very few electrons can remain on the island at the same time, due to the electrostatic repulsion of the electrons on the island. When extra electrons are prevented from tunneling, no current is

able to flow. The phenomenon whereby electrons are prevented from tunneling is called a "Coulomb blockade."

It is useful to control the number of electrons that enter and exit the island. This can be done by a metal gate electrode near the island. When the voltage of the gate electrode is increased, an additional electron is able to tunnel onto the island from the source. From there, the extra electron tunnels off the island onto the drain. This electron flow continues and creates a measurable current flow through the island.

An increase in the gate voltage will cause the number of electrons on the island to stabilize and no current will flow. However, a further increase in voltage will cause more electrons to migrate onto the island. This results in a step-wise function in the current flow. One limitation of the SET is the fact that at high temperatures, the thermal energy of the electrons may be able to overcome the Coulomb blockade. Ultimately, the thermal energy may be sufficient to allow the electrons to tunnel onto the island. In this case, current will flow regardless of the gate voltage. Ultimately, more work needs to be done in order for the SET to be fully operational at room temperatures. Once this occurs, SETs may be usable in the creation in nanocomputers.

QUANTUM DOT CELLS

A final potential technology that may assist in the development of nanocomputers is a quantum dot cell. A quantum dot is a small potential well or box that is able to electrostatically isolate a single or few electrons. The electrostatic field of the quantum dot can be changed in order to determine the number of electrons in the quantum dot. Quantum dots are made from tiny insulated regions of

conducting material. They are capable of holding zero to hundreds of electrons. Quantum dots rely on specific quantum effects, unlike SETs. Another difference is that quantum dots cannot be used to store and retrieve information because the exact number of electrons in the device is unknown, due to the low resistivity of the device. However, interactions between quantum dots can retrieve and store information. Quantum dots can affect one another even if they are not wired together. Two dots can influence one another due to long-range electrostatic interactions. One dot's electric field of electrons can change the number of electrons in another nearby quantum dot.

The way that this works is that the addition of an electron to one quantum dot causes an electron to vacate a nearby dot. This is contingent on the electron having a vacancy to jump to. Quantum dots can be lined up to cause movement of electrons. The first quantum dot in this series must have an electron to start the movement of electrons in quantum dots. This electron will have to tunnel into the first quantum dot from a nearby reservoir. Since the quantum dots can be used this way, they can be two-state devices. The two states will correspond to occupancy of the dot by zero or one electron. When two devices are set next to each other, they will obtain opposite states due to the electrorepulsion between electrons. Quantum dots can be arranged in a certain series and can serve as electron reservoirs.

Quantum dot cells communicate through their electric field (exchange of photons), rather than through the flow of current (exchange of electrons). Since this is the case, quantum dots can give rise to wireless computation. The major obstacle in

this type of technology is the fabrication of the quantum dot. It is also imperative that the location and size of the dots is precisely controlled. Finally, as with SETs, quantum dots have a maximum temperature of operation, especially as the size of the dot decreases.

These are issues that need to be addressed before the use of quantum dots becomes a reality in the construction of nanocomputers.

FUTURE OF TECHNOLOGY

It is exciting to look into the future of technology. In an age of continuous innovation and invention, when the discovery of today loses its sheen tomorrow, it is not easy to pinpoint technologies that will transform our future. Engineering and technical developments are everyone's concern, as they will not be confined to industry, university classrooms, and R&D labs.

Instead, they will make a tremendous difference in our day-to-day lives. Here I will attempt to identify some of the technologies that will revolutionize our lives and our values in the coming years.

QUANTUM COMPUTERS

Unlike current PCs, quantum computers will have switches that can be in an on or off state simultaneously. The mechanism that will make this possible is known as superposition, and the switches are referred to as quantum bits. The system will make quantum computers operate very fast. A basic quantum computer is likely to be operational by 2020.

PROGRAMMABLE MATTER

Scientists are in the midst of creating a substance that can take a specific shape to perform a specific task. The substance is known as claytronics, and it consists of catoms. Individual catoms are programmed to move in three dimensions and position themselves so that they assume different shapes.

This technology is likely to have numerous applications ranging from medical use to 3D physical rendering. It may take around two decades to become a reality.

TERASCALE COMPUTING

Techies are working on a project that would make our PCs able to contain tens to hundreds of parallel working cores. The device will have the capability to process huge amounts of information. To create this technology, Intel is exploring the possibility of using nanotechnology and allowing for billions of transistors.

REPLIEE ROBOTS

Repliee is one of the most advanced life-like robots ever created. Repliee, an android, is covered with a substance which is very similar to human skin. Sensors placed inside the robot control its movements and enable it to respond to its environment. Astonishingly, the robot can flutter its eyelids and replicate breathing. Repliee operates best in a static condition.

ORGANIC COMPUTERS

To further advance the computing realm, techies need to create a hybrid CPU that is silicon based but contains organic

parts as well. The most promising progress in information processing concerns a neurochip that places organic neurons onto a network of silicon or other materials.

Future computers will be able to bridge the silicon and organic spheres to utilize processors that incorporate both of these elements. When we speak of organic, we're usually talking about health, sustainable farming methods, food, a way of life.

Now, let's talk about computers. Our computers today are based on silicon transistors. Now, I'm not an electrical engineer so I don't pretend to understand exactly how they work, but here's the central point: Silicon chips have tiny, tiny grooves in them that serve as on/off switches for electronic signals. Each on/off pathway is a transistor.

The more transistors you can fit on a chip, the more calculations it can make. Since Intel developed the first silicon-based computer chip in the late 60?s, technology has advanced in making these chips chock full of many more transistors, jacking up computer speed as we get better at cramming more into the chip. The problem with this approach is, on the scale that we are currently functioning in, we can only fit in so much on the chip until there is simply no space left and we hit a limit.

Enter organic nanoscale electric circuits. A nanoscale is scale that is molecular in size, using the actual organic molecules to convey the electric signals instead of having them travel through a small, but still not nearly molecular, sized groove in the pathways of a silicon chip.

Researchers of this new technology say that they have succeeded in coupling together several contacts in an electric

circuit, and doing so has enabled them to produce prototype computer electronics on the nanoscale. "We have succeeded in placing several transistors consisting of nano-wires together on a nano device.

It is a first step towards realization of future electronic circuitry based on organic materials - a possible substitute for today's silicon-based technologies.

This offers the possibility of making computers in different ways in the future," said Thomas Bjørnholm, Director of the Nano-Science Centre,

Department of Chemistry at University of Copenhagen. The revolutionary capabilities of this type of technology could be enormous, as well as dangerous.

You can get much smaller than the molecular level, except if you go quantum, which we're really far from figuring out how to do.

Never mind that though. Just know that, with organic nanocomputers, computing speed could go up by a factor of thousands.

SPRAY-ON" NANOCOMPUTERS

The "spray-on" nanocomputer would consist of particles that can be sprayed onto a patient. It would monitor the patient's medical condition and communicate wirelessly to other machines.

CARRIER ETHERNET

Carrier Ethernet is a business service/access technology. It can serve as a transport method for both business and residential service. Ethernet will dominate the metro space in the future and will slowly displace SONET/SDH over the

next 10 to 20 years. Development sustains life. However, techies cannot afford to forget that technological advancement will remain inadequate in the absence of contributions from all branches of knowledge and will not flourish if it does not benefit society.

2

Developments in Nanocomputing

There have been countless approaches for handling the advent of molecular computing, and Computer scientists today have come up with several proposals to build components of nanocomputer. Some of these approaches will be summarized below, making note of the advantages and disadvantages in their implementation.

DNA CARBON NANOTUBES AND NANOWIRES

We know that a computer's building blocks can be a key component to system efficiency and speed. An organic carbon based nanowire structure has been proposed by a group of researchers from Duke, NC State, and The University of North Carolina. These researchers have focused on using the building block of the human body as the basis of building a computer. In this proposal, the top-down assembly approach of lithography would be replaced by DNA guided self-

assembly. The DNA assembly uses strands of DNA to create tiles which in turn create a lattice structure.

They focus on a waffle-like lattice of 16x16 nanometre size, and they attach carbon nanotube devices to the DNA tile scaffolding through a process they call functionalization, which is the process of adding a molecule to another material through a chemical reaction. This circuit architecture does have a few shortcomings inherent in it. The self-assembly of the DNA lattices proposed does not produce a regular pattern which would insure a low level of defects.

They have found that by minimizing the number of tags and producing repetitive structures is desired. However, by minimizing the number of tags, there is a limit to the amount of complexity a circuit can then have. The DNA self assemblies will not be able to have as much complexity as a CMOS transistor as a result.

In a shift away from the CMOS circuit, the DNA self assemblies are designed to have some level of defect tolerance. This is mostly due to the fact that the error rate in photolithography is quite low. However, self-assemblies will inherently have a higher number of defects in production. This particular proposal does not outline a method for handling defects, and that is a shortcoming of this presentation.

The proposed architecture for DNA self assemblies must handle limited node connectivity and random node placements. These limitations affect the performance in large scale control of the nodes. An accumulator based instruction set architecture is proposed to deal with the control and communications limitations.

The accumulator-based ISA reduces the need for widespread coordination and communication among many components, since the only data dependence involves the accumulator, and instructions are processed in order. Each instruction within the accumulator makes up an execution packet containing operands, operations, and the accumulator value to form a sequence.

This approach supports parallelism through the independence of data, and by initiating several execution packets at once. This proposal also puts forth a networking architecture for the execution packets. Three logical networks are proposed to handle three different types of information: memory requests, memory responses, and execution packets. Memory and execution are separated onto two different physical networks to avoid problems with deadlocking. This research has mapped out some of the major hurdles they must clear and are thinking ahead to putting their fabrication together to make a system, but they are still in the beginning phases.

Another paper produced by this group dealing with current research topics outline their experimental progress with carbon based nanotubes and nanowires. A 4x4 DNA tile has been produced for a variety of application computations, and they have demonstrated highly parallel computation via a set of DNA crossover tiles. [LABEAN] They hope to fashion a simple circuit with nanotube control in the near future.

CHEMICALLY ASSEMBLED ELECTRONIC NANOTECHNOLOGY (CAEN)

This approach is presented by Seth Copen Goldstein at Carnegie Mellon University. Goldstein believes that CAEN

devices will be the cornerstone of the nanocomputer and replace the dominant silicon transistor. The devices he proposes are small, dense, low power and reconfigurable. The proposed fabrication of these devices makes switches and wires using chemical self-assembly as opposed to the traditional lithography method which is currently used for CMOS chips. As stated by Goldstein, "The active devices are created where two appropriately functionalized wires intersect.

By choosing the molecules that coat the wires appropriately, the intersection points can be ohmic contacts, diodes, configurable switches, etc. The important point to note is that no precise alignment is required to create the devices." After proposing how these devices would be manufactured, Goldstein goes on to explain how a computer using these devices might be architected with his current fabrication plan.

Cost and fault tolerance is the most important aspects of CAEN devices, and Goldstein explains that the proposed manufacturing process will result in the following architectural characteristics:

- Transistors, or other three terminal devices, will probably not be part of CAEN devices.
- Any deterministic aperiodic circuit will have to be created after the device is fabricated through reconfiguration.
- Chips will have to work in the presence of a significant number of defects

Unfortunately these CAEN devices cannot create three terminal devices because precise alignment is required and

this is not economically viable. CAEN devices would have to be limited to two terminals, and any processes which required three terminals (such as an inverter) would have to be built in conjunction with a CMOS device.

Goldstein believes that CAEN based devices can replace the silicon chip. These devices have the possibility of being reconfigurable, testable, and can be fault tolerant. A device called a negative differential resistance (NDR) can replace the traditional transistor circuit.

Unfortunately, current research requires a great deal of voltage in order for this device to work. Further study will be required in order to produce a low power and low static device. Despite its current limitations, the NDR is a key piece in Goldstein's proposed resonant tunneling diodes which are the building block of the molecular latch which will in his mind ultimately replace the transistor as we know it.

NANOBLOCKS AND NANOFABRICS

Goldstein has built upon the concept of Chemically Assembled Electronic Nanotechnology (CAEN) by outlining a reconfigurable architecture called a Nanofabric. This Nanofabric architecture could be used for factory-programmable devices configured by the manufacturer to emulate a processor or other computing device, or for reconfigurable computing devices.

A Nanofabric is a two dimensional mesh of interconnected NanoBlocks. A NanoBlock is a logic block that can be programmed to implement three-bit input to three-bit output Boolean function and its complement, as well as for switches to route signals.

This paper outlines how to build an AND gate, an OR gate, and XOR gate, and an ADDER. These "gates" as we traditionally are used to calling them, exist in what they call the molecular logic array (MLA) where all of the functionality of the block is located. The upside to the method outlined in the paper is that it is similar to commercial Field Programmable Gate Arrays (FPGAs), allowing these nano-gates to leverage current FPGA tools.

The MLA would be created using "directed assembly", which means that the atoms are arranged one by one. While this would allow computer makers to place components in a precise manner, the signals, which are routed using diode-resistor logic, would be degraded every time it goes through a configurable switch.

While a plan for defect and fault tolerance is presented, the inherent problem is with the nanowires lacking resistance. Goldstein contends that while this is a problem, it is still easier to implement defect tolerance than with CMOS because wires can be swapped and rows can be moved anywhere within the NanoBlock without affecting the circuit. CMOS currently has little defect tolerance, and chips with a high error rate are discarded. The cost of directed assembly may be prohibitive as well and not economically viable. Testing the Nanofabric proved to be difficult because it is not possible for the components to be tested in isolation. After production of the fabric, a defect map is created to make the fabric into a viable circuit. The fabric is self-testing, and will provide fault free regions of the fabric.

Other researchers at the University of Massachusetts are looking to use Nanofabric to create integrated circuits which

can be tailored towards the application domain. They are named NASICs, which stands for Nanoscale Application-Specific Integrated Circuits.

This research is built upon the Nanofabric proposal by Goldstein. In an effort to build these NASIC tiles, the researchers at UMass have presented a solution to a difficult problem. As they state, "A key issue when building NASIC tiles is the low effective density of latch circuits.

This is due to the difficulty of building sequential circuits on a nano-grid." In order to get around this limitation, they have developed a nano-latch, which allows for temporary storage along a nano-wire in a circuit. This research gives promise for improving nano-device density, scalability and reliability

QUANTUM CELLULAR AUTOMATA

A Quantum Cellular Automata (QCA) is another plausible nanodevice. This is a design which is based upon a theory of using electrons and their energy states for computing. This is a topic which has been studied for over 40 years, back to the work of von Neumann. Rather than being an individual device of a computer circuit scaled to a nanometre size, the QCA is a complete entity.

As described by researchers at Notre Dame, "quantum-dot cellular automata (QCA), parallels conventional computing in using a binary representation of information (not qubits) yet realises functionality in ways which are well-suited to the properties of single molecules - using molecules not as current switches but as structured charge containers."[LENT] The QCA is a single molecule which will contain a representation of the binary values 0 or 1 based

upon the static electric charge of the electron. The cells have not been found at this point in time to be viable at room temperature, but rather only at very cold degrees in the Kelvin range. Needless to say, operating at such low temperatures is not practical for any working device, and this is a serious limitation to the usefulness of QCAs.

The QCA fabrication does differentiate from other devices that we have seen because the fabrication uses lithography. "Molecular QCA cells must be attached to surfaces in ordered arrays. We are using electron-beam lithography to burn narrow "tracks" along surfaces."

The temperature limitations of Quantum Cellular Automatas have been given a new perspective through a study of Magnetic QCA described by Cowburn and Welland. Their presentation uses interacting submicron magnetic dots to perform logic operations and propagate information. [BECKETT] Although these devices will not be in the nanometre range, they will use less power than current CMOS.

CMOL

Some researchers are presenting a possible bridge from CMOS to the next generation of molecular computing by formulating a hybrid circuit. One such hybrid circuit is dubbed a CMOL, a play on the combination of CMOS and molecular. The CMOL is "a circuit [which] combines an advanced CMOS subsystem with two, mutually perpendicular, arrays of parallel nanowires and similar molecular devices formed at each crosspoint of the nanowires." [LIKHAREV]

This approach is important in several ways. First, it's unlikely that the computing industry will not be able to make the leap from its top-down lithography methodology to a bottom up approach using nanowires and integrating with some molecular components without some intermediate step. The CMOL maintains the stability of the silicon chip and takes a small step forward by using bottom up fabrication. The CMOS circuit maintains the level of functionality that we are used to. As mentioned, the system bus can be an area in which slowness can occur. Experiments with the CMOL circuits have operated faster than regular computers with better power dissipation.

The creator of the CMOL hybrid circuit claims, 'The development of CMOL technology (especially the high-yield molecular self-assembly) will certainly require a major industrial effort and substantial time period, probably not less than 10 to 15 years. However, this timing may be still acceptable to prevent the impending crisis of Moore's law, provided that we start right now."

MARKOV RANDOM NETWORK

Researchers at Brown University have a proposed architecture based on Markov Random Fields (MRF). A MRF is a concept based on Markov chain, which is a "graph of stochastic variables, where each variable has the property that is independent of all the others (future and past) given its two neighbors" A Markov Random Field network allows circuits to behave in a relatively independent fashion, allowing for re-configurability and a good level of fault tolerance.

The researchers feel that carbon nanotubes are one of the most promising devices which have been built at the

Nanoscale. Some issues are inherent with these nanodevices though; these researchers point out that the use of carbon nanotubes circuits will increase failure rates and bring about heat dissipation issues.

In order to deal with possible failure rates upwards of 10% and heating issues at the thermal limit, it is necessary to configure a network of circuits which can handle a high fault level and can somehow dissipate heat throughout the circuit network.

It has been found through experiments that the use of a Markov Random Field Network in conjunction with the Gibbs formulation for dissipation of heat can be a viable solution, because as the authors state, "its operation does not depend o perfect devices or perfect connections. [...] Successful operation only requires that the energy of correct states is lower than the energy of errors"

CELL MATRIX

The Cell Matrix architecture is presented by the Cell Matrix Corporation in Salt Lake City. This piece of research differs from the previous sections in the sense that this is a proposal of a true architecture as opposed to a design of a nano-sized component.

This architecture is currently being designed and tested in the silicon chip domain, but all research, configuration and design is for the coming of nanostructures. Currently, the Cell Matrix Corporation is developing an atomic cell unit, which is repeated to form a multidimensional matrix of cells to make a highly scalable architecture, which differs from other proposals discussed thus far.

The architecture is similar to the field programmable gate arrays (FPGAs) like the Nanofabric described in an earlier section, and it also has similarities to the cellular automaton proposed by Von Neumann. Several aspects of a cell matrix distinguish it from a Von Neumann cellular automaton. First, the cell matrix has the same three dimensional nature.

The cell matrix can only be in a certain number of states it can be in at any given time, just like a cellular automaton. Second, a cell matrix is fine grained and reconfigurable, and third it is programmed like a digital circuit, and not like a cellular automaton. The benefits which cell matrix architecture provides are many.

It can efficiently handle a large number of switches due to the fact that its controlling system is highly scalable, and can handle any addition of cells to the system. This architecture also promises to provide highly parallel large scale computing. This is the result of the fact that each cell contains all necessary functionality, and can compute across the matrix at any point in a parallel fashion.

This group has demonstrated this highly parallel computing ability through an experiment with a search space problem. The cell matrix architecture relies on a very simple hardware definition mixed with a complex programming of each individual cell. Simple logic functions can be handled by individual cells, whereas more complex functions are handled by a collection of cells, which are spatially close to each other. In these collections, cells are set up to perform a subset of the work of the entire circuit.

These cells communicate with one another in an asynchronous fashion. Each cell also has the ability to

reconfigure itself, allowing for self-testing and fault tolerance, although experimentation and testing in this arena is still taking place.

The Cell Matrix Corporation believes that their design can provide dynamic circuitry, allowing each of the cells to change their original behaviour or the behaviour of cells around them, or to migrate into new areas within the matrix. They also see this as being a mechanism to handle faulty cells in order to make their circuits fault tolerant.

Defect tolerance is always an important issue with any computer architecture due to the cost which defects add to the final cost of a circuit. Due to the fact that silicon chips are discarded if their fail rate is too high, a new architecture can find cost savings in total manufacturing costs if the new circuitry can fin a way to be fault tolerant by working around bad cells. The Cell Matrix has implemented defect tolerance and self testing in their architecture.

They have designed a test driver which has the following goals:

- Permit reporting of faults with a high resolution
- Permit access to a region despite failed regions near it.
- It should be easy to extend to a decentralized, parallel, distributed fault testing process with as small a footprint as possible.
- Have a driver which can share the hardware with other tasks so it can perform its functions on subcomponents of a critical system while that system is running. [DURBECK01]

The Cell Matrix architecture outline is well planned and several experiments on digital circuits have been promising. However, no experiments have yet been performed on any microscopic components. In theory the layout may be scalable and successful; more incorporation of current nano-component research should be considered.

NANOPRISM

It has already been determined that fault tolerance, while something not so prevalent in the construction of silicon chips, will need to be an important part of any architecture presented for a nanocomputer. One way in which many manufacturers provide better reliability is by implementing techniques to increase redundancy.

However, it is known that adding redundancy cannot always increase reliability of a device due to the fact that the redundant device can also contain faults. A paper from Virginia Tech is attempting to determine the level at which fault tolerance and redundancy can produce a reliable nano-architecture.

As stated by the researchers, "The questions we try to answer in this paper is, what level of granularity and what redundancy levels result in optimal reliability for specific architectures. In this paper, we extend previous work on evaluating reliability-redundancy trade-offs for NAND multiplexing to granularity vs. redundancy vs. reliability trade-offs for other redundancy mechanisms, and present our automation mechanism using the probabilistic model checking tool PRISM."

The NANOPRISM tool is based on another probabilistic model checking tool called PRISM built at the University of

Binghamton, which is designed for conventional CMOS architectures. The NANOPRISM tool will be a valuable resource for building a new architecture for a nanocomputer. Redundancy can be a very important yet delicate technique to employ, due to the fact that employing too much can cause a lessening of reliability, not to mention the fact that there are several levels of granularity at which redundancy can be implemented. Because this tool automates the checking of a defect tolerant architecture, areas in which trade off levels are achieved can be identified much more quickly. The number of circuits and cells are going to increase exponentially as nanodevices scale up to the level they are supposed to be at. If we currently have almost 1 billion transistors on a single silicon chip, the number of devices we can just imagine the numbers present in a nanocomputer.

PROPOSED NANOCOMPUTER ARCHITECTURE DESIGN

After looking at all of these different nanodevices and proposed networks and architectures, how can we determine what is the best direction to go in order to create the optimal computer architecture for the next generation? The first thing to look at is what will be the major issues and limitations with the devices we wish to use as the basic building blocks for our computer.

The focus will not necessarily be on how these devices will be built whether that is through lithography or self assembly, it is immaterial, and concern about cost will be factored in, but not necessarily a focal point. Rather, we are more

concerned with what qualities the finished product and the surrounding architecture will need to possess in order to be successful.

What are some qualities are interested in?:

- Fault Tolerance,
- Self testing/Re-configuration,
- Heat Tolerance and Power Dissipation,
- Independent and parallel processing,
- Improved communication, handling current pervasive bottlenecks in interconnections,
- Scalability.

We will use these important qualities as a guide to what each nanocomputer component will contain, as we lay out our proposed nanocomputer architecture from the bottom up.

CHIP DESIGN

The chip design of the nanocomputer would be best suited to have a carbon base which is both self-testing and reconfigurable. Because we know that faults will be at levels upwards of 10% for any chip built in the nanometre range, it will be necessary for these features to be part of any solution we choose. Also, it's important to remember the breadth of our failure rate. A Nanocomputer will contain not just one billion but several billion chips. If we have a computer with 5 billion chips, we will have 500 million chips which are defective with a 10% error rate. It will be an important cost factor to be able to repair these chips after they have been fabricated.

Of the research we have gone over thus far, the best proposed replacement for the CMOS chips is the NanoBlock,

based on the FPGA, the Field Programmable Gate Array. These blocks are carbon based. These devices have been created from a bottom up assembly; they are fault tolerant and reconfigurable.

Bad devices or blocks of devices can be swapped out if necessary Another advantage to the NanoBlock is that it is chemically assembled, allowing more flexibility with reconfiguration. The devices can be manipulated so that changes can be made to improve any faults which may make an entire block unusable.

The argument for NanoBlocks also stems from the need to insure heat and power dissipation. The NanoBlocks are low power devices, and this makes it a good candidate as a base for building a nanocomputer.

INSTRUCTION SET DESIGN

Very little research in nanotechnology has addressed the future of instruction set design. We know that our hardware systems will become more complex and more powerful. We can infer that the instruction set design should then move back from a reduced instruction set (RISC) design to a more powerful complex instruction set (CISC).

This way, we can take advantage of the improvements in the hardware and write less complex compilers. The instructions can also perform more work in a single clock cycle, improving the speed operations are performed in the system. However, the one architecture that we have looked at which took instruction set design into consideration only used a simple accumulator. One would think that a more complex design would be beneficial and more productive,

but it is possible that a simple approach will be the best way to manage so many million components.

RISC was chosen as the instruction set for this generation of components due to cost, and it's conceivable that the next generation will choose a similarly simple instruction set due to size constraints and sheer numbers.

The instruction set will presumably be one of the last design components to consider, but designers must take it into consideration when building components.

INTERCONNECTIONS AND COMMUNICATION POINTS

Interconnections are another vital piece to consider. If there can be devices in the multi-billion in a nanocomputer, these nano-connectors will also number in the multi-billion and will need to be efficient and allow for both local and global communication.

The development of nanowires as communication pathways between neighboring cells is the most well developed form of nano-sized interconnects. The carbon based nanowires are the most promising format because of their stability.

These devices also allow for self-testing, reconfiguration, and self-assembly. Each of these qualities makes it an attractive component in terms of costs, reliability, and practicality. It is also presumed that the reduction in size of these wires can improve density and we can move more data through interconnections at a faster rate. This can reduce starvation which can sometimes occur in a system due to bus latency and lack of bandwidth.

While carbon nanowires have been produced in the lab, the nanowires which we must use in our nanocomputer must

improve their resistance. Arranging these nanowires in a Markov Random Circuit as one group of researchers has proposed would be a valuable arrangement, and would be an important algorithm to employ in a nanocomputer.

MAIN MEMORY, MULTIPROCESSING AND INSTRUCTION LEVEL PARALLELISM

Memory management is and will continue to be a very difficult topic facing all computer designers. Virtual memory demands which are currently limited by space will be unlimited in a nanocomputer architecture. Memory will be more plentiful, cheaper and faster in a nanocomputer, and we can expect to move from the 1 gigabyte of memory standard on modern computers to numbers two to five times more powerful. However, these advantages will bring a new host of issues.

Instruction level parallelism may be more difficult because there will be so much more data to handle. Algorithms will need to be developed to insure that critical sections of code are executed properly. However, we do know that we will have many more CPUs which we can cluster together with much less space and cost requirements. It is likely that instruction level parallelism will fall away to a new paradigm of handling large scale multiprocessing.

Perhaps machines will be split apart so that certain CPUs are specifically for certain jobs, or certain tasks within a computer. Nanocomputer research has not reached the point of coming up with a memory management scheme, but the chances that it is exactly as we handle memory today is unlikely.

STORAGE

A nanocomputer will have the possibility of an unlimited amount of storage space. Currently we have computers which can hold up to 150 gigabytes of information. We expect a nanocomputer to be able to hold terabytes of information. These storage disks will also be easily searchable and can also be partitioned.

Presumably we can hold volumes of data, and backup tapes or requiring repositories to be placed on a network of computers to make up a server can be a thing of the past. Smaller embedded devices will be able to hold much more data as well.

Our entire medical history can exist on a key, or entire music collection can exist on a device the size of a credit card or smaller.

CONCLUSION

The future of nanotechnology will bring exciting change to the computing industry. We are promised machines which are faster, cheaper, and smaller. This promise will come with incredible challenges and sacrifice, and the computing industry must be flexible and agile enough to meet the demands of manufacturing these new devices and designing the machines which will contain them.

The industry must also change the current manufacturing processes and current design paradigms to meet the differences between this new generation of devices and our generation.

These new devices will require a completely new approach in manufacturing, moving from lithography top down design

to bottom up design. It may also require self-assembly of components in order to be economically viable. This will require a new set of manufacturing tools and processes. Computer manufacturers will have to change their plants and assembly lines completely in order to build these new components.

A nanocomputer will also require a new instruction set, and innovative ways to handle the new challenges when dealing with molecular components must be considered. Hardware will be less reliable and self-testing and self-configuration must be built in to any nanocomputer. This may make manufacturing more expensive at first, and more time consuming.

The end of hardware improvements predicted by Gordon Moore is upon the computing industry, and lab experiments to build the next generation of hardware must take shape to keep pace with the demanding computing needs of the world. Valuable research has been made in many areas, but we are still in the primitive phases.

The industry has about ten or twenty years left before this paradigm shift must occur. It can be gleaned that we could experience some setbacks in the new generation of devices were they may not perform as reliably or as quickly as we are used to with CMOS based computing devices.

However, one could say that it would be worth taking a few steps back in order to move many hundreds of steps forward, which is what nanotechnology is promising us. Below are some pictorial representations of some of the nanodevices and architectures we have discussed in this paper.

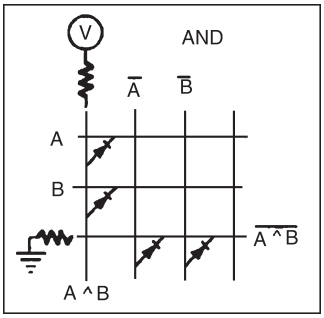


Fig. A Two Input and Gate Implemented in a CAEN Grid

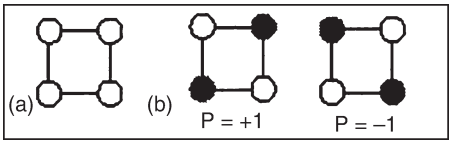


Fig. QCA Four Dot cell

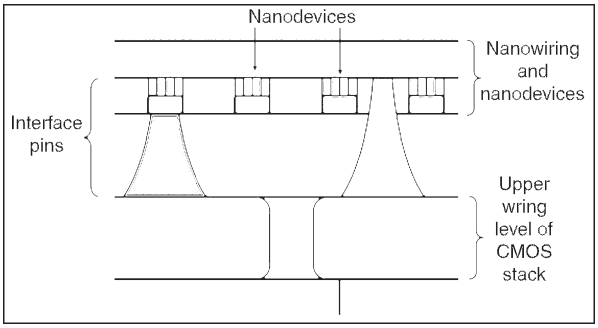


Fig. Structure of a CMOL Circuit

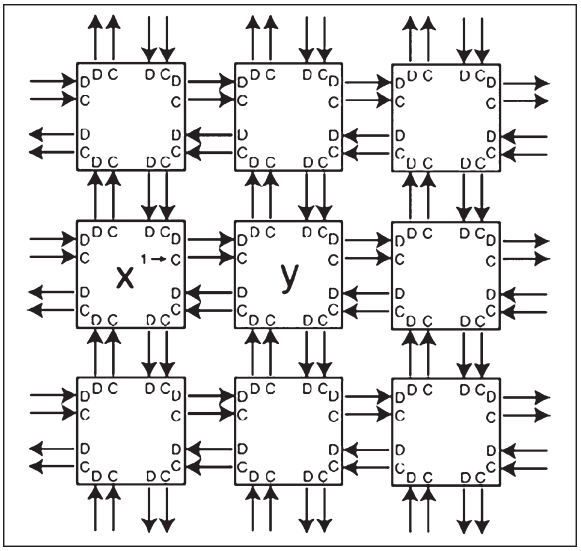


Fig. 3 x 3 Cell Matrix

QUANTUM DOTS

The use of semiconductors has greatly increased in the last century. As new technologies start to rely more and more on semi-conductors, their shortcomings are more and more apparent. Traditional semi-conductor devices have been found to be too big and too slow.

As engineers search for a faster and more adaptable alternative to conventional semiconductors they have discovered quantum dots, a new form of semiconductors that model atoms. Being only nanometres in size, these pseudo-atoms take semi-conductors to a whole new level and can allow devices to work almost at the speed of light. Furthermore, quantum dots have numerous applications in optical technologies, mediums, and industries. This paper seeks to introduce the principle of quantum dots, their creation methods, and their applications. Modern electronics, as well as many other fields of science, rely on the use of semi-conductors. Quantum dots (QDs) are particles that hold a droplet of free electrons which simulates "the ultimate miniaturized semiconductor." Any material that can conduct electricity better than an insulator but not as well as a conductor is considered a semi-conductor. What makes semi-conductors so important is that their unique structure allows different semi-conductors to carry current under different circumstances. This gives the user more control over the flow of current. Most semi-conductors are crystalline substances such as germanium and silicon. We can see its use from the basis of electronic parts such as diodes and transistors to biomedical processes.Conventional semi-

conductors are used often in electrical circuits. However, they have limited ranges of tolerance for the frequency of the current they carry. The low tolerance of traditional semi-conductors often poses a problem to circuits, and many of its other applications. This is what makes the use of quantum dots so important. As they are fabricated artificially, different quantum dots can be made to tolerate different current frequencies through a much larger range than conventional ones.

The use of quantum dots as semi-conductors offers more freedom to just about everything involving the use of semi-conductors. Quantum dots can best be described as false atoms. The primary material that a quantum dot is made out of is called a "hole", or a substance that is missing an electron from its valence band giving it a positive charge. The primary material is extremely small, which is why it is called a dot, and at that size, electrons start to orbit it. Since quantum dots do not have protons or neutrons in the centre, their mass is much smaller. Since the mass at the centre is smaller than that of an atom, quantum dots exert a smaller force on the orbiting electrons causing an orbit larger than that of a regular atom.

Daneshvar, personal communication, Jul 15, 2005). With a mass that small, scientists are able to precisely calculate and change the size of the band-gap of the quantum dot by adding or taking electrons. The band-gap of a quantum dot is what determines which frequencies it will respond to, so being able to change the band-gap is what gives scientists more control and more flexibility when dealing with its applications.

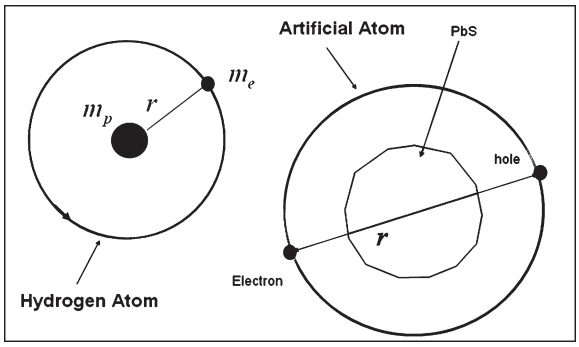


Fig. The Orbit of a Hydrogen Atom to that of a Quantum Dot.

One way to synthesize quantum dots is through molecular beam epitaxy. In this process, certain chemicals are evapourated and then sprayed to condense into small objects on a substrate. The condensation of the chemical on the substrate is similar to water on glass. If someone drops water on glass the water condenses into many balls.

As more layers are sprayed onto the substrate the size of the balls starts to build up into pyramid-shaped objects. Eventually, the balls build up to a specific size and they're quantum dots. This process has some downsides though. It is much harder to use quantum dots while they are still attached to the substrate. While they are all attached together on the substrate they act as one solid which almost defeats the purpose of creating the quantum dots.

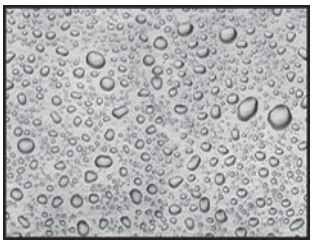


Fig. The Different Chemicals, once Sprayed onto the Substrate, Acts Almost like Water and form Together in "Balls", or Quantum Dots.

Another way to form quantum dots is through electron beam lithography. This process is a little like etching a chip.

A mask is created with an electron beam that has many tiny holes in it. Then evapourated chemicals, similar to the ones used in epitaxy, are sprayed through the mask onto a substrate, creating many little balls.This process has some of the same shortcomings as epitaxy, mainly that the quantum dots are still connected to the substrate after synthesis. Additionally, scientists have found it difficult to create such small masks that need to have holes just nanometres in diametre. Lithography was originally a very popular process for creating quantum dots; however, this process creates many defects and is slow compared to the other processes.

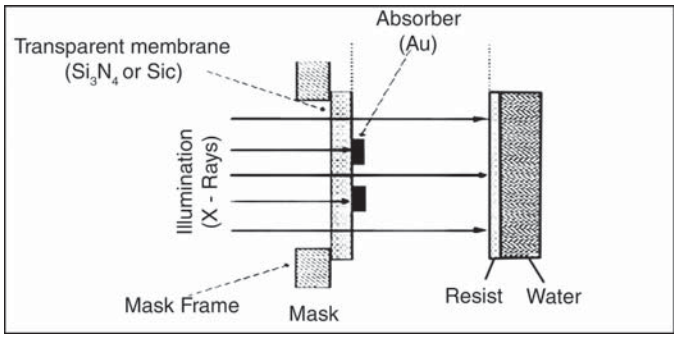


Fig. The X-ray Light Reacts with the Photo-resist on the Water to Create Quantum Dots.

Colloidal synthesis is a process that involves creating quantum dots in a liquid. This is by far the best technique for the formation of quantum dots because the process can occur under "benchtop conditions," or in a normal laboratory setting.

When certain materials, primarily those from periodic groups two through four, are dissolved in a certain type of polymer solute the solution can enter into a phase where particles can come together to form quantum dots. Since the size is dependent on time, the longer the dots are left in

the solution the bigger they get. This is in part what makes colloidal synthesis the most popular method. Scientists can use time to change the properties of the quantum dot, engineering it for certain light frequencies. This process, unlike lithography and epitaxy, synthesizes the quantum dots in such a way that they are suspended individually, making it easier for use in applications.

As mentioned, QDs have many interesting characteristics that not only contribute to numerous applications but also display phenomena that are consequential to the fundamental study of Physics. Typical dimensions of QDs are between nanometres to a few microns. Due to their extremely small size, QDs can be controlled by a few electrons and this provides many advantages that can optimize devices. In addition to the nanoscale size that they exhibit, QDs enable superior transport and optical properties that has beneficial applications in the fields including immunocytoc-hemistry and the study of lasers.

In the biological field of science, QDs have become known to be very useful. Recent studies of QDs have resulted in developing new fluorescence immunocytochemical probes. A probe is a substance that is radioactively labeled or otherwise marked and used to detect or identify another substance in a sample.

A fluorescence immunocytochemical probe is usually used to detect antigens in tissues. In contrast to organic fluorophores, which are not photostable, QDs have properties of high brightness, photostability, narrow emission spectra and an apparent large Stokes' shift, thus they can replace the usage of organic fluorophores. The current mode of

detecting the antigens which takes from two to six days can speed up to a matter of hours using quantum dots.

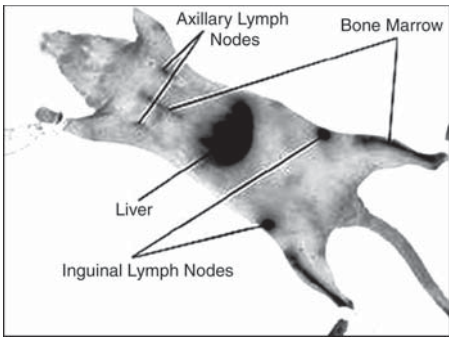


Fig. Immunocytochemical Probes are used in the Dead Rodent. The Probes in the Body have Circulated and now Show up under Florescent Light, Creating a much Safer Alternative to the X-ray.

Prior to the introduction of the QD, microelectronic technology has focused on reducing the size of transistors to produce increasingly smaller, faster and more efficient computers (Shrinking Information Storage to the Molecular Level). However, this method is reaching its physical limit due to the restrictions placed by the laws of physics that do not allow these devices to operate below a certain size. With this advantageous feature of QDs, information storage can be brought down to the molecular level.

Since no flow of electrons to transmit a signal is needed, electric current does not need to be produced and heat problems are avoided. Also, the quantum dot devices are sensitive enough to and can make a usage of the charges of single electrons. With improvements in quantum-dot ordering and positioning, it is possible for us to hope in the near future to address and store information optically in a single quantum dot, thus opening the possibility of ultrahigh-density memory devices.

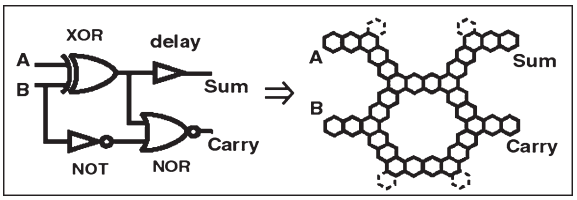


Fig. Nanocomputers might have a Completely new type of Structure made up of 'Cells'. One way of Building this
Structure would be using Quantum-Dots.

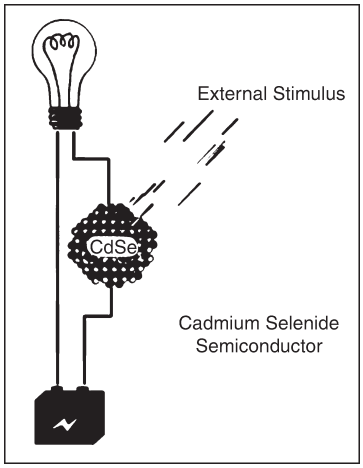
QDs also have other applications like quantum dot lasers which promises far more great advantages than quantum well lasers.Because QD lasers are less temperature-dependent and less likely to degrade under elevated temperature, it allows more flexibility for lasers to operate more efficiently.

Other beneficial features of QD lasers include low threshold currents, higher power, and great stability compared to the restrained performance of the conventional lasers. Respectively, the QD laser will play a significant role in optical data communications and optical networks. Optical switches have been a major research objective in the scientific community.

The use of optical switches would increase the rate at which data can be transferred. With regular switches, data can only travel as fast as the electrical current can, but optical switches can travel,

Almost as fast as the speed of light. The principle of optical switches is that semi-conductors will only allow certain levels of energy to pass through it. So if we place a quantum dot semi-conductor in a circuit, but supply a voltage below the acceptable range, current will not flow. However, if we shine a light on the quantum dot semi-conductor, it would put enough energy into the semi-conductor that it will allow

current to flow. This idea is mainly for powering electronic devices, but using quantum-dots as receivers for electrical data is just a step up.



Quantum dots have applications outside of biology and engineering. An idea that may be instituted in the future would be against counterfeiting money. The treasury could engineer quantum dots to be responsive to a specific frequency of light and suspend them in ink that they would print onto money. Shining light with the same frequency that the ink solution has been engineered for would reveal whether or not the money is real or counterfeit. This idea can be used for just about any substance that could be illegally duplicated. Another security application involves attaching quantum dots to dust. QDs can be engineered so that they have the same properties as dust and give off infrared radiation. In hostile areas, this "quantum dust" can be used to track wanted criminals or the movement of hostile activity. In urban areas, "quantum dust" can be used as a security device to set off alarms if the infrared radiation is detected. Though QDs are still under research for other possible applications and need more technological

advancement in order to be put into use, the features introduced will grant far better optical communication, significant change in electronic devices, and even detection of antigens in the body tissues.

HAZY SHELLS OF COMPUTRONIUM THAT RING THE SUN

Concentric clouds of nanocomputers the size of rice grains, powered by sunlight, orbiting in shells like the packed layers of a Matrioshka doll – are still immature, holding barely a thousandth of the physical planetary mass of the system, but they already support a classical computational density of 10^{42} MIPS; enough to support a billion civilizations as complex as the one that existed immediately before the great disassembly. The conversion hasn't yet reached the gas giants, and some scant outer-system enclaves remain independent – Amber's Ring Imperium still exists as a separate entity, and will do so for some years to come – but the inner solar system planets, with the exception of Earth, have been colonized more thoroughly than any dusty NASA proposal from the dawn of the space age could have envisaged.

From outside the Accelerated civilization, it isn't really possible to know what's going on inside. The problem is bandwidth: While it's possible to send data in and get data out, the sheer amount of computation going on in the virtual spaces of the Acceleration dwarfs any external observer. Inside that swarm, minds a trillion or more times as complex as humanity think thoughts as far beyond human imagination as a microprocessor is beyond a nematode worm.

A million random human civilizations flourish in worldscapes tucked in the corner of this world-mind. Death is abolished, life is triumphant. A thousand ideologies flower, human nature adapted where necessary to make this possible. Ecologies of thought are forming in a Cambrian explosion of ideas: For the solar system is finally rising to consciousness, and mind is no longer restricted to the mere kilotons of gray fatty meat harbored in fragile human skulls.

PROJECT

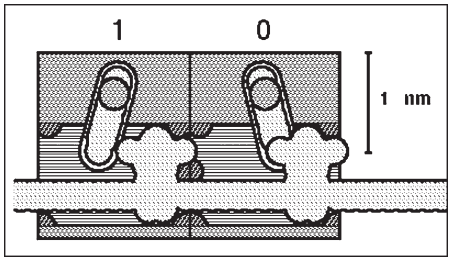
It has to be clearly stated that current operating speeds of nano-electromechanical single electron transistors (NEMSETs) are of the order of 1 GHz, which is not competitive with standard complimentary metal oxide semiconductors (CMOS_ As we have found in recent measurements self-excitation can be exploited to generate mechanical oscillations without any ac excitation. Hence, dc voltages are sufficient to operate the NMC. Basically, a dc voltage creates an electric field to support mechanical oscillations of the nanopillars. A classical example is straightforward to construct. It has to be noted that onset of the mechanical oscillations is induced by a thermal fluctuation, which is found to be enhanced, if the electrical field is inhomogeneous.

The current work that is described as nanomechanical, will still be using DC current. However, a mechanical piece the pillar controls the flow of current. We propose a fully mechanical computer based on nanoelectro-mechanical elements. Our aim is to combine this classical approach with modern nanotechnology to build a nanomechanical computer (NMC) based on nanomechanical transistors.

The main motivation behind constructing such a computer is three fold:

1. Mechanical elements are more robust to electromagnetic shocks than current dynamic random access memory (DRAM) based purely on complimentary metal oxide semiconductor (CMOS) technology,
2. The power dissipated can be orders of magnitude below CMOS and
3. The operating temperature of such an NMC can be an order of magnitude above that of conventional CMOS.

Drexler's work on nanomechanical computer concepts is not mentioned. They do discuss the potential for reversible computing implementation. A summary of Nanosystems is here There was an analysis and simulation of the Drexler Nanocomputer architecture by Bryan Wagner The Drexler idea was based on nanoscale rod logic



Mechanism for two nanocomputer gates, initial position. One control rod with two gate knobs is seen laterally; two more rods with knobs are seen end on. Each rod with associated knobs is a single molecule.

Drexler chose to model this cruder system to show that even simple and easy to define mechanical processes could have interesting performance at the nanoscale Robert Freitas's Nanomedicine book describes nanomechanicaland

nanoelectronic computers, biocomputers and briefly examines the ultimate limits to computation including reversible and quantum computing.

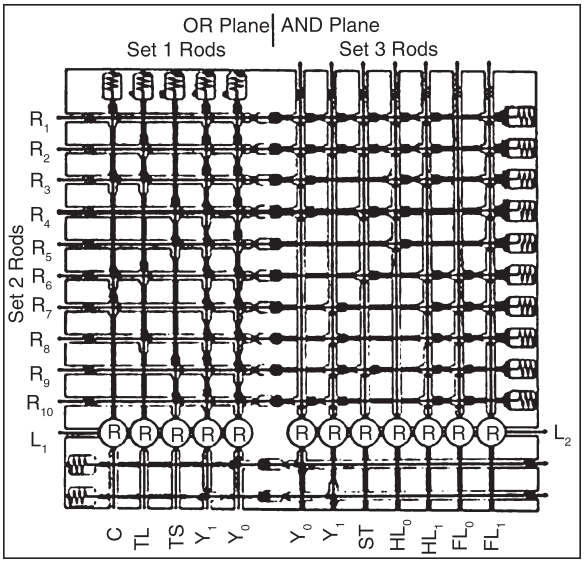


Fig. Schematic of Programmable Logic Array (PLA) Finite State Machine Implementing Rod Logic for a Nanomachanical Central Processing unit.

3

Universal Computing

The concept of a "Universal Constructor" is perhaps more recent. The concept was well understood by von Neumann in the 1940's, who defined a "Universal Constructor" in a two-dimensional cellular automata world. Such a model is a mathematical abstraction something like an infinite checkerboard, with several different types of "checkers" that might be on each square. The different pieces spontanesouly change and move about depending on what pieces occupy neighbouring squares, in accordance with a pre-defined set of rules.

Von Neumann used the concept of a Universal Constructor in conjunction with a Universal Computer as the core components in a self-replicating system. The possibility of fabricating structures by putting "...the atoms down where the chemist says..." was recognized by Feynman in 1959. Drexler recognized the value of the "assembler" in 1977.

The assembler is analogous to von Neumann's Universal Constructor, but operates in the normal three dimensional world and can build a large, atomically precise structures by manipulating atoms and small clusters of atoms. He published the concept in 1981 and in subsequent work has proposed increasingly detailed designs for such devices.

The basic design of Drexler's assembler consists of:

- A molecular computer,
- One or more molecular positioning devices (which might resemble very small robotic arms), and
- A well defined set of chemical reactions (perhaps one or two dozen) that take place at the tip of the arm and are able to fabricate a wide range of structures using site-specific chemical reactions.

It is now common for forecasts of future technical abilities to include the ability to fabricate molecular devices and molecular machines with atomic precision. While there continues to be debate about the exact time frame, it is becoming increasingly accepted that we will, eventually, develop the ability to economically fabricate a truly wide range of structures with atomic precision.

This will be of major economic value. Most obviously a molecular manufacturing capability will be a prerequisite to the construction of molecular logic devices. The continuation of present trends in computer hardware depends on the ability to fabricate ever smaller and ever more precise logic devices at ever decreasing costs.

The limit of this trend is the ability to fabricate molecular logic devices and to connect them in complex patterns at the molecular level. The manufacturing technology needed will,

almost of necessity, be able to economically manufacture large structures (computers) with atomic precision (molecular logic elements). This capability will also permit the economical manufacture of materials with properties that border on the limits imposed by natural law.

The strength of materials, in particular, will approach or even exceed that of diamond. Given the broad range of manufactured products that devote substantial mass to load-bearing members, such a development by itself will have a significant impact. A broad range of other manufactured products will also benefit from a manufacturing process that offers atomic precision at low cost.

Given the promise of such remarkably high payoffs it is natural to ask exactly what such systems will look like, exactly how they will work, and exactly how we will go about building them. One might also enquire as to the reasons for confidence that such an enterprise is feasible, and why one should further expect that our current understanding of chemistry and physics (embodied in a number of computational chemistry packages) should be sufficient to explain the operating principles of such systems. It is here that the value of computational nanotechnology can be most clearly seen. Molecular machine proposals, provided that they are specified in atomic detail (and are of a size that can be dealt with by current software and hardware), can be modeled using the tools of computational chemistry. There are two modeling techniques of particular utility. The first is molecular mechanics, which utilizes empirical force fields to model the forces acting between nuclei. The second is higher order ab initio calculations, which will be discussed in a few paragraphs.

MOLECULAR MECHANICS

More complex analyses, particularly analyses that involve searching through large configuration spaces, can limit the size of system that can be effectively handled. As will be discussed later the need to search through large configuration spaces (as when determining the native folded structure of an arbitrary protein) can be avoided by the use of relatively rigid structures (which differ from relatively floppy proteins and have few possible configurations).

The modeling of machine components in vacuum reduces the need to model solvation effects, which can also involve significant computational effort. In molecular mechanics the individual nuclei are usually treated as point masses. While quantum mechanics dictates that there must be a certain degree of positional uncertainty associated with each nucleus, this positional uncertainty is normally significantly smaller than the typical internuclear distance. Bowen and Allinger provide a good recent overview of the subject.

While the nuclei can reasonably be approximated as point masses, the electron cloud must be dealt with in quantum mechanical terms. However, if we are content to know only the positions of the nuclei and are willing to forego a detailed understanding of the electronic structure, then we can effectively eliminate the quantum mechanics.

For example, the H_2 molecule involves two nuclei. While it would be possible to solve Schrodinger's equation to determine the wave function for the electrons, if we are content simply to know the potential energy contributed by the electrons (and do not enquire about the electron

distribution) then we need only know the electronic energy as a function of the distance between the nuclei.

That is, in many systems the only significant impact that the electrons have on nuclear position is to make a contribution to the potential energy E of the system. In the case of H_2, E is a simple function of the internuclear distance r. The function E(r) summarizes and replaces the more complex and more difficult to determine wave function for the electrons, as well as taking into account the inter-nuclear repulsion and the interactions between the electrons and the nuclei. The two hydrogen nuclei will adopt a position that minimizes E(r). As r becomes larger, the potential energy of the system increases and the nuclei experience a restoring force that returns them to their original distance. Similarly, as r becomes smaller and the two nuclei are pushed closer together, we also find that a restoring force pushes them farther apart, again restoring them to an equilibrium distance.

More generally, if we know the positions r_1, r_2.... r_N of N nuclei, then E(r_1, r_2.... r_N) gives the potential energy of the system. Knowing the potential energy as a function of the nuclear positions, we can readily determine the forces acting on the individual nuclei and therefore can compute the evolution of their position over time. The function E is a newtonian potential energy function (not quantum mechanical), despite the fact that the particular value of E at a particular point could be computed from Schrodinger's equation. That is, the potential energy E is a newtonian concept, but the particular values of E at particular points are determined by Schrodinger's equation.

While it would in principle be possible to determine E by solving Schrodinger's equation, in practice it is usual to use available experimental data and to infer the nature and shape of E by interpolation. This approach, in which empirically derived potential energy functions are created by interpolation from experimental data, has spawned a wide range of potential energy functions, many of which are sold commercially. Because the gradient of the potential energy function E defines a conservative force field F, molecular mechanics methods are also called "force field" methods.

While it is common to refer to "empirical force field" methods, the more recent use of ab initio methods to provide data points to aid in the design of the force fields makes this term somewhat inaccurate, though still widely used. The utility of molecular mechanics depends crucially on the development of accurate force fields. Good quality force fields have been developed for a fairly broad range of compounds including many compounds of interest in biochemistry.

While we will not attempt to survey the wide range of force fields that are available, one particular subset of compounds for which good quality force fields are available involve H, C, N, O, F, Si, P, S, Cl when they are restricted to form chemically uncomplicated structures (*e.g.*, bond strain is not too great, dangling bonds are few or absent, etc). Many atomically precise structures which should be useful in nanotechnology fall in this class and can be modeled with an accuracy adequate to determine the behaviour of molecular machines. The, given essentially arbitrary coordinates of the carbon and hydrogen atoms in a system, Brenner's potential will return the energy of the system. This permits molecular

dynamical modeling of arbitrary hydrocarbon systems, including systems which use synthetic reactions involved in the synthesis of diamond.

A particular reaction of interest is the selective abstraction of a chosen hydrogen atom from a diamond surface. See *Surface patterning by atomically-controlled chemical forces: molecular dynamics simulations* by Sinnott *et al.*, Surface Science 316 (1994), L1055-L1060; see also the work of Robertson *et al.* for an illustration of the use of Brenner's potential in modeling the behaviour of graphitic gears, including failure modes).

Brenner has commented on the utility of this potential for modeling proposed molecular machine systems. Of course, the "accuracy" of the force fields depends on the application. A force field which was accurate to (say) 10 kcal/mole would be unable to correctly predict many properties of interest in biochemistry.

For example, such a force field would lead to serious errors in predicting the correct three-dimensional structure of a protein. Given the linear sequence of amino acids in a protein, the protein folding problem is to determine how it will fold in three dimensions when put in solution. Often, the correct configuration will have an energy that differs from other (incorrect) configurations by a relatively modest amount, and so a force field of high accuracy is required. Unlike the protein folding problem, where an astronomical range of configurations of similar energy are feasible, the bearing basically has only one configuration: A bearing.

While the protein has many unconstrained torsions, there are no unconstrained torsions in the bearing. To significantly

change any torsion angle in the bearing would involve ripping apart bonds. Thus, the same force field which is of marginal utility in dealing with strands of floppy protein is quite adequate for solid blocks of stiff diamondoid material. Not only will small errors in the force field still result in an accurate prediction of the global minima, (which in this case will be a single large basin in the potential energy surface) but also the range of possible structures is so sharply limited that little or no computational effort need be spent comparing the energies of different configurations. Long computational runs to evaluate the statistical properties of ensembles of configurations are thus eliminated.

This style of design has been called, only half in jest, "molecular bridge building" because bridges are also designed with large safety margins. This observation, that the same force field that is inadequate for one class of structures is quite adequate for the design and modeling of a different class of structures, leads to a more general principle.

Computational experiments generally provide an answer with some error distribution. If the errors produced by the model are of a similar size to the errors that would result in incorrect device function, then the model is unreliable. On the other hand, if the errors in the model are small compared with the errors that will produce incorrect device function, then the results of the model are likely to be reliable. If a device design falls in the former category, *i.e.*, small errors in the model will produce conflicting forecasts about device function, then the conservative course of action is to reject the proposed design and keep looking. An even stronger (although somewhat more subtle) statement is possible. The

bearing illustrated in figures is simply a single bearing from a very large class of bearings. The strain in the axle and the sleeve is proportional to the diameter of the bearing. By increasing the diameter, we can reduce the strain. Thus, we can design bearings in this broad class in which the strain can be reduced to whatever level we desire. Because we are dealing with what amounts to a strained block of diamond, the fact that we can reduce strain arbitrarily means that we can design a bearing whose local structure bears as close a resemblance to an unstrained block of diamond as might be wished.

We can therefore be very confident indeed that some member of this class will perform the desired function (that of a bearing) and will work in accordance with our expectations.

Given that we have developed software tools that are capable of creating and modeling most of the members of a broad class of devices, then we can investigate many individual class members.

This investigation then lets us make rather confident statements about the functionality that members of this class can provide, even if there might be residual doubts about individual class members.

A moments reflection will show that the class of objects which are chemically reasonably inert, relatively stiff (no free torsions), and which interact via simple repulsive forces that occur on contact; can describe a truly vast class of machines. Indeed, it is possible to design computers, robotic arms and a wide range of other devices using molecular parts drawn from this class.

FORCE FIELD

While empirical force fields are sufficiently accurate to model the behaviour of chemically stable stiff structures interacting with other chemically stable stiff structures, they do not (at present) provide sufficient accuracy to deal with chemical transitions. Thus, if we wish to model the manufacture of a molecular part then we must use higher order ab initio techniques.

These techniques impose severe constraints on the number of atoms that can be modeled (perhaps one or two dozen heavy atoms, depending on the hardware, software, and specific type of modeling being attempted), but can provide an accuracy sufficient to analyse the chemical reactions that must necessarily take place during the synthesis of large, atomically precise structures.

An analysis of the abstraction of a hydrogen atom from various structures, including isobutane (which serves as a model of the diamond (111) surface), has been carried out to illustrate the kind of reactions that are of interest. More generally, higher order ab initio techniques are sufficient to analyse the addition or removal of a small number of atoms from a specific site on a work piece.

Synthesis of a large object would then consist of repeated site specific applications of a small number of basic operations, where each basic operation changed the chemical structure of only a small number of atoms at a time. Provided that we reject reaction mechanisms where the result predicted depends on errors that are smaller than can reasonably be modeled, the analysis of these basic operations can be satisfactorily carried out with current methods and hardware.

CURRENT MODELING METHODS

Clearly, not all molecular machines satisfy these constraints. Is the range of molecular machines which do satisfy these constraints sufficiently large to justify the effort of designing and modeling them? And in particular, can we satisfactorily model The fundamental purpose of an assembler is to position atoms. To this end, it is imperative that we have models that let us determine atomic positions, and this is precisely what molecular mechanics provides. Robotic arms or other positioning devices are basically mechanical in nature, and will allow us to position molecular parts during the assembly process.

Molecular mechanics provides us with an excellent tool for modeling the behaviour of such devices. The second fundamental requirement is the ability to make and break bonds at specific sites. While molecular mechanics provides an excellent tool for telling us where the tip of the assembler arm is located, current force fields are not adequate to model the specific chemical reactions that must then take place at the tip/work-piece interface involved in building an atomically precise part (though this statement must be modified in light of the work done with Brenner's potential, discussed above). For this, higher order ab initio calculations are sufficient.

The glaring omission in this discussion is the modeling of the kind of electronic behaviour that occurs in switching devices. Clearly, it is possible to model electronic behaviour with some degree of accuracy, and equally clearly molecular machines that are basically electronic in nature will be extremely useful. It will therefore be desirable to extend the range of computational models discussed here to include

such devices. For the moment, however, the relatively modest inclusion of electrostatic motors as a power source is probably sufficient to provide us with adequate "design space" to design and model an assembler.

While it might at first glance appear that electronics will be required for the computational element in the assembler, in fact molecular mechanical logic elements will be sufficient. Babbage's proposal from the 1800's clearly demonstrates the feasibility of mechanical computation (it is interesting to note that a working model of Babbage's difference engine has been built and is on display at the British Museum of Science.

They were careful to use parts machined no more accurately than the parts available to Babbage, in order to demonstrate that his ideas could have been implemented in the 1800's). Drexler's analysis of a specific molecular mechanical logic element and further analysis of system design issues also make it clear that molecular mechanical computation is sufficient for the molecular computer required in an assembler. Molecular electronic proposals for computation have not been worked out as clearly at the system level as the molecular mechanical concepts. Thus, at the moment, molecular mechanical proposals are better understood, at least in this particular context. This situation is likely to change, and when it does the electronic designs could be incorporated (where appropriate) into the design and modeling of an assembler. However, it is not entirely obvious that electronic designs will prove superior to molecular mechanical designs, particularly when device parameters such as size and energy dissipation are considered.

While it seems virtually certain that electronic devices will prove faster than molecular mechanical devices, it is less obvious that electronic devices must necessarily be either smaller or dissipate less energy (though confer more recent work on reversible logic). Indeed, given the difficulty of localizing individual electrons, it seems possible that electronic devices will prove to be inherently larger than molecular mechanical devices. This might provide a long term role for molecular mechanical devices as high density memory elements, or as logic elements when space (or atom count) is particularly constrained.

Whether the assembler is designed and built with an electronic or mechanical computer is less significant than designing and building it. By way of example, a stored programme computer can be built in many different ways. Vacuum tubes, transistors, moving mechanical parts, fluidics, and other methods are all entirely feasible.

Delaying the development of the Eniac because transistors are better than vacuum tubes would have been a most unwise course of action. Similarly, as we consider the design, modeling, and construction of an assembler, we should not hesitate to use simpler methods that we can understand in favour of better methods that are not yet fully in hand.

The methods of computational chemistry available today allow us to model a wide range of molecular machines with an accuracy sufficient in many cases to determine how well they will work. We can deliberately confine ourselves to the subset of devices where our modeling tools are good enough to give us confidence that the proposals should work. This range includes bearings, computers, robotic arms, site-

specific chemical reactions utilized in the synthesis of complex structures, and more complex systems built up from these and related components. Drexler's assembler and related devices can be modeled using our current approaches and methods.

MOLECULAR COMPILERS

Computational nanotechnology includes not only the tools and techniques required to model proposed molecular machines, it must also include the tools required to specify such machines. Molecular machine proposals that would require millions or even billions of atoms have been made.

The total atom count of an assember might be roughly a billion atoms. While commerically available molecular modeling packages provide facilities to specify arbitrary structures, it is usually necessary to "point and click" for each atom involved. This is obviously unattractive for a device as complex as an assembler with its roughly one billion atoms. It should be clear that molecular CAD tools will be required if we are to specify such complex structures in atomic detail with a reasonable amount of effort.

An essential development is that of "molecular compilers" which accept, as input, a high level description of an object and produce, as output, the atomic coordinates, atom types, and bonding structure of the object.

Using this approach, it will be possible to substantially reduce the development time for complex molecular machines, including Drexler's assembler. This approach is similar in spirit to the computer aided design and modeling used to speed the development of many products today. The

author was part of a general purpose computer start up which successfully designed and built a new computer from scratch. This included the hardware, software, compilers, operating system, etc. During this process, extensive use was made of computational models to verify each level of the design. The operating system was written in Pascal, and checked out on another computer. The compilers were written in Pascal, and also checked out on another computer. The code produced by the compilers was checked out on an instruction set simulator. The microcode was checked out on a micro-instruction simulator. The logic design was checked out with logic level simulation tools, and circuit simulation packages were used to verify the detailed electronic design. All this work was done at the same time.

The software was written and debugged even though the machine on which it would eventually be executed didn't exist. The microcode was written and debugged before the hardware was available. When the hardware was finally made available, system integration went relatively rapidly. Imagine, for a moment, how long it would have taken to develop this system had we carried out the development in the obvious sequential fashion. First, we would have implemented the hardware and only then begun work on the microengine.

Later, with the hardware and microengine fully checked out and working, we could have developed and debugged the microcode. With this firmly in hand, we could then have written the compilers and verified the code they produced. Finally, we could have implemented and checked out the operating system. Needless to say, such a strategy would have been very slow and tedious.

Doing things in the simple and most obvious way often takes a lot longer than is needed. If we were to approach the design and construction of an assembler using the simple serial method, it would take a great deal longer than if we systematically attacked and simultanesouly solved the problems that arise at all levels of the design at one and the same time. That is, by using methods similar to those used to design a modern computer, including intensive computational modeling of individual components and sub-systems, we can greatly shorten the time required to design and build complex molecular machines.

This can be further illustrated by considering the traditional manner of growth of our synthetic capability over time. Today, we find we are able to synthesize a certain range of compounds and structures. As time goes by, we will be able to synthesize an ever larger range of structures. This growth in our ability will proceed on a broad front, and reflects the efforts of a broad range of researchers who are each pursuing individual goals without concern about the larger picture into which they might fit. As illustrated in figure, given sufficient time we will eventually be able to synthesize complex molecular machines simply because we will eventually develop the ability to synthesize just about anything.

However, if we wish to develop a particular kind of device, *e.g.*, an assembler, then we can speed the process up by conducting computational experiments designed to clarify the objective and to specify more precisely the path from our current range of synthetic capabilities to the objective. Such computational experiments are inexpensive, can provide very

detailed information, are possible for any structure (whether we can or cannot synthesize it) and will, in general, reduce the "time to market" for the selected product.

Computational experiments let us examine structures quickly and easily, rejecting those which have obvious defects (a precursor to the bearing shown in figure, for example, was too strained. By modifying the design and again minimizing the structure, we found a design with an acceptable strain).

This kind of examination of the "design space" is impossible with physical experiments today, but is easily done with computational experiments.

Computational experiments also provide more information. For example, molecular dynamics can literally provide information about the position of each individual atom over time, information which would usually be inaccessible in a physical experiment. Of course, the major advantage of computational experiments over physical experiments in the current context is the simple fact that physical experiments aren't possible for molecular machines that we can't make with today's technology.

By using computational models derived from the wealth of experimental data that is available today, we can (within certain accuracy bounds) describe the behaviour of proposed systems that we plan to build in the future. If we deliberately design systems that are sufficiently robust that we are confident they will work regardless of the small errors that must be incurred in the modeling process, we can design systems today that we will not be able to build for some years, and yet still have reasonable confidence that they will work.

By fully utilizing the experience that has been developed in the rapid design and development of complex systems we can dramatically reduce the development time for molecular manufacturing systems. It is possible to debate how long it will be before we achieve a robust molecular manufacturing capability. However, it is very clear that we'll get there sooner if we develop and make intelligent use of molecular design tools and computational models. These will let us design and check the blueprints for the new molecular manufacturing technologies that we now see on the horizon, and will let us chart a more rapid and more certain path to their development.

NANOSENSOR

Nanosensors are any biological, chemical, or physical sensory points used to convey information about nanoparticles to the macroscopic world. Though humans have not yet been able to synthesize nanosensors, predictions for their use mainly include various medicinal purposes and as gateways to building other nanoproducts, such as computer chips that work at the nanoscale and nanorobots.

Presently, there are several ways proposed to make nanosensors, including top-down lithography, bottom-up assembly, and molecular self-assembly.

Medicinal uses of nanosensors mainly revolve around the potential of nanosensors to accurately identify particular cells or places in the body in need. By measuring changes in volume, concentration, displacement and velocity, gravitational, electrical, and magnetic forces, pressure, or temperature of cells in a body, nanosensors may be able to distinguish between and recognize certain cells, most notably

those of cancer, at the molecular level in order to deliver medicine or monitor development to specific places in the body. In addition, they may be able to detect macroscopic variations from outside the body and communicate these changes to other nanoproducts working within the body. One example of nanosensors involves using the fluorescence properties of cadmium selenide quantum dots as sensors to uncover tumors within the body. By injecting a body with these quantum dots, a doctor could see where a tumor or cancer cell was by finding the injected quantum dots, an easy process because of their fluorescence.

Developed nanosensor quantum dots would be specifically constructed to find only the particular cell for which the body was at risk. A downside to the cadmium selenide dots, however, is that they are highly toxic to the body. As a result, researchers are working on developing alternate dots made out of a different, less toxic material while still retaining some of the fluorescence properties.

In particular, they have been investigating the particular benefits of zinc sulfide quantum dots which, though they are not quite as fluorescent as cadmium selenide, can be augmented with other metals including manganese and various lanthanide elements. In addition, these newer quantum dots become more fluorescent when they bond to their target cells. Potential predicted functions may also include sensors used to detect specific DNA in order to recognize explicit genetic defects, especially for individuals at high-risk and implanted sensors that can automatically detect glucose levels for diabetic subjects more simply than current detectors.

DNA can also serve as sacrificial layer for manufacturing CMOS IC, integrating a nanodevice with sensing capabilities. Therefore, using proteomic patterns and new hybrid materials, nanobiosensors can also be used to enable components configured into a hybrid semiconductor substrate as part of the circuit assembly. The development and miniaturization of nanobiosensors should provide interesting new opportunities. Other projected products most commonly involve using nanosensors to build smaller integrated circuits, as well as incorporating them into various other commodities made using other forms of nanotechnology for use in a variety of situations including transportation, communication, improvements in structural integrity, and robotics. Nanosensors may also eventually be valuable as more accurate monitors of material states for use in systems where size and weight are constrained, such as in satellites and other aeronautic machines.

Existing Nanosensors

Currently, the most common mass-produced functioning nanosensors exist in the biological world as natural receptors of outside stimulation. For instance, sense of smell, especially in animals in which it is particularly strong, such as dogs, functions using receptors that sense nanosized molecules.

Certain plants, too, use nanosensors to detect sunlight; various fish use nanosensors to detect minuscule vibrations in the surrounding water; and many insects detect sex pheromones using nanosensors. Certain electromagnetic sensors have also been in use in photoelectric systems. These work because the specific sensors called, aptly, photosensors are easily influenced by light of various wavelengths.

The electromagnetic source transfers energy to the photosensors and energizes them into an excited state which causes them to release an electron into a semiconductor. At that point, it is relatively easy to detect the electricity coming from the sensors, and thus easy to know if the sensors are receiving light. Though more advanced uses of photosensors incorporating other forms of nanotechnology have yet to be implemented into consumer society, most film cameras have used photosensors at the nano size for years. Traditional film uses a layer of silver ions that become excited by solar energy and clump into groups, as small as four atoms apiece in some cases, that scatter light and appear dark on the frame.

Various other types of film can be made using a similar process to detect other specific wavelengths of light, including x-rays, infrared, and ultraviolet. One of the first working examples of a synthetic nanosensor was built by researchers at the Georgia Institute of Technology in 1999. It involved attaching a single particle onto the end of a carbon nanotube and measuring the vibrational frequency of the nanotube both with and without the particle. The discrepancy between the two frequencies allowed the researchers to measure the mass of the attached particle.

Chemical sensors, too, have been built using nanotubes to detect various properties of gaseous molecules. Carbon nanotubes have been used to sense ionization of gaseous molecules while nanotubes made out of titanium have been employed to detect atmospheric concentrations of hydrogen at the molecular level. Many of these involve a system by which nanosensors are built to have a specific pocket for

another molecule. When that particular molecule, and only that specific molecule, fits into the nanosensor, and light is shone upon the nanosensor, it will reflect different wavelengths of light and, thus, be a different colour.

Production Methods

There are currently several hypothesized ways to produce nanosensors. Top-down lithography is the manner in which most integrated circuits are now made. It involves starting out with a larger block of some material and carving out the desired form. These carved out devices, notably put to use in specific microelectromechanical systems used as microsensors, generally only reach the micro size, but the most recent of these have begun to incorporate nanosized components.

Another way to produce nanosensors is through the bottom-up method, which involves assembling the sensors out of even more minuscule components, most likely individual atoms or molecules. This would involve moving atoms of a particular substance one by one into particular positions which, though it has been achieved in laboratory tests using tools such as atomic force microscopes, is still a significant difficulty, especially to do en masse, both for logistic reasons as well as economic ones.

Most likely, this process would be used mainly for building starter molecules for self-assembling sensors. he third way, which promises far faster results, involves self-assembly, or "growing" particular nanostructures to be used as sensors. This most often entails one of two types of assembly. The first involves using a piece of some previously created or

naturally formed nanostructure and immersing it in free atoms of its own kind. After a given period, the structure, having an irregular surface that would make it prone to attracting more molecules as a continuation of its current pattern, would capture some of the free atoms and continue to form more of itself to make larger components of nanosensors.

The second type of self-assembly starts with an already complete set of components that would automatically assemble themselves into a finished product. Though this has been so far successful only in assembling computer chips at the micro size, researchers hope to eventually be able to do it at the nanometer size for multiple products, including nanosensors. Accurately being able to reproduce this effect for a desired sensor in a laboratory would imply that scientists could manufacture nanosensors much more quickly and potentially far more cheaply by letting numerous molecules assemble themselves with little or no outside influence, rather than having to manually assemble each sensor.

NANOROBOTS

Nanorobotics is emerging as a demanding field dealing with miniscule things at molecular level. Nanorobots are quintessential nanoelectromechanical systems designed to perform a specific task with precision at nanoscale dimensions. Its advantage over conventional medicine lies on its size. Particle size has effect on serum lifetime and pattern of deposition. This allows drugs of nanosize to be used in lower concentration and has an earlier onset of therapeutic action.

It also provides materials for controlled drug delivery by directing carriers to a specific location. The typical medical nanodevice will probably be a micron-scale robot assembled from nanoscale parts. These nanorobots can work together in response to environment stimuli and programmed principles to produce macro scale results.

ELEMENTS OF NANOROBOTS

Carbon will likely be the principal element comprising the bulk of a medical nanorobot, probably in the form of diamond or diamondoid/fullerene nanocomposites. Many other light elements such as hydrogen, sulfur, oxygen, nitrogen, fluorine, silicon, etc. will be used for special purposes in nanoscale gears and other components. The chemical inertness of diamond is proved by several experimental studies. One such experiment conducted on mouse peritoneal macrophages cultured on DLC showed no significant excess release of lactate dehydrogenase or of the lysosomal enzyme beta N-acetyl-D-glucosaminidase an enzyme known to be released from macrophages during inflammation.

Morphological examination revealed no physical damage to either fibroblasts or macrophages, and human osteoblast like cells confirming the biochemical indication that there was no toxicity and that no inflammatory reaction was elicited in vitro. The smoother and more flawless the diamond surface, the lesser is the leukocyte activity and fibrinogen adsorption. Interestingly, on the rougher "polished" surface, a small number of spread and fused macrophages were present, indicating that some activation had occurred.

The exterior surface with near-nanometer smoothness results in very low bioactivity. Due to the extremely high

surface energy of the passivated diamond surface and the strong hydrophobicity of the diamond surface, the diamond exterior is almost completely chemically inert.

DESIGN OF NANOROBOTS

Nanorobots will possess full panoply of autonomous subsystems whose design is derived from biological models. Drexler evidently was the first to point out, in 1981, that complex devices resemble biological models in their structural components. The various components in the nanorobot design may include onboard sensors, motors, manipulators, power supplies, and molecular computers.

Perhaps the best-known biological example of such molecular machinery is the ribosome the only freely programmable nanoscale assembler already in existence. The mechanism by which protein binds to the specific receptor site might be copied to construct the molecular robotic arm.

The manipulator arm can also be driven by a detailed sequence of control signals, just as the ribosome needs mRNA to guide its actions. These control signals are provided by external acoustic, electrical, or chemical signals that are received by the robot arm via an onboard sensor using a simple "broadcast architecture a technique which can also be used to import power. the biological cell may be regarded as an example of a broadcast architecture in which the nucleus of the cell send signals in the form of mRNA to ribosomes in order to manufacture cellular proteins.

Assemblers are molecular machine systems that could be described as systems capable of performing molecular manufacturing at the atomic scale which require control

signals provided by an onboard nanocomputer This programmable nanocomputer must be able to accept stored instructions which are sequentially executed to direct the manipulator arm to place the correct moiety or nanopart in the desired position and orientation, thus giving precise control over the timing and locations of chemical reactions or assembly operations.

CONSTRUCTION

There are two main approaches to building at the nanometer scale: positional assembly and self-assembly. In positional assembly, investigators employ some devices such as the arm of a miniature robot or a microscopic set to pick up molecules one by one and assemble them manually. In contrast, self-assembly is much less painstaking, because it takes advantage of the natural tendency of certain molecules to seek one another out. With self-assembling components, all that investigators have to do is put billions of them into a beaker and let their natural affinities join them automatically into the desired configurations. Making complex nanorobotic systems requires manufacturing techniques that can build a molecular structure via computational models of diamond mechanosynthesis.

DMS is the controlled addition of carbon atoms to the growth surface of a diamond crystal lattice in a vacuum-manufacturing environment.

Covalent chemical bonds are formed one by one as the result of positionally constrained mechanical forces applied at the tip of a scanning probe microscope apparatus, following a programmed sequence.

Recognition of Target Site by Nanorobots

Different molecule types are distinguished by a series of chemotactic sensors whose binding sites have a different affinity for each kind of molecule. The control system must ensure a suitable performance. It can be demonstrated with a determined number of nanorobots responding as fast as possible for a specific task based scenario. In the 3D workspace the target has surface chemicals allowing the nanorobots to detect and recognize it. Manufacturing better sensors and actuators with nanoscale sizes makes them find the source of release of the chemical. Nanorobot Control Design (NCD) simulator was developed, which is software for nanorobots in environments with fluids dominated by Brownian motion and viscous rather than inertial forces.

First, as a point of comparison, the scientists used the nanorobots' small Brownian motions to find the target by random search. In a second method, the nanorobots monitor for chemical concentration significantly above the background level. After detecting the signal, a nanorobot estimates the concentration gradient and moves towards higher concentrations until it reaches the target.

In the third approach, nanorobots at the target release another chemical, which others use as an additional guiding signal to the target. With these signal concentrations, only nanorobots passing within a few microns of the target are likely to detect the signal. Thus, we can improve the response by having the nanorobots maintain positions near the vessel wall instead of floating throughout the volume flow in the vessel from monitoring the concentration of a signal from others; a nanorobot can estimate the number of nanorobots

at the target. So, the nanorobot uses this information to determine when enough nanorobots are at the target, thereby terminating any additional "attractant" signal a nanorobot may be releasing. It is found that the nanorobots stop attracting others once enough nanorobots have responded. The amount is considered enough when the target region is densely covered by nanorobots. Thus these tiny machines work at the target site accurately and precisely to that extent only to which it is designed to do.

Nanorobots Evading the Immune System

Every medical nanorobot placed inside the human body will encounter phagocytic cells many times during its mission. Thus all Nanorobots, which are of a size capable of ingestion by phagocytic cells, must incorporate physical mechanisms and operational protocols for avoiding and escaping from phagocytes. The initial strategy for medical nanorobots is first to avoid phagocytic contact or recognition.

To avoid being attacked by the host's immune system, the best choice is to have an exterior coating of passive diamond. The smoother and flawless the coating, the lesser is the reaction from the body's immune system. And if this fails then to avoid it's binding to the phagocyte surface that leads to phagocytic activation. If trapped, the medical nanorobot can induce exocytosis of the phagosomal vacuole in which it is lodged or inhibit both phagolysosomal fusion and phagosome metabolism. In rare circumstances, it may be necessary to kill the phagocyte or to blockade the entire phagocytic system.

The most direct approach for a fully functional medical nanorobot is to employ its motility mechanisms to locomote

out of, or away from, the phagocytic cell that is attempting to engulf it. This may involve reverse cytopenetration, which must be done cautiously (*e.g.*, the rapid exit of nonenveloped viruses from cells can be cytotoxic). It is possible that frustrated phagocytosis may induce a localized compensatory granulomatous reaction.

Medical nanorobots therefore may also need to employ simple but active defensive strategies to forestall granuloma formation. Metabolizing local glucose and oxygen for energy can do the powering of the nanorobots. In a clinical environment, another option would be externally supplied acoustic energy. When the task of the nanorobots is completed, they can be retrieved by allowing them to exfuse themselves via the usual human excretory channels or can also be removed by active scavenger systems.

Nanorobots in Cancer Detection and Treatment

The development of nanorobots may provide remarkable advances for diagnosis and treatment of cancer. Nanorobots could be a very helpful and hopeful for the therapy of patients, since current treatments like radiation therapy and chemotherapy often end up destroying more healthy cells than cancerous ones. From this point of view, it provides a non-depressed therapy for cancer patients. The Nanorobots will be able to distinguish between different cell types that is the malignant and the normal cells by checking their surface antigens (they are different for each type of cell).

This is accomplished by the use of chemotactic sensors keyed to the specific antigens on the target cells. Another approach uses the innovative methodology to achieve decentralized control for a distributed collective action in the

combat of cancer. Using chemical sensors they can be programmed to detect different levels of E-cadherin and beta-catenin in primary and metastatic phases. Medical nanorobots will then destroy these cells, and only these cells.

The following control methods were considered:

- *Random*: Nanorobots moving passively with the fluid reaching the target only if they bump into it due to Brownian motion.
- *Follow gradient*: Nanorobots monitor concentration intensity for E-cadherin signals, when detected, measure and follow the gradient until reaching the target. If the gradient estimate subsequent to signal detection finds no additional signal in50ms, the nanorobot considers the signal to be a false positive and continues flowing with the fluid.
- *Follow gradient with attractant*: As above, but nanorobots arriving at the target, they release in addition a different chemical signal used by others to improve their ability to find the target. Thus, a higher gradient of signal intensity of E-cadherin is used as chemical parameter identification in guiding nanorobots to identify malignant tissues. Integrated nanosensors can be utilized for such a task in order to find intensity of E-cadherin signals. Thus they can be employed effectively for treating cancer.

Nanorobots in Cancer Detection and Treatment

Pharmacyte is a self-powered, computer controlled medical nanorobot system capable of digitally precise transport, timing, and targeted-delivery of pharmaceutical agents to specific cellular and intracellular destinations within the

human body. Pharmacytes escape the phagocytic process as they will not embolize small blood vessels because the minimum viable human capillary that allows passage of intact erythrocytes and white cells is 3–4 micronmeter in diameter, which is larger than the largest proposed Pharmacyte. Pharmacytes will have many applications in nanomedicine such as initiation of apoptosis in cancer cells and direct control of cell signaling processes.

Pharmacytes could also tag target cells with biochemical natural defensive or scavenging systems, a strategy called "phagocytic flagging". For example, novel recognition molecules are expressed on the surface of apoptotic cells. In the case of T lymphocytes, one such molecule is phosphatidylserine, a lipid that is normally restricted to the inner side of the plasma membrane [1m] but, after the induction of apoptosis, appears on the outside.

Cells bearing this molecule on their surface can then be recognized and removed by phagocytic cells. Seeding the outer wall of a target cell with phosphatidylserine or other molecules with similar action could activate phagocytic behaviour by macrophages, which had mistakenly identified the target cell as apoptotic substances capable of triggering a reaction by the body Pharmacytes would be capable of carrying up to approximately 1cubicmeter of pharmaceutical payload stored in onboard tanks that are mechanically offloaded using molecular sorting pumps operated under the control of an onboard computer.

Depending on mission requirements, the payload can be discharged into the proximate extracellular fluid or delivered directly into the cytosol using a transmembrane injector

mechanism. If needed for a particular application, deployable mechanical cilia and other locomotive systems can be added to the Pharmacyte to permit transvascular and transcellular mobility, thus allowing delivery of pharmaceutical molecules to specific cellular and even intracellular addresses with negligible error.

Pharmacytes, once depleted of their payloads or having completed their mission, would be recovered from the patient by conventional excretory pathways. The nanorobots might then be recharged, reprogrammed and recycled for use in a second patient who may need a different pharmaceutical agent targeted to different tissues or cells than in the first patient.

Nanorobots in the Diagnosis and Treatment of Diabetes

Glucose carried through the blood stream is important to maintain the human metabolism working healthfully, and its correct level is a key issue in the diagnosis and treatment of diabetes. Intrinsically related to the glucose molecules, the protein hSGLT3 has an important influence in maintaining proper gastrointestinal cholinergic nerve and skeletal muscle function activities, regulating extracellular glucose concentration. The hSGLT3 molecule can serve to define the glucose levels for diabetes patients.

The most interesting aspect of this protein is the fact that it serves as a sensor to identify glucose. The simulated nanorobot prototype model has embedded Complementary Metal Oxide semi-conductor (CMOS) nanobioelectronics.

It features a size of ~2 micronmeter, which permits it to operate freely inside the body. Whether the nanorobot is

invisible or visible for the immune reactions, it has no interference for detecting glucose levels in blood stream. Even with the immune system reaction inside the body, the nanorobot is not attacked by the white blood cells due biocompatibility.

For the glucose monitoring the nanorobot uses embedded chemosensor that involves the modulation of hSGLT3 protein glucosensor activity. Through its onboard chemical sensor, the nanorobot can thus effectively determine if the patient needs to inject insulin or take any further action, such as any medication clinically prescribed. The image of the NCD simulator workspace shows the inside view of a venule blood vessel with grid texture, red blood cells (RBCs) and nanorobots. They flow with the RBCs through the bloodstream detecting the glucose levels.

At a typical glucose concentration, the nanorobots try to keep the glucose levels ranging around 130 mg/dl as a target for the Blood Glucose Levels (BGLs). A variation of 30mg/dl was adopted as a displacement range, though this can be changed based on medical prescriptions. In the medical nanorobot architecture, the significant measured data can be then transferred automatically through the RF signals to the mobile phone carried by the patient. At any time, if the glucose achieves critical levels, the nanorobot emits an alarm through the mobile phone.

Controlling Glucose Level using Nanorobots

In the simulation, the nanorobot is programmed also to emit a signal based on specified lunch times, and to measure the glucose levels in desired intervals of time. The nanorobot can be programmed to activate sensors and measure

regularly the BGLs early in the morning, before the expected breakfast time. Levels are measured again each 2 hours after the planned lunchtime. The same procedures can be programmed for other meals through the day times.

A multiplicity of blood borne nanorobots will allow glucose monitoring not just at a single site but also in many different locations simultaneously throughout the body, thus permitting the physician to assemble a whole-body map of serum glucose concentrations.

Examination of time series data from many locations allows precise measurement of the rate of change of glucose concentration in the blood that is passing through specific organs, tissues, capillary beds, and specific vessels. This will have diagnostic utility in detecting anomalous glucose uptake rates which may assist in determining which tissues may have suffered diabetes-related damage, and to what extent.

Other onboard sensors can measure and report diagnostically relevant observations such as patient blood pressure, early signs of tissue gangrene, or changes in local metabolism that might be associated with early-stage cancer. Whole-body time series data collected during various patient activities levels could have additional diagnostic value in assessing the course and extent of disease.

This important data may help doctors and specialists to supervise and improve the patient medication and daily diet. This process using nanorobots may be more convenient and safe for making feasible an automatic system for data collection and patient monitoring. It may also avoid eventually infections due the daily small cuts to collect blood samples, possibly loss of data, and even avoid patients in a busy week

to forget doing some of their glucose sampling. These Recent developments on nanobioelectronics show how to integrate system devices and cellular phones to achieve a better control of glucose levels for patients with diabetes.

Respirocyte

The artificial mechanical red cell, "Respirocyte" is an imaginary nanorobot, floats along in the blood stream. These atoms are mostly carbon atoms arranged as diamond in a porous lattice structure inside the spherical shell. The Respirocyte is essentially a tiny pressure tank that can be pumped full of oxygen (O_2) and carbon dioxide (CO_2) molecules.

Later on, these gases can be released from the tiny tank in a controlled manner. The gases are stored onboard at pressures up to about 1000 atmospheres. Respirocyte can be rendered completely nonflammable by constructing the device internally of sapphire, a flameproof material with chemical and mechanical properties otherwise similar to diamond.

There are also gas concentration sensors on the outside of each device. When the nanorobot passes through the lung capillaries, O_2 partial pressure is high and CO_2 partial pressure is low, so the onboard computer tells the sorting rotors to load the tanks with oxygen and to dump the CO_2. When the device later finds itself in the oxygen-starved peripheral tissues, the sensor readings are reversed.

That is, CO_2 partial pressure is relatively high and O_2 partial pressure relatively low, so the onboard computer commands the sorting rotors to release O_2 and to absorb CO_2.Respirocytes mimic the action of the natural hemoglobin-

filled red blood cells. But a Respirocyte can deliver 236 times more oxygen per unit volume than a natural red cell.

This nanorobot is far more efficient than biology, mainly because its diamondoid construction permits a much higher operating pressure. So the injection of a 5 cm^3 dose of 50% Respirocyte aqueous suspension into the bloodstream can exactly replace the entire O_2 and CO_2 carrying capacity of the patient's entire 5,400 cm^3 of blood. Respirocyte will have pressure sensors to receive acoustic signals from the doctor, who will use an ultrasound-like transmitter device to give the Respirocyte commands to modify their behaviour while they are still inside the patient's body.

Artificial Phagocytes

A microbivore has been described, whose primary function is to destroy microbiological pathogens found in the human bloodstream, using the "digest and discharge" protocol. Nanorobotic artificial hypothetical phagocytes called "microbivores" could patrol the bloodstream, seeking out and digesting unwanted pathogens including bacteria, viruses, or fungi. Microbivores when given intravenously (I.V) would achieve complete clearance of even the most severe septicemic infections in hours or less.

This is far better than the weeks or months needed for antibiotic-assisted natural phagocytic defences. The nanorobots do not increase the risk of sepsis or septic shock because the pathogens are completely digested into harmless simple sugars, monoresidue amino acids, mononucleotides, free fatty acids and glycerol, which are the biologically inactive effluents from the nanorobot.

A Hypothetical Mobile Cell-Repair Nanorobot

Another nanorobot, the Chromallocyte would replace entire chromosomes in individual cells thus reversing the effects of genetic disease and other accumulated damage to our genes, preventing aging.

Chromallocyte is a hypothetical mobile cell-repair nanorobot capable of limited vascular surface travel into the capillary bed of the targeted tissue or organ, followed by extravasation, histonatation, cytopenetration, and complete chromatin replacement in the nucleus of one target cell, and ending with a return to the bloodstream and subsequent extraction of the device from the body, completing the cell repair mission."

Inside a cell, a repair machine will first size up the situation by examining the cell's contents and activity, and then take action. By working along molecule-by-molecule and structure-by-structure, repair machines will be able to repair whole cells. By working along cell-by-cell and tissue-by-tissue, they (aided by larger devices, where need be) will be able to repair whole organs. By working through a person organ by organ, they will restore health. Because molecular machines will be able to build molecules and cells from scratch, they will be able to repair even cells damaged to the point of complete inactivity.

Further Applications of Nanorobots

Nanorobots could be used to maintain tissue oxygenation in the absence of respiration, repair and recondition the human vascular tree eliminating heart disease and stroke damage, perform complex nanosurgery on individual cells, and instantly staunch bleeding after traumatic injury.

Monitoring nutrient concentrations in the human body is a possible application of nanorobots in medicine. One of interesting nanorobot utilization is also to assist inflammatory cells (or white cells) in leaving blood vessels to repair injured tissues. Nanorobots might be used as well to seek and break kidney stones. Nanorobots could also be used to process specific chemical reactions in the human body as ancillary devices for injured organs. Nanorobots equipped with nanosensors could be developed to deliver anti-HIV drugs.

Another important capability of medical nanorobots will be the ability to locate stenosed blood vessels, particularly in the coronary circulation, and treat them mechanically, chemically, or pharmacologically.

To cure skin diseases, a cream containing nanorobots may be used. It could remove the right amount of dead skin, remove excess oils, add missing oils, apply the right amounts of natural moisturizing compounds, and even achieve the elusive goal of 'deep pore cleaning' by actually reaching down into pores and cleaning them out. The cream could be a smart material with smooth-on, peel-off convenience. A mouthwash full of smart nanomachines could identify and destroy pathogenic bacteria while allowing the harmless flora of the mouth to flourish in a healthy ecosystem. Further, the devices would identify particles of food, plaque, or tartar, and lift them from teeth to be rinsed away. Being suspended in liquid and able to swim about, devices would be able to reach surfaces beyond reach of toothbrush bristles or the fibres of floss.

As short-lifetime medical nanodevices, they could be built to last only a few minutes in the body before falling apart

into materials of the sort found in foods. Medical nanodevices could augment the immune system by finding and disabling unwanted bacteria and viruses. When an invader is identified, it can be punctured, letting its contents spill out and ending its effectiveness. If the contents were known to be hazardous by themselves, then the immune machine could hold on to it long enough to dismantle it more completely.

Devices working in the bloodstream could nibble away at arteriosclerotic deposits, widening the affected blood vessels. Cell herding devices could restore artery walls and artery linings to health, by ensuring that the right cells and supporting structures are in the right places.

This would prevent most heart attacks. Nanorobots could be used in precision treatment and cell targeted delivery, in performing nanosurgery, and in treatments for hypoxemia and respiratory illness, dentistry, bacteremic infections, physical trauma, gene therapy via chromosome replacement therapy and even biological aging. It has been suggested that a fleet of nanorobots might serve as antibodies or antiviral agents in patients with compromised immune systems, or in diseases that do not respond to more conventional measures. There are numerous other potential medical applications, including repair of damaged tissue, unblocking of arteries affected by plaques, and perhaps the construction of complete replacement body organs.

Nanoscale systems can also operate much faster than their larger counterparts because displacements are smaller; this allows mechanical and electrical events to occur in less time at a given speed. Nanotechnology as a diagnostic and treatment tool for patients with cancer and diabetes showed

how actual developments in new manufacturing technologies are enabling innovative works which may help in constructing and employing nanorobots most effectively for biomedical problems.

Nanorobots applied to medicine hold a wealth of promise from eradicating disease to reversing the aging process (wrinkles, loss of bone mass and age-related conditions are all treatable at the cellular level); nanorobots are also candidates for industrial applications. The advent of molecular nanotechnology will again expand enormously the effectiveness, comfort and speed of future medical treatments while at the same time significantly reducing their risk, cost, and invasiveness.

CELLULAR BIOSCANNING

The goal of cellular bioscanning is the noninvasive and non-destructive in vivo examination of interior biological structures. One of the most common nanomedical sensor tasks is the scanning of cellular and subcellular structures. Such tasks may include localization and examination of cytoplasmic and nuclear membranes, as well as the identification and diagnostic measurement of cellular contents including organelles and other natural molecular devices, cytoskeletal structures, biochemical composition and the kinetics of the cytoplasm.

The precision and speed of medical nanodevices is so great that they can provide a surfeit of detailed diagnostic information well beyond that which is normally needed in classical medicine for a complete analysis of somatic status.

Except in the most subtle cases a malfunction in any one component of the cellular machinery normally provokes a

cascade of pathological observables in many other subsystems. Detection of any one of these cascade observables, if sufficiently unique and well-defined, may provide adequate diagnostic information to plan a proper reparative procedure.

DNA is one of the few cellular components that is regularly inspected and repaired. Most homeostatic systems adopt a more simple philosophy of periodic replacement of components regardless of functionality. Nanomedicine allows the philosophy of inspection and repair to be extended to all cellular components. The following discussion briefly describes a few of the sensory techniques that may be useful in achieving this objective.

CELLULAR TOPOGRAPHICS

Tactile topographic scanning provides the most direct means for examining cellular structures in vivo. For example, ex vivo live cell scanning in air by commercially available atomic force microscopy using a 20–40 nm radius tip allows nondestructive feature resolutions of ~50 nm across the top 10 nm of the cell.

Investigators have used the scanning tip to punch a hole in the cell, then pull back and scan the breach, observing the membrane heal itself via self-assembly in real time. Up to ~KHz scanning frequencies are possible, with up to 1024 data points per scan line. Assuming a scan rate of ~10^6 pixels/sec, a micron-scale nanodevice, once securely anchored to the surface of a (~20 micron)^3 human cell, could employ a tactile scanning probe to image the 0.1% of plasma membrane lying within its (1.4 micron)^2 vicinity in ~2 sec to

~1 nm^2 resolution (~1 mm/sec tip velocity), or ~50 sec to ~0.2 nm (*e.g.*, atomic) resolution (~0.2 mm/sec tip velocity).

Inside the cell, and again post-anchoring, an entire 6 $micron^2$ mitochondrial surface could be imaged to atomic resolution in ~100 sec; the surface of a 100-nm length of 25-nm diameter microtubule could be atomically resolved in 0.2 sec—though of course in a living cell these structures may be changing dynamically during the scanning process.

From Eqn., continuous power dissipation of a $(0.1\ micron)^2$ scan head moving at 1 mm/sec through water is ~0.002 pW (~kT/pixel at ~10^6 pixels/sec), though of course the energy cost of sensing, recording, and processing each pixel must be at least ~10 kT/pixel (ignoring the possibility of pre-computational image compression), so the total scanner power draw could be as high as 0.02–0.1 pW in continuous operation. Special scanning tips and techniques should allow topographic, roughness, elastic, adhesive, chemical, electrostatic, conductance, capacitance, magnetic, or thermal surface properties to be measured.

For large (~0.1–1 micron) cellular components, identification and preliminary diagnosis of improper structure may be possible using measured surface characteristics which may be matched to entries in an extensive onboard library, perhaps combined with dynamic monitoring of anomalies in cross-membrane molecular traffic using chemical nanosensors.

Most membranes are self-sealing, so it should also be possible to gently insert a telescoping member into the target organelle, which member then slowly reticulates and extrudes smaller probes with sensory tips in a fixed pattern and step

size, allowing the acquisition of detailed internal structural and compositional information. Small (~1–10 nm) protein-based components of the cell such as enzymes, MHC carriers, and ribosomes are self-assembling or require only modest assistance to self-assemble. Reversible mechanical denaturation of these smaller proteins, using diamondoid probe structure to displace water molecules and by using specialized handling tools akin to functionalized AFM tips and molecular clamps, may be followed by precise nondestructive amino acid sequencing to allow identification and diagnostic compositional analysis.

Afterward, the protein molecule may be refolded back into its original minimum-energy conformation, possibly with the assistance of chaperone-like structures in some cases.

Acoustic Microscopy

Acoustic microscopy is another noninvasive scanning technique that can provide nanomedically useful spatial resolutions. Frequencies are in the GHz range, far exceeding the relaxation time of the protoplasm. A cryogenic acoustic microscope operated at 8 GHz has demonstrated 20 nm lateral resolution in liquid helium. By 1998, the best resolution achieved with water as the coupling fluid has been 240 nm at a frequency of 4.4 GHz. Operated in water at 310 K, a nanomechanical 1.5 GHz acoustic microscope would achieve a far-field minimum lateral resolution $l/2 \sim 500$ nm, $\sim 10^5$ voxels per human cell, sufficient to locate and count all major organelles and a few intermediate-scale structures.

Scanning acoustic microscopes (SAM) operated at 2 GHz in reflection mode (which allows detection of interference effects) achieve 30–50 nm resolution in the direction of the

acoustical axis, and the Subtraction SAM approach reveals topographical deviations of 7.5 nm at 1 GHz.

It has been proposed that picosecond ultrasonics could be used to obtain an image of the cytoskeleton with detail comparable to that of conventional X-ray images of a human skeleton. Power requirements are a significant constraint on acoustic reflection microscopy (echolocation).

However, cells and body tissues are mesoscopic "junkyards"—highly heterogeneous media which may produce large numbers of nontarget scattering events, thus increasing the difficulty of extracting signal from noise.

Additional complications arise due to:

- Scattering on rough surfaces;
- Rapid pressure variations with range in the Fresnel zone or near field of the transmitted signal;
- Cytoplasmic viscosity inhomogeneities due to the asymmetric arrangement of cytoskeletal structure, granules, vacuoles, and the endomembrane system; and
- Variation in cytoplasmic Young's modulus due to time-varying tensions in semirandomly-distributed cytosolic fibrillar elements which alter elasticity and thus the local speed of sound.

Magnetic Resonance Cytotomography

Electrostatic scanning is largely ineffective because of Debye-Huckel shielding. Magnetic stray field probes allow resolution of 10 nm physical features, but only for materials with substantial magnetic domains—which most biological substances lack. But nuclear magnetic resonance (NMR) imaging may allow cellular tomography by creating 3-D

proton (hydrogen atom) density maps. Atomic density maps of other biologically important elements with nonzero nuclear magnetic moments (including D, Li^6, B^{10}, B^{11}, C^{13}, N^{14}, N^{15}, O^{17}, F^{19}, Na^{23}, Mg^{25}, P^{31}, Cl^{35}, K^{39}, Fe^{57}, Cu^{63} and Cu^{65}) may also be compiled. For example, sodium imaging is already used clinically to assess brain damage in patients with strokes, epilepsy, and tumors. In a hypothetical NMR cytotomographic nanoinstrument, a large permanent magnet is positioned near the surface of the cell or organelle to be examined. This creates a large static background magnetic field that polarizes the protons. The spatial gradient of this field establishes a unique resonant frequency, called the Larmor frequency, within each isomagnetic surface throughout the test volume.

A second time-varying magnetic driver field is then scanned through the full range of resonant frequencies, exciting into resonance the protons in each isomagnetic surface, in turn. Depending on the sensor implementation chosen, each resonance detected may cause absorption of driver field energy, an increase in measured impedance, or even a return echo of magnetic energy as the excited protons relax to equilibrium in ~1 sec. The large polarizing magnet is then rotated to a new orientation, moving the isomagnetic surfaces to new positions within the test volume, and the scan is repeated. A 3-D map of the spatial proton distribution may be computed after several scan cycles.

Near-Field Optical Nanoimaging

Electromagnetic waves of optical wavelength λ that interact with an object are diffracted into two components, called

"far-field" and "near-field." The propagation of electromagnetic radiation over distances $z > l$ acts as a spatial filter of finite bandwidth, resulting in the familiar diffraction-limited resolution $\sim\lambda/2$.

Classical optics is concerned with this far-field regime with low spatial frequencies $< 2/l$, and conventional optical imaging will be difficult at the cellular level in vivo. (Short-wave X-rays will damage living biological cells.) Information about the high spatial frequency components of the diffracted waves is lost in the far-field regime, so information about sub-wavelength features of the object cannot be retrieved in classical microscopy.

However, for propagation over distances $z << \lambda$, far higher spatial frequencies can be detected because their amplitudes are then of the same order as the sample ($z = 0$). This second diffraction component is the "near-field" evanescent waves with high spatial frequencies $>2/\lambda$. Evanescent waves are confined to subwavelength distances from the object. Thus a localized optical probe, such as a subwavelength aperture in an opaque screen, can be scanned raster fashion in this regime to generate an image with a resolution on the order of the probe size.

The original Near-field Scanning Optical Microscope (NSOM) surpassed the classical diffraction limit by operating an optical probe at close proximity to the object. The NSOM probe uses an aluminum-coated "light funnel" scanned over the sample. Visible light emanates from the narrow end (~20 nm in diameter) of the light funnel and either reflects off the sample or travels through the sample into a detector, producing a visible light image of the surface with ~12 nm

resolution at λ = 514.5 nm provided the distance between light source and sample is very short, about 5 nm, with signal intensities up to 10^{11} photons/sec (~50 nanowatts).

This represents a resolution of ~λ/40. The near-field acoustic equivalent is found in the medical stethoscope, which exhibits a resolution of ~λ/100.Applications include dynamical studies at video scanning rates, low-noise high-resolution spectroscopy, and differential absorption measurements.

Optical imaging of individual dye molecules has already been demonstrated, with the ability to determine the orientation and depth of each target molecule located within ~30 nm of the scanned surface. Molecules with nanometer-scale packing densities have been resolved to ~0.4-nm diameters using STMs to create photon emission maps, and laser interferometric NSOMs have produced clear optical images of dispersed oil drops on mica to ~1 nm resolution.

Live specimen 250-nm optical sectioning for three-dimensional dynamic imaging has also been demonstrated. Submicron laser emitters have been available since the late 1980s. It should be possible to use NSOM-like nanoprobes to optically scan the surfaces of cells or organelles to <~1 nm resolution, mapping their topography and spectroscopic characteristics to depths of tens of nanometers without penetrating the surface.

However, a proper membrane-sealing invasive light funnel ~20 nm in diameter might not seriously disrupt some cellular or organelle membranes and thus could be inserted into the interiors of these bodies or through the cytoskeletal interstices to permit deeper volumetric scanning.

The thermal conductivity of water at 310 K is 0.623 watts/m-K and the energy per 500-nm photon is 400 zJ, so for 1 micron^3 of watery tissue the maximum scan rate is ~35 micron^3/sec-K, if e^{SNR} photons are used to image each 1 nm^3 voxel with SNR = 2. Thus a 1-sec volumetric optical scan of an aqueous ~1-micron^3 sample volume to 1 nm^3 resolution requires a ~1 nanowatt scanner running at ~GHz bit rates, raising sample volume temperature by ~0.03 K.

NSOM permits the determination of five of the six degrees of freedom for each molecule, lacking only the optically inactive rotation around the dipole axis. In principle, it should be possible to produce <~1 nm resolution near-field optical scans of in situ protein molecules, since with atomically precise fabrication and single lines of atoms as conductors a minimum light guide (metal-dielectric-metal) is 3 atoms wide.

Given a molecular laser and adequate collimation, photons passing through folded proteins will scatter according to the molecular structure. Detection of sufficient photons comprising these scattering patterns should allow the noninvasive determination of protein structure; polarized photons provide information on chirality. Absorption and fluorescent signals will be visible from phenyl rings, tryptophan, and bound cofactors such as ATP and adenine (which is fluorescent). Positions of monoclonal antibodies on virus surfaces are now identifiable experimentally using NSOM; it should be possible to map binding sites on virus and cell surfaces using fluorescently labelled antibodies.

Single molecule detection has been proposed as a tool for rapid base-sequencing of DNA. Other optically-based cellular imaging techniques must be distinguished. Optical

Coherence Tomography (OCT) uses Michelson interferometry to achieve ~10 micron spatial resolutions over tissue depths of 2-3 mm in nontransparent tissue in the near infrared.

However, OCT requires numerous physical components not easily implemented on a micron-size detector, femtosecond pulse shaping, and illumination power levels of ~10^7 watts/m^2 >> 100 watts/m^2 "safe" continuous limit in tissue. Bioluminescence techniques in which the light source is placed inside the tissue has a spatial resolution limited to ~10% of the depth, or ~10 microns resolution at a ~100 micron depth.

Coherent anti-Stokes Raman scattering (CARS) already permits organelle imaging in living cells and can in theory create a point-by-point chemical map of a cell using two intersecting lasers. Three-dimensional observation of microscopic biological nonliving structures by means of X-ray holography requires a high degree of spatial coherence and good contrast between target and surroundings. Good contrast may be achieved in the wavelength range between the K absorption edges of carbon (l = 4.37 nm) and oxygen (l =2.33 nm), the "water window" where carbon-containing biological objects absorb radiation efficiently but water is relatively transparent. A 50-micron diameter emitter has been tested that uses near-IR 5-femtosecond laser pulses impinging upon a helium gas target to create a well-collimated (<1 milliradian) beam of coherent soft X-rays at a 1 KHz repetition rate producing a brightness of 5×10^8 photons/ mm^2-$milliradian^2$-sec, a peak X-ray intensity of >10^{10} watts/ m^2 on the propagation axis behind the He target.

Cell Volume Sensing

Many cellular parameters need not be measured directly in order to be detected by a medical nanodevice. Cell volume sensing is a case in point. An intracellular nanodevice can indirectly monitor changes in the volume of the cell in which it resides by one of two methods.

First, measurements of the mechanical deformation of the cellular membrane, stretch-activated channels, or cytoskeletal strains and structural changes are quite sensitive to alterations in total cell volume.

Second, concentration or dilution of the cytoplasmic environment through cell shrinkage or swelling leads to the activation of various volume-regulation responses which may be detected by the nanorobot. Changes in the concentration of soluble cytosolic proteins may nonspecifically affect enzyme activity via "macro- molecular crowding". Minor changes in cell volume can cause severalfold changes in ion transport.

Cellular signalling entities that have been linked to the transduction and amplification of the primary volume signal include Ca^{++} transients, phosphoinositide turnover, eicosanoid metabolism, kinase/phosphatase systems such as JNK and p38, cAMP, and G-proteins. These signal amplification pathways may be monitored using chemical concentration sensors aboard the medical nanodevice, allowing the nanorobot to eavesdrop on the natural sensory channel traffic of the cell.

NONINVASIVE NEUROELECTRIC MONITORING

Noninvasive measurement of axonal traffic within nerve bundles will require multiple sensors and greater sensitivity

to compensate for shielding by the perineurium, a tight resistive sheath enclosing the bundle ~20 microns thick with resistivity ~4000 ohm-cm.

Electric Field Neurosensing

The "typical" ~20 micron human neuron discharges 5-100 times per second, moving from -60 mV potential to +30 mV potential in ~10^{-3} sec. Thus the variation in electric field at the axonal surface is ± 4500 volts/m. Since electric field sensors can detect 100 volt/m fields up to GHz frequencies, an electric sensor attached by circumaxonal cuff or pressed against the axonal surface (possibly at the node of Ranvier) should readily detect each action potential discharge. (The diameter of human nerve axons is 0.1-20 microns.)

By 1998 silicon-to-neuron extracellular junctions already permitted direct stimulation of individual nerve cells in vitro without killing the cells, and extracellular electrodes were commonly used to detect neuronal electrical activity noninvasively, both in vivo and in vitro.

Artificial electric fields may also be employed to trigger or moderate neural signals. Cell membrane capacitance is typically ~0.01 picofarads/micron^2, and varies with the state of the health of the cell.

Magnetic Field Neurosensing

The magnetic flux density caused by a single action potential discharge is ~0.1 microtesla at the axonal surface, which may be detected by a "compass oscillator" type magnetosensor with a ~KHz maximum sampling rate. It might be possible for artificial magnetic fields to directly influence

neural transmissions. Even a static 65 millitesla field has been shown to reduce frog skin Na^+ transport by 10-30%.

Each neuronal discharge develops an electrical energy of ~20 picojoule (~10^{10} kT), far smaller than the magnetic energy stored in a B = 1.4 tesla field of a permanent micromagnet traversing an L^3 = (20 micron)^3 volume which from Eqn. is $B^2 L^3/2 \mu_0$ ~ 6000 pJ. If properly manipulated, such a field may be sufficient to enhance, modulate, or extinguish a passing neural signal.

Neurothermal Sensing

While neurons come in many shapes and sizes, our "exemplar" ~14,000-micron^3 neuron discharging ~90 mV into an input impedance of ~500 Kohms produces ~0.2 microampere current per pulse and generates a continuous (average) 100-300 pW of waste heat as measured experimentally. The discharge rate of 5-100 Hz can produce brief surges up to ~2000 pW during a high-frequency train, but the duty cycle of such trains is far less than 100%, reducing time-averaged dissipation to the observed 100-300 pW range.

Single impulses are measured experimentally to produce 2-7 microkelvin temperature spikes in cold or room-temperature mammalian non-myelinated nerve fibres and ~23 microkelvins in non-myelinated garfish olfactory nerve fibres at an energy density ranging from 270-1670 joules/ m^3-impulse from 0-20°C.

In non-myelinated fibres the initial heat occurs in two temperature-dependent phases: a burst of positive heat, followed by rapid heat reabsorption. The positive heat derives

from the dissipation of free energy stored in the membrane capacity, and from the decrease in entropy of the membrane dielectric with depolarization.

An L ~ 20-micron neuron in good thermal contact with an aqueous heat sink at 310 K has thermal conductance L K_t ~ 10^{-5} watts/K, so trains of 5-100 Hz impulses lasting 1 second should raise cellular temperature by 10-30 microkelvins; up to ~200 microkelvin thermal spikes from such trains have been observed experimentally.

These events are easily detectable by nanoscale thermal sensors capable of ~1 microkelvin sensitivity up to ~1 KHz. The ~microkelvin heat signature of individual impulses or very short pulse trains can probably be temporally resolved because the minimum pulse repetition time is ~10 millisec (at 100 Hz) which is much longer than the thermal time constant for an L ~ 20-micron neuron which is t_{EQ} ~ L^2 C_V/ K_t ~ 3 millisec.

Direct Synaptic Monitoring

The synaptic cleft between the axonal presynaptic terminal and the dendritic postsynaptic membrane is 10-20 nm in most synapses, although in the vertebrate myoneural junction it may be as large as 100 nm. Contact area per bouton is ~1 $micron^2$, giving a total gap volume of ~10^7-10^8 nm^3. The density of acetylcholine receptors is highest in muscles along the crests and upper thirds of the junctional folds (~10,000/$micron^2$), and is lowest in the extrasynaptic regions (~5/$micron^2$).

Each action potential discharge triggers the release of ~10^4-10^5 molecules of acetylcholine into the gap volume of an

active neuromuscular junction (diffusion time ~1 microsec), raising c_{ligand} from near zero to ~3 × 10^{-4} molecules/nm^3 (~0.0005 M) in ~1 millisec, followed by near-complete hydrolyzation by acetylcholinesterase during the 1-2 millisec refractory period.

Into the gap volume may easily be inserted a ~10^5 nm^3 neurotransmitter concentration sensor able to measure ~100 acetylcholine molecules in ~1 millisec, thus detecting pulses at the fastest discharge rate. A similar device could be used to precisely regulate neuro-transmitter concentration at the junction, and hence the neural signal itself, under nanodevice control. Simple electrochemical and mechanochemical artificial synapses have been demonstrated.

RF AND MICROWAVE OSCILLATIONS

Interestingly, this frequency span is very close to the maximum trigger/reset frequency for bioelectronic molecules. Note that (~100 mV) (1.6 x 10^{-19} coul) ~ 4 kT, so a membrane molecule with a single charge on either end should be reliably reoriented by a depolarization wave, coupling pressure waves and electrostatic field fluctuations. However, the direct detection of 10-100 GHz millimeter radiation by non-nanotechnological means is experimentally difficult and controversial because the tests must be performed in vivo in close proximity to an actively metabolizing cell in water—and water strongly absorbs microwaves over macroscale ranges. Nevertheless, active cells have shown enhanced Raman anti-Stokes scattering, an effect ascribed to the converse of the Frohlich oscillations. In one study, the normalized growth rate of yeast cultures was enhanced or

inhibited when irradiated by CW microwave fields of ~30 watts/m^2 of various frequencies; growth rate data spanning 62 separate runs revealed a repeatable frequency-dependent spectral fine structure with six distinct peaks of width ~10 MHz near 42 GHz.

Investigations of related phenomena are ongoing and voluminous; the interested reader should peruse *Bioelectromagnetics*, the archival journal of this field. 100 GHz waves attenuate only ~1% after passing through ~3 microns of soft tissue. A single electron injected into an integral membrane protein could act as an oscillating dipole, making a 300 volt/m signal 10 nm from the protein antenna with an energy transfer of ~0.004 kT per cycle; ~1000 oscillating electrons could produce a measurable field. A 20-micron diameter cell modeled as a nonuniform spherical dipole layer with transmembrane dipoles located 10 nm apart and embedded in a dissipative medium could produce 10^{2}–10^{5} volt/m microwave fields 1-10 microns from the cell surface.

Thus, a variety of rf and microwave electromagnetic emanations may in theory be detectable both within and nearby living cells which could prove diagnostic of numerous internal states.

Such states may include cytoskeletal dynamics, metabolic rates, plasmon-type excitations due to the collective motion of ions freed in chemical reactions, positional, rotational or conformational changes in biological macromolecules and membranes, internal movements of organelles and nerve traffic conduction, cellular pinocytosis, cellular reproduction events, cell membrane identity, and cell-cell interactions.

MACROSENSING

Macrosensing is the detection of global somatic states (inside the human body) and extrasomatic states (sensory data originating outside the human body). While the treatment here is necessarily incomplete, the discussion nevertheless gives a good feel for the kinds of environmental variables that internally-situated nanodevices could sense.

Not all capabilities outlined here need be available on every nanorobot, since injection of a cocktail of numerous distinct but mutually cooperative machine species allows designers to take full advantage of the benefits of functional specialization.

In many cases, a given environmental variable can be measured by several different classes of sensor device. However, since these devices are microscopic it is in theory possible to operationalize almost all of the macrosensing capabilities described below in one patient using just a billion devices (~10 mm^{-3} whole-body deployment density), a total volume of ~1 mm^3 of nanorobots or ~0.1% of the typical ~1 cm^3 therapeutic dose.

BLOOD PRESSURE

Blood pressure ranges from 0.1– 0.2 atm in the arteries to as low as 0.005 atm in the veins. The systolic/diastolic differential ranges from 0.05–0.07 atm in the aorta and 0.01–0.02 atm in the pulmonary artery, falling to 0.001–0.003 atm in the microvessels, or 0.003–0.005 atm if the precapillary sphincter is dilated.

In venous vessels, pulse fluctuations are 0.002–0.010 atm in the superior vena cava, 0.004–0.006 atm in the subclavian vein, ~0.004 atm in venules generally, and ~0.0005 atm in

the brachial vein. There is also a ~0.05 Hz random fluctuation in the microvessels with amplitude on the order of 0.004–0.007 atm. Both blood pressure and pulse rate can be reliably monitored by a medical nanodevice virtually anywhere in the vascular system using a $(68\ nm)^3$ pressure sensor with ~0.001 atm sensitivity.

Pulse propagation through body tissue is somewhat muted due to absorption in compressible fatty membranes, but most cells lie within 1–3 cell-widths of a capillary so the cardiac acoustic signal should still be measurable using more sensitive detectors. The time-averaged interstitial pressure in subcutaneous tissue is 0.001–0.004 atm. Arterial pulse waves (vascular oscillations) carry subtle messages about the health of internal organs and the arterial tree.

The idea of using pulse waves for diagnosis has a long history dating back 2000 years in China. Waves detected by manual probing are classified using such subjective and qualitative descriptors as floating, deep, hidden, rapid, slow, moderate, feeble, replete, full, thready, faint, weak, soft, slippery, hesitant, hollow, firm, long, short, swift, running, intermittent, uneven, taut, string-tight, gigantic, or tremulous.

Abnormal waves were empirically related to disease states. Wave data gathered by nanodevices could make possible a theoretically sound, quantitative system of noninvasive observation, classification, and diagnosis as a supplement to other nanomedical tools.

Respiratory Audition

The variation of mechanical pressure over a complete respiratory cycle is ~0.003 atm in the pleura, ~0.002 atm at

the alveoli, detectable by nanomedical pressure sensors positioned in the vicinity of the respiratory organs. Holding a deep breath further stretches the pulmonary elastic tissue, up to 0.02 atm. However, turbulent flows at Reynolds number $N_R > 2300$ in the trachea, main bronchus and lobar bronchus produce a whooshing noise that may be the loudest noncardiac sound in the human torso during conventional auscultation.

The energy dissipation for turbulent flow in a tube is:

$$P_{turb} = P_{lam} Z = 8 \pi_{air} v^2 L Z \text{ (watts)}$$

where P_{lam} is the dissipation for laminar (Poiseuillean) flow in a long circular cylindrical tube of length L, v is the mean flow velocity, and turbulence factor $Z = 0.005 (N_R^{3/4} – (2300)^{3/4})$, a well-known empirical formula.

For $\eta_{air} = 1.83 \times 10^{-5}$ kg/m-sec for room-temperature air (20°C), $P_{turb} = 0.87$ milliwatts for the trachea; $P_{turb} = 0.66$ milliwatts for the main bronchus (L = 0.167 m, v = 4.27 m/sec and $N_R = 3210$); and $P_{turb} = 0.09$ milliwatts for the lobar bronchus (L = 0.186 m, v = 4.62 m/sec and $N_R = 2390$), totalling ~1.6 milliwatts acoustic emission from a ~120 cm^3 upper tracheobroncheal volume.

This is a power density of 13 watts/m^3 corresponding to a pressure of 4×10^{-5} atm assuming a 300 millisecond measurement window at the maximum respiration rate. The amplitude of an acoustic plane wave propagating through tissue attenuates exponentially with distance due to absorption, scattering and reflection. The amplitude is given approximately by,

$$A_x = A_D e^{-\alpha F x}$$

where A_0 is the initial wave amplitude in atm, A_x is the amplitude a distance x from the source, and α is the amplitude absorption coefficient. The function F expresses the frequency dependence of the attenuation. For pure liquids, $F = F_{liq} = \nu^2$ (Hz^2); for example, $\alpha_{liq} = 2.5 \times 10^{-14}$ sec^2/m for water at room temperature. However, for soft tissues, $F = F_{tiss} \sim \nu$ (Hz).

Mechanical Body Noises

Many other mechanical body noises should be globally audible to properly instrumented medical nanodevices. If normal chewing motions (of hard foods) release 1–10 milliwatts in a ~100 cm^3 oral volume with a ~1 sec jawstroke, power density is ~10 watts/m^3 or ~10^{-4} atm of toothcrunching noise. A stomach growl registering 45 dB (vs. 30 dB whisper, 60 dB normal conversation) at 2 meters has a source power of 160 milliwatts; released from a 10 cm^3 gastric sphincter volume gives a ~2×10^{-6} atm acoustic wave, detectable throughout the body.

Walking and running releases 20-100 joules/footfall for a 70 kg man; assuming the energy is absorbed within a ~1 cm thickness or within ~1 second by the sole of the foot, Eqn. implies an upward-moving planar compression wave of 0.4-2.0 atm, easily detectable by acoustically instrumented nanodevices body-wide. Hand-clapping generates 0.02-0.2 atm pulses, also easily detectable.

Lesser noises including ~30 millisec hiccups at (4–60)/min, intestinal and ureteral peristalsis, sloshing of liquid stomach contents, heart murmurs, a tap on the shoulder by a friend, nasal sniffling and swallowing, clicks from picking

or drumming fingernails, crepitations, manustuprations and ejaculations, the rustling noise of clothing against the skin, flapping eyelids, anal towelling, bruits (including murmurs and thrills) due to vascular lesions, dermal impact of water while showering, copulatory noises, urethral flow turbulence during urination, transmitted vibrations from musical instruments, creaking joints, and squeaking muscles can be detected locally if not globally.

Implantation of significant interconnected in vivo diamondoid structures may produce increased sensitivity to internal noises, due to the extremely low acoustic absorption coefficient of diamond.

Vocalizations

Average source power for conversational speech in air is ~10 microwatts at the vocal cords (60 dB), up to ~1000 microwatts for shouting (90 dB) and as little as 0.1 microwatts (30 dB) for whispering. Vocal cord surface area ~1 cm^2, giving an acoustic intensity I ~ 0.001–10 watts/m^2. (Using the decibel notation, dB = 10 $\log_{10}$ (I/I_0), where $I_0 \sim 5 \times 10^{-13}$ watts/m^2 in air, $I_0 \sim 1 \times 10^{-16}$ watts/m^2 in water.) In a planar traveling wave, pressure amplitude A_p (N/m^2) is related to power intensity I by,

(N/m^2)

For water at 310 K, $\rho = 993.4$ kg/m^3 and $v_{sound} = 1500$ m/sec, therefore A_p = 0.0005–0.05 atm for speech, detectable by nanodevices throughout the body due to minimal attenuation at audible frequencies. Other easily detectable vocalizations include whistling, humming, coughing, sneezing, rales, wheezing, expectorating, eructations, flatus, vomiting, hawking and noseblowing.

4

Nano Networks

CONCEPTS

PASSIVE AND ACTIVE

Passive DRM protects its content from onlookers who do not have a DRM-enabled client. Encryption is generally used for Passive DRM, so that the content is meaningless garbage unless you have the right bits in your client. I consider this 'passive' protection, because the data is inaccessible by default and only becomes accessible if you have the right kind of client, with the right key.

Active DRM, then, would be a scheme where protection is only provided if the client in use is one that is correctly coded to block access where it has not been specifically granted. This is a scheme in which the data is readily accessible to most normal viewers/players, but has a special

code that tells a DRM-enabled viewer/player to hide the content from people who haven't been approved.

The whole problem is his two categories are a false distinction. You can't arbitrarily draw a line through a system and say "this is passive, this is active." For your CSS example, if you consider a given player's decryption code along with an arbitrary encrypted DVD, you have a system with both active and passive elements. If you leave out either of those elements, you have a disc that won't play or a player with no disc, the only perfectly secure system (assuming your cryptography is good.) When judging the efficiency of new compression schemes, the size of the decoder is added to the size of the compressed data to get a fair assessment of its efficiency. Otherwise you could win contests with a one-byte file and a 10 GB decoder programme that simply contains all the actual data. Whichever way you design a system, complexity is being pushed from one party to another but never eliminated. For DVD, where most of the complexity is in the player, there is a huge variety of player implementations that each have their own bugs. The author of every disc needs to test against many combinations of players because of that problem.

Likewise, if you push the complexity onto the disc by including executable code there, the player gets simpler but the disc could be buggy. However, in that case, the content author will get a bad reputation for the buggy disc (see the Sony rootkit fiasco he mentions).

This doesn't just apply to DRM. While he might consider a MPEG4-AVC video file as "passive" in his terminology, it is really a complex series of instructions to the decoder.

Look at the number of different but valid ways to encode video and you'll see it's closer to a programme than to "passive" data. Now in his definition for "Active DRM", he is not actually describing the general class of software protection techniques. He is describing a system that is poorly-designed, often due to an attempt to retrofit DRM onto an existing system without it. Ofcourse it makes sense that if you have two ways to access the content, one with DRM and the other without, the additional complexity makes no sense to the end-user or mass copiers. It may make economic sense to the content author, but they have to weigh the potential risks to their business also (annoying users vs. stopping some casual copying.)

Even assuming his terminology makes sense, the Windows Media Center system he references is actually a combination of "active" and "passive". The cable video stream is encrypted ("passive"), and the Windows DRM component is "active". In particular, it has a "black box" DLL that checks the host environment and hashes various items to derive a key, hence the problem.

No one, especially content producers, asks for an unprotected format and assumes they'll bolt on something later. All you're seeing is the fact that some formats made it out of the gate without protection (CD) or were weak and eventually broken (DVD-CSS). In those cases, some people make the effort to add optional "assassin" software, even if it only slows down some small percentage of users. You and I both agree that this is ineffective, easily bypassed, and can interfere with legitimate use. But some companies have analysed the cost/benefit tradeoff and based on that

information continue to deploy this approach rather than accept the alternative (do nothing).

Side note: the reason why this subject is particularly important to me is that some people are already using the concepts "active" and "passive" in terms of DRM. The claim is that crypto-based systems (*i.e.*, AACS) are "passive" and software protection (*i.e.*, BD+) are "active". This is nonsense since you have to consider the software that decrypts the disc part of the system, and thus there are no true "passive" systems. It's effectively already happened–while multi-region DVD players are common and inexpensive in the UK, in the US, I find it difficult to find a player that is advertised as multi-region. While they're clearly not illegal, the marketplace here has somehow been coerced into making them unavailable.

Maybe my proposed two types should be "DRM" and "not DRM":) Well, it's some kind of restriction, although it's certainly not in the same class as something designed to be DRM from the beginning. It's similar to people saying XOR with a constant is "encrypting the data"—nowhere near accurate. In that case, I call it "masking the data" since it is slightly transformed. So what's the term for "unprotected, but with a couple speedbumps bolted on the side"?

DUAL-FREQUENCY

A dual-frequency-band patch antenna (200) capable of resonating at two different operating frequencies. The dual-frequency-band patch antenna (200) includes two conductive plates (202, 204) that are placed side by side, with one conductive plate sized to resonate at a first frequency and the other conductive plate sized to resonate at a second frequency. The conductive plates (202, 204) are electrically

joined by an electrical connection (210), and share a single feed (208) from a transceiver. The conductive plates (202, 204) are separated from a ground plane (206) by a dielectric medium (212), which according to the present invention is very thin in terms of the wavelength at which the antenna operates ($i(*i.e.*), 'low-profile').

When the antenna is excited at the first frequency, the first conductive plate (202) resonates like a conventional patch antenna. The second conductive plate (204), however, acts like a parasitic patch, thereby increasing the overall bandwidth and efficiency. Conversely, when the antenna (200) is excited at the second frequency, the second conductive plate (204) resonates and the first conductive plates (202) acts like a parasitic patch.

VOLTAGE AND CURRENT SOURCES

Real voltage sources can be represented as ideal voltage sources in series with a resistance r, the ideal voltage source having zero resistance. Real current sources can be represented as ideal current sources in parallel with a resistance r, the ideal current source having infinite resistance.

WHAT IS AN IDEAL VOLTAGE SOURCE

Looking at voltage sources first, let's consider the concept of an ideal voltage source. When you buy a battery you buy a nine (9) volt battery or a 1.5 v battery, or a twelve (12) volt car battery or some other battery that has a specified voltage. Clearly those batteries - electrical sources - are assumed to

have a certain voltage - whatever it is - that doesn't change much. When you buy a power supply for your calculator or a telephone answering machine or some other device you need to look at the voltage you need for the power supply. They usually come in a few specified voltages. It seems clear that many sources are designed to give you some particular voltage and to attempt to maintain a constant voltage.

Below we'll consider simple models for voltage sources that maintain a constant voltage and we'll take a look at how you can represent that kind of device. A voltage source that can maintain a constant voltage - no matter what you do to it, like drawing a lot of current, or putting it in a situation where current flows through it - is an ideal voltage source.

That's what we are going to examine in this lesson. We'll leave it for another lesson to look at sources that are not ideal. So, let's answer that question above. What is an ideal voltage source? And there's another question implied. Why worry about an ideal voltage source since nothing like that exists in the real world. We'll take those questions in order.

IDEAL VOLTAGE SOURCES

The concept of an ideal voltage source is pretty simple, and it was really embedded in the previous discussion.

- An ideal voltage source is a voltage source that maintains the same voltage across the source's terminals no matter what current is drawn from the terminals of the source or what current flows into the terminals.
- That's it in a nutshell. If the source is a DC Source, we can plot a voltage current plot for an ideal voltage

source. The plot is shown below. However, we need to define terms. Here is a circuit symbol for an ideal voltage source.

In this symbol, we assume the following:

- The voltage across the terminals is denoted as V_t.
- The load current flowing from the source to a load (presumably a load is attached when the source is in a circuit) is denoted as I_L.
- With those definitions, here is the source symbol. It's just a circle with polarity indicated.

And, here is the plot of terminal voltage against load current.

Given the discussion above, we can say:

- V_t = constant, no matter what the load current is. That's pretty much the description of the ideal voltage source. It's not too complex, but it is an important concept. In the next section we'll look at how you can put this concept to use. For the rest of this section we'll look at ideal current sources starting next.

WHAT IS AN IDEAL CURRENT SOURCE

An ideal current source is a simple model for many current sources. It is reminiscent of the ideal voltage source but with voltage and current interchanged. Here is the story.

- There is a special circuit symbol for an ideal current source. See below.
- I_L = constant, no matter what the terminal voltage is.
- The plot of load current against terminal voltage is similar to the plot for an ideal voltage source, but voltage and current are interchanged. Here is the plot.

Notice that an ideal current source is somewhat similar to an ideal voltage source. However, when you use an ideal source - usually when doing circuit analysis- there is a significant difference in the analysis. However, that's getting ahead of the story. We first have to worry about how you would "use" an ideal source, when we know that there is no such thing as an ideal source, *i.e.* a source that is "perfect" in some way.

USING IDEAL SOURCES

The idea of using ideal sources is something that you may rebel at. After all, there is no such thing as an ideal source anywhere in the world. You can't pull an ideal source off the shelf in the lab, so why are we even talking about them? The answer to that question is that you use ideal sources when you have a non-ideal (a real source) source in a circuit. There are two important things to note.

- There are some sources that are very good sources and that can be modelled as ideal sources. (And when that happens, be grateful.) Some situations like that include the following.
  - A power supply in the lab. Many times you connect a power supply to some electronic circuit, for example, and when you connect the circuit you find that the output voltage from the power supply doesn't change measurably. (After all, power supply designers try to make that happen!) In that case, the power supply might be considered to be an ideal source at least as long as you are working on that particular circuit.
- There are many sources that do not perform ideally.

However, it has proven to be possible to construct models of real sources, and those models often contain ideal source in combination with other ideal elements (like resistors, etc.).

- The vinin and Nortonequivalent circuits are examples of models of real sources that can account for loading effects (*i.e.* drawing enough current from the source to change the output voltage) and they are widely used in circuit analysis. You will even find that manufacturers give you parameter values for Thevinin and Norton equivalents on the front panel of many instruments like function generators.

You often have situations in which the sources that you use can be approximated with ideal sources. Shown below is a bridge circuit powered by a battery. Often a battery maintains a pretty constant voltage across the terminals, so you may be able to replace the battery with an ideal voltage source when you analyse the circuit.

Here's the circuit with an ideal voltage source substituted for the battery. At this point, you may know how to do the analysis so you're ready to go. If you don't know how to do the analysis, you'll get there in these lessons. to go to the first lesson on circuit analysis. However, you may want to spend more time looking at sources, particular sources that are not ideal.

PC Controlled Voltage or Current Source

This project was begun as a means to charge and cycle NiCad batteries but has become a versatile tool for experiments requiring either a controlled voltage up to +30V and/or a current of +/- 2.5 Amps. If all you want to do is

charge and recycle NiCad batteries, then, with hindsight, I'd advise you to buy one of the commerical units available. The circuit I will describe will do the job but it is experimental and has many short comings. It is more the beginning of a useful tool than one that has been fully developed. This circuit and the related software have taken me a good 2 years to reach even a marginal degree of reliability despite it's apparently simple actions.

My complete assembly uses the Discovery Series K-2805 Parrallel Port Interface Kit as distributed by Dick Smith Electronics as the source of control voltages and, if used as a charger, the Battery Cell measurements. Any other PC interface with at least 4 analogue inputs and 2 analogue outputs, both with 0 to +5V range, could be put to use. Below is the circuit for one of the two identical voltage/current sources included in my construction.

The components are:

Vrawpos/neg is supplied in my construction by a seperate power supply module which comprises a centre tapped transformer, a 4amp rectifier and two 10000uF capacitors supplying unregulated 18/36V peak DC.

Typically I use this circuit with Prawpos at +18V and Prawneg and Ground connected to the centre tap. More typically Prawpos would be connected to +18V, Prawneg to -18V and Ground to the centre tap.

For higher voltages, Ground can be connected to the -18V supply but great care must be taken this is not connecting a -18V voltage to the PC interface ground which may be earthed through the PC.

"With versitility comes confusion!" - R.Parker 2001AD

The load in my case is a NiCad battery of between 1.2 and 24V, but it could really be anything; active loads like batteries are just more of a challenge to control.

Note: to discharge a 1.2V battery will required a negative power source, but for larger voltages, Prawneg may be connected to ground as described above.

Q1 is sources voltage to the load and Q2 sinks voltage from the load. Naturally they get hot and so need good heat sinking and some thermal isolation from other components and sensitive fingers. I was lucky to find a preloved aluminium box with an internal frame on which I mounted the transistors. Air holes either end have further helped to keep the box reasonably cool.

Source or Sink actions can be disabled by opening S2 or S3 respectively. R2 and R3 then clamp the transistor base to prevent transistor operation. Even then, some minor leakage will be apparent through R2 & R3 themselves; a better solution may be to have S2 & S3 in circuit to the load but their rating would need to be increased accordingly.

Transistors Q3 & Q4 buffer the controlling action of opamp OA1 which, throught the magic of feedback, eliminates any dead zone due to the Vbe voltage drops in Q3, Q4 or Q1 (*i.e.* without feedback, any control voltage would first have to overcome these voltage drops before current could flow to or from the load).

The link between the base of Q3 and the emitter of Q4 is to enable lower discharge voltages to ground when Vrawneg is to ground (enabling discharge to +0.7V with S2 closed, +1.1V with S2 open). This will only work with D1 there to allow the emitter of Q4 to rise above ground. D1 is also there

to prevent Base to Emitter breakdown of Q4 (which occurs when the output of OA1 exceeds +6V); it doesn't really prevent it, but it does limit the current from OA1s output through the base to the emitter. The same it true for Q3 if the output of OA1 goes below -6V.

Trying to understand this apparently simple circuit has caused me a few headaches. As mentioned already, to reliably discharge a 1.2V cell a split supply is really necessary but it is a bit scary letting a PC connect a -15V to a +1.2V cell.

A voltage between 0 and +5V is fed to OA1 either from a PC interface or, when the interface is unplugged, from the manually operated VR1 (it assumes that the impedance of the Analogue Voltage from the PC interface is much lower than R12). I have used this circuit to charge batteries using only the manual contol but VR1 does not provide a stable control for low currents and frequent adjustment is necessary (not to mention a good alarm clock to remind you to turn it off). Still the manual control is particularly useful for quick diagnostics.

With SW1 in position A, Current Source Mode, negative feedback to OA1 comes from a current measurement taken across R1 via OA2. Since the OA2 is referenced to +2.5V, 0V on the control voltage should result in -2.5Amps (sinking) through R1 (assuming a suitable load is connected). If the control voltage rises to +5.0V, then +2.5Amps (sourcing) through R1 should result. If the load imedance is too high, the output voltage at the load will rise or fall towards the supply rail voltages (more on this in a moment).

With SW1 in position B, Voltage Source Mode, negative feedback to OA1 comes from the output voltage, divided by a

factor of 6 by R9 and R10. A control voltage of 0V will result in 0V at the output, +5V will (should) result in +30V (if Vpos is > +30V) at the output. I am thinking R10 could be connected to Vneg to allow negative outputs but it seemed more useful to have the output voltage controlled relative to ground and I haven't tried it out.

For the purposes of Battery charging, SW1 remains in position A. With the opamp used (LM324) and the the supply conditioning I provide for the opamp (see below), the maximum supply voltage in current source mode is Vrawpos - 4.5V (13.5V for 18V supply). As such, running this system off a car battery will only allow you to charge a 5 cell NiCad. Using an OpAmp which can source and sink closer to supply voltage and possibly connecting the OpAmp supply directly to Vrawpos (quite okay for a 13.5V car battery supply) may allow you to charge 7 cell packs off a car battery but I haven't tried it. I think we'll leave fast charging off car batteries to a more dedicated unit.

I originally thought it would be useful to measure both output voltage and output current and since both feedback signals should be in the 0 to 5V range, these can be measured with the PC Interface that I use which has 10 analogue inputs. Current measurement turns out to be of no particular use in battery charging as any errors in the control of the source or sink current arise from the +2.5V reference voltage which in turn are reflected in the measurement with a few extra errors thrown in.

The current measurement may be of more use where this device is being used as a voltage source or in analysing someone elses charger (I have done this to determine the

parameters for an unlabelled NiCad pack that came with a toy car and plug pack charger - see External Logging). Current can also be measured by a sensitive voltmeter across R1 using test poit TP1.

Below is the circuit I used to supply the I.C.s and voltage references:

The components are:

Label	**Component**
D1	Rectifier Diode rated over 30V and 100 mA
ZD1	Zener Diode 30V 1W
ZD2	Zener Diode 5.1V 400mW
Q5	BC557 PNP
RD1	LM336-2.5V Voltage Reference
D5	Small Rectifier Diode (~30mA duty)
LED1	Whatever
R13	330R 1W
R14	2.2K 1/4W
R15	110R 1/4W
R16	1K 1/4W
C2	10uF 50VW

The LM324 is happy with voltages up to 30V, so this supply is clamped at 30V with ZD1 and C2 removes a good deal of ripple from the unregulated supply. The value of R13 was found by trial and error to match the I.C. supply current. D4 is there to prevent the inevidable mistake.

The voltage reference circuit draws more curreThe LM324 is happy with voltages nt than I would have liked so I have used a current source from the unregulated supply to feed 5V Zener ZD2 which in turn feeds the 2.5V reference circuit. Originally I used a simple resistor to feed ZD2 but with the supply varying from 10 to 36V, the current to ZD2 also varied greatly and all sorts of problems followed. Q5 controls the

current to approx. 20mA by using the voltage drop through LED1 as a reference voltage. The diode is there to drop the "5V" level to about 4.7V. This is connected in my unit to the 5V supply in the PC interface (through a 1K resistor) so that both operate at the exact same level. This was a good idea when the 2.5V reference was created by a simple divider, but now that I use a 2.5V reference diode, this link and diode D5 are nolonger necessary and I have only shown it as this is what is in my construction.

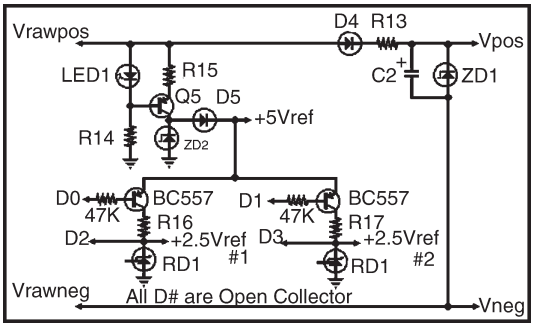
While I have shown only one 2.5V ref. circuit (R16 & RD1), I have in my construction 2 circuits, one for each voltage current source circuit. This was necessary before I fitted Q5 to provide a constant current to ZD2. I suspect two 2.5V reference circuits are nolonger necessary but I haven't tried it yet.

One of the problems in using this Current Source for battery charging is that, until the software has taken control, this unit will initially attempt to sink 2.5Amps from the battery (control voltage at 0V). As such, either the battery must be left disconnected or S3 left open until the software has got itself together. If the computer crashes for some reason the results would be unpleasant.

The simplest solution I have come up with so far is to disable the +2.5V reference circuit with the following alternative supply circuit.

While D0 to D3 are open collector outputs from the PC Interface which must be enabled by the software before they sink any current. As such, while D0 and D1 are not enabled following boot up, the 2.5V reference will float between 0V (as will occur with zero control volts) and about 0.8V. On

my unit at boot up, this results in about 100mA discharge of the connected battery. Not exactly passive, but not disasterous either.



As it happens, this modification provides additional versatility. By connecting D2 and D3 to the 2.5V reference voltages, these can be clamped to rise no greater than 0.9V. When this is so, 5V control voltage will now correspond to 5Amps source. Given the right power supply (mine could not handle this load for long) and the right load, this could be useful. There would be no affect in the voltage source mode.

This is the end of description of the circuit regarding voltage and current source control. The software (charger) I've written for an IBM PC compatible (based on Intel Microprocessor 8086 series and above) computer and the Discovery Series PC Interface, has a manual test mode and a battery charger mode.

Future modes intended for the voltage/current source will be added as they occur to me (or you). I will be considering release of the source code on a GNU licence with the hope that useful modification would be released back to me. In the mean while, if you would like the source code, my current e-mail address is "rpact" with an ISP called "dingoblue.net.au" (I'm trying to avoid automatic spam so put an @ in between and mail me).

CONCEPT OF LINEARITY AND LINEAR NETWORK

LINEARITY PROPERTY

The growth in areas of application of electric circuits has led to an evolution from simple to complex circuits. To handle the complexity, engineers over the years have developed some theorems to simplify circuit analysis. Such theorems include Thevenin's and Norton's theorems. Since these theorems are applicable to linear circuits, we first discuss the concept of circuit linearity. In addition to circuit theorems, we discuss the concepts of superposition, source transformation, and maximum power transfer in this chapter. The concepts we develop are applied in the last section to source modeling and resistance measurement.

Linearity is the property of an element describing a linear relationship between cause and effect. Although the property applies to many circuit elements, we shall limit its applicability to resistors in this chapter. The property is a combination of both the homogeneity (scaling) property and the additive property. The homogeneity property requires that if the input (also called the excitation) is multiplied by a constant, then the output (also called response) is multiplied by the same constant. For a resistor, for example, Ohm's law relates the input I to the output v,

$$v=iR$$

If the current is increased by constant k, then the voltage increases correspondingly by k, that is,

$$kiR=kv$$

The additive property requires that the response to a sum of inputs is the sum of the responses to each input applied separately. Using the voltage current relationship of a resistor, if,

$$v_1=i_1R$$

and,

$$v_2=i_2R$$

then, applying (i_1+i_2) gives,

$$v=(i_1+i_2)R=i_1R+i_2R=v_1+v_2$$

We say that a resistor is a linear element because the voltage-current relationship satisfies both the homogeneity and the additive properties. In general, a circuit is linear if it is both additive and homogeneous. A linear circuit consists of only linear elements, linear dependent sources, and independent sources.

A linear circuit is one whose output is linearly related (or directly proportional) to its input.note that since $p=i^2R=v^2/R$ (making it a quadratic function rather than a linear one), the relationship between power and voltage (or current) is nonlinear. Therefore, the theorems covered in this chapter are not applicable to power.

To illustrate the linearity principle, consider the linear circuit shown in Figure. The linear circuit has no independent sources inside it.

It is excited by a voltage source vs, which serves as the input. The circuit is terminated by a load R. We may take the current i through R as the output. Suppose $v_S=10V$ gives i = 2 A. According to the linearity principle, v_S=1V will give i = 0.2 A. By the same token, i = 1 mA must be due to v_S=5mV.

SUPERPOSITION

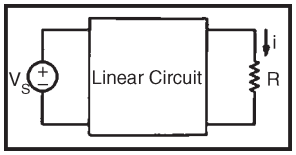


Fig. A linear circuit with input Vs and output

If a circuit has two or more independent sources, one way to determine the value of a specific variable (voltage or current) is to use nodal or mesh analysis as in. Another way is to determine the contribution of each independent source to the variable and then add them up. The latter approach is known as the superposition.

The idea of superposition rests on the linearity property. The superposition principle states that the voltage across (or current through) an element in a linear circuit is the algebraic sum of the voltages across (or currents through) that elements due to each independent source acting alone.

The principle of superposition helps us to analyse a linear circuit with more than one independent source by calculating the contribution of each independent source separately. However, to apply the superposition principle, we must keep two things in mind.

- We consider one independent source at a time while all other independent sources are turned off. This implies that we replace every voltage source by 0 V (or a short circuit), and every current source by 0 A (or an open circuit). This way we obtain a simpler and more manageable circuit.
- Dependent sources are left intact because they are controlled by circuit variables.

With this in mind, we apply the superposition principle in three steps:

Steps to apply superposition principle:

- Turn off all independent sources except one source. Find the output (voltage or current) due to that active source using the techniques covered in.
- Repeat step 1 for each of the other independent sources.

Find the total contribution by adding algebraically all the contributions due to the independent sources.

Analyzing a circuit using superposition has one major disadvantage: it may very likely involve more work. If the circuit has three independent sources, we may have to analyse three simpler circuits each providing the contribution due to the respectively individual source. However, superposition does help reduce a complex circuit to simpler circuits through replacement of voltage sources by short circuits and of current sources by open circuits.

Keep in mind that superposition is based on linearity. For this reason, it not applicable to the effect on power due to each source, because the power absorbed by a resistor depends on the square of the voltage or current. If the power value is needed, the current through (or voltage across) the element must be calculated first using superposition.

SOURCE TRANSFORMATION

We have noticed that series-parallel combination and wye-delta transformation help simplify circuits. Source transformation is another tool for simplifying circuits. Basic to these tools is concept of equivalence. We recall that an equivalent circuit is one whose v-i characteristics are identical

with the original circuit. In Section, we saw that node-voltage (or mesh current) equations can be obtained by mere inspection of a circuit when the sources are all independent current (or all independent voltage) sources. It is therefore expedient in circuit analysis to be able to substitute a voltage source in series with a resistor for a current source in parallel with a resistor, or vice versa, as shown in Figure Either substitution is known as source transformation.

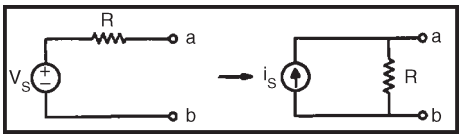


Fig. Transformation of independent sources.

A source transformation is the process of replacing a voltage source vs in series with a resistor R by a current source is in parallel with a resistor R, or vice versa.

The two circuits in Figure are equivalent–provided they have the same voltage–current relation at terminals a-b. It is easy to show that they are indeed equivalent. If the sources are turned off, the equivalent resistance at terminals a-b in both circuits is R. Also, when terminals a-b are short circuited, the short circuit current flowing from a to b is $i_{SG}=v_S/R$ in the circuit in the left hand side and $i_{SG}=i_S$ for the circuit on right hand side. Thus, $v_S/R=i_S$ in order for the two circuits to be equivalent.

Source transformation also applies to dependent sources, provided we carefully handle the dependent variable. As shown in Figure, a dependent voltage source in series with a resistor can be transformed to a dependent current source in parallel with the resistor or vice versa where we make sure that Equation is satisfied.

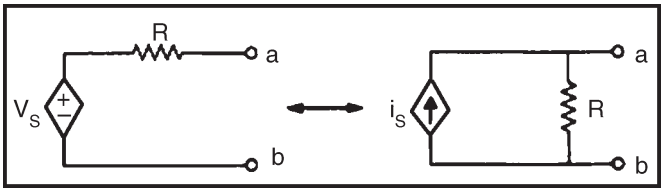


Fig. Transformation of dependent sources.

Like wye-delta transformation we studied in a source transformation does not affect the remaining part of the circuit. When applicable, source transformation is a powerful tool that allows circuit manipulations to ease circuit analysis. However, we should keep the following points in mind when dealing with source transformation.

- Note from that the arrow of the current source is directed towards the positive terminal of the voltage source.
- Note from Equation that source transformation is not possible when R = 0, which is the case with an ideal voltage source. However, for a practical, nonideal voltage source, $R \neq 0$. Similarly, an ideal current source with $R = \infty$ cannot be replaced by a finite voltage source.

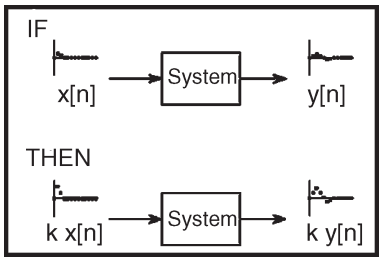
REQUIREMENTS FOR LINEARITY

A system is called linear, if it has two mathematical properties:homogeneity (hôma-gen-â-ity) and additivity. If you can show that a system has both properties, then you have proven that the system is linear. Likewise, if you can show that a system doesn't have one or both properties, you have proven that it isn't linear. A third property, shift invariance, is not a strict requirement for linearity, but it is a mandatory property for most DSP techniques.

When you see the term linear system used in DSP, you should assume it includes shift invariance unless you have reason to believe otherwise.

These three properties form the mathematics of how linear system theory is defined and used. Later in this chapter we will look at more intuitive ways of understanding linearity. For now, let's go through these formal mathematical properties.

As illustrated in Figure; homogeneity means that a change in the input signal's amplitude results in a corresponding change in the output signal's amplitude. In mathematical terms, if an input signal of *x[n]* results in an output signal of *y[n]*, an input of *kx[n]* results in an output of *ky[n]*, for any input signal and constant, *k*.

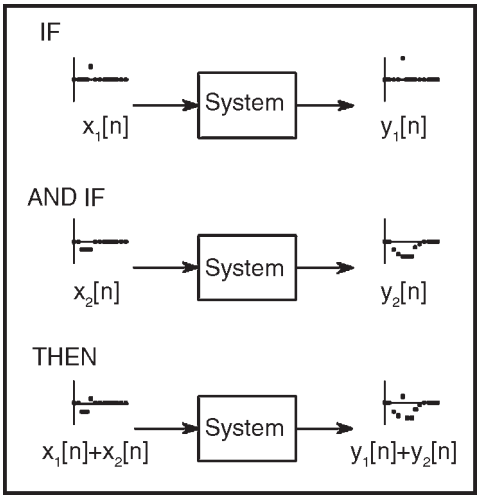


A simple resistor provides a good example of both homogenous and non-homogeneous systems. If, the input to the system is the voltage across the resistor, *v(t)*, and the output from the system is the current through the resistor, *i(t)*, the system is homogeneous. Ohm's law guarantees this; if the voltage is increased or decreased, there will be a corresponding increase or decrease in the current.

Now, consider another system where the input signal is the voltage across the resistor, *v(t)* but the output signal is the power being dissipated in the resistor, *p(t)*. Since power is proportional to the square of the voltage, if, the input signal

is increased by a factor of two, the output signal is increase by a factor of four. This system is not homogeneous and therefore cannot be linear.

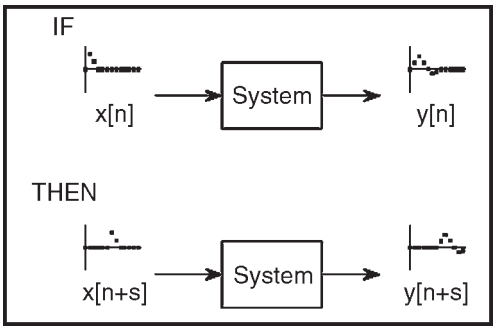
The property of additivity is illustrated in Figure; Consider a system where an input of $x_1[n]$ produces an output of $y_1[n]$. Further suppose that a different input, $x_2[n]$, produces another output, $y_2[n]$. The system is said to be additive, if an input of $x_1[n] + x_2[n]$ results in an output of $y_1[n] + y_2[n]$, for all possible input signals. In words, signals added at the input produce signals that are added at the output.



The important point is that added signals pass through the system without interacting. As an example, think about a telephone conversation with your Aunt Edna and Uncle Bernie. Aunt Edna begins a rather lengthy story about how well her radishes are doing this year. In the background, Uncle Bernie is yelling at the dog for having an accident in his favourite chair. The two voice signals are added and electronically transmitted through the telephone network. Since this system is additive, the sound you hear is the sum of the two voices as they would sound if transmitted individually. You hear Edna and Bernie, not the creature, Ednabernie.

A good example of a nonadditive circuit is the mixer stage in a radio transmitter. Two signals are present: an audio signal that contains the voice or music, and a carrier wave that can propagate through space when applied to an antenna. The two signals are added and applied to a nonlinearity, such as a pn junction diode. This results in the signals merging to form a third signal, a modulated radio wave capable of carrying the information over great distances. As shown in Figure; shift invariance means that a shift in the input signal will result in nothing more than an identical shift in the output signal. In more formal terms, if an input signal of *x[n]* results in an output of *y[n]*, an input signal of *x[n + s]* results in an output of *y[n + s]*, for any input signal and any constant's.

Pay particular notice to how the mathematics of this shift is written, it will be used in upcoming chapters. By adding a constant's to the independent variable, *n*, the waveform can be advanced or retarded in the horizontal direction. For example, when $s = 2$, the signal is shifted left by two samples; when $s = -2$, the signal is shifted right by two samples.



Shift invariance is important because it means the characteristics of the system do not change with time (or whatever the independent variable happens to be). If a blip in the input causes a blop in the output, you can be assured

that another blipwill cause an identical blop. Most of the systems you encounter will be shift invariant. This is fortunate, because it is difficult to deal with systems that change their characteristics while in operation.

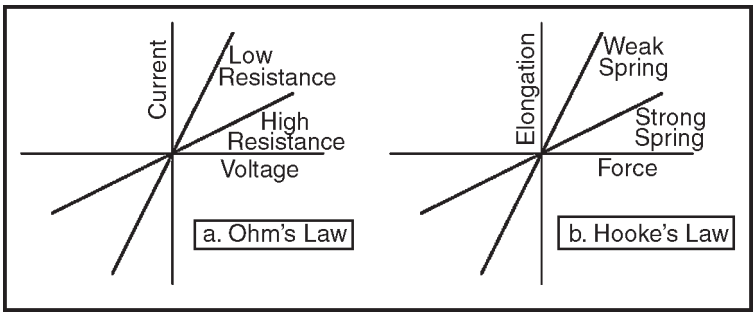
For example, imagine that you have designed a digital filter to compensate for the degrading effects of a telephone transmission line. Your filter makes the voices sound more natural and easier to understand. Much to your surprise, along comes winter and you find the characteristics of the telephone line have changed with temperature. Your compensation filter is now mismatched and doesn't work especially well. This situation may require1a more sophisticated algorithm that can adapt to changing conditions. Why do homogeneity and additivity play a critical role in linearity, while shift invariance is something on the side? This is because linearity is a very broad concept, encompassing much more than just signals and systems. For example, consider a farmer selling oranges for $2 per crate and apples for $5 per crate. If the farmer sells only oranges, he will receive $20 for 10 crates, and $40 for 20 crates, making the exchange homogenous.

If he sells 20 crates of oranges and 10 crates of apples, the farmer will receive: This is the same amount as if the two had been sold individually, making the transaction additive. Being both homogenous and additive, this sale of goods is a linear process. However, since there are no signals involved, this is not a system and shift invariance has no meaning. Shift invariance can be thought of as an additional aspect of linearity needed when signals and systems are involved.

STATIC LINEARITY AND SINUSOIDAL FIDELITY

Homogeneity, additivity, and shift invariance are important because they provide the mathematical basis for defining linear systems. Unfortunately, these properties alone don't provide most scientists and engineers with an intuitive feeling of what linear systems are about. The properties of static linearity and sinusoidal fidelityare often of help here. These are not especially important from a mathematical standpoint, but relate to how humans think about and understand linear systems. You should pay special attention to this section.

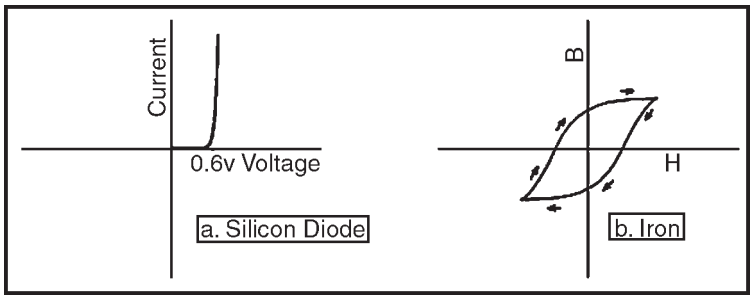
Static linearity defines how a linear system reacts when the signals aren't changing, *i.e.*, when they are *DC* or static. The output is the input multiplied by a constant. That is a graph of the possible input values plotted against the corresponding output values is a straight line that passes through the origin. This is shown in Figure for two common linear systems: Ohm's law for resistors and Hooke's law for springs. For comparison, Figure shows the static relationship for two nonlinear systems: a pn junction diode and the magnetic properties of iron.



All linear systems have the property of static linearity. The opposite is usually true, but not always. There are systems that show static linearity, but are not linear with respect to changing signals. However, a very common class of systems

can be completely understood with static linearity alone. In these systems it doesn't matter if the input signal is static or changing. These are called memoryless systems because the output depends only on the present state of the input, and not on its history.

For example, the instantaneous current in a resistor depends only on the instantaneous voltage across it, and not on how the signals came to be the value they are. If a system has static linearity and is memoryless then the system must be linear. This provides an important way to understand (and prove) the linearity of these simple systems.



An important characteristic of linear systems is how they behave with sinusoids a property we will call sinusoidal fidelity: If the input to a linear system is a sinusoidal wave, the output will also be a sinusoidal wave, and at exactly the same frequency as the input. Sinusoids are the only waveform that have this property. For instance, there is no reason to expect that a square wave entering a linear system will produce a square wave on the output. Although a sinusoid on the input guarantees a sinusoid on the output, the two may be different in amplitude and phase. This should be familiar from your knowledge of electronics: a circuit can be described by its frequency response, graphs of how the circuit's gain and phase vary with frequency.

Now for the reverse question: If a system always produces a sinusoidal output in response to a sinusoidal input, is the system guaranteed to be linear? The answer is no, but the exceptions are rare and usually obvious. For example, imagine an evil demon hiding inside a system, with the goal of trying to mislead you. The demon has an oscilloscope to observe the input signal and a sine wave generator to produce an output signal. When you feed a sine wave into the input, the demon quickly measures the frequency and adjusts his signal generator to produce a corresponding output.

Of course, this system is not linear because it is not additive. To show this, place the sum of two sine waves into the system. The demon can only respond with a single sine wave for the output. This example is not as contrived as you might think phase lock loops operate in much this way.

To get a better feeling for linearity, think about a technician trying to determine if an electronic device is linear. The technician would attach a sine wave generator to the input of the device and an oscilloscope to the output. With a sine wave input, the technician would look to see if the output is also a sine wave.

For example, the output cannot be clipped on the top or bottom, the top half cannot look different from the bottom half, there must be no distortion where the signal crosses zero, etc. Next, the technician would vary the amplitude of the input and observe the effect on the output signal. If the system is linear, the amplitude of the output must track the amplitude of the input. Lastly, the technician would vary the input signal's frequency, and verify that the output signal's frequency changes accordingly.

As the frequency is changed, there will likely be amplitude and phase changes seen in the output, but these are perfectly permissible in a linear system. At some frequencies, the output may even be zero, that is, a sinusoid with zero amplitude. If the technician sees all these things, he will conclude that the system is linear. While this conclusion is not a rigorous mathematical proof, the level of confidence is justifiably high.

5

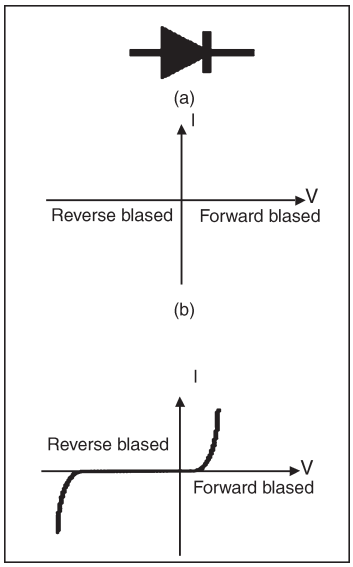
Molecular Electronic Digital Logic in Nanocomputer

Molecular electronic digital logic circuits are built up simply from two or three or a few molecular electronic diode switches. Several simple molecular electronic digital logic circuits are AND, OR, XOR gate, which are three fundamental logic gates. Logic circuits for digital systems may be combinational, such as half adder. The area of the molecular electronic logic structures is one million times smaller than analogous logic structures that currently are implemented in micro-scale, solid-state semiconductor integrated circuits.

BACKGROUND

A diode is simply a two-terminal switch. It can turn a current on or off as it attempts to pass through the diode from the "in" to the "out" terminal. The diode is said to be "on" when there's current flow, and it is said "off" when there's no current flow. When the voltage difference across the diode

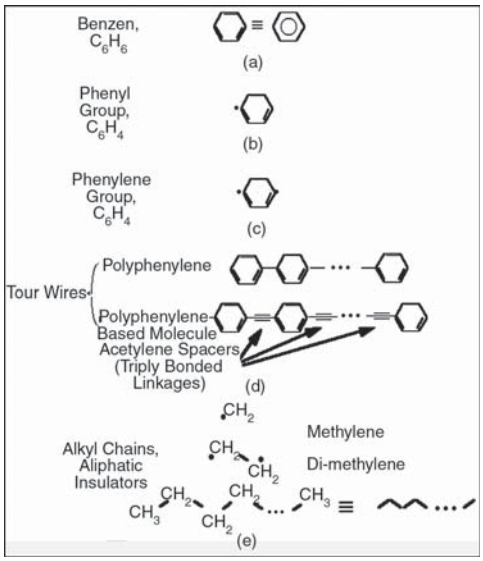
across the diode is greater than or equal to zero, it is the forward bias; reversely, if the voltage difference across the diode across the diode is less than zero, it is reverse bias.



The and gate with a logic-1 output if and only if all the input signals are logic-1; otherwise, the output products logic-0. The OR gate with a logic-1 when any input is a logic-1. Its output becomes logic-0 if all input signals are logic-0. The XOR (exclusive-OR) with a logic-0 if all the input signals are logic-1 or if all the input signals are logic-0; otherwise, the output products logic-1. A half adder is a combinational circuit that forms the arithmetic sum of two input bits. It consists of two binary inputs and two binary outputs. The input variables designate the augend and addend bits; the output variables produce the sum and carry.

The carry output is 0 unless both inputs are 1. The output S represents the least significant bit of the sum. Polyphenylene-based chains and carbon nanotubes are two primary types of molecules that have been proposed and confirmed for use as the potential basis or backbone for

current-carrying, molecular-scale electronic devices. Polyphenylene-based molecular wires and switches involve chains of organic aromatic benzene rings, shows in Figure. They serve as a conductors or wires. "Aliphatic" organic molecules serve as insulators due to not easy transport electron current, shows in Figure. They formed by single bond molecules that contain only sigma bonds that lie along the exes of the atoms that are joined to form the backbone of the molecule.



Carbon nanotube-based molecular scale electronic devices is fullerene structure which consist of graphite cylinder — curved, a honeycomb lattice rolled into a cylinder, it only consisting carbon; and closed at either end with caps containing pentagons.

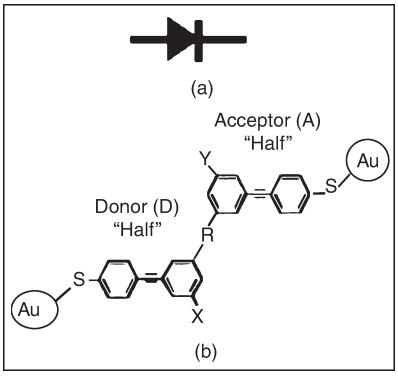
They can make an extremely conductive wire. It is complicated to develop carbon nanotube logic architectures due to their complex chemistry structures. On the other hand, polyphenylene-based molecules are much easier to propose the more complex molecular electronic structures.

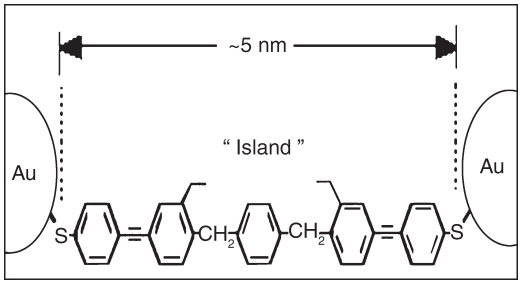
Additionally, these polyphenylene-based molecules are much smaller than carbon nanotubes provides better molecular circuit designs.

PROPOSED POLYPHENYLENE-BASED MOLECULAR RECITYING DIODES SWITCH DESIGN

The molecular structure of polypheneylene-based rectifying diode switch shown in Figure has two intramole-cular dopant groups, which are electron donating and electron withdrawing, substitute as X, and Y, respectively. The electron donating is separated by a semi-insulating group R from an electron withdrawing (electron acceptor). The common semi-insulating substituents R are aliphatic groups, such as sigma-bonded methylene groups, R = $–CH_2–$ or dimethylene groups, R = $–CH_2CH_2–$.

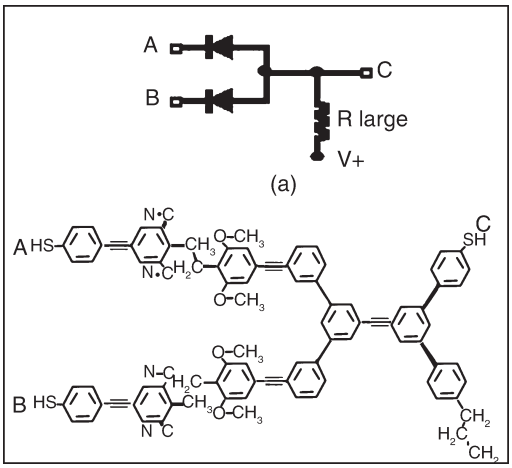
The aliphatic methylene group is used as the central bridging group for the polyphenylene-based rectifying diode designs because it is the smallest nonconducting. The common electron donating substituents X are $–NH_2$, –OH, $–CH–_3$, and $–CH_2CH_3$. The common electron withdraw subsitituents Y are $–NO_2$, –CN, –CHO, and –COR', where R' is an aliphatic chain.

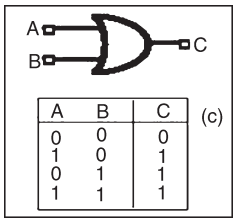




Novel Electronic Logic Gates

Diode-diode logic structures are used in molecular AND and OR gates. Both circuit for AND and OR digital logic gates based upon the diodes. The schematic diagrams for these two circuits.

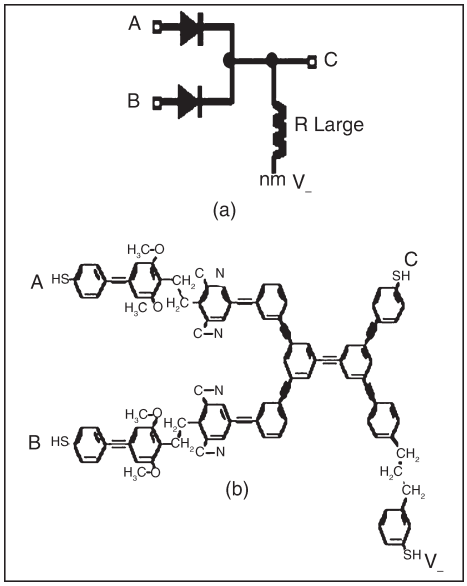


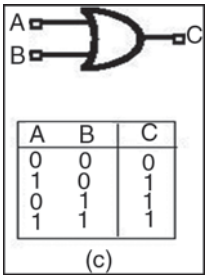


A	B	C
0	0	0
1	0	1
0	1	1
1	1	1

The operation of the diode-diode AND gate, the potential at C of the AND gate circuit is V+ or 1 when both inputs A and B are switched to "on", the both inputs are "1". The diode is reverse biased since cathode greater than anode. The diode behaves like an open switch that no current flows.

If either input A or input B is set to "off" and the other to "on". The input presents the binary number "1" corresponds to the setting of "on".





A	B	C
0	0	0
1	0	1
0	1	1
1	1	1

(c)

Since one of the two inputs is set to "off", the diode behaves as a closed switch, and it is forward biased which allows current to flow. Therefore, the potential at C is zero volts, which is "0". When both inputs A and B are switched to "off", the diode behaves as a closed switch for both inputs, then both of diodes are forward biased in circuit allowing current to flow. The potential at C drops to 0, the output of the entire gate is 0. See Figure.

The operation of the diode-diode OR gate, the potential C is "0" when both inputs A and B are switched to "off". The

diode acts as open switch, and it is reverse biased, which prevents current flow. If either input A or input B is set to "off" and the other to "on". The setting potential applied in the "on" is corresponds to binary "1". Since one of the input is set to "1", there's a forward biased diode in the circuit allowing current to flow.

Then the potential at C is "1". If both inputs A and B are switched to "on", both the diodes behaves as closed switch, and both of the diodes are forward biased, which allow current flows in the circuit. The potential at C then is "1". The dimension of these molecular logic gates each only about 3 x 4 nm^2, which is approximately one million times smaller than the logic element fabricated on a semiconductor chip. The primary different between the AND gate in Figure and OR gate in Figure is that the orientation of the molecular diodes is reversed. Figure 4c, d and Figure 6c, d show the symbol of gate and its truth tables of the inputs and outputs for each gate.

Molecular XOR uses diode-based logic. A schematic diagram for its circuit is shown in Figure 8a. The symbol of an "N", or a "Z" lying on its side represents an RTD (Resonant Tunneling Diodes), in the schematic circuit. The molecular implementation of this diode-based XOR is shown in Figure 8b. The area of this molecular logic gate is around 5 x 5 nm^2. One of the Reed-Tour molecular RTDs and two of the polyphenylene-based Tour wires form the molecular XOR gate. It is very similar to that for the OR gate in Figure, except for the addition of the molecular RTD.

There are two main cases to be considered for the operation of the diode-based molecular XOR gate. First case is when

both inputs A and B are "0" or both inputs are "1", where the RTD is "off" and the output voltage of the gate is therefore "0". The second case if only one of the inputs is "on", for example, input A is "on", 1, and input B is "off", 0. Then RTD becomes "on", thus the output of the XOR gate is "1".

When combine the bond together with several fundamental molecular logic gates will make larger molecular structures. The most common combinational logic circuit for a binary half adder design is shown in Figure. The novel design for the molecular structure is displayed in Figure. The size of the molecular electronic half adder is approximately 10 x $10nm^2$. The symbol of A and B represent the one bit binary inputs to the adder, while S and C represent the one bit outputs, the sum and the carry bits, respectively.

FABRICATION

For the past few decades, electronic computers have become tremendously powerful as their basic structure, the transistor, has become smaller and smaller. However, the laws of quantum physics, quantum mechanics and the technology to fabricate may be soon prevent any more reduction in the dimension and size of the traditional JFET and MOSFET. Experts have predicted that the next-generation electronics during the next 10 to 15 years will become very difficult and more costly to fabricate, as the smallest features on mass-produced transistors reduce its size from their present lengths of 250 nm to100 nm and below.

Also, it is doubtful that it will work and operate effectively in integrated electronic circuits. In order to build and fabricate

circuit elements in the nanometre scale, perhaps even to the molecular scale, we will have to come up with new technology to produce transistor consistently for ultra-dense circuitry. These new nanometre-scale electronic devices will need to perform not only as switches and but also as amplifiers, similar to the transistors that we use today.

INTRODUCTION

The question in not whether we have the best technology to construct a nanocomputer that carries out the optimal computational performance. It is whether the nano technologies and designs will function effectively and also fabricated economically, consistently, and reliably. It is very important to improved fabrication technologies.

Fabrication technologies control the progress in nanotechnology and nanoelectronics. In other words, without fabrication technologies, we would not even allow ourselves to think and propose the next generation electronics. It does not matter how small an electronic device can be built in theory or how tiny it could be.

In production, fabrication processes determines the limitation on how small the device can be built. In recent years, a significant amount of effort and resources have been applied to advance techniques for the fabrication of nanometre-scale structures.

There has been a major increase in the research and development of the techniques in which nanometre-scale structures can be fabricated. With the great improvement in fabrication, we can now start developing theories on how nano electronics could be made, and perhaps use our imagination to picture the next generation computer.

Present techniques for the fabrication can be broken down into four categories:

1. Lithography
2. Molecular Beam Epitaxy (MBE)
3. Mechanosynthesis
4. Chemosynthesis

Lithography and MBE are more of traditional techniques to fabricate nano-electronic devices; and are mostly employed by the semiconductor industries. On the other hand, mechanosynthesis and chemosynthesis are both emerging fabrication techniques.

LITHOGRAPHY

Lithography is a technique that uses a beam of light or matter to create a pattern on a surface. There are several lithography techniques that are currently being used in the industry; including UV lithography, X-ray lithography, atom lithography, and Electron-beam lithography.

UV Lithography

Most modern integrated circuits are produced by photolithography. Photolithography is a process that beams visible or ultraviolet light through a reusable mask and onto a thin coating of photoresistive material covering a silicon wafer.

X-ray Lithography

X-ray lithography is a further refinement of lithographic techniques using ultraviolet light. This refinement provides a more precise tool with which to carve out a pattern on a substrate. The smaller wavelengths of X-rays allow feature sizes from 500 to 30 nm to be attained.

Electron-beam Lithography

Electron-beam lithography replaces the light beam and masks used in photolithography with a direct beam of electrons. It works well with for high resolution features because electrons have much shorter wavelengths than light and can be focused very precisely using electric field.

Atom Lithography

Atom lithograph actually writes the atom directly onto the substrate. It uses the standing wave of light as mask to guide a beam of atoms to desired resting places on the surface of a wafer.

MBE

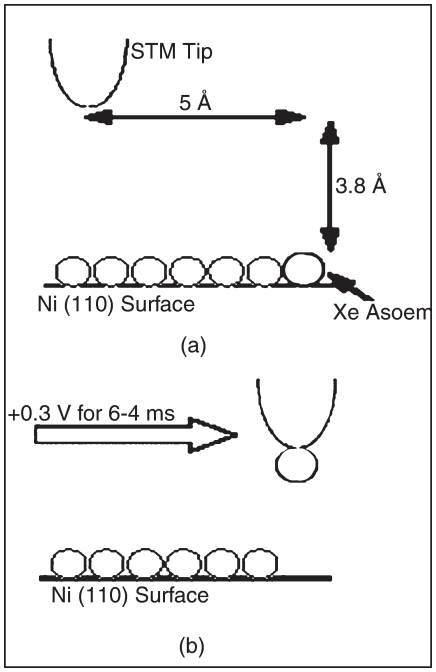
MBE is an advanced fabrication technique for creating layered surfaces. Molecular beam epitaxy uses a beam of molecules under low pressure that collides with a heated single-crystal surface to create epitaxial layers of molecules.

Mechanosynthesis

Nano electronic devices maybe one day assembled by the mechanical positioning of atoms or molecular building blocks, one atom or one molecule at a time, a process known as mechanosynthesis. Quantum mechanics assures that the molecular-scale moving parts should not be subject to the large frictional effects that defeated earlier attempts to build complex macroscopic mechanical computers. However, there are near-term drawbacks. One such drawback is that the fabrication of such nanomechanical devices is likely to require "hand-made" parts assembled one atom or molecular subunit at a time using STMs in processes that are relatively slow.

While this might be done, it would be tedious work to move even a few atoms into a specific position this way, and it would be increasingly more difficult to manufacture reliably the many precision parts for the computer.

It is possible, though, that this problem might be alleviated, somewhat, by the perfection and evolution of recently developed STM arrays that could build many nanoscale components in parallel. Stereospecific chemical reactions and chemical self-assembly also might be applied to help realise a mechanical nanocomputer.



CHEMOSYNTHESIS

Chemosynthesis is also an emerging fabrication of the components for nano-scale electronics. Chemical self assembly is a form of chemosynthesis. Chemical self assembly is the spontaneous orientation of a number of molecules. It usually occurs in non-covalent bonding among molecules. One advantage of this method is the error

correction process. It corrects the wrong type of molecules, and wrong positioned molecules in the assembly process. Another type of chemosynthesis is Hybrid Chemosynthesis, it combines the use of atom beams with some techniques of self-assembly.

Future Challenges

Demonstration of a Molecular Electronic Rectifier or Transistor

We need to increase the density and raise the temperature in which nanoelectronic devices can operate above the cryogenic range, it is very important to fabricate nanoelectronic devices on the same scale as a single molecule. One proposed method is to design and synthesis of single molecule.

Fabricate Working Electronic Device from Molecular Transistors

Even if we know how to make molecular transistors, the assembly of these components into a working logic structure still presents a problem. One possible method to the assemble such a device is to use a scanning-tunneling electron microscope to arrange the molecular components on a surface

Demonstration of a Nanoscale Silicon Quantum Heterojunction

For us to reduce the size of modern electronic devices down to the nanometre scale, it is apparent that we need to construct quantum wells of that dimension. Knowing that, we must build very tiny layers of solid structures, where

each layers are made of different semiconductors with different energies. These layered structures as we know are semiconductor heterojunctions. It is very difficult to make them reliable on the nanometre scale. It is even more difficult to make them on the nanometre scale out of silicon compounds. This is critical is we were to continue the trend towards fast reduction in the size of solid-state electronic devices.

Demonstration of Nanometre-scale Quantum dot Cells and Wireless Logic

The design for constructing wireless quantum dot computer logic is a very promising idea for implementing nanoelectronic computers. In order to make nanometre-scale devices of this type, we need to come up with a method to fabricate and test this device.

Demonstration of Terabit Quantum-effect Electronic Memory "Chip"

If we were to build nanoelectronic logic devices, it is very possible to assemble from them is terabit (10^12 bit) memory array. With terabit memory array, we would have a much larger storage. Also, we will have a much faster access and no moving mechanical parts. Storage of a movie on a such chip is on example.

Nanofabrication with a Micro-STM or Micro-AFM

It is very difficult to mechanically assemble nanoscopic structures and devices with macroscopic probes. Using microelectromechanical systems (MEMS) devices will permit

more efficient mechanical manipulation of nanometre-scale structures. We will need to apply micro-STMs and micro-AFMs to practical nanofabrication.

Parallel Nanofabrication with a Micro-STM or Micro -AFM Arrays

For one thing, if nanoelectronics is to become practical and reliable, we must fabricate nanometre-scale structures by the billions and with high effieniency. Now, we fabricate nanostructures one at a time with a micromechanical STM or AFM is simply not enough.

Responsive Virtual Environment for Realistic, Stimulated Nanomanipulation

We need to be able to simulate nanometre-scale experiment in real time on a digital computer, including all the key quantum mechanical effects, and then use that computer simulation to generate a virtual environment. The quantum simulations required for this type of simulated virtual environments are well beyond our current quantum simulation technology. We need to work and address this challenge.

The Interconnect Problem

Even all the other challenges to fabricate nanometre-scale electronic devices are overcome. We still need to find a way to get information in and out of a dense computational structure with trillions of electrical elements.

Nanocomputers will store a tremendous amount of information in a very tiny and limited space, and the computer will generate information extremely fast. We will

need to control and coordinate the elements of the computer. New methods are necessary for input/output and for control.

CONCLUSION

It is evident that the conventional semiconductor technology and photographic etching techniques will reach its theoretical limits. It is necessary to come up with new approaches to build the computers of next generation. Whether or not nanocomputers can be built will depend upon several factors; including device speed, power dissipation, reliability, and ease of fabrication.

Applying the methods of quantum dots, single electron transistor, and resonant tunneling devices, and the method of fabrication techniques, we should be able to achieve the high expectation for the next generation nanocomputers.

COMPILING SCHEME TO NANO-COMPUTERS

NANO-COMPUTERS

Our objective in these notes will be to identify some very simple machine models which are nevertheless powerful enough to run Scheme. Since even real RISC micro-processors are vastly more complex than the simple machines we consider, we'll refer to our machines as nano-computers. Of course even though in theory Scheme can be compiled into nano-computer code, we are not suggesting that this would be a sensible thing to do.

The resulting compiled code is spectacularly inefficient, and there would be no point in trying to run it. Rather, what

we learn from this reduction of Scheme to a nano-computer is that apparently very limited nano-computers are as resistant to uniform analysis as much richer Scheme expressions. For example, the compilation process transforms the Halting Problem for Scheme into the Halting Problem for nano-computers. Hence the Nano-Computer Computer Halting Problem is "just as hard" as that for Scheme. In particular, the Nano-Computer Computer Halting Problem is undecidable. It is worth emphasizing that the definition of undecidability refers to

Scheme procedures. Undecidability of the Nano-Computer Halting Problem means that SCHEME programmes are incapable of deciding whether nano-computer programmes halt. So even using a richly featured language, it is not possible to predict the behaviour of extremely limited programmes.

FOOTNOTE

For certain kinds of "simple" machines, it is possible to show that the Halting Problem for Simple Machines cannot be decided BY "simple" machines. For example, we might choose simple machines to be the kind of "pushdown-store" machines used to parse inputs according to a BNF grammar.

These machines are too limited to include a decider for their own halting problem. But undecidability BY simple machines need not imply that a problem is really hard. For example, there is a rather efficient Scheme procedure which decides the Halting Problem for Pushdown-Store machines.)

END-FOOTNOTE

A result like this provides some valuable guidance about where the "hard part" of programme analysis resides. For

example, an important part of real compiler technology involves programme analyses which are far more detailed than just determining the halting property.

How can compilers do this analysis, since we know such analysis is undecidable? The answer is that compilers DON'T fully analyse ALL possible source programmes—they do the best they can. For some classes of programmes, compilers can use procedures which are guaranteed to provide complete analyses; for programmes which fall outside of the classes which the compiler can handle, there is no guarantee of the quality of analysis or of the resulting compiled code. It would be a waste of time for compiler designer to try developing perfect analysis methods for some unanalysable subclass of programmes.

From the compilation of Scheme to nano-computers, we learn, for example, that simply cutting down on the built-in operations in a language cannot be expected to yield an analysable class of programmes. In this way, identifying simple unanalysable/undecidable subclasses of programmes gives useful guidance on where NOT to waste time. The insight that even very simple programmes have undecidable properties plays an additional important role in the Theory of Computation.

It is a key technical fact used in establishing the pervasiveness of undecidability in subject areas far removed from Computer Science. So far we have seen only "incestuous" undecidable problems: problems about programmes which cannot be decided by programmes. Later we will discover undecidable problems about polynomials, simple algebras, and simple grammars, for example.

The proofs will be based on learning to "programme" with polynomials, algebraic equations, etc., so they simulate more familiar kinds of programmes. But these problems from outside Computer Science were not intended to embody programming, and it is generally very difficult to programme with them.

For example, writing algebraic equations which, in a suitable sense, "simulate," Scheme programmes would be an extremely cumbersome exercise. On the other hand, as should be expected, it is much easier to write equations which simulate nano-computer programmes. But simulating nano-computers is sufficient, since nano-computers in turn can simulate Scheme.

One of simplest nano-computers we consider is the n-Counter machine. This is a machine with n registers for some $n > 0$, each register capable of holding a nonnegative integer of arbitrary size. The only operations of the machine are to increment (add one) to any register, decrement any register (leaving it unchanged if the register contains zero), and branch on whether a register contains zero. The registers are called "counters" since in a single step they can only be incremented and decremented.

For example, in 2-counter machines, the only instructions are:

Instruction	**Informal meaning**
Inc1	Increment first register
Inc2	Increment second register
Dec1	Decrement first register, except leave it alone if it is 0.
Dec2	Decrement second register, except leave it alone if it is 0.
Ifr1 n1 n2	If register 1 is zero, go to line n1,

	otherwise go to line n2.
Ifr2 n1 n2	If register 2 is zero, go to line n1, otherwise go to line n2.

A 2-counter machine, M, is a finite sequence of such instructions, numbered consecutively starting at zero. For example, the machine M1 below, adds the contents of its first counter to its second counter and then halts (by transfering to an instruction number larger than any in the programme).

2-counter Machine M1

0: ifr1 4 1	Line 4 means "halt"
1: dec1	
2: inc2	
3: ifr1 0 0	Goto line 0

The "configuration" of a 2-CM is defined to be a triple (k, l, m) of nonnegative integers. The intended interpretation of the triple is that k is the number of the next instruction to be executed, and l and m are the contents of counters 1 and 2.

Formally, we associate with 2-CM, M, a partial function step_M from configurations to configurations, where step_m(k, l, m) is the configuration, if any, after execution of M's k^th instruction when counter 1 contains l, and counter 2 contains m.

For example, for the machine above:

- Step_M1(0, 2, 0) = (1, 2, 0),
- Step_M1(1, 2, 0) = (2, 1, 0),
- Step_M1(2, 1, 0) = (3, 1, 1),
- Step_M1(3, 1, 1) = (0, 1, 1).

Exercise 1: Give a precise definition of the function step_M for an arbitrary 2-CM, M.

Now given the limited structure of counter machines, it is not even apparent what it would MEAN to compile Scheme onto one of them. Basic Scheme printable values such as symbols, lists or even negative integers are not available in counter machines, so we cannot expect a counter machine programme to produce the same printable values as a Scheme source programme.

The obvious thing to do would be to define some representation of Scheme values in terms of counter machine values, namely, nonnegative integers. We would then require that for any Scheme "source" expression, S, the compiled code—a counter machine we will call M_S—halts with the contents of some designated "output" counter equal to the nonnegative integer representation of the value of S.

However, for our purposes it is not necessary to develop such a representation of Scheme values. Namely, to conclude that the halting problem for various nano-computers is undecidable, it is enough to simulate halting behaviour only. We say a 2-CM HALTS iff it halts when started in the "standard" initial configuration (0, 0, 0). Our main result is:

THEOREM

Scheme->2-CM) There is a procedure for translating any Scheme expression, S, into a 2-counter machine, M_S, such that S halts iff M_S halts. From this we immediately conclude

NANO-COMPUTING

The history of computer technology has involved a sequence of changes from gears to relays to valves to transistors to integrated circuits and so on. Today's techniques can fit logic

gates and wires a fraction of a micron wide onto a silicon chip. Soon the parts will become smaller and smaller until they are made up of only a handful of atoms. At this point the laws of classical physics break down and the rules of quantum mechanics take over, so the new quantum technology must replace and/or supplement what we presently have.

It will support an entirely new kind of computation with new algorithms based on quantum principles. Presently our digital computers rely on bits, which, when charged, represent on, true, or 1. When not charged they become off, false, or 0.

A register of 3 bits can represent at a given moment in time one of eight numbers (000,001,010,...,111). In the quantum state, an atom (one bit) can be in two places at once according to the laws of quantum physics, so 3 atoms (quantum bits or qubits) can represent all eight numbers at any given time.

So for x number of qubits, there can be 2^x numbers stored. (I will not go into the logic of all this or this paper would turn into a book!). Parallel processing can take place on the 2^x input numbers, performing the same task that a classical computer would have to repeat 2^x times or use 2^x processors working in parallel. In other words a quantum computer offers an enormous gain in the use of computational resources such as time and memory.

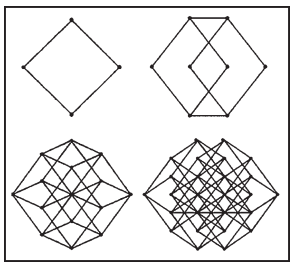
This becomes mind boggling when you think of what 32 qubits can accomplish. This all sounds like another purely technological process. Classical computers can do the same computations as quantum computers, only needing more

time and more memory. The catch is that they need exponentially more time and memory to match the power of a quantum computer.

An exponential increase is really fast, and available time and memory run out very quickly. Quantum computers can be programed in a qualitatively new way using new algorithms. For example, we can construct new algorithms for solving problems, some of which can turn difficult mathematical problems, such as factorization, into easy ones. The difficulty of factorization of large numbers is the basis for the security of many common methods of encryption. RSA, the most popular public key cryptosystem used to protect electronic bank accounts gets its security from the difficulty of factoring very large numbers. This was one of the first potential uses for a quantum computer. "Experimental and theoretical research in quantum computation is accelerating world-wide.

New technologies for realising quantum computers are being proposed, and new types of quantum computation with various advantages over classical computation are continually being discovered and analysed and we believe some of them will bear technological fruit. From a fundamental standpoint, however, it does not matter how useful quantum computation turns out to be, nor does it matter whether we build the first quantum computer tomorrow, next year or centuries from now. The quantum theory of computation must in any case be an integral part of the world view of anyone who seeks a fundamental understanding of the quantum theory and the processing of information."(Centre for Quantum Comput-ation)

HYPERCUBES COULD BE BUILDING BLOCKS OF NANOCOMPUTERS



Multi-dimensional structures called hypercubes may act as the building blocks for tomorrow's nanocomputers – machines made of such tiny elements that they are dominated not by forces that we're familiar with every day, but by quantum properties.

As Samuel Lee and Loyd Hook from the University of Oklahoma explain, microelectronic devices are continually getting smaller and faster, in accordance with Moore's Law. Already, integrated circuits and transistors are reaching the nanometre scale, although they still operate based on the physical properties on the macro-scale. True nanoelectronics, the researchers explain, are not just scaled down microelectronics, but devices that will be dominated by quantum properties, and will therefore require new architectures and novel structures.

"Compared to today's microcomputers, the main advantages of future nanocomputers are higher circuit density, lower power consumption, faster computation speed and more parallel and distributed computing capabilities. For example, today's integrated circuits process information in the form of a continual flow of electrons.

Nano integrated circuits, however, may process individual electrons, reducing the scale and power consumption. Such circuits would require that nano logic devices be able to count single electrons, as well as the ability for parallel computing, reversibility, locality, and a three-dimensional architecture.

To address these challenges, Lee and Hook have investigated hypercubes, which researchers have previously considered as elements of nanocomputers. In their study, which will be published in a future issue of IEEE Transactions on Computers, Lee and Hook propose a variant of the classic hypercube called the "M-hypercube" that could provide a higher-dimensional layout to support the three-dimensional integrated circuits in nanocomputers.

The M-hypercube has a structure similar to a classic hypercube, which basically extends from a square to a cube to increasingly complex M-dimensional shapes. M-hypercubes (of any dimension) are composed of nodes and links. The nodes act as gates, receiving and passing electrons through, while the links act as the paths that electrons travel along.

"The unique structure of hypercubes, including M-hypercubes, has been shown to be effective in parallel computing and communication networks and provides a unique ideal intrinsic structure which fulfills many of the needs of future nanocomputing systems," Lee said. "These needs include massively parallel and distributed processing architecture with simple and robust communication linkages."

ATOMIC-SCALE SIMULATIONS OF NANOELECTRONIC DEVICES

Unlike in classic hypercubes, M-hypercubes contain two types of nodes: state nodes, which are embedded on the

"joints" of the M-hypercubes; and transmission nodes, which are embedded in the middle of the links between state nodes. In one arrangement, the researchers embedded two state nodes on each joint, both representing a single state. Each node can be turned on or off, with the transmission nodes having the ability to isolate parts of the cube from other parts when in the off state. Depending on the number of states required by an operation, the M-hypercube can be expanded by adding extra dimensions (which contain more nodes) or constricted by reducing its dimensions. For example, if only four states are required, the logic architecture would be a 2-D hypercube (a square), which has four state nodes. In general, the number of state nodes in a hypercube is 2m, with m being the M-hypercube's dimensionality.

"We might construct M-hypercubes of dimensions greater than three in three-dimensional space if we allow the communication linkages at the nodes of M-hypercubes to not be mutually perpendicular," Lee explained. For logic operations that require many states, the researchers propose a method that could reduce the dimensions of the M-hypercube by essentially decomposing the hypercube into two lower-dimensional M-hypercubes, connected in parallel. If needed, these two M-hypercubes could themselves be decomposed into still less complex M-hypercubes, reducing the number of state nodes required per state.

In another arrangement, Lee and Hook combined an M-hypercube with an N-hypercube, resulting in what they call an "MN-cell." Due to its versatility, the device could serve as a building block for designing sequential nano logic gates of any size and complexity.