NASA Contractor Report 198478

Technology Directions for the 21st Century
Volume I

Giles F. Crimi, Henry Verheggen, William McIntosh, and Robert Botta
Science Applications International Corporation
McLean, Virginia

May 1996

Prepared for
Lewis Research Center
Under Contract DAEA32–96–D–0001

National Aeronautics and
Space Administration

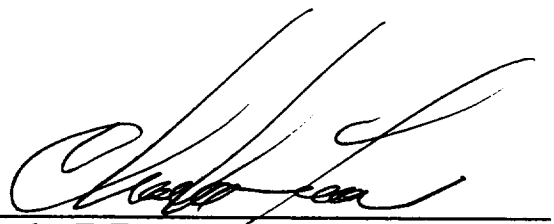
PREFACE

It has been said that the only thing that is constant is change. This adage is well borne out in both the understanding of space and the pursuit of that understanding. In the last twelve months, the scientific studies of our universe, made possible by spectacularly successful flight programs such as the Hubble Space Telescope, Galileo, and Magellan, almost daily have expanded our understanding of the universe and the planets in our solar system. Here on Earth, the space business has also undergone a similar revolutionary process. In the ever complex requirement to balance the pursuit of meaningful science programs with the shrinking availability of funding, NASA has entered the era of having to do more with less. Our focus has shifted to designing faster, cheaper, and better missions without compromising safety or scientific value. Therefore, our technology strategies must correspondingly change to accommodate this new approach. Technology application and systems acquisition must emphasize standardized, multi-use services and operations concepts, leveraged with those of our industrial partners, international counterparts, and academia, as appropriate. Movement toward privatization of space business will be required to better share rewards, exchange technical knowledge, and increase operational efficiencies. These changes are no longer options, but are essential to sustaining our economy and assuring the continued technological leadership of our Nation.

One common thread to the success of our flight programs has been the reliable communication of each bit of information from the spacecraft to the ultimate user. The responsibility for this critical task primarily has rested within NASA's Office of Space Communications (OSC). OSC provides standardized, multi-use communications, data processing, navigation, and mission operation services to both NASA and non-NASA flight programs. NASA's Office of Advanced Concepts and Technology is chartered to develop and advance technology. OSC is a technology user, which adapts existing technology and applications to meet specific program requirements. While OSC's success has greatly benefited from the explosive growth of communications and computing technologies, this new paradigm of "faster-cheaper-better" will require an even greater ability to predict where technology is going.

These reports focus on supporting one element of this new paradigm: the continuing effort to predict and understand technology trends to better plan development strategies. The objectives of this series of documents are to: (1) validate, revise or expand relevant technology forecasts; (2) develop applications strategies to incorporate new technologies into our programs; and, (3) accomplish Agency goals, while stimulating and encouraging development of commercial technology sources. Volumes I and II, together with a future Volume III, summarize several studies relevant to communication and data processing technologies which were selected for their appropriateness to help us meet our challenges.

Comments from interested parties would be especially appreciated, and may be directed to NASA Headquarters, Dr. Albert R. Miller, at (202) 358-4804, or to NASA Lewis Research Center, Ms. Denise S. Ponchak, at (216) 433-3465.



Charles T. Force
Associate Administrator for
Space Communications

ACKNOWLEDGMENT

This work was performed by Science Applications International Corporation (SAIC), Technology Applications Group, in McLean, Virginia for the National Aeronautics and Space Administration (NASA). This report summarizes a series of investigations of various technologies which are important to long range planning for the operational systems and capabilities for which NASA's Office of Space Communications is responsible. An earlier version of this report was published in March, 1993. This current version updates the prior material, and includes some new material for computer processing related technologies.

NASA is pleased to acknowledge the following SAIC personnel who contributed to this effort: Dr. Giles Crimi (Program Manager), Mr. Henry Verheggen (Principal Investigator), Mr. William McIntosh (volume update and Technical Editor), and Dr. Robert Botta (volume update and Technical Editor.)

The technical direction for this work was provided by Ms. Denise S. Ponchak, NASA Lewis Research Center, and Dr. Albert R. Miller, NASA Headquarters, Office of Space Communications.

TABLE OF CONTENTS

	Page
PREFACE	i
ACKNOWLEDGMENT	ii
LIST OF FIGURES	v
FOREWORD	vi
CHAPTER 1 MOORE'S LAW AND SEMICONDUCTOR TECHNOLOGY TRENDS	1-1
SUMMARY	1-1
1. INTRODUCTION	1-1
2. LONG-RANGE TECHNOLOGY TRENDS	1-1
2.1 ANALYSIS OF TRENDS	1-1
2.2 FUTURE APPLICATIONS	1-5
2.3 EMERGING TECHNOLOGIES	1-6
3. RESULTS	1-9
3.1 IMPACT OF MOORE'S LAW AND SEMICONDUCTOR TECHNOLOGY ON NASA	1-9
3.2 TECHNOLOGY ROADMAP	1-10
3.3 FEASIBILITY AND RISK	1-11
CHAPTER 2 COMPUTING TECHNOLOGY TRENDS	2-1
SUMMARY	2-1
1. INTRODUCTION	2-1
2. LONG-RANGE TECHNOLOGY TRENDS	2-2
2.1 ANALYSIS OF TRENDS	2-2
2.2 FUTURE APPLICATIONS	2-7
2.3 EMERGING TECHNOLOGIES	2-9
3. RESULTS	2-9
3.1 IMPACT OF COMPUTING TECHNOLOGY ON NASA	2-9
3.2 TECHNOLOGY ROADMAP	2-12
3.3 FEASIBILITY AND RISK	2-12

TABLE OF CONTENTS (Continued)

Page

CHAPTER 3 STORAGE TECHNOLOGY TRENDS **3-1**

SUMMARY **3-1**

1. INTRODUCTION **3-1**

2. LONG-RANGE TECHNOLOGY TRENDS **3-2**

2.1 ANALYSIS OF TRENDS 3-2
2.2 FUTURE APPLICATIONS 3-5
2.3 EMERGING TECHNOLOGIES 3-5

3. RESULTS **3-8**

3.1 IMPACT OF STORAGE TECHNOLOGY ON NASA 3-8
3.2 TECHNOLOGY ROADMAP 3-9
3.3 FEASIBILITY AND RISK 3-9

CHAPTER 4 PHOTONICS TECHNOLOGY TRENDS **4-1**

SUMMARY **4-1**

1. INTRODUCTION **4-1**

2. LONG-RANGE TECHNOLOGY TRENDS **4-2**

2.1 ANALYSIS OF TRENDS 4-2
2.2 FUTURE APPLICATIONS 4-4
2.3 EMERGING TECHNOLOGIES 4-5

3. RESULTS **4-6**

3.1 IMPACT OF PHOTONICS TECHNOLOGY ON NASA 4-6
3.2 TECHNOLOGY ROADMAP 4-7
3.3 FEASIBILITY AND RISK 4-7

ANNEX 1. ACRONYMS AND ABBREVIATIONS **A1-1**

LIST OF FIGURES

		Page
1-1	Transition Between Technologies - Maintaining Exponential Growth	1-2
1-2	Development of Lithography	1-3
1-3	Evolution of CMOS Feature Size, 1960-2030	1-4
1-4	Submicron Technologies	1-8
2-1	Long-Range Computer Technology History and Projection	2-6
2-2	Integrated Circuit Device Density	2-7
2-3	Comparison of Technology Projections	2-8
3-1	First Law in Disk Density	3-3
3-2	DRAM Density Trend	3-4
3-3	Comparison of Secondary Storage Technologies	3-6
3-4	Hypothetical Effects of Dissimilar Doubling Rates Over a Decade	3-8
4-1	Photonic Device Capabilities	4-2
4-2	Comparison of Secondary Storage Technologies	4-4

FOREWORD

The National Aeronautics and Space Administration (NASA) is a major user of communications and information systems technology. NASA requirements are growing rapidly as missions and instruments become more complex and data processing intensive. NASA information systems typically have long life cycles not only because of the nature of the science missions and the realities of the funding process, but because missions are delayed and expected lifetimes are exceeded. Systems development is complicated by the fast pace of technological change, exemplified by the doubling in performance of microprocessors as quickly as every 18 months, popularly referred to as Moore's Law. Systems may become obsolete early in the life cycle, unless the capability to upgrade is designed in from the beginning. Waiting, for example, 10 years to upgrade a system because of budget constraints may no longer be affordable, since the changes in technology may make it nearly impossible to salvage existing software, a major investment cost. In order to be able to plan for upgrades to major systems, data are needed to project the future performance levels of communications and data processing systems and related components. Ultimately, such data may be used to build a model for technological change that would allow planners to make informed estimates of system cost many years in advance. This would permit a "just-in-time" strategy for technology insertion at the moment in time when it is still at the state of the art.

The Office of Space Communications (OSC) is tasked by NASA to conduct this planning process to meet NASA's science mission and other communications and data processing requirements. A set of technology trend studies was undertaken by Science Applications International Corporation (SAIC) for OSC to identify quantitative data that can be used to predict performance of electronic equipment in the future to assist in the planning process. Only commercially available, off-the-shelf technology was included. For each technology area considered, the current state of the technology is discussed, future applications that could benefit from use of the technology are identified, and likely future developments of the technology are described. The impact of each technology area on NASA operations is presented together with a discussion of the feasibility and risk associated with its development. An approximate timeline is given for the next 15 to 25 years to indicate the anticipated evolution of capabilities within each of the technology areas considered. The steep rate of change projected in some technology areas underscores the need for such data for the purpose of planning information systems. Another beneficial result of the data collection effort is the insight it provides to areas where breakthroughs are likely to occur in key technologies.

The study findings are organized into two volumes. The current volume consists of four chapters covering computers, data storage, and photonics. Two chapters are devoted to computers: one examining the microscopic level of chips and molecular computing; and the other concerned with the macroscopic level, including classes of processors, system architectures, and applications. One chapter each is devoted to data storage and photonics. Volume II also contains four chapters, one each on database systems, computer software, neural and fuzzy systems, and artificial intelligence. Within each volume, the principal study observations are summarized at the beginning of each chapter.

CHAPTER 1. MOORE'S LAW AND SEMICONDUCTOR TECHNOLOGY TRENDS

SUMMARY

For several decades, semiconductor device density and performance have been doubling about every 18 months (Moore's Law). With present photolithography techniques, this rate can continue for only about another 10 years. Continued improvement will need to rely on newer technologies.

Transition from the current micron range for transistor size to the nanometer range will permit Moore's Law to operate well beyond 10 years. The technologies that will enable this extension include:

- Single-electron transistors
- Quantum well devices
- Spin transistors
- Nanotechnology and molecular engineering

Continuation of Moore's Law will rely on huge capital investments for manufacture as well as on new technologies. Much will depend on the fortunes of Intel, the premier chip manufacturer, which, in turn, depend on the development of mass-market applications and volume sales for chips of higher and higher density. The technology drivers are seen by different forecasters to include video/multimedia applications, digital signal processing, and business automation.

Moore's Law will affect NASA in the areas of communications and space technology by reducing size and power requirements for data processing and data fusion functions to be performed onboard spacecraft. In addition, NASA will have the opportunity to be a pioneering contributor to nanotechnology research without incurring huge expenses.

1. INTRODUCTION

Moore's Law, coined in 1965 by Gordon Moore, the current chairman of Intel Corporation, states that semiconductor device integration improves at a rate of doubling every 18 months. It is arguable that this formula captures the most important technological and economic phenomenon of our time. Computer and electronics technology is one of the driving forces in the US economy, if not the world. Moore's Law also forms the basis of any quantitative model of technological change. The rapid pace of change shown by Moore's Law is a strong argument for the need for planning for change. Therefore it is critical that those in a position of policy-making and planning understand the workings and implications of Moore's Law. There is a need to find out how far Moore's Law can be extrapolated, and how far it can be relied on for planning.

2. LONG-RANGE TECHNOLOGY TRENDS

2.1 ANALYSIS OF TRENDS

Moore's Law seems to depict a phenomenon with indefinite exponential growth, but is this an illusion? John Walker of Autodesk recently gave a talk on this subject. He stated that for a given technology, the rate of improvement is actually not exponential. The true curve is S-shaped. The first part of the S-curve approximates an exponential curve, but then levels out. That is, a technology starts improving slowly, then accelerates, and finally reaches maturity, at which point no further improvement is possible. When maturity is reached, a new technology must replace the old one. Exponential growth can be effectively maintained if the transition to the new technology is smooth, as illustrated in Figure 1-1.

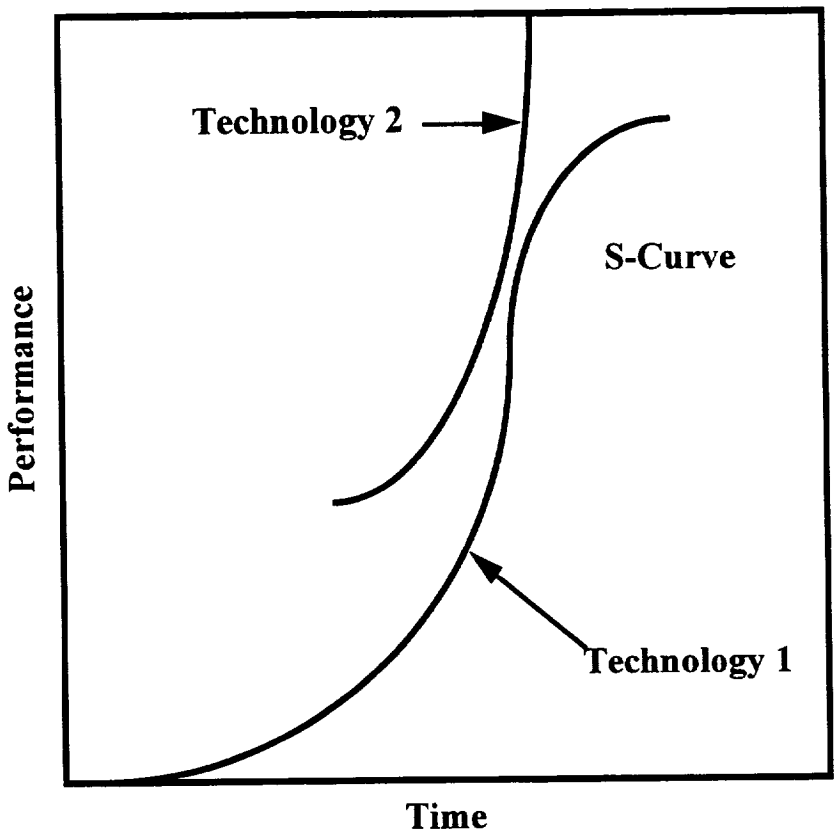


Figure 1-1. Transition Between Technologies - Maintaining Exponential Growth

Moore's Law has accurately characterized technology evolution over a period of several decades now. However, the scaling down of metal-oxide semiconductor (MOS) devices must inevitably be limited. For example, if Moore's Law is extrapolated to 2030, a single logic gate will be the size of a molecule [1]. But molecular-scale logic cannot be built using photolithography and MOS transistors. Current estimates are that MOS transistors can be shrunk to the 0.1 micron level while staying within the photolithographic domain. This will take Moore's Law into the next decade [2]. Actual photolithographic technology life cycles are shown in Figure 1-2. In a recent interview, Moore also expressed the opinion that the improvement rate will eventually slow [3]. Therefore, a new technology will be required for the transition from 1 micrometer to 1 nanometer. That new technology will involve engineering at the molecular scale.

Current technology is exemplified by Intel's 0.35 micron process that is being used for both late-1995 Pentiums and the first version of the P6. (The P6 will have twice the computing power of the Pentium when built with the same process.) [4,5] Another example is Hitachi's 0.35 micron process which will allow the design of a custom 5-million gate, 1280-pin chip. A third example is Digital's Alpha 21164, which has a clock rate of 300MHz, and perform 1200 million instructions per second (MIPS) [6].

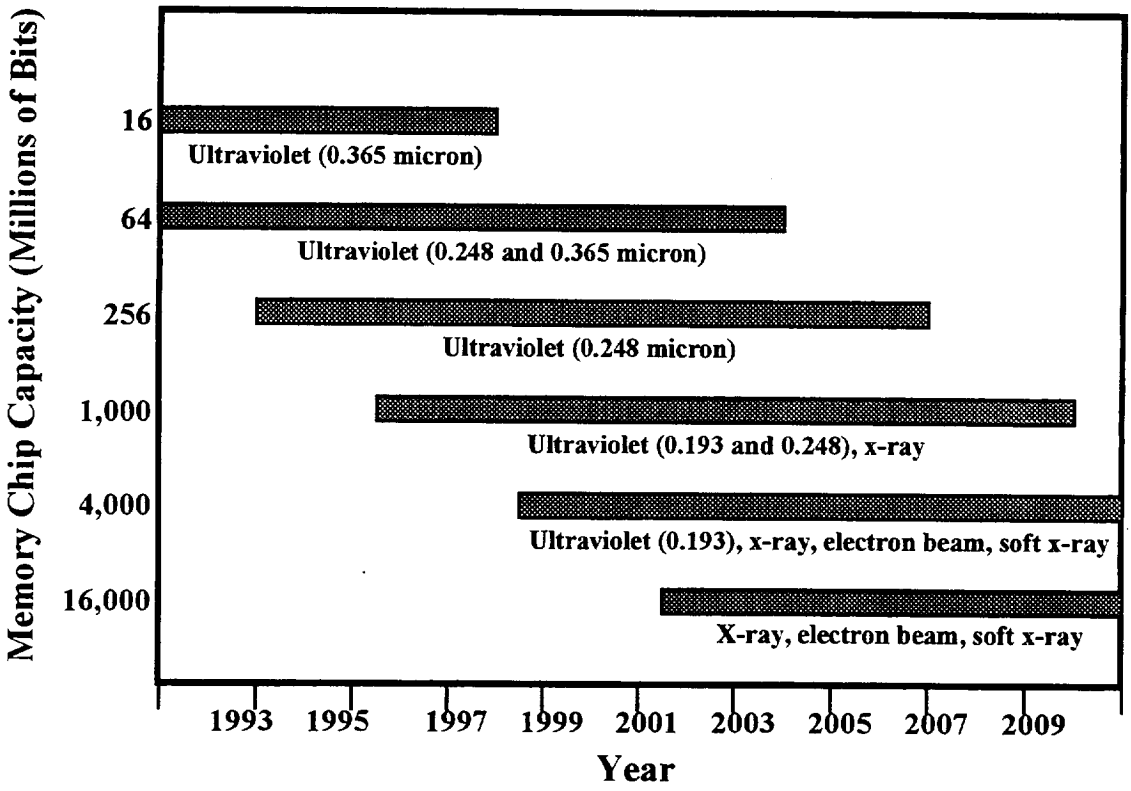


Figure 1-2. Development of Lithography

Source: Scientific American, February, 1995, p. 92.

J.D. Meindl has recently given a detailed analysis and projection of future complementary MOS (CMOS) technology [7]. Meindl's projection of Moore's Law is reproduced in Figure 1-3. Meindl considers several categories of limiting factors on the future improvement of CMOS. The kinds of limits associated with device density are:

- Fundamental limits, associated with the laws of physics independent of material and device design. These are:
  - A thermodynamic limit, the minimum usable signal power due to thermal noise
  - A quantum mechanical limit, the minimum signal power according to Heisenberg's uncertainty principle
  - The speed of light limit for signal propagation
- Material limits, associated with the material properties, such as thermal conductivity and carrier mobility.
- Device limits, associated with the device type. Specifically, for the MOS field-effect transistor, the critical limit is the minimum effective channel length.
- Circuit limits, associated with the logic gate design, such as the circuit gain with respect to threshold voltage and equivalent circuit capacitances.

- System limits, associated with the overall chip architecture, the heat removal mechanism, and many others.
- Practical limits, primarily the limits of the manufacturing technique and the return on capital investment in the manufacturing plant.

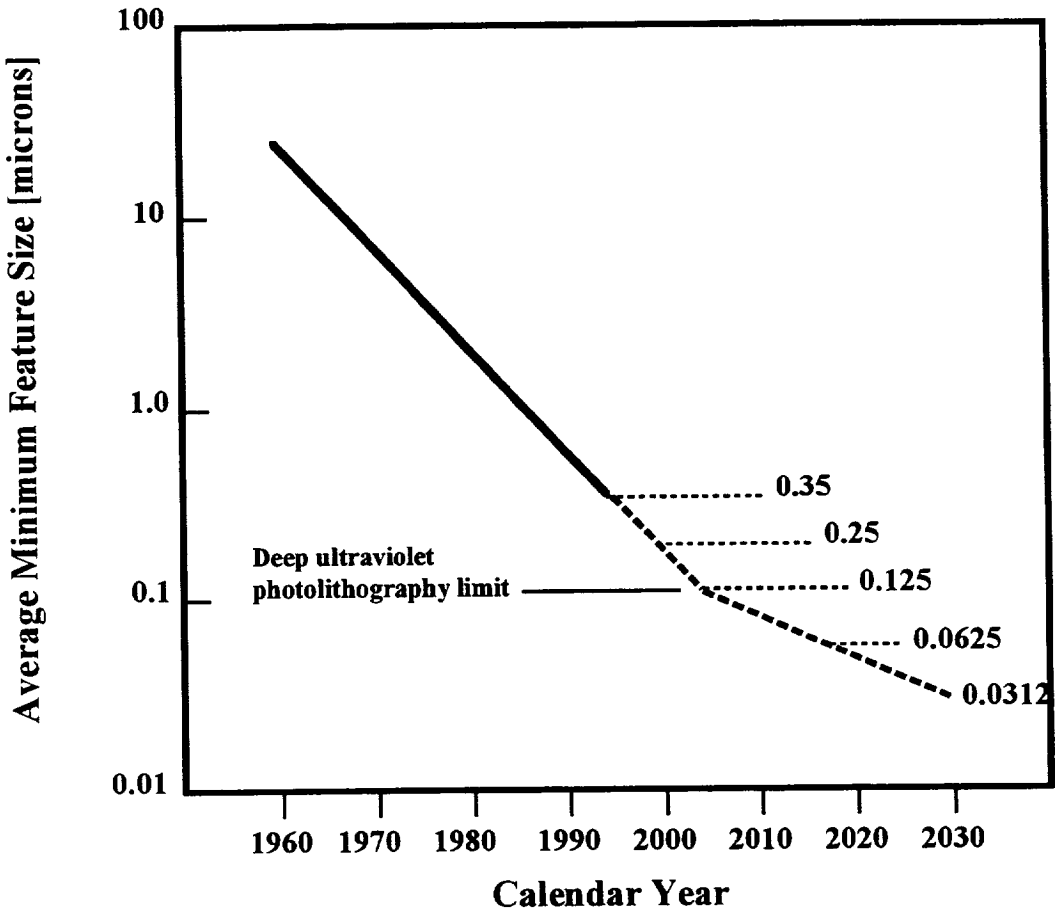


Figure 1-3. Evolution of CMOS Feature Size, 1960 - 2030

If these limits are plotted in the parameter space, the results show that the severity of constraint is in the order listed above. The most limiting parameter is the wavelength of the light used in the photolithographic process. For deep-ultraviolet light it is about 0.1 micron. Return on investment is also a severe constraint, since the cost of photolithographic manufacturing follows Rock's Law (see below). Huge expansions in the size of the market must be achieved to profitably deploy new generations of photolithography and X-ray lithography manufacturing below the present 0.35 microns indicated in Figure 1-3.

Looking beyond the looming crisis in photolithography, there is yet "plenty of room at the bottom", as Richard Feynman's famous article says [8]. There are many new device concepts that point the way toward higher device densities in the 21st century. Some newer technologies are quantum effect transistors, single-electron transistors, spin transistors, molecular switches, and

nanotechnology, to be discussed below. Nanotechnology, or molecular-scale engineering, is touted as the eventual path to the nanometer domain.

Not all exotic technologies will see smooth sailing, however. Many non-linear switching devices have been tried in the laboratory, and have been found deficient, especially with regard to gain. Transistors will probably continue to be the device of choice for as long as possible, because of their high gain [9].

Ultimately, the continuation of Moore's Law depends on economic factors such as the growth of population, the growth of per capita wealth, the needs of the mass market, and the point of diminishing return on investment. The cost for each new design supposedly follows "Rock's Law", which states that halving the device feature size requires a 5-fold increase in manufacturing cost [3, 10]. Therefore, in order to raise the capital needed, the market for computers must continue to grow. However, there is no obvious reason why the mass consumer market should continue to require ever-greater performance. Market growth depends on the ingenuity of the industry in conceiving new attractively priced applications that the consumer will be induced to buy.

The economics of Moore's Law seems to depend largely on the fortunes of one company, Intel. The quasi-monopoly that Intel holds seems to be a key ingredient for success. Monopoly is a feature of a new theory of economics by Paul Romer [11]. Although Romer is a free-market economist his new theory paradoxically shows the importance of monopoly and "barrier to entry" for the growth of technology, because of the huge amounts of capital needed. Romer sees technology as a fundamental force in economics, rather than being a merely serendipitous effect, as it is treated in standard economic theories. Perhaps it is the law-like nature of semiconductor improvement that is forcing economists to adopt this view.

Since mass-market applications and volume sales will be the only way to pay for the huge capital investments in mass-production chip factories, the processors used in these applications will no longer lag behind the state-of-the-art generation processors, but will be those processors [12]. The consumer market will henceforth drive computer technology, assuming attractive pricing is achievable.

Another driving force for Moore's Law is business automation. The U.S. economy is seeing a wave of automation in the wake of recent corporate downsizing. There are enormous markets for technology in the automation area, and the degree of automation is relatively simple. That is to say, it doesn't require sophisticated artificial intelligence. But in the future, progress in automation will depend on progress in artificial intelligence. This field appears to have its own unique limitations that will limit the application of computer technology, as discussed below.

2.2 FUTURE APPLICATIONS

Intel foresees video/multimedia as the driving application for increased mass consumption of higher-MIPS central processing unit (CPU) chips. "Multimedia" applies to the application domain and to the communication domain. Applications must be able to display multiple media types, of which video is the most computationally intensive. They must also be able to search and retrieve databases with multiple associated file types including video files. In the Intel scheme, the CPU must also handle the coding/decoding, compression/decompression, and protocols for communications of high bandwidth media, including video. This is called 'native signal processing'. Therefore, Intel is seeking to make its future market from communications functions.

Texas Instruments (TI) views things differently, since its specialty is digital signal processing (DSP). TI's new TMS320C80 DSP chip has 2 billion operations per second performance [13] for computationally intensive applications such as real-time video compression and decompression. TI's argument is that such specialized chips will always stay well ahead of general purpose CPUs. Therefore, the bandwidth-intensive signal processing should be done outside the CPU. TI's

strategy is also oriented toward the emerging markets. For these markets, it will focus on 486 technology, which it has licensed from Intel.

Some futurists such as George Gilder foresee yet a different scenario [14]. Gilder maintains that since the bandwidth capacity of communications is improving faster than Moore's Law, eventually bandwidth available at the desktop will be more than sufficient, so that the need for sophisticated compression, decompression, and other signal processing will be reduced. Also, the need for storage may be reduced, since everything could be retrieved from remote databases as quickly as from local storage. Gilder believes that bandwidth competes against MIPS. At the current time, the infrastructure is not available to have the impact that Gilder foresees.

Automation is a key application of higher computing power for NASA. However, the likely future rate of advance for artificial intelligence is much slower than the available hardware computing power. This is not because higher MIPS will not bring new capabilities, such as faster searches and larger neural networks. Certain kinds of real-time neural networks are beyond the capabilities of current hardware, and will only be implementable when hardware has improved. Rather, the problem is a fundamental impasse in the theory of artificial intelligence: no one knows how intelligence works. While it is clear that some activities of humans and animals are susceptible to computer simulation, it is not clear that computers, either sequential or parallel, are the best model for what it is that humans and animals do in general. This impasse could eventually slow the need for higher-power computers by closing off currently envisaged applications that require higher power, by restricting the kinds of automation that can be achieved. This may not be an undesirable outcome, since it would mean that there would remain a need for human beings!

2.3 EMERGING TECHNOLOGIES

2.3.1 Single-Electron Transistors

A type of transistor has been developed that can reliably operate using current flow of a single electron (electron charge quantization) [15]. Whereas standard transistors are degraded by quantum mechanical tunneling, this device uses tunneling as the basis of its operation. The single-electron transistor is capable of using a single electron to represent one bit of information. The single-electron flow phenomenon is clear-cut down to about 10 nanometers (0.01 micron) of device scale. At this scale, an improvement of a factor of 1000 in device density would be achieved. Below 10 nanometers, the energy levels of the electrons become quantized because of the electron's confinement. The effects of charge quantization and energy quantization are mixed at this scale. This does not preclude the single-electron flow phenomenon however. The physics of these smaller devices is only now being explored. While single-electron transistors are a clear candidate for future improved integrated circuit density, the manufacturing problems for under 0.1 microns are the main obstacle to mass production.

2.3.2 Quantum Well (QW) Devices

QW devices use potential energy wells to confine electrons to fewer than the normal three dimensions [16]. Semiconductors have energy levels determined by the material properties. QW devices have energy levels defined by the geometry of the device. The first QWs were fabricated as thin films, with the electrons confined to a plane (two-dimensional confinement). Transistors (high electron mobility transistors) and laser diodes built with this approach have become commonplace. More recent research has focused on quantum wires (one-dimensional confinement) and quantum dots (zero-dimensional confinement). Quantum dots behave like custom-tailored atoms, since the energy levels are determined by geometry. These devices do not have the physical limitations that ordinary transistors do when scaled to 100 or even 10 nanometers. As with other novel approaches, however, fabrication is the major problem.

2.3.3 Spin Transistors

The spin transistor is an all-metal transistor that operates on the basis of the flow of polarized electrons [17]. In a semiconductor, two separate populations of charge carriers (electrons and holes) are sustained due to the fact that they do not annihilate each other instantaneously. They have a finite lifetime before recombination. Similarly, it is possible to sustain two distinct populations of charge carrier in a metal, as counterintuitive as this may seem. The two populations are spin-up and spin-down electrons, and by creating non-uniform spatial distributions of these populations, an electric field arises. This can be used to build an active device.

Spin transistors scale very well to a size of 100 nanometers. This would correspond to a packing density of 100 times that of the current state of the art. The switching times of the spin transistor are about ten times slower than silicon field effect transistors, however, but have greatly reduced power dissipation. Spin transistors have an advantage when scaling to 10 nanometers, namely, several orders of magnitude greater charge carrier density. This is important, since at this scale semiconductors have only a few charge carriers per active junction.

The most advantageous near-term application is for nonvolatile random access memories. Nonvolatile storage is accomplished by indefinite storage of magnetic states.

2.3.4 Optical Computing

Are photons better than electrons for the purposes of higher computing performance? Photonic devices depend on the same photolithographic technology that semiconductor devices do. Furthermore, the input and output arrays for photonic computers cannot in principle have feature sizes smaller than the wavelength of light being used by the device. So the scaling limits on photonic devices follow the limits of photolithographically manufactured devices in general. However, there is an advantage for massively parallel architectures, since photons can be propagated through free space in many parallel channels to achieve massive interconnectivity.

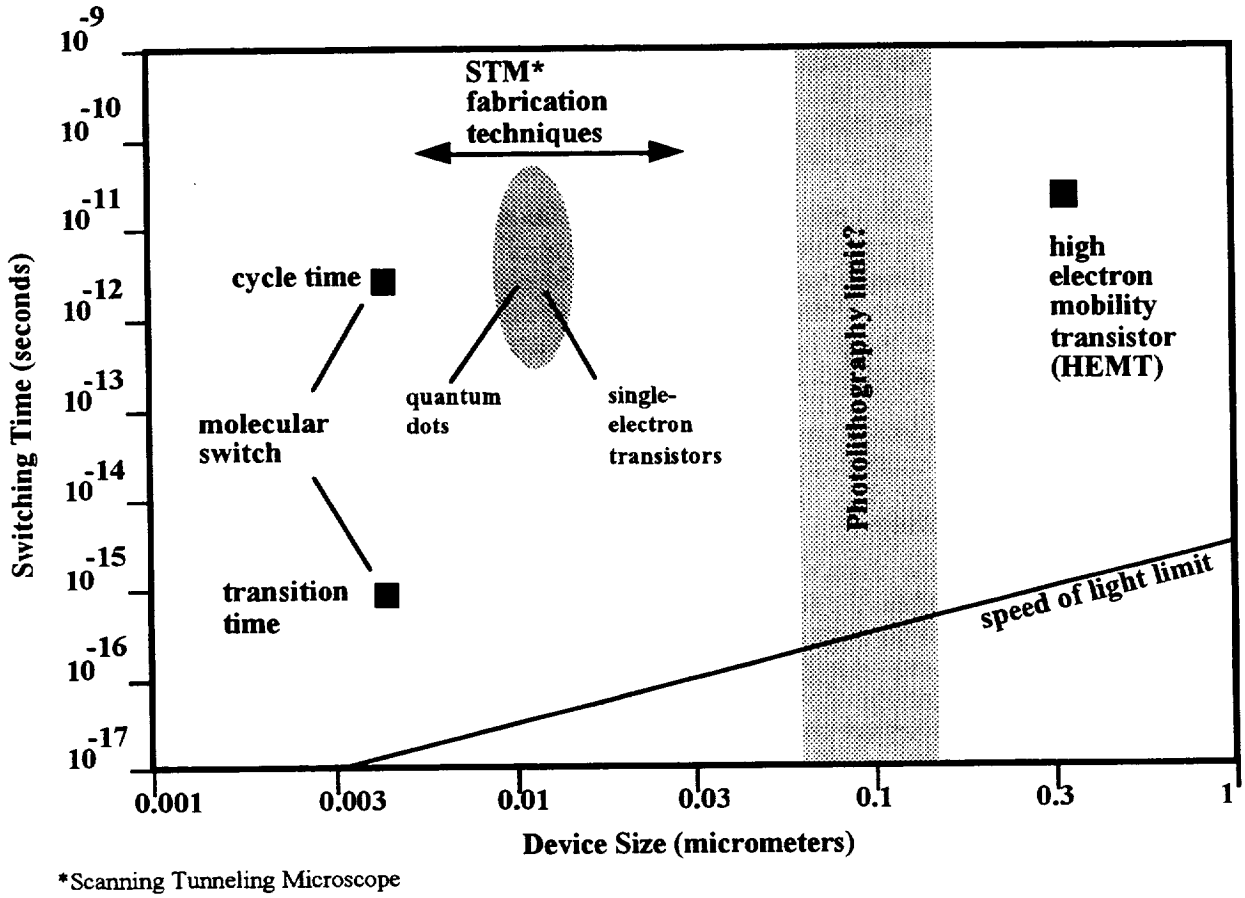
2.3.5 Nanotechnology and Molecular Engineering

The term nanotechnology refers to a range of innovative concepts involving fabrication of devices at the scale of 1 - 100 nanometers. This includes electronic devices, electromechanical devices, and organic molecular devices. Recently, a great deal of interest has been expressed in these ideas for space applications because of the potential for improvements in both electronic and materials aspects of space system design.

Rock's Law presents a major problem for the future of technology. One way around it has been proposed: self-manufacturing devices. This concept is based on von Neumann's theory of self-replicating machines. Fabrication of arrays of molecular structures at the nanometer scale may require this sort of self-replication. Since self-replication is an exponential phenomenon, the cost of mass production will be low. Also, the costly fine tolerances of present-day IC fabrication equipment would not be needed, since the molecular processes would have built-in reliability and precision of assembly based on their molecular structure.

Another way of accomplishing mass production at the nanometer scale, beyond photolithography, is to build large arrays of micromechanical assemblers such as the scanning tunneling microscope (STM). Such arrays of STMs can be built photolithographically. A single silicon substrate could contain thousands of STMs. Each STM in turn would manufacture thousands of nanometer-scale devices by means of molecule-by-molecule assembly. In this way, photolithography could be used to bootstrap beyond itself. STMs have been used to manipulate individual atoms, and at Cornell, an array of micro-STMs has been built [18]. The Cornell array consists of micromachine STMs, each one of which is supposed to manipulate individual atoms within its domain. With thousands of STMs operating in parallel, and at the high speed allowed by

their small size, a mass storage device could be built in which individual bits would be represented by individual atoms. A summary of advanced technologies is illustrated in Figure 1-4.



*Scanning Tunneling Microscope

Figure 1-4. Submicron Technologies

2.3.6 Quantum Computation

There are several senses in which quantum mechanics comes into play in computational devices. In one sense, all semiconductor transistors are quantum devices, since they are based on solid state physics, which is quantum mechanical. In another sense, a quantum computational device is simply a logic gate that is so small that its operation must take into account quantum effects such as source-to-drain tunneling. The single-electron transistor is another example, in which several quantum effects are found to coexist, but which can be harnessed to good effect. However, all of these devices are used to build computers that are mathematically equivalent to a Turing machine, which is a classical-physics machine. There is a more radical sense of quantum computation, such as in a system that uses quantum behavior to achieve computations that cannot be performed by a Turing machine.

A Turing machine has states which are definite and singular, whereas a quantum computer could have a complex superposition of many states (complex in the sense of complex numbers). This is because the quantum state of a system propagates according to the Schrödinger equation as a complex linear superposition of states which are possible solutions of the equation, until a measurement or other outside influence destroys the coherence of the system. This property will not allow quantum computers to compute faster than classical computers, nor will it allow the

computation of functions that classical computers cannot calculate. However, it will perform one type of computational task by a factor arbitrarily faster than a classical computer. This is the type of computation in which a large number of independent function evaluations f(0), f(1), f(2),...f(N) must be done, but the resulting series of values is not part of the answer. Only a distilled answer such as "there are not exactly N odd values" is required as the output. A quantum computer can evaluate all the f(•) values simultaneously. Furthermore N can be arbitrarily large [19, 20].

What could such a computer usefully do? One recently identified task is factoring of large primes. Thus, if anyone knew how to build a quantum computer, it might be possible to break cryptographic systems that use a public-key design [21]. Other practical applications will undoubtedly soon be discovered.

3. RESULTS

Moore's Law can be extrapolated for about 10 more years. Beyond 10 years deep ultraviolet photolithography will likely falter around 0.1 microns. Thus, within 10 years a new manufacturing technology must begin to be deployed in order to sustain Moore's Law. It is too early to say what that technology will look like, other than that ultimately it will involve molecular engineering. There may be an intermediate technology based on X-ray lithography, or electron beam fabrication. The development of an X-ray lithography manufacturing infrastructure will be enormously expensive. Likewise, a nanotechnology infrastructure will require a vast amount of additional research and development, but there can be little doubt that it will happen. There is also little doubt that Moore's Law will continue to apply beyond 10 years.

3.1 IMPACT OF MOORE'S LAW ON NASA

3.1.1 Communications

The communications technology trend is toward bandwidth-on-demand. This goal permits the optimum allocation of resources. In a sense, this is a strategy which minimizes infrastructure in exchange for computational (protocol) complexity. This strategy is very general and has broad applications. For example, it also applies to radio spectrum reuse. Through the use of advanced semiconductor technology, radio spectrum is used only when it is needed, and is otherwise available to other users.

The capacity of fiber-optic technology has been growing at a rate faster than Moore's Law. This could have an impact on computer architectures, as George Gilder believes. On the other hand, the bandwidth available at the desktop has not expanded nearly as much as is technically feasible. This is because the installation of the appropriate communications infrastructure has lagged, because of its enormous cost.

Many third world countries are using wireless technology to reduce the cost of local telecommunications infrastructure. But can wireless technologies compete with wireline approaches for a high-bandwidth infrastructure? The limits on radio spectrum would seem to rule out wireless as a solution to the "last-mile" infrastructure problem when it comes to high bandwidth. But Moore's Law provides some mitigation of spectrum scarcity, in ways that would not have been obvious in the past. Limited spectrum becomes less of a barrier to expansion since low-cost radios can be proliferated in microcellular networks. Software radios will dramatically bring down the cost of cellular base stations. The concept of a software radio is a classic trade of hardware for software.

3.1.2 Space Technology

The exponential rate of Moore's Law means that there is a high "opportunity cost" to building large, expensive space systems with long development cycles, wherein, by the time a project is fielded, the technology it uses is obsolete. The more current the technology is, the more can be accomplished by a project within a given budget. There are two alternative strategies: simply waiting, or reducing the development cycle time. By simply waiting, more advanced technology will

become available relatively quickly. However, this is not a very realistic alternative, since logically, one would always be waiting. The only reasonable strategy is to reduce the development cycle. Thus, there is an opportunity cost incentive to reduce project complexity, and to deploy designs with short development cycles.

Since the microelectronics industry is giving us enormously powerful computers in tiny packages, there is an incentive to translate spacecraft functions that traditionally required bulky hardware into computational functions. Such a translation does not always exist. For example sensor functions don't translate to computation. Point sensors can be easily miniaturized, but area sensors cannot. Imaging optics are limited by the Rayleigh criterion and received signal level requirements. Increasing the resolution or signal level requires increasing the optics diameter. Similar considerations hold for radio antennas. On the other hand, software radios are an example of how functions that required a lot of bulky analog components can now be reduced to computational functions. This is a fruitful new area for exploitation.

More sophisticated spectrum management (a computation/protocol problem) can also improve available radio spectrum. For example, if all spacecraft in low earth orbit shared spectrum, each could use the spectrum of others when the others are not using it or are not within interference range. This would require not only spectrally agile radios, but radios that maintain databases on each other's locations, such as through the exchange of global positioning system data in a global network.

3.1.3 Research and Development

Photolithographic and X-ray lithographic manufacturing technology are capital-intensive endeavors which will continue to be the exclusive domain of the major semiconductor businesses. Electron beam lithography is a continuing viable method for small-quantity, research-oriented fabrication of state-of-the-art circuitry. However, nanotechnology appears to be a field in which a research institution such as NASA can be a pioneer, with relatively low investment. There is still much "low-hanging fruit" in this broad-ranging field. For example, a scanning tunneling microscope can be built by a graduate student for a few thousand dollars. Photolithography will continue to be important for manufacturing micromachines that may be used to build structures at an even smaller scale. The Advanced Research Projects Agency is currently setting up a low-cost prototyping service for micromachine manufacture. This will lower the cost of research in this area.

In the long-range future, as nanotechnology matures, there may be a synergy between research on electronic devices and mechanical devices. Nanotechnology could generate innovations in computers, materials, and all forms of manufacturing.

Likewise, quantum computing is a new field that NASA could benefit from. At this time, quantum computation has only been theorized. No quantum computers have yet been built, but there are several research groups pursuing its development.

3.2 TECHNOLOGY ROADMAP

A technology roadmap, or projected timeline, for semiconductors device development over the next 25 years follows.

1996-2000: CMOS transistor technology will continue to be standard. Nanotechnology will be in an early research stage.

2000-2010: CMOS technology will continue on the path of Moore's Law, but a new manufacturing technology, perhaps X-ray lithography, will be deployed. Nanotechnology will begin to show practical results. Perhaps a usable molecular mechanical computer, such as theorized by Eric Drexler [22], will be built.

2010-2020: CMOS technology may be replaced by one of many possible nanotechnologies, including mechanical, electronic, or bio-chemical techniques.

3.3 FEASIBILITY AND RISK

There is a consensus that Moore's Law applied to CMOS transistor technology, can be confidently extrapolated to 2005-2010. There is also a consensus, based on the laws of physics, that photolithographic manufacturing techniques will falter. There is also a question as to whether the mass market can continue to provide the required capital investments required for future technology growth. Although the amount of investment required to invent the new manufacturing technologies of the 21st century is huge, there is clearly a role for smaller R&D projects in advanced topics such as nanotechnology.

References

1. Birge, Robert, "Protein-Based Computers", *Scientific American*, March 1995, pp. 90-95.
2. Lenzner, Robert, "Whither Moore's Law?", *Forbes Magazine*, Sept. 11, 1995, pp. 167-168.
3. Lipman, Jim, "Submicron EDA tools help tackle tough designs", *EDN*, June 8, 1995, pp. 44-62.
4. Clark, Don, "A Big Bet Made Intel What It Is Today; Now It Wagers Again", *Wall Street Journal*, June 7, 1995, p. A1.
5. Intel World Wide Web Home Page
6. Geppert, Linda, "Solid State", *IEEE Spectrum*, January, 1995, pp. 35-39.
7. Meindl, James D., "Low Power Microelectronics: Retrospect and Prospect", *Proceedings of the IEEE*, Vol. 83, No 4, April 1995, pp. 619-635.
8. Feynman, Richard, "There's Plenty of Room at the Bottom", *IEEE Journal of Microelectromechanical Systems*, Vol. 1, No. 1, March 1992.
9. Keyes, Robert W., "The Future of the Transistor", *Scientific American*, June 1993, pp. 70-78.
10. Stix, Gary, "The Wall", "Science and Business", *Scientific American*, July 1994.
11. Robinson, Peter, "Paul Romer", *Forbes ASAP*, June 5, 1995, pp. 67-72.
12. Stork, Johannes M.C., "Technology Leverage for Ultra-low Power Information Systems", *Proceedings of the IEEE*, April 1995, Vol. 83, No. 4, pp. 607-618.
13. Texas Instruments World Wide Web Home Page
14. Gilder, George, "Telecosm: The Bandwidth Tidal Wave", *Forbes ASAP*, December, 5, 1994, pp. 163-177.
15. Likharev, K.K., and Tord Claeson, "Single Electronics", *Scientific American*, June 1992, pp. 80-85.
16. Reed, Mark A., "Quantum Dots", *Scientific American*, Jan. 1993, pp. 118-123.
17. Johnson, Mark, "The All-Metal Spin Transistor", *IEEE Spectrum*, May 1994, pp. 47-51.

18. "Scoping for Data", *Popular Science*, August 1995, p. 34

19. Deutsch, D., "Quantum Computation", *Physics World*, 5, pp. 57-61, 1992.

20. Deutsch, D. and R. Jozsa, "Rapid solution of problems by quantum computation", *Proc. Royal Society of London*, A439, pp. 553-558, 1992.

21. Deutsch, D., and A. Ekert, "Quantum communication moves into the unknown", *Physics World*, 6, pp. 22-23, 1993.

22. Drexler, Eric, Foresite Institute, World Wide Web Home Page.

Bibliography

Corcoran, Elizabeth, "Diminishing Dimensions", *Scientific American*, Nov. 1990, pp. 122-131.

Horgan, John, "Gravity quantized? A radical theory of gravity weaves space from tiny loops", *Scientific American*, Sept. 1992.

Lewis, Ted, "Where is computing headed?", *IEEE Computer Magazine*, August 1994.

Stix, Gary, "Toward Point One", *Scientific American*, Feb. 1995, pp. 90-95.

"What cognition might be, if not computation", The Journal of Philosophy, XCII, No. 7, July, 1995, pp. 345-381.

CHAPTER 2. COMPUTING TECHNOLOGY TRENDS

SUMMARY

A unique characteristic of computing technology is the fast rate of change, with central processing unit (CPU) performance doubling every 18 months. With specialized accelerators, desktop computers can achieve near supercomputer performance for particular applications.

The current trend to replace sequential computing with parallel computing complicates the system design process, since it will take time for the new industry to stabilize, for a supporting software culture to develop, and for standards for open systems to be developed. The near term trend toward the symmetric multi-processor (SMP) allows increased parallelism while maintaining the simpler sequential processing model.

Waiting 10 years to upgrade a computer system may no longer be realistic. Pre-planned upgrades must become part of the system life-cycle design and budget process.

The conversion of systems to a client-server architecture has significant life-cycle cost benefits. This trend extends itself through an increase in bandwidth availability that will allow for further distribution of computer resources that include the Internet centered computer.

For applications such as pattern recognition and image processing, parallel, connection-intensive architectures are seen as the path to orders-of-magnitude gains over today's high-end computers and as a bridge to advanced technologies, such as optical and molecular computing.

1. INTRODUCTION

The performance advantage of emitter-coupled logic (ECL) technology has historically been the justification for the higher price of mainframes and minicomputers compared with the price of microcomputers and workstations. But in recent years, the performance of complementary metal oxide semiconductor (CMOS) technology has been rapidly overtaking the domain of ECL while being significantly lower in cost. For example, standard Pentium microprocessors can be run at 250 megahertz (MHz) clock rates under certain conditions. Digital Equipment Corporation's (DEC's) Alpha reduced instruction set computing (RISC) processor is a 300 MHz CMOS chip. The result is that there has been a trend toward replacing minis and mainframes with microcomputers and workstations, a process sometimes called downsizing. At the same time, the low cost of high-performance CMOS processor chips is the driving force behind the move toward parallel architectures for higher-performance computers. It is in principle easier to achieve high levels of performance by using multiple parallel processors than by pushing the envelope of integrated circuit device density. (A new integrated circuit fabrication line can cost as much as $1 billion.) These developments all point to the fact that the dominant trend for the rest of the 1990s will be parallelism using CMOS.

NASA, as a major user of computers of all classes, will be affected by this trend. The long life cycle of many NASA systems makes it imperative that NASA maintain a current, quantitative picture of the long-range trends in computer performance for the various computer classes and maintain an awareness of new technologies on the horizon. This report attempts to bring together some of this kind of data to assist in the planning process.

Although this chapter covers primarily commercial off-the-shelf technology, in some cases no mass market has yet developed. In particular, the advanced parallel computers that form the basis for emerging trends have been funded mainly by Advanced Research Projects Agency (ARPA), while commercial funding is lacking. Many agencies have contributed to the ARPA funding, primarily the Department of Defense (DOD), NASA, the Department of Energy (DOE), and the intelligence agencies. Advanced technologies, such as optical and molecular computing, are

researched at a broad range of organizations, including universities, corporations, and research consortia. In Japan, there is a strong research program in molecular electronics. In the commercial world, use of parallel computers has been limited, although recently, some large retail businesses have begun to purchase them for database applications.

2. LONG-RANGE TECHNOLOGY TRENDS

A defining trend in computing in the 1990s is parallelism, at the chip level, the architecture level, and at the level of internetted architectures. The most significant trend in computer applications will be the proliferation of parallel machines in business applications. This trend will be driven by price-performance ratio, since this architectural approach is more cost competitive than traditional architectures. The older distinctions between computer classes have been replaced by the general category of "server", a scalable parallel computer. A sampling of server vendor catalogs reveals that most product lines consist of a low-end single-CPU offering, followed by a series of higher-performance multiple-processor offerings, with typically from two to twelve processors.

A negative aspect of the trend toward parallelism is that a new software approach is required just at a time when industry is struggling to adopt open systems standards for traditional architectures. Parallelism may bring a new phase of confusion and lack of standards. However, a shakeout in the parallel computing market continues through the 1990s, and the great diversity in commercial parallel architectures continues its rapid evolution. The availability and cost of software appears to be a major driving force in the success of these architectures. SMP machines have made gains in the commercial market because of their simpler, sequential, processing model. The massively parallel processor (MPP) lags in usability because of the difficulty and cost of software development for the architecture. Through the rest of the decade, the MPP will deliver the highest processing performance at a premium price while SMP computers increase their capabilities. The MPP architecture will overtake SMP as the software issues associated with this architecture are resolved.

Beyond the year 2000, we anticipate that the trend in improvements for integrated circuit technology will continue, enabled by new electronic devices such as quantum well devices and single-electron transistors. We also expect that photonic and molecular scale devices will play a much more visible role in high-end computing applications, spurred by connection-intensive parallel architectures, such as neural networks.

2.1 ANALYSIS OF TRENDS

2.1.1 High-End Machine Technology

Information for this section was drawn primarily from Kelly [1] and recent Intel data [2]. In the mid- to high-price classes, three parallel architecture types predominate:

- Multiple instruction, multiple data (MIMD)/shared memory: This model includes the SMP architecture. It works best when requirements for passing of data and synchronization are minimal. Easy to program, hard to build.
- MIMD/distributed memory: Easy to build, hard to program. Good for less regular problems.
- Single instruction, multiple data (SIMD): Works well for large, regular, data-parallel problems. Uses simple, inexpensive processors.

The following are examples of systems representative of these architectures:

- MIMD/shared memory
  - SEQUENT—2–30 processors, 32-bit, 100 MHz, Pentium, 200 million to 3 billion instructions per second, $50,000 to $2.5 million.

  - T3E Cray MPP—Up to 2048 of DEC Alpha microprocessors (64-bit, 600 million instructions per second [MIPS] sustained), 1.2 trillion floating point operations per second (tera-FLOPS)estimated peak performance.
- MIMD/distributed memory
  - Intel Touchstone SIGMA—2048 processors (i860), 128 Mbytes of memory each, 150 billion FLOPs (giga-FLOPS) [2].
  - NCube—8–8192 processor, custom 64-bit very large scale integration (VLSI), 4 to 16 megabytes of memory each, 8192 processors at 27 giga-FLOPS
- SIMD
  - Thinking Machines CM-5—32 processors, 128 mega-FLOPS, RISC with 4 vector pipelines, 4 giga-FLOPS, $1.4 million scalable to 16000 processors
  - MASPAR—1024 to 16384 processors, 32-bit, 1024 at 75 mega-FLOPS, $110,000, 16384 processors at 1.2 giga-FLOPS, $1.3 million.

It is estimated that sustained 1 trillion FLOPS performance will be achieved before the end of the decade in the price range of $40 million to $70 million. Both Cray and Intel have announced machines that are projected to deliver on this goal. Performance of 250 giga-FLOP is available today from the Intel Paragon supercomputer. The tera-OP barrier should fall to a new Intel machine with over 9000 P6 processors which is funded by DOE for nuclear simulation and is expected to be deployed late in 1996. Gordon Bell, the former vice president for research and development at DEC, questions the usefulness of the government's quest for this goal, however [3]. He believes that the government is focusing too narrowly and too quickly on computing power and not enough on allowing the supporting user and programmer culture to mature.

Performance prediction is a big issue. The performance of sequential machines can vary by 2 or 3 times, perhaps as much as 10 times, between predicted and actual performance. In the case of parallel machines, the difference between predicted and actual performance can be as much as a factor of 100 to 1,000 because of the problem of parallelization. The performance is much more dependent on the nature of the data and the algorithm and the degree to which they can be partitioned for efficient parallel execution. The problems of evaluating vendor claims and of system integrators in accurately pricing their bids are greatly increased. The success of the effort to migrate parallel computers to the wider business market will depend on the success in developing automatic parallelizing compilers to facilitate the transfer of standard applications to the new architecture. (Parallelizing compilers automatically convert sequential programs for running on massively parallel computers.)

2.1.2 Symmetric Multi-Processor

Large scale business applications have adopted SMP architectures because of the price performance advantages of using multiple microprocessor chips to deliver the processing power of a mainframe. The programming model for this architecture maintains the illusion of a sequential computer, and therefore allows straight-forward porting of code from single processor computers to SMP machines. The operating system performs the job of load balancing and coordination between processors.

The capabilities of this architecture continue to increase along the Moore's Law (Gordon Moore, Intel chairman) trend line as the three limiting factors of operating system parallelism, processor speed, and system bus bandwidth continue to be overcome by advances in technology. All the vendors of this architecture are improving the capabilities of their operating system to handle the rigors of parallel processors. Hewlett-Packard has recently increased the maximum number from eight to twelve processors in the latest release of their SMP operating system[4]. Sequent has announced a new system bus technology that uses fiber optics to carry the system bus signals. This product, named NUMA-Q, allows the system bus bandwidth to increase along with the number of processors[5]. NUMA-Q allows a variety of interconnect architectures that include a

hypercube. This technology will start to bridge the architectural differences between SMP and MPP system and lead to more sophisticated software solutions for the MPP architecture early in the 21st century.

2.1.3 Workstation and Microcomputer Technology

The performance dividing line between personal computers (PCs) and workstations is gradually disappearing in favor of a continuously scalable line of machines. The P6 is a PC architecture designed to penetrate the server and workstation market, and merging these three types of computers into a common architecture. By 1998, desktop computers will be at the level of 1 giga-FLOPS, priced between $7,000 and $15,000, and will have a flat screen, a color liquid crystal display with 4 megapixels, 2 Gbytes of main memory, and an optical "hard disk" holding 100 Gbytes [1].

Microprocessors, the building block for low- to high-end computers, today have clock rates of 150 MHz, 200 MIPS of performance, 5 million transistors, and are fabricated using 0.35-micron design technology [6]. By the year 2000, the corresponding numbers will be 500 MHz, 1,000 MIPS, 40 million transistors, 0.2-micron bipolar CMOS, 4 megabytes of cache memory, and a 64-bit architecture using on-chip parallelism [1]. The distinction between workstations and PCs will fade. A range of performance will be available by means of plug-in options. Upgrades will be possible by means of plug-in cards, using multi-chip module packaging techniques instead of printed circuit boards. An issue affecting the future of open system designs is whether the Windows 95 and Windows NT operating systems will crowd out Unix and other operating systems as PCs encroach on workstation and server performance. Unix has a large following in the scientific and engineering world, but Windows 95 and Windows NT have a marketing edge because of Microsoft's overwhelming dominance of the PC market.

The contest between RISC and complex instruction set computer (CISC) microprocessors has clearly been won by the RISC designs. The new Intel P6 is optimized for 32 bit code and executes the older ipx86 16 bit code at rate slower than a Pentium [7]. The P6 has a special execution unit for ipx86 instructions that are too slow to run a one instruction per clock cycle. 32 bit code will always run at RISC speeds of one instruction per cycle. The next version of microprocessor will be designed in conjunction with Hewlett Packard and will be based on a very long instruction word (VLIW) architecture and marks the abandonment of the ipx86 instruction set. The fastest microprocessor today is the RISC design DEC Alpha with a clock speed of 300 Mhz [6].

Memory chips will continue to progress, and dynamic random access memory (DRAM) chips of at least 1 gigabit capacity will be the standard by the year 2000. Device density will be further improved using 3-dimensional packaging techniques such as stacks of chips in cubic arrangements. Memory technology is discussed in more detail in Chapter 3, Storage Technology Trends.

Silicon-based semiconductors will keep improving in performance and will remain the chief material in chip making through the decade. Bijan Davari [8] shows the continued downward scaling of silicon technology over the next ten years. At this point, channel lengths will have shrunk to below 0.1 micron and operating voltages will be reduced to 1 V. "...Speed improvement of about 7x, (and) density improvement of about 20x ... are expected relative to today's 5 V technology at 0.6 microns."

With clock rates in the hundreds of megahertz, packaging becomes a principal driver of performance growth. The P6 marks the first mass market use of multi chip module (MCM) technology. This technology will become more important as clock frequencies continue to increase because of the associated reduction in length and increase in transmission effects on intrachip wiring. MCMs will account for at least one-third of all packages by the end of the decade.

2.1.4 The Information Superhighway

The explosive growth of the Internet and especially the World Wide Web will further drive the growth in the electronics industry. The Web is currently based on hyper text markup language (HTML). Using a Web browser, a user generates a simple and singular transaction with the HTML server over the Web. The server responds by returning the requested resource. Sun's Java language for secure multimedia client server systems has recently emerged to contend for position as the new de-facto standard for Web application development. Growing use of the Web will necessitate an increased demand for transaction processors as Web servers.

The expanded use of the Web has led to a further increase in the demand for bandwidth to carry the traffic. Web pages now incorporate graphics as a matter of course, and there in an increase in the use of audio and video data. HTML and Java allow any type of data: text, graphics, audio, and video, to be requested and transferred over the Web. Graphics data currently demands the most bandwidth. The transfer of audio and video data will necessitate improvements in network protocol because of the isochronous nature of this data. Audio and video files could be transferred under the current packet switching protocol for later replay, but Internet phone or video phone would require that data be delivered on a time schedule and in order. The asynchronous transfer mode (ATM) is the protocol which is being touted as the solution to this isochronous requirement. ATM is a protocol based on short, fixed-length packets of 53 bytes. It can be switched very quickly in hardware because of its minimal overhead processing. It is suitable for very high bandwidth data and for on-demand bandwidth allocation. However, it is an untested solution which is in the early phase of implementation.

Several vendors are making Internet-ready multimedia workstations which are capable of handling audio and video data from the Internet and which cost just over $10,000. The prices for these machines will fall as volume shipments increase and cutting edge technology becomes mature and has commodity availability and price. The growth of the Internet increases the demand for capability in all three areas for the client server architecture: more server capability due to HTML Web transactions; more network bandwidth because of more traffic and the move from text to graphics, audio, and video over the network; and more client capability to receive, transform, and display real-time audio and video.

The Internet gives rise to a new and expanded view of computing and further propels the diversification of resources that started with the PC revolution. George Gilder [9] foresees a day when the Internet will not only link computers but also allow the further distribution of system resources. Using the ATM protocol, data can be transferred at up to 155 megabits per second and will soon deliver 2.4 Gbits per second [10]. With these sorts of data bandwidths available, computing resources can be further diversified away from the desktop. The PC represents a small version of a mainframe on the desk. The network computer would allow computing resources to be deployed in a more distributed fashion. The only element of the computer that need to be on the desktop is the user interface.

This concept is seen as competition to the traditional Intel/Microsoft PC [11]. The network computer is downplayed by Bill Gates, who describes it as a terminal or phone. But it is really an extension of client server computing with a lean, standard and cheap client. PCs and workstations are quickly converging into a high powered desktop computer. The Internet computer cuts out expensive and unnecessary capability and uses networked resources to perform these functions at a reduced price. The key enabling resource is large amounts of communications bandwidth. By using large amounts of cheap bandwidth, computing resources can be fundamentally redeployed. The main issue is who will pay for the high bandwidth infrastructure. If these costs are included, the network computer may not look so cheap. On the other hand, not everyone needs a desktop supercomputer, so at some point consumer PC performance might plateau. The concepts of network computer and PC may then be not so far apart.

2.1.5 Summary of Computing Trends Data

Several decades of history have accumulated for computer technology, and attempts have been made to collate some of these data. For example, Moravec [12] gives a historical trendline that traces computer development throughout the 20th century. The latter half is shown in Figure 2-1. This part, from 1950 on, shows an improvement in throughput per unit cost of about 45 percent per year compounded. This is close to Moore's Law, which predicts a doubling every 2 years.

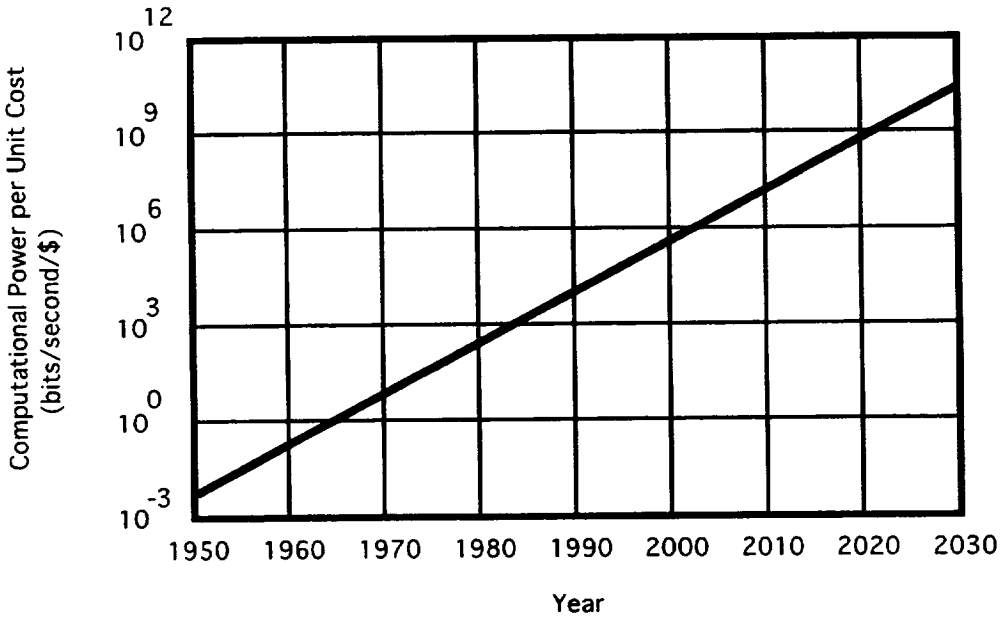


Figure 2-1. Long-Range Computer Technology History and Projection. (Hans Moravec)

Source: Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, 1988

A historical trendline for the device density of Intel microprocessors [13] is shown in Figure 2-2. This shows a rate of improvement of 40 percent per year, also following Moore's Law.

The two curves together demonstrate that improvement measures per device follow improvement measures per unit cost, or in other words, the cost per manufactured unit remains constant. Therefore, cost projections can be derived from performance improvement projections.

Figure 2-3 shows a composite of several recent projections of computer performance [1, 14, 15]. The two upper trendlines are projections taking parallelism into account. As can be seen, predictions for single processors, the true measure of underlying technology, fall mostly in the range of 100-to-1 to 1,000-to-1 in the 15-year interval shown, or 36 percent to 60 percent per year improvement. The 40 percent per year figure (Moore's Law) is a reasonable, perhaps conservative, estimate. It should also be stated that predictions of improvement due to parallelism are inherently less reliable than for single processors since, according to Kelly [1], performance of a parallel computer can vary by two to three orders of magnitude for a given algorithm, depending on how the algorithm is fine-tuned for the particular parallel computer.

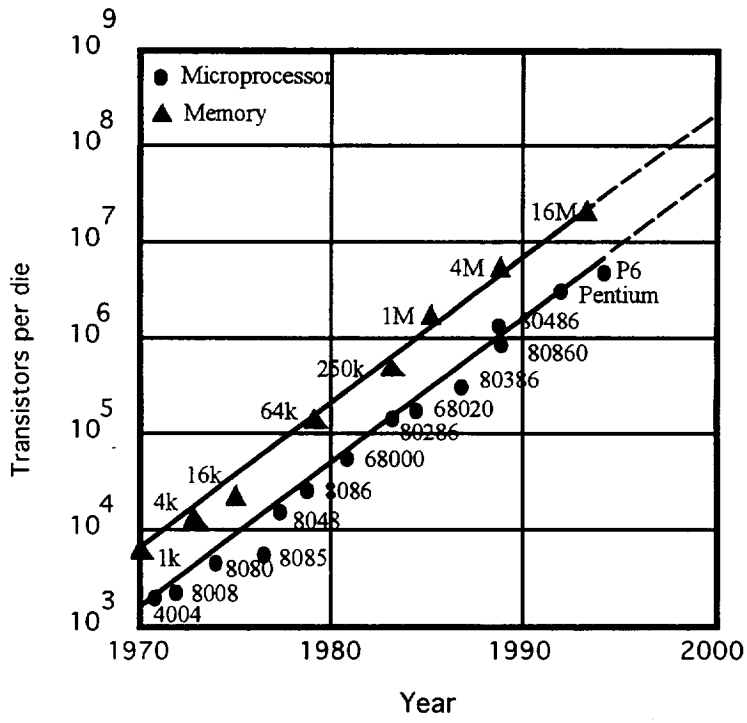


Figure 2-2. Integrated Circuit Device Density

Source: Intel Corporation

In order to keep the progress of technology moving along these trend lines, corporations in the forefront of technology development have found it necessary to form partnerships with their competitors. Examples of these partnerships include IBM, Siemens and Hitachi in memories, IBM and Motorola in microprocessors, Intel and Hewlett-Packard in microprocessors, and DEC and Microsoft in server operating systems[16]. This trend will continue as the cost of introducing new technology generations grows at an exponential rate. The profits from advanced technology continue to justify the amount of investment needed to fund the succeeding generation of technology. The size of the investment, one billion dollars for a modern chip plant, means that the risk of failure can be catastrophic. These companies are spreading the risk by teaming with competitors. This trend will follow the growing cost of technology development.

2.2 FUTURE APPLICATIONS

Mass transaction processing in government, insurance, banking, and the Internet, and economies of scale derived from these applications have historically pushed down price-performance ratios. In the future, computing technology will be driven by new economies of scale: document imaging, image processing, storage and retrieval, database search, remote database communications, image communications, and graphical user interfaces. These applications will provide opportunities for technology insertion. Since many of these applications are well-matched to NASA needs, there will be great benefits accruing from technology insertion.

A mix of sophisticated graphical user interfaces, simulation, and expert systems should allow less-skilled workers to perform sophisticated tasks currently performed by highly educated professionals. As these techniques migrate to mass market price levels, by virtue of low-cost hardware, they become available to lower-paid workers. This has obvious implications for the cost of labor. User interfaces are an important component. Advanced user interfaces—such as virtual

reality, a form of 3-dimensional visual display imitating real-world views—will require enormous processing power to generate a believable real-world appearance to the user. Current graphics workstations can perform 2 million polygon renderings per second. Virtual reality will require about 2 billion polygons per second.

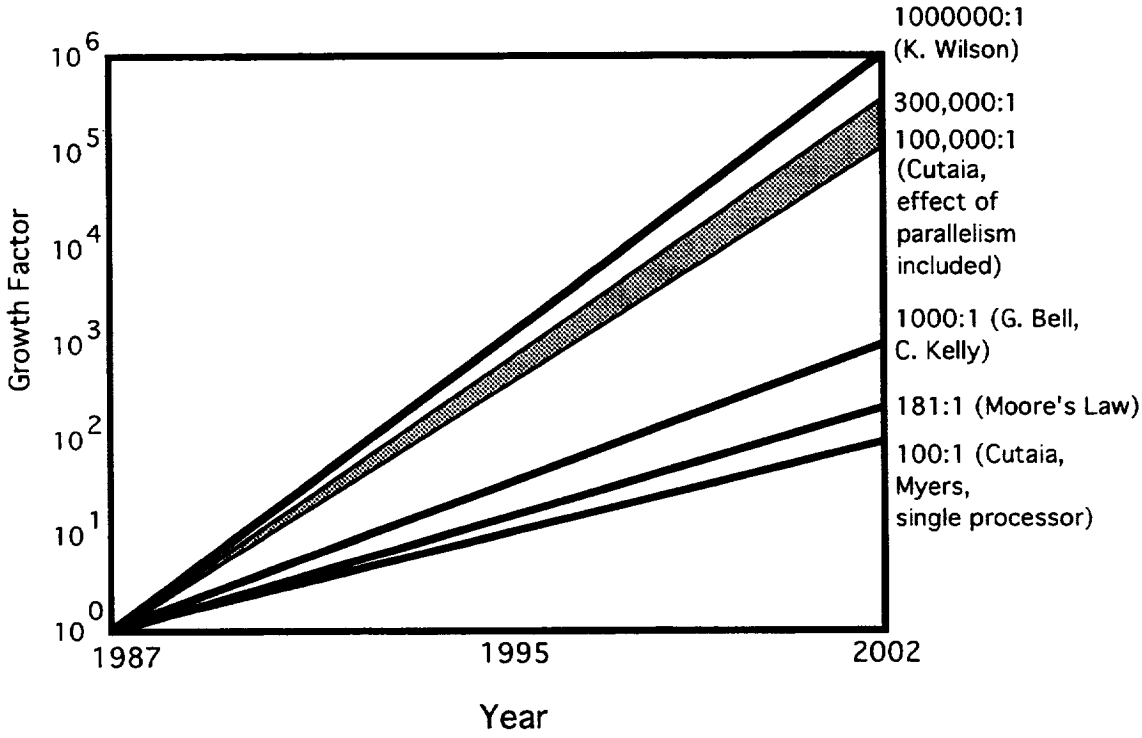


Figure 2-3. Comparison of Technology Projections

Sources: E. Yourdon, *Decline & Fall of the American Programmer*, Englewood Cliffs: Prentice Hall, 1992
C. Kelly, "Information Technology; Computer Hardware and Architectures", SAIC Memo randum
R. Katz, et al., "Disk Architectures for High Performance Computing", Proc. IEEE, Vol. 77, No. 12, December 1989, p. 1842

For space applications, the key parameter is functionality per unit weight. Increasing performance trends for electronics are directly coupled to miniaturization. There is therefore a cascade of effects, since more functionality per unit weight leads to lower launch costs, which leads satellites on orbit, which results in more communications traffic. Ultra-lightweight spacecraft developed by the Ballistic Missile Defense Office indicate the potential for low-cost satellites for university research and private commercial applications.

Communications bandwidth and transmit signal power can be traded for processing power when on-board processing can be used for data compression. Imagery can theoretically be compressed in a lossy fashion to as high as 1,000-to-1. The high payoff of compression of data for which lossy compression is acceptable is an incentive to develop high-performance, space-qualified processors. Similarly, using on-board processing, transmit signal power can be traded for error correction coding and advanced waveform design to reduce weight.

2.3 EMERGING TECHNOLOGIES

Perhaps the most active area of alternative computer technologies is the field of optical computing. We discuss this field in more detail in Chapter 4, Photonics Technology Trends, but will give a brief overview here.

Optical methods are under active investigation as a better way to do on-chip and chip-to-chip interconnections. For high clock rates, optical broadcast of the clock signal from a point above the chip avoids the severe clock skew problems of electronic distribution methods.

Optical techniques have been used for many years for special-purpose computing elements, for example, acousto-optic modulators for instantaneous radio frequency (RF) spectrum analysis and Fourier transform lenses and coherent light to do very high speed spatial Fourier transforms and correlations of images. The number of applications is growing, and the technology is improving. For example, a number of innovations make likely the practical realization of "integrated optics", or optical and electronic devices on a common substrate: surface-emitting laser diodes, planar optical waveguides, diode lasers and optical photodetectors on the same chip with electronics, holographic optics, and fast spatial light modulators. Special-purpose analog optical processors are likely to be in routine use by the end of the decade. However, widespread use of all-optical digital computers is likely still decades away.

A compelling case can be made for a good match between optical computing and artificial neural networks [1]. Software-based sequential simulations do not take full advantage of this algorithm. The characteristics of an ideal hardware implementation of a neural network are speed, parallelism, and massive interconnectivity. Optical computers have these features. For large-scale networks, it is impractical to build wire interconnects. In optical computers, no wiring is required between nodes of the network.

Neural network architectures will become the solution of choice over many classical approaches to image processing, signal processing, real-time control, and optimization. Hardware currently lags software in this field, creating a strong motivation for pushing the boundaries of hardware implementations. Analog computing applied to neural networks, in particular, can for certain problems give large increases in computing power. The most famous of these implementations is Carver Mead's artificial neural retina, which executes 100 million analog operations per second on a chip [17]. Mead claims that his chip exhibits a 10,000-to-1 advantage in power consumption per operation over an equivalent digital device.

3. RESULTS

Computer hardware technology is the fastest evolving technology in history, opening up new opportunities, but creating new problems for information systems builders, operators, and owners.

3.1 IMPACT OF COMPUTING TECHNOLOGY ON NASA

3.1.1 System Architectures

The momentum for "downsizing" will continue during the remainder of the 1990s. A recent report indicates a potential 30 percent cost saving in downsizing from mainframes to microcomputers over 5 years [18]. Another report notes that there are hidden costs, such as integration costs, to be taken into account, but that the savings are still significant [19]. The impact on "legacy code" (existing software that is to be re-implemented on a new generation of platforms) will be that advanced software re-engineering and reverse engineering techniques will need to be applied to port the existing code to downsized, open systems.

The speed of local area networks (LANs) will be equivalent to that of mainframe backplanes in the late 1990s. For example, Fiber Channel and the High Performance Parallel Interface (HIPPI) operate at 800 megabits per second (Mbps) in a point-to-point mode. These non-protocol-based

techniques, with their low software overhead, are preferred for high-end applications. Therefore, the LAN becomes the backplane; modularity can be achieved through servers plugged into the LAN. Small, networked computers facilitate rapid prototyping. Prototypes can be scaled up to full operational capability by duplicating hardware modules using a parallel distributed architecture. OSC's telemetry applications are well-suited to parallel distributed partitioning. The telemetry applications are also efficiently scalable. There is therefore a potential benefit to this kind of architecture for image telemetry networking and processing.

Hardware accelerators are sometimes overlooked as an approach to "downsizing" for well-defined but computationally intensive applications. It may take less time and money to build a specialized accelerator than to fine-tune code to run on vector supercomputers or massively parallel computers. For example, a recently built accelerator for a simulated annealing algorithm used to solve complex scheduling problems was built on a single circuit board that plugged into a personal computer backplane, yet outperformed a Cray Y-MP by between 30 and 80 times, while costing only $700 in volume production [20]. In this case, users were able to drop down at least four levels of price-performance, from supercomputer to personal computer.

3.1.2 Reliability

Hardware advances can open up new possibilities for reliability engineering, including higher integration levels, on-chip redundancy, failure-prediction circuitry, and others. A study by Cutaia [14] indicates that computer chips in the year 2000 will be more reliable by a factor of 100 than those of today. A representative of the Air Force Electronic Systems Division recently stated, "We now have irrefutable data from industry showing that commercial components and open-system architectures are less expensive, more reliable, and of higher quality than their JAN [a military high-reliability parts standard] and Military Standard equivalents" [21]. According to DOD opinion, it is the use of military specification parts that will have to be justified in the future.

3.1.3 Software

Advances in hardware can have a major impact on software-related issues. For example, the advent of powerful personal computers in the 1980s allowed software developers to use sophisticated interactive development environments at their desktops and to debug and re-compile their programs on a much shorter turn-around cycle than previously. On the other hand, the same desktop computing power has permitted the use of graphical user interfaces which require many more lines of code than the earlier non-graphical applications.

Yourdon [14] suggests several changes in the software field as a result of the vastly more powerful computers of the future. First, he points out that the greatly increased throughput capabilities of future computers will change the nature of the data being processed from predominantly text to a mix of voice-data, text with images, and pure imagery. This capability will drive the development of new applications. Second, he sees the possibility that automated code generation techniques and fourth generation languages could become more widespread, although he notes that the typical programmer's psychological resistance to this trend may not allow it to become widespread in the next 10 years. Third, he states that a new approach to software testing is the use of simulation—an approach that will require greatly enhanced hardware capabilities. Fourth, he surmises that much of the newly available CPU power will be absorbed by more sophisticated user interfaces, such as pen-based computing and data visualization.

Hardware and software vendors are teaming up to deliver solutions to particular marketplaces. This trend runs counter to the concept of developing open systems. These solution providers are driven by increased competition to fine tune the solution for the best performance at the lowest cost with the highest immediate payoff. This trend means that there will continue to be "flavors" of cross-platform systems, and an associated porting cost between the different platforms that run

these systems. There will be flavors of Windows NT and flavors of Oracle, just like there are flavors of UNIX.

3.1.4 Information Overload

One of the big problems faced by NASA is the large amount of imagery data that will result from the Earth Observing System (EOS) and the Space Station. While DOD's use of imagery is different, an experience base has been built within DOD that confirms that there is indeed a very big problem. DOD estimates that only 1 to 10 percent of imagery from high-data-rate sensors can be evaluated by analysts. This implies that a lot of money is being spent on what is essentially a "brute-force" method, i.e., collect as much data as possible and hope that the analysts will look at the right 10 percent. Automation of the analysis function requires mainly the invention of new algorithms, not something that hardware can help with. However, a recent paper reports on DOD thinking as to how neural network hardware might help [22]. The paper describes six potential applications of advanced hardware implementations of neural networks in order of ascending complexity, along with quantitative estimates as to the required neural network performance:

Application	Number of Interconnects	Number of Interconnects/second
Sonar array processing	$4x10^4$	$4x10^4$
Electronic intelligence (radar signal ID)	$3x10^4$	10^9
Synthetic aperture radar ground surveillance	10^9	10^9
Infrared ground surveillance	10^9	10^{10}
Look-down infrared aircraft detection	$4x10^{11}$	$4x10^9$
Multisensor fusion	10^{11}	10^{12}

Except for the sonar case, the processing required is beyond the supercomputer level. It is clear that special-purpose neural network processors will be required to achieve these performance levels. The point to be made is that hardware can be used to attack the problem of automatic imagery analysis, but that it requires massive parallelism on a scale that is beyond the state of the art. This indicates that there is a possible payoff for investments in hardware implementations of neural networks. Jet Propulsion Laboratory, for example, is actively pursuing this.

3.1.5 Organizational Impacts

Better price-performance redistributes power to lower organizational levels. Smaller groups of scientists, or even individual scientists, will be able to accomplish that which currently requires large organizations in which resources are pooled. Increased personal computing power will lead to more "personalized" science. This is illustrated by the "big science/small science" debate. Another illustration of the phenomenon is the current wave of enthusiasm for the prospects for using small, lightweight satellites for low-budget space science projects. NASA policy reflects this thinking when it advocates "cheaper, smaller and faster" spacecraft. The advancement in digital electronics, with its decrease in size and power consumption, and its increase in processing power, represents an enabling technology for smaller and cheaper yet very capable spacecraft. In communications systems, advances in mixed-signal integrated circuits such as analog-to-digital converters can move more signal processing functions from hardware to software, for example.

This philosophy has already had an impact on NASA. However, this line of thinking can only be carried so far. Many spacecraft functions will be unaffected by advances in semiconductors. For example, sensors depending on surface area for energy collection, such as antennas and optics, cannot be miniaturized.

The simplified, low-cost approach is being tried in the commercial world in the new low-earth-orbit satellites for mobile communications. The Orbcomm system will use 26 low-cost satellites,

with a cost of less than $200 million for the system, but the communications functions are rather limited. To achieve ubiquitous coverage for voice communications, more complex and expensive satellites are required, such as in the Iridium system. This system will cost several billions of dollars.

3.2 TECHNOLOGY ROADMAP

To summarize, a technology roadmap, or hypothetical timeline, will be described that illustrates the significant technology developments as they might occur over the next 25 years.

1995–2000: PC s and workstations quickly converge to a high powered desktop computer. Scalable servers replace mainframes at the high end of the market. Massively parallel super computers deliver tera-FLOP performance for scientific applications. Increase availability of bandwidth allows further distribution of computing resources and gives rise to the network computer.

2000–2010: In this time period, the need to exploit the power of ever-greater parallelism for high-end applications may lead to the use of more specialized computers. Parallelism requires optimizing the mapping of algorithm to architecture, a process that is easier when using specialized computers. For image processing, an area of interest to NASA, computers will begin to make use of hardware neural networks, cellular automata, or other parallel architectures implemented using both optical components and quantum-well semiconductor chips. These computers would be used as accelerators for more conventional systems. By the end of this period, it should be clear which of the alternative technologies will replace conventional integrated circuits.

Miniaturization will be such that robotic spacecraft for scientific research will be very small, and launch costs will be reduced accordingly. It could therefore be expected that a wide variety of commercial groups will be putting spacecraft in orbit, creating an incentive for the development of cheap, small, portable or rapidly deployable automated space-ground communications links.

2010–2020: In this period, computers will approach the density of biological systems, i.e., the molecular or nanometer scale. The appearance of molecular computers will require the use of architectures and "programming" techniques not familiar today. Biological models of computation will probably be important, along with programming through learning, evolution, and self-organization. Internal to these computers, signals will be conveyed by chemical and photonic means as well as by electrons.

3.3 FEASIBILITY AND RISK

When commercial off-the-shelf technology is applicable, as is often the case for ground-based systems, computer hardware development risk is not a NASA concern. NASA has typically been more involved with pioneering the application of high-performance computers to difficult problems. Recently, NASA has become a leading purchaser of massively parallel computers. NASA has also been a leader in research on advanced specialized computers, such as hardware neural networks and optical computers.

Gordon Bell believes that acceleration of computer technology development by direct funding of particular computer designs is a high-risk undertaking [3]. He observes that so far, the companies that have been created by ARPA funding to build particular computer architectures have not had commercial success. Military computer development efforts have often resulted in the military being stuck with service-unique, obsolete computers. Companies that were indirect beneficiaries of ARPA funded university research have much better commercial track records. Bell makes a strong case for the government to fund basic research and not product development. Government-funded product development has historically not been able to compete in the commercial market; hence, the development funding is usually wasted. The current push to fund particular massively parallel computers is focused on raw peak power, and this results in

underfunding the programming infrastructure needed to support widespread acceptance of massively parallel computers. NASA's emphasis on applications (software) for these systems will reduce the risk for all users.

In the area of emerging technologies discussed earlier, optical computing has made advances, but will remain a laboratory demonstration for the near-term, except for certain highly specialized applications. Nanotechnology and molecular computing offer a path to new levels of miniaturization, but will require the development of new engineering infrastructures to become mainstream. The first benefits of molecular-scale computing will probably come in storage devices.

References

1. Kelly, C., "Trends in Computing Architectures", internal SAIC memorandum, May 1992.
2. Intel World Wide Web Home Page, December 1995.
3. Bell, G., "Ultracomputers; A Teraflop Before Its Time", *Communications of the ACM*, Vol. 35, No. 8, August 1992, p. 27.
4. Hewlett-Packard system benchmarks, September, 1995.
5. Sequent Computer Systems, Inc., "Sequent's NUMA-Q Architecture", October 1995.
6. Pountain, D., et al., "CPU Scorecards", *BYTE*, November 1995, p. 179.
7. Halfhill, T., "Intel's P6", *BYTE*, April 1995, p. 42.
8. Davari, B., et al., "CMOS Scaling for High Performance and Low Power - The Next Ten Years", *Proceedings of the IEEE*, Vol. 83, No. 4, April 1995, p. 595.
9. Gilder, G., "The coming Software Shift", *Forbes ASAP*, August 1995.
10. Gilder, G., "The Bandwidth Tidal Wave", *Forbes ASAP*, December 1994.
11. Silverthorne, S., "Dubious Extinction", *PC Week Inside*, November 13,1995.
12. Moravec, H., *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, 1988.
13. Gelsinger, P., et al., "Microprocessors Circa 2000", *IEEE Spectrum*, October 1989, p. 43.
14. Yourdon, E., *The Decline and Fall of the American Programmer*, Englewood Cliffs, Prentice-Hall, 1992.
15. Katz, R., et al., "Disk System Architectures for High Performance Computing", *Proc. IEEE.* Vol. 77, No. 12, December 1989, p. 1842.
16. Choi, A., "DEC Chips Away at Critics with New Microprocessor", *The Wall Street Journal, December 11, 1995.*
17. Mead, C., and M. A. Mahowald, "The Silicon Retina", *Scientific American*, May 1991, p. 76.
18. Eckerson, W., "Hidden Costs May Come with Downsizing Savings", *Network World*, June 29, 1992, p. 41.
19. Pitta, J., "Nature Abhors a Monopoly", *Forbes*, July 6, 1992, p. 100.
20. Abramson, D., "A Very High Speed Architecture for Simulated Annealing", *IEEE Computer*, May 1992, p. 27.
21. Kachmar, M., "US Air Force Advances Goals of RISE Effort", *Microwaves & RF*, June 1992, p. 42.
22. Lupo, J. C., "Defense Applications of Neural Networks", *IEEE Communications Magazine*, November 1989, p. 83.

CHAPTER 3. STORAGE TECHNOLOGY TRENDS

SUMMARY

Magnetic disk capacity will continue to improve at a rate of at least doubling every 3 years over the next 10 years. Recent laboratory results demonstrate that densities of 1 gigabit per square inch are feasible. Access times improve much more slowly and are limited by the mechanical nature of the system.

Magneto-optical disk systems will continue to improve as well. Recent laboratory results demonstrate that densities of 45 Gbits per square inch are feasible.

Random access memory (RAM) capacity will continue to improve at a rate of at least quadrupling every 3 years. Laboratory results demonstrate that densities of 4 gigabits per chip are feasible.

There is a basic trade off between storage capacity, speed of access and cost. Large storage capacity must be cheap because of its size, but size and cost are achieved by sacrificing speed. Fast access memory is expensive, so its size must be limited. This relationship has led to the layering of storage systems with static random access memory (SRAM) used for cache, dynamic random access memory (DRAM) making up main memory, hard disk employed for bulk storage, and tape used for archival storage. Optical storage has replaced floppy disk as a distribution media, but has lagged in its promise to replace hard disk or tape in their primary functions. Silicon carbide non-volatile RAM devices are capable of higher speeds and greater densities than present-day non-volatile memory devices.

Holographic storage is the most promising advanced technology for achieving random access volumetric storage. Volumetric storage offers orders of magnitude greater density than surface storage. Prototype systems have been announced and these systems should make it to market during the next five years.

Molecular-scale storage devices are theoretically capable of millions of times greater storage density than any other technique. However, the achievement of practical devices is limited by the interconnection problem and is, in any case, perhaps two decades from realization.

1. INTRODUCTION

NASA will be at the forefront of users of large-scale mass storage and archive systems in the next 10 years. NASA needs to be able to deploy systems that make use of the latest capabilities of storage technology in order to handle the terabyte-per-day data volumes that will be generated by planned near-earth missions. Technology changes and breakthroughs in the storage field can have major consequences for an archival type of storage system. While costs can be lowered in the long run if a system is upgraded, time and funds must be expended to transfer the data to the new medium. It is therefore critical to be aware of the possible breakthroughs sufficiently in advance to be able to plan for a relatively smooth transition.

In the last decade or two, several trends have emerged in the field of storage technology. For example, storage technology has improved at a rate slower than that of microprocessors. This has led to what is called the "input/output (I/O) gap." The I/O gap refers to the fact that the throughput of computers is limited by memory technology. This problem has been exacerbated by the increase in microprocessor clock frequency, because it allows less time to access information in memory. This has a negative effect on computer system evolution. It also has an impact on networks, since storage is required for staging of file transfers over networks. The I/O gap is expected to be mitigated dramatically within a decade. A large amount of research has gone into this area and many specialty memories have been developed to address this gap such as enhanced

DRAM, synchronous DRAM, and multibank DRAM.[1] These new memories are, of course, more expensive than simple DRAMs.

Another trend is that magnetic storage has been able to keep pace with optical storage, when in earlier years, it had been predicted that optical would replace magnetic. Optical suffers from a fundamental data access speed problem. It takes too long to write to optical media when compared to magnetic disk. A third significant trend is that solid-state storage technology is approaching the density of magnetic mass storage. This report will quantify these observations to the extent possible.

Because memory devices are generally simpler than other computing circuitry, memory technology advances often precede and enable computer advances. This plus NASA's very large storage requirements are reasons for expecting very high potential payoff from advances in this field.

2. LONG-RANGE TECHNOLOGY TRENDS

2.1 ANALYSIS OF TRENDS

2.1.1 Magnetic (Winchester) Disk Technology

Magnetic disk density has increased at a rate of approximately doubling every 3 years, as shown in Figure 3-1 [2]. This shows the evolution of large-format Winchester disks. Recent laboratory results at IBM proved that 10^9 bits per square inch can be achieved [3]. "Areal density (the combination of linear and track density), which used to grow at a 25 percent annual rate, has been growing 60 percent annually since 1990, Industry observers predict it will continue to do so for the foreseeable future."[4] Beyond the year 2005, it is possible that new technologies may begin to replace mechanically driven magnetic systems. However, the impressive track record of magnetic-device storage density suggests that this technology may continue to be the mass storage standard for at least the next 10 years. Beyond the 10-year timeframe, the picture is less clear because of the possible appearance of holographic storage and other advanced media.

At the same time, the increasing density is producing newer and smaller disk formats. While the 5.25-in. size was standard in the 1980s, there has been a rapid proliferation of smaller formats in recent years, including 3.5 in., 2.5 in., 1.8 in., and 1.3 in. The last is almost the size of large solid-state chips. An interesting fact about these smaller formats is that the price per bit has gone down and the reliability has gone up with each new format. This has led to the development of the redundant array of inexpensive disks (RAID) concept. According to this concept, parallel arrays of small disks have several advantages: first, they are theoretically less expensive for the same amount of storage; second, they provide redundancy and error correction capability when a disk drive fails, increasing availability by orders of magnitude; third, they have potentially increased transfer rates because data are accessed in parallel across multiple disks, rather than in series. These benefits have fueled an increasing demand for RAID products and a corresponding diversity of RAID offerings from a variety of suppliers.[5] The fall in RAID prices to about one dollar per megabyte has increased its popularity to the point where RAID is used for a substantial segment of large data storage applications.

2.1.2 Optical and Magneto-Optical Disk Technology

The typical erasable optical disk product of today carries about 4.6 Gbytes in a single removable 5.25-in. format and costs $1,700 [6]. The disks cost an additional $200. It is used primarily for backup and long-term storage because of its longevity (the length of time it retains data) and its portability feature. Removable Winchesters do not have as high a capacity on a single disk. Winchesters have faster access times, but the gap is narrowing. Optical disk drives have access times as fast as 35 milliseconds, and in the near future, new techniques are expected to boost the speed. Optical disks are relatively easily adapted for parallel access.

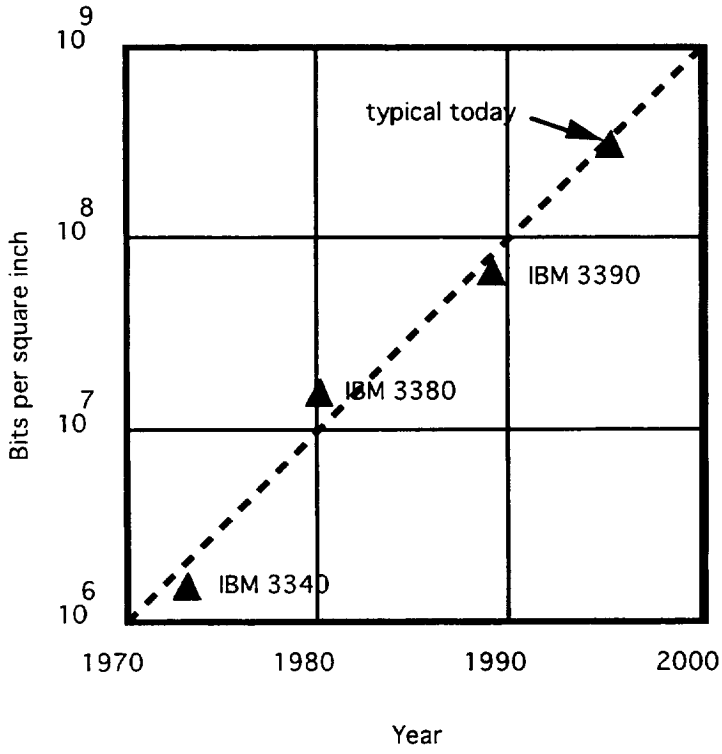


Figure 3-1. First law in Disk Density

Sources: Katz, R., et al., "Disk System Architectures for High Performance Computing", *Proc. IEEE*, Vol. 77, No. 12, 1989, pp. 1842-1858.
Merz, J., "Rigid Disk Systems", National Media Laboratory Home Page.

Because of the nature of the laser transducer, a read/write head can be built containing thousands of lasers and detectors to access thousands of recording tracks simultaneously. CD-ROMs are now writeable and have been forecast by Disk/Trend Inc. to capture 89 percent of the Gbyte storage market [7].

Optical disks have densities on the order of $4x10^8$ bits per square inch, whereas magnetic disks are at about $6x10^7$ bits per square inch. However, the IBM results (10^9 bits per square inch) show that in the laboratory at least, magnetic technology has actually surpassed current optical products. On the other hand, other new developments show that an optical transducer can be used to modulate a magnetic surface to surpass the density reported by IBM. AT&T recently announced a prototype magneto-optical disk technology in which laser light funneled through an ultra-fine-tipped optical fiber was able to change the polarity of magnetic particles on the disk surface at a density of 45 gigabits per square inch, more than an order of magnitude greater than the IBM results and two orders of magnitude greater than current optical disk products. This density is also greater than permitted by the diffraction limit of light, a fact which contradicts earlier predictions that optical methods would top out at the diffraction limit [8].

2.1.3 Solid State (RAM) Technology

The storage density of RAM chips has increased at a rate of quadrupling every 3 years, or about 60 percent improvement per year on average. This is shown by Figure 3-2 [2,9]. This parallels the growth in density of the CPU chip, but the relative speed improvements do not follow each other.

There is an I/O speed gap between the CPU and the RAM. The speed of the fastest RAM chip, static RAM, has improved at 40 percent per year, while dynamic RAM has improved more slowly. Main memory has been able to keep pace with CPU performance improvement by the use of static RAM cache memory in front of a larger but slower dynamic RAM main memory. (Static RAM cache memory is a temporary high-speed buffer for holding data that are expected to be requested by the CPU before the request arrives, in order to avoid wasting time accessing the slower dynamic RAM.)

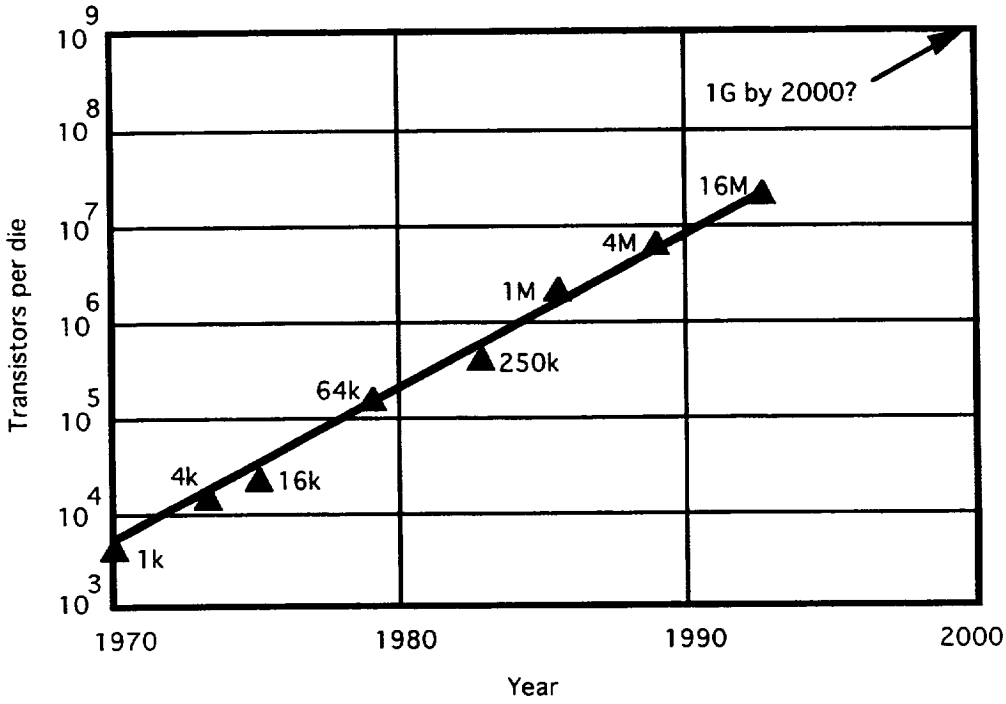


Figure 3-2. DRAM Density Trend

Source: Intel Corp.

Recent developments in RAM technology demonstrate that the growth in density will continue at the same pace or better through the end of the century. This is represented in Figure 3-2 by the data point for the year 2000 at the 1 gigabit per chip density level. Laboratory versions of these chips have been fabricated by NEC and Hitachi using 0.2 micron and 0.16 micron lithography respectively [10]. IBM has demonstrated transistor technology capable of a density of 4 gigabits per die using electron beam lithography [11]. Many years of work will be required to bring this technology to the mass-production stage, hence our estimate for the year 2000.

It is of interest to compare the growth rates of RAM and Winchester technologies. RAM technology will likely surpass magnetic disk technology around the turn of the century in terms of density of stored information. This convergence has major implications, since RAM is so much faster than disk memory but at a greater cost. The move from magnetic media to solid state media will occur when the trade off between speed of memory access and cost per byte dictate the change. This change has already occurred in two applications; mobile computing and satellite recorders. Both of these applications have special requirements of power and durability that have driven the change, but additional applications will make the switch when the speed of access is worth the extra cost of storage.

A increasingly popular approach to reducing the I/O gap between the CPU and the RAM has been introduced by Rambus [12]. They have developed a 500 Mbps bus for interfacing RAM chips directly to the CPU that can be implemented on the RAM chips themselves. The aim of this concept is to eliminate the need for RAM cache. The Rambus technology has been adopted in high bandwidth applications such as full-motion video and 3-D imaging. The fastest implementation is 1000 Mbytes/s where two 8-bit channels are run in parallel.[13] Rambus has licensed this technology to many of the large DRAM manufactures.

2.2 FUTURE APPLICATIONS

Some future applications in the commercial market will generate economies of scale. For example, high definition television (HDTV) over fiber-optic cable will require huge amounts of high speed storage at the transmit site. There will therefore presumably be a commercial development effort to build a cost-effective system for doing this. A similar incentive exists for medical imaging networking applications, although the funding source for this is less obvious. A third major driver for smaller, faster, and cheaper storage is the rapidly growing document imaging and storage industry. The latter application is the nearest-term. To date, government applications have driven the development of large scale electronic document and retrieval systems, but in future more commercial development funding will likely be directed to this application.

In the area of advanced technology, holographic storage, described in the next section, is applicable to fast database search and retrieval. Holographic memory facilitates massively parallel search and retrieval because it can be randomly addressed, can be scanned at high speed because light beams are switched rather than mechanical heads, and can read out entire pages in parallel.

2.3 EMERGING TECHNOLOGIES

2.3.1 Holographic Volumetric Storage

The alternative technology that has seen the largest amount of research over several decades is optical holographic storage, sometimes called holostore or optical RAM (ORAM). The slow progress in this field has been due primarily to the shortcomings of the available materials used as the medium for making erasable volumetric phase holograms. The ideal material must obviously be erasable. It must also be transparent as well as being photosensitive. If the material were not transparent, the holograms could not be 3-dimensional. The storage technique must achieve non-destructive read operations, a feat which is very difficult with most materials. Recent progress in non-destructive read operations in photorefractive crystals has made practical, cost-competitive holographic storage appear to be within reach. The system built by the Microelectronics and Computer Technology Corp. (MCC) [14] uses strontium barium niobate in the form of an array of 2,500 crystal rods (crystallites) packed into a volume of 5 cm by 5 cm by 0.5 cm.

This group has left MCC to form Tamarack Storage Devices which has announced a prototype system. The initial Tamarack system will store up to 20 Gbytes. The company announced the completion of a optical head which integrates the components necessary to store and read data [15]. This product is targeted at video applications where a large bandwidth and volume of data are required. Observers predict that follow-on systems will target main stream computer applications with the first system being a write-once read-many (WORM) system. The system can be configured for large storage or high bandwidth [16]. This gives rise to three configurations: a high capacity system with over 350 Gbytes of storage capacity, a high bandwidth system with 327 Mbytes/s of I/O bandwidth, and a general purpose system with 2.3 Gbytes of storage and 80 Mbytes/s of data transfer. Each of these systems represents a trade-off between cost, storage capacity and I/O bandwidth.

In the Tamarack prototype, data are stored as 2-dimensional hologram planes stacked in 3 dimensions. Each plane constitutes a page of data containing up to 100 Kbits. Each crystal module contains about 30 pages per stack and 2,500 stacks, for a total of about 1 Gbyte. All the data in

each page are read out simultaneously in 10 microseconds, resulting in maximum readout rate on the order of 300 Mbps. This system uses multiple crystals in a jukebox arrangement.

Several other companies are pursuing research in this area including Storage Technology, Hughes Aircraft, Rockwell International, and IBM. Storage Technology has announced a breakthrough that will allow 100 Mbits to be stored in a cubic centimeter [17]. This is an impressive result for a single crystal. A major drawback of the technology is that the crystals are expensive and have a limited number of writes before wearing out.

A comparison of holographic storage to existing technologies is shown in Figure 3-3.

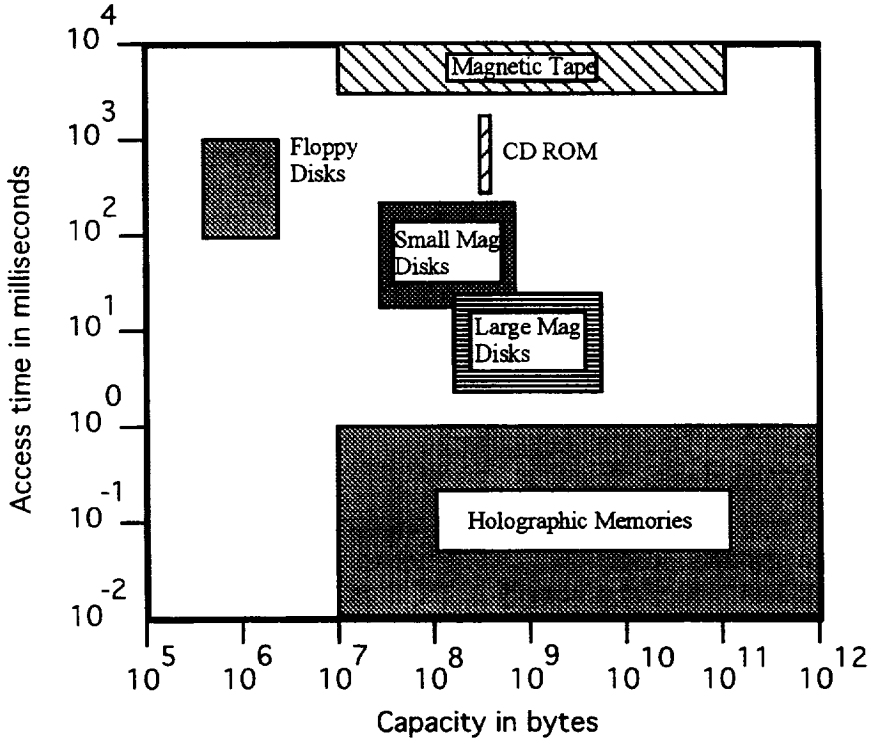


Figure 3-3. Comparison of Secondary Storage Technologies

Source: Berra, P.B., et al., "The Impact of Optics on Data and Knowledge-Based Systems", *IEEE Trans. KDE*, March 1989

Observers envision a phased approach to introducing this technology. In order of increasing implementation complexity, some applications are disk replacement, disk caching, front-end processor caching, system bus interfacing, and direct connection to the CPU. The easiest and earliest potential application is archival storage using a WORM configuration. As crystals become more robust these systems will be able to replace a hard-disk drive, since minimal changes would be required to existing computer architectures. In this application, performance improvements of workstations would be from 2 to 30 times, because of reduced seek and latency times. (The seek time is the time required for the disk transducer head to reach the data on the disk. The latency time is the total delay from the time data are requested to the time the data can be used by the CPU.) It would also eliminate mechanically moving parts, thereby increasing reliability. For large minicomputer and storage server applications, a holographic device could serve to boost disk farm (a disk farm is a group of disk drives) performance as a non-volatile, fast disk cache, leaving the existing investment in disks intact. In the future are implementations in which the ORAM connects to the system bus as the main memory, or as advances in integrated optoelectronics are achieved, the ORAM could be even more closely coupled to the CPU.

The large amount of commercial and university research going into this technology indicates that the prospects are good that product will appear over the next five years. The impact of these product will increase after the year 2000 because of the increased durability of the storage crystals. This technology will represent a major advancement of storage technology between 2000 and 2010.

2.3.2 Molecular Storage Devices

This technology uses single molecules to store the information and presents the ability to greatly increase the density of stored data. According to a article on molecular-scale electronic devices [18], the area storage density of a molecular-scale memory could be as high as 10^{16} bits per square cm, compared to $4x10^{8}$ bits per square cm for optical storage, or a factor of 25 million difference. While the feasibility of using complex molecules as electron gating devices has been experimentally verified, an obstacle to achieving theoretical density limits is the connection problem. That is, how does one get signals into and out of such tiny devices? Some researchers believe that optical techniques such as are used in optical computing experiments offer the best chance. Others believe that molecular devices will not be built for their high-density advantage, but for their speed advantage. Switching speeds for molecular gates are on the order of 3 picoseconds (1 picosecond is 10^{-12} seconds).

2.3.3 Silicon Carbide Devices

Flowing out of work of the Department of Defense, research on silicon carbide semiconductor devices has yielded the possibility of new types of non-volatile RAM devices that are capable of operating at higher speeds than current non-volatile devices. Non-volatility is achievable by virtue of the fact that current leakage in a silicon carbide memory cell is 100,000 times less than in conventional DRAMs. Higher speeds are achievable because the devices can tolerate higher temperatures and can therefore operate at higher clock speeds. Silicon carbide memory cells can also be made 3 times smaller than DRAM cells, because of silicon carbide's higher dielectric constant [19].

2.3.4 Exotic Solid-State Storage

There are several non-volatile solid-state technologies which hold promise as mass storage devices. Among the advantages of these solid-state technologies are size, weight , power, and speed. Vertical Bloch line, holographic, and magnetoresistive RAM offer the promise of near term improvements in cost and storage density [20]. Other technologies with long range prospects are Josephson junction, and persistent spectral hole burning.

Josephson junction technology uses a superconductor material to store data. This technology was heavily researched in the 1970's and early 1980's for extremely fast computer technology. The switching component is based on the magnetic flux in a loop. The memory access time for this technology is on the order of 35 picoseconds (one trillionth of a second) although the fabrication technique is similar to that for semiconductors and share their limits.

Persistent spectral hole burning uses impurities in solids at liquid helium temperatures and high resolution spectroscopy. The storage material is a polymer or glass which is transparent in visible light, while the impurities have an absorption band in the visible wavelength region. If a narrow beam of light from a laser is focused onto a small region of the structure, a resulting change in the electronic states of the molecules absorbing at that wavelength modifies the absorptivity at the laser wavelength producing a hole in the absorption band [21].

2.3.5 Summary

Differing rates of technological change among the different types of devices that make up a computer system have led to the so-called I/O gap. One author's attempt to quantify the I/O gap is shown in Figure 3-4. Closing the I/O gap is a major objective of advanced technologies that are

poised for deployment in the early 21st century. It appears to be possible that ultimately, solid-state technologies will replace moving disk systems. Competing alternatives include ultra-small transistors, holographic storage, and molecular-scale devices. The differing rates of change suggest that solid-state electronic technology will overtake Winchester disk technology in terms of surface density. The cost per byte of storage will determine the applications which adopt solid-state storage over magnetic media.

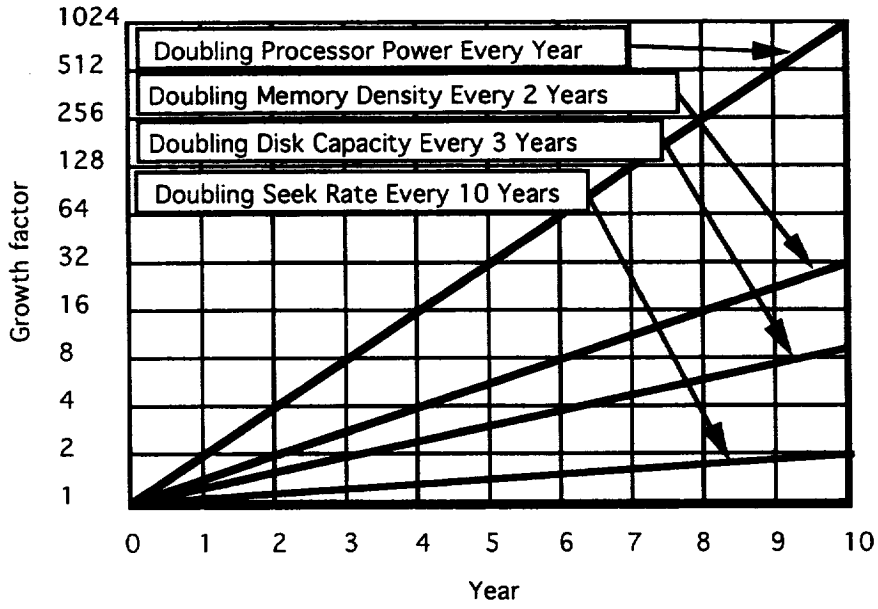


Figure 3-4. Hypothetical Effects of Dissimilar Doubling Rates Over a Decade

Source: Katz, R., et al., "Disk System Architectures for High Performance Computing", *Proc. IEEE*, Vol. 77, No. 12, 1989, pp. 1842-1858

Projections for magneto-optical, holographic, and molecular storage are more difficult to make. It seems reasonably certain that mass-market products using holographic volumetric storage will make an appearance by the turn of the century, but will not replace older methods for many years beyond. The ability to do random memory access at high speed in a volumetric storage medium will represent a major breakthrough because of the enormously greater density of volumetric media.

Molecular-scale devices, while having vastly higher theoretical surface densities than optical or magnetic surface storage, may be limited in practice by the connection problem. If molecular-scale technology moves in a direction of using optics for interconnection, the distinction between magneto-optical, holographic, and molecular technologies could be viewed as a question of the choice of medium, since they all involve modification of material states on a microscopic scale by means of light. The corollary is that they are all limited by the optical transducer technology.

3. RESULTS

3.1 IMPACT OF STORAGE TECHNOLOGY ON NASA

The implementation of a technology such as holographic storage would eliminate the distinction between present-day primary and secondary (mass) storage, since it has the speed and random accessibility of RAM and the capacity of magnetic hard disks. This would greatly boost the performance of all I/O-intensive applications, including high-speed networking, since

networking requires large buffers for staging file transfers. The definitions of primary and secondary storage will, however, evolve over time, so that there will undoubtedly always be media that sacrifice speed for higher density and will serve as secondary storage.

Another obvious advantage in replacing magnetic disks for large storage arrays with either semiconductor RAM or ORAM will be the gain in reliability. This will result in large operations and maintenance cost savings for large buffers such as may be required in level-zero processing of 300 Mbps data streams from projects such as the Earth Observing System.

High-speed random access mass storage is also needed for improving database access. Random access and low latency times are synonymous with fast database search. This capability is lacking in present day large databases. Fast transfer rates from mass storage are also beneficial for long-distance file transfers, a capability that will be in increasing demand in the missions' Earth Observing Mission period.

3.2 TECHNOLOGY ROADMAP

The technology roadmap for storage technology describes a hypothetical timeline for major system impacts and changes over the next 25 years.

1996–2000: In this period, prototypes of petabyte (10^{15} bytes) archival systems will be built using conventional magnetic and optical storage media. Magnetic tape will continue to be the medium of choice for high-speed I/O and volumetric storage. These systems will be large and costly because existing technology is not well-suited to this scale of storage. Optical storage systems will begin to compete with magnetic storage devices.

2000–2010: In this period, there may be the first commercial offering of non-contact, non-mechanical mass storage systems. This may include holographic volumetric devices and, perhaps, non-volatile RAM with mass-storage densities. This development would greatly increase the reliability and reduce the operating costs of large archival systems.

2010–2020: During this period, the first commercial offering of molecular scale storage devices may appear. Optical research will pay off in its application to molecular storage. Optical methods will most likely be used for I/O to molecular storage media. There will also be a possible use of neural-style computers in which the memory and the computational elements are indistinguishable.

3.3 FEASIBILITY AND RISK

Implementation risk can be reduced through the use of commercial technology. Some of the opportunities which take advantage of commercial technology developments include television over fiber cable to home subscribers, medical imaging, and electronic document storage, all of which require massive amounts of storage.

There appears to be little risk in magnetic disk, optical disk, and solid-state storage technology development. The major companies in the computer field have already announced laboratory prototypes of storage technologies for deployment at the turn of the century. Apparently, the delay in getting these technologies to the field is primarily one of developing mass-production manufacturing techniques. It would therefore appear that small quantities of these devices could be obtained much earlier, if cost were no object.

The advanced and exotic technologies discussed in this report represent considerably greater risk. Holographic storage is still limited by the speed of the spatial light modulator required in its operation. Spatial light modulator technology has been slow to develop. Given its many applications, not only in optical storage but also in optical computing, it would appear that there would be a high payoff to accelerating the development of spatial light modulators. The

development of a volumetric random access storage technique would be truly revolutionary and would greatly improve the performance of high data rate networks and remote database access. It is too early to assess the development risk of the exotic technologies mentioned.

References

1. Weber, S., "Variety Extends DRAM's Reach", *EE Times*, August 5, 1994, p. 58.
2. Katz, R., et al., "Disk System Architectures for High Performance Computing", *Proc. IEEE*, Vol. 77, No. 12, 1989, pp. 1842–1858.
3. Wood, R., "Magnetic Megabits", *IEEE Spectrum*, May 1990, p. 32.
4. Merz, J., "Rigid Disk Systems", National Media Laboratory Home Page.
5. Carr, E., "RAID Gets High Grades for Capacity and Performance", *Network Computing*, October 15, 1995, p.92
6. Advertisement by Pinnacle Micro, *BYTE Magazine*, November, 1995.
7. Costlow, T., "In the Race for more Capacity, Hard-disk, Tape and Optical Devices must Turn into...Superdrives", *OEM*, October 1, 1995, p. 62.
8. "War and Peace on a Pinhead", *Network World*, August 10, 1992, p. 2.
9. Gelsinger, P., et al., "Microprocessors Circa 2000", *IEEE Spectrum*, October 1989, p. 43.
10. Geppert, L., "Solid State", *IEEE Spectrum*, January, 1995, p.35.
11. Andrews, D., "IBM Unveils Tiny Little Transistors", *BYTE Magazine*, August, 1992, p. 28.
12. Andrews, D. L., "Rambus's New Memory Architecture Could Put More Video PCs on Desktops", *BYTE Magazine*, April 1992, p. 26.
13. Kao, T., "Rambus is the Choice for Bandwidth", *EE Times*, November, 13, 1995, p. 107.
14. Parish, T., "Crystal Clear Storage", *BYTE Magazine*, November 1990, p. 283.
15. "Tamarack Reckons it has Holographic Memory Taped", *PowerPC News*, August 5, 1994.
16. Lorentz, R., "Holographic Storage Review", National Media Laboratory Home Page.
17. "Storage Technology Looks to Holographic Storage", *PowerPC News*, May 23, 1995.
18. Clarkson, M., "The Quest for the Molecular Computer", *BYTE Magazine*, May 1989, p. 268.
19. Ackerman, R., "Silicon Carbide Advances Promise Versatile Circuits", *Signal*, June 1992, p. 39.
20. Ashton, G., "NML Solid State Memory Study", National Media Laboratory Home Page.
21. Callaby, D., "Persistent Spectral Hole Burning", National Media Laboratory Home Page.

CHAPTER 4. PHOTONICS TECHNOLOGY TRENDS

SUMMARY

Because of the wide bandwidth capability of light signals, photonic device technology will continue to be developed for both guided-wave and free-space communications. It is expected that in the 21st century, space communications will follow the lead of commercial guided-wave networks, which are in the process of being converted to optical frequencies.

Optical methods will increasingly be used for storage, such as holography for high-density, high-speed volumetric storage. Another key application of optical technology is in high-speed switching of communication signals. Self-electro-optic-effect devices (SEEDs) hold the promise of ultimately being able to switch signals at terabit per second rates.

Optical methods will also be used more frequently in computers, first for backplane communications for chip-to-chip and on-chip communications, and later for the CPU itself. Optical computing is still in an immature state. Its success depends on the development of high-speed, low-cost spatial light modulators, currently still in the research phase. There is a recent trend toward hybrid optical/electronic computing which permits a reduction in required optical input energy coupled with increased processing flexibility.

Finally, optics will come into its own for connection-intensive architectures, such as neural networks, and for interfacing to molecular-scale electronics.

1. INTRODUCTION

Photonics is the application of light to signal processing, computing, switching, and communications. Research in the application of photonics to these areas has been ongoing for more than 30 years. The first fruitful application to find widespread use was fiber-optic communications, but new applications are on the horizon. Photonics has a complementary role with respect to electronics. For certain specialized applications in signal and data processing, it offers orders of magnitude more potential throughput than electronic systems. The Department of Defense, and the Air Force in particular, have decided to place emphasis on photonics for weapon systems of the 21st century [1]. Besides their huge bandwidth, photonic devices have advantages for aerospace applications, such as a much reduced power dissipation requirement, and greater immunity to electromagnetic interference (EMI).

The classical optical processing element is the lens, which can perform a 2-dimensional Fourier transform (FT) using coherent laser light. The FT is accomplished by virtue of the fact that a lens causes a plane wave (a single spatial frequency) to be focused to a point. The time to perform the FT is the time it takes for the light to traverse the distance between the object plane and the image plane of the lens, about 1 nanosecond per foot. The key factor is the 2-dimensional nature of the mathematical function and its transform—all points of the 2-dimensional field are processed in parallel. Optical systems are inherently massively parallel processors. Furthermore, no interconnection wiring is required. Theoretical performance approaches 10^{21} multiplications per second.

Unfortunately, engineering obstacles prevent the throughput rate from reaching the theoretical limit. The throughput is limited by the time it takes to form an input image and read out the output image. This is called the frame time. These functions are performed by electro-optical devices whose electronic bandwidths fall far short of the ideal. The input/output (I/O) devices are sequentially scanned, like the raster scan of a cathode ray tube. This sequential operation negates the inherent parallelism of the optical channel. Typical sequential I/O devices used in the past had characteristics similar to video devices, frame times of milliseconds, and bandwidths in the megahertz range. The situation is thus analogous to that of fiber optics—the inherent bandwidth of

the optical channel is orders of magnitude greater than that of the I/O electronics. But this situation is changing. Advances in fast spatial light modulators have the potential to break the I/O logjam to the point where optical techniques may be able to surpass digital electronic computing. Other recent advances, such as surface-emitting laser diodes, integrated planar optical and electronic components on the same chip, holographic optics, and volumetric holographic storage, bring the goal of true optical computing closer to reality [2].

2. LONG-RANGE TECHNOLOGY TRENDS

2.1 ANALYSIS OF TRENDS

The capabilities of photonic devices for switching applications in comparison with electronic ones are shown in Figure 4-1 [3]. This shows the potential data throughput of these devices plotted against the connectivity, or potential numbers of interconnects per device. This plot captures a figure of merit—the time-space bandwidth consisting of the throughput multiplied by the connectivity—that measures suitability for switching, since connectivity refers to the number of parallel inputs and outputs. This plot also indirectly indicates the advantages of optical systems for parallel processing, since parallel processors require a high degree of interconnectivity between elements. Also, the devices in the figure, such as the optical logic etalon and the SEEDs, can be used as computing elements as well as switching elements. The SEED was used as the computing element in Alan Huang's optical computer built at AT&T Bell Laboratories [4, 5].

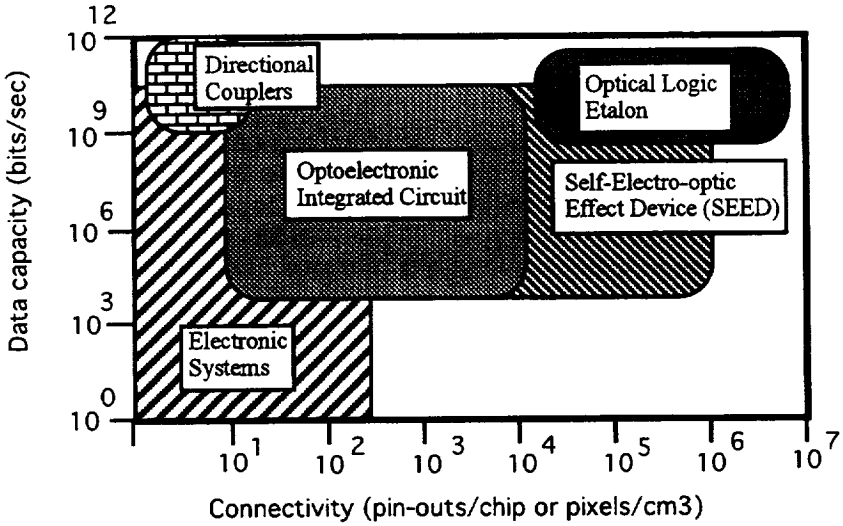


Figure 4-1. Photonic Device Capabilities

Source: Berra, P.B., et al., "The Impact of Optics on Data and Knowledge-Based Systems", *IEEE Trans. KDE*, March 1989

2.1.1 SEED Arrays

The SEED is a bistable device first built by AT&T Bell Laboratories in the mid-1980s. These devices can both receive and generate photons, supporting a number of logical functions that would be difficult to realize with silicon electronics. Although a general purpose optical architecture is still years away, AT&T Bell Laboratories in 1994 developed 4x4 bistable SEED arrays that can be concatenated to produce a high-density switching network. Driven by the high-speed switching requirements of ATM networks running at 622 Mbps, Bell Labs has continued to seek improvements in optical switching. They are now building a switch that will have optical-interconnection density of 16,000 inputs/outputs per square centimeter. This compares with

competitive electronic technology that achieves only one-tenth the interconnection density. Ultimately, all-optical switches may be able to route data at rates up to terabits per second [6, 7].

2.1.2 Optical Computing

Optical computing offers inherently parallel processing and is promising for realizing image-based multimedia computing. For optical computing, the key parameter is the performance of spatial light modulators, the limiting element in most optical computers. The figure of merit is the time-space bandwidth. In the past, the performance of spatial light modulators has made them competitive with the best electronic supercomputers in terms of number of operations per second [8]. However, electronic computers are advancing very rapidly, making it difficult for optical device research to keep pace. It is also very difficult to project the future rate of progress for spatial light modulators. It may be supposed that a time-space bandwidth of 10^{12} will be achieved by the year 2000.

To take advantage of the ultimate potential of optical computers, researchers will have to eliminate the sequential scanning that occurs in the spatial light modulator. Ideally, to do this, electronic-to-photonic conversions would be eliminated, and images would be communicated in parallel directly to a 2-dimensional sensor array mated to the spatial light modulator input. Advances in 3-dimensional device integration and packaging would be required for this.

Because of the difficulty of constructing an all-optical computer, a recent trend is away from purely optical systems toward hybrid systems that balance the use of optics and electronics. Such an approach permits the optical input energy to be decreased and electronic processing allows greater complexity at each smart pixel in the optical array. Bell Labs is considering the production of custom smart pixel arrays that would permit researchers to integrate as much electronic complexity as they desire with the SEED optical elements [9].

Despite the advances that have been made, optical computing is still in a rather immature state. It is anticipated that optical computers will continue to be used primarily for specialized applications for at least the next decade, but that general-purpose optical computers will not reach the commercial market within that period. Special-purpose analog optical computers will be used whenever they can outperform electronic computers by at least a factor of 10, or when the required connectivity is impractical for electronics. A general-purpose digital optical computer with a performance of 1 tera-operations per second was built by OptiComp Corp. for SDIO [10], but it is not clear yet whether this computer is transferable to commercial applications. AT&T has a rivalry with IBM to prove that photonics is a valid approach, and has both the incentive and financial resources to prove it. Looking ahead 20 years, it appears likely that optical computers or hybrid optical-electronic computers will come into their own when applied to advanced connection-intensive architectures for which there do not as yet appear to be any electronic equivalents because of the interconnectivity problem.

2.1.3 Holographic Storage

Holographic storage is a technique for storing information throughout a volume (i.e., in three dimensions) rather than in a planar array. The relative performance of holographic storage compared to other technologies is shown in Figure 4-2 [11]. The principles of this technique have been known for several decades and recent advances show that commercial development is possible, perhaps within 10 years. A commercially available holographic storage device would be a revolutionary breakthrough in storage, since, as can be seen from the figure, it combines the volumetric density of magnetic tape with the access times of RAM. It is expected that holographic storage will be commercially developed within the next 10 years. Further discussion is provided below in section 2.3.

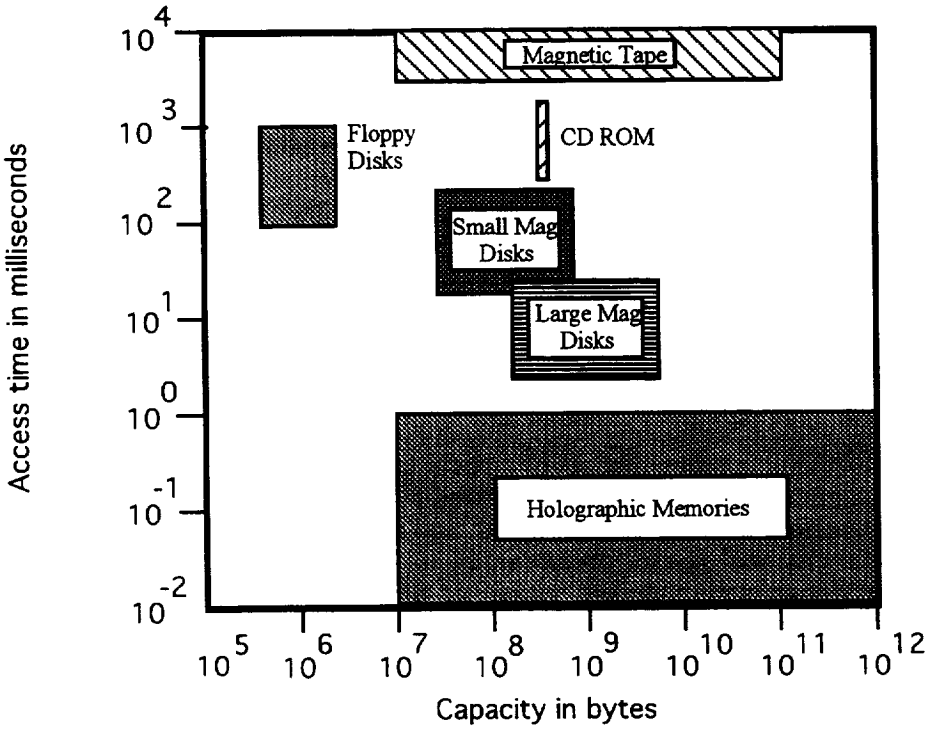


Figure 4-2. Comparison of Secondary Storage Technologies

Source: Berra, P.B., et al., "The Impact of Optics on Data and Knowledge-Based Systems", *IEEE Trans. KDE,* March 1989

2.2 FUTURE APPLICATIONS

We mentioned above two key advantages of optical methods—parallelism and connectivity without wires. Optical systems have a higher potential bandwidth than electronic ones, and the bandwidth is not degraded by resistance, capacitance, or inductance. The lack of properties characteristic of electronic conductors also means no crosstalk or EMI, at least not to the degree found in electronics. Historically, parallelism and connectivity were paramount. The first optical processors were used for image reconstruction for synthetic aperture radar (SAR) at a time when electronic computers could not compete. In the future, other properties may be important. Ballistic missile defense researchers are interested in size, weight, power, robustness, and radiation hardening for space-borne applications. Optics can improve these characteristics. On the other hand, photons have certain disadvantages when compared with electrons. Their lack of charge makes it difficult to construct all-optical modulators—essential elements in computer systems. Electronic computers are based on the transistor, a three-port non-linear device with gain. Such a device is much more difficult to build for photons, although they are being built today.

Currently, optics are used in computers mainly for interconnection. Optics are also used in certain highly specialized analog signal processing applications, such as real-time spectral analysis. In the future, it is the hope of many researchers that an all-optical supercomputer will be built that will far surpass electronic ones. However, this latter goal is at least two decades away, and it is as yet uncertain that optics will ever be able to compete with electronics. Electronic computers have proven to be more flexible, convenient, accurate, and cost-effective, even in an application in which optics once excelled, such as SAR processing, which is now done again by an electronic computer. For the foreseeable future, optical computing will be targeted at special algorithms in which optics excel.

Optical methods are emerging in electronic computers. Fiber optics for communications is a well-established application, but fibers are also being used for backplane interconnections in some computers, such as AT&T's telephone switches. Other optical devices, such as holographic crossbar switches and planar waveguides embedded in printed circuit boards, are being developed for backplane interconnection of multiple processors in parallel supercomputers. Such applications, however, may not be commercially available for five or 10 years [9]. In some cases, this intra-computer network is implemented by means of a free-space broadcast technique [12]. In the near future, optical crossbar switches will be very important in the high-end networking applications, such as supercomputer-LAN file transfers and digital video multimedia [13]. Optical communication between chips is also being explored to speed up the data rates and reduce the number of I/O pins. On-chip, optical broadcast methods are being developed that use planar diffractive optics to distribute clock signals without skew to all parts of the chip.

Analog computing using light is being done for certain specialized problems in which the algorithm is fixed, the throughput is high, but the accuracy required is low, and the number of sequential steps is also low [14]. An example is high-speed Fourier transforms for instantaneous RF spectrum analyses for electronic warfare. Another example is fast parallel image analyses, such as for pattern recognition and automatic target recognition. Since images are 2-dimensional intensity functions, they are well-suited to the 2-dimensional mathematical properties of optical systems. Other applications include linear and non-linear algebraic equation-solving, eigenvector problems, matrix inversion, and linear programming. [15]

2.3 EMERGING TECHNOLOGIES

2.3.1 Connection-Intensive Architectures

The 2-dimensional property of optical systems is well-matched to the massive interconnectivity requirements of neural network architectures. Electronic hardware is far behind the needs of neural network algorithms. Researchers such as Demetri Psaltis and Nabil Farhat are working on optical implementations of neural networks that could provide a quantum leap ahead [16]. And, neural networks are now seen as the preferred front-end analysis tool in "random problem" [17] domains, such as image recognition, image feature extraction, and associative memory. Neural network execution of the simulated annealing algorithm is one of the most important applications. Simulated annealing is a fast algorithm for reaching approximate solutions to the traveling salesman problem and its variants. When the Jet Propulsion Laboratory compared a neural net approach with a classical procedure programmed on an Intel Hypercube for such a problem, the neural net solution had an accuracy of 0.999 while taking only 0.0001 of the time required by the Hypercube [18].

Demand for neural network-based computation will increase dramatically in future, but currently, neural networks are typically implemented in a serial algorithm, which limits performance. In future, the algorithms will be executed in parallel hardware to take better advantage of the inherent parallelism and speed of the neural net architecture. For large neural networks (containing thousands of nodes), optical computing methods may provide the best way to efficiently achieve massive interconnectivity at high speed [19]. At very high levels of connection density, optical methods may be the only feasible approach, because it becomes impractical to build physical wire interconnects. For example, a fully connected network of 1E+6 nodes has 1E+12 interconnections.

There is also mutual benefit in coupling optics to molecular-scale computing. Molecular-scale electronic devices are so small, e.g., four nanometers [20], that metallic interconnections are too difficult to fabricate. Optical interconnection is one logical alternative. In fact, molecular computing devices such as the one discussed in Clarkson [20] bear a strong resemblance to optical computing devices.

2.3.2 Optical Storage

Perhaps the earliest high-payoff application of photonics will be holographic storage. By the end of the decade, practical working devices will be available that have the mass storage capacity of high-speed tape recorders (terabits), have the access time of solid state RAM, have no moving parts, and will fit into the space of a Winchester disk drive. In comparison with a 2-dimensional optical storage medium, which has a theoretical storage density of $4x10^8$ bits per square centimeter, 3-dimensional holographic storage has a density of $8x10^{12}$ bits per cubic centimeter, a factor of 20,000 difference [21].

Holographic storage would also have a major impact on high-speed database search, since it combines the mass-storage capabilities of a magnetic disk with the fast access times of solid-state RAM. The I/O gap between mass storage devices and computers is, in effect, a bottleneck in large database systems. This bottleneck could be overcome by a fast volumetric storage technique using the holographic approach. Interestingly, holographic memories also have an associative memory property deriving from the holographic principle itself. According to many experts, associative memory or content-addressable memory is the preferred mechanism for high-speed database look-up systems since memory contents do not have to be searched. [11]

3. RESULTS

3.1 IMPACT OF PHOTONICS TECHNOLOGY ON NASA

Optical processing devices have general applicability to NASA problems as well as being well-matched to particular applications. In a general sense, optical methods will enhance computer technology, a good thing for NASA in light of its high data rate processing requirements. Particular specialized applications include ground-based processing of very high data rate multiplexed communications, an application well suited to the high parallelism of optics. High-speed calculations involving matrix mathematics, such as Reed-Solomon decoding or compression algorithms using 2-dimensional transforms, would be suited to optical computers. Another example would be an ultra high-speed dynamically re-configurable crossbar switch for packet switching and synchronous frame demultiplexing. Optical or hybrid electro-optical processors may be able to reduce size, weight, and power for computers, storage, and communications components onboard spacecraft. Imaging instruments, the primary sources of high-data-rate communications requirements, could especially benefit as could control systems that require the instantaneous adjustment of hundreds of system parameters in response to like numbers of sensor inputs. Using advanced processors onboard spacecraft could greatly reduce the bandwidth of communications and radically alter the architecture of a space-to-ground communications network.

The main impact of optical methods on processing, storage, and communications will be increased performance and miniaturization, along with reduced weight and power. Of direct interest is the miniaturization afforded by optical communications elements. For high-bandwidth communication links requiring high directionality, antenna size is greatly reduced by using optical or near-optical frequencies. A consequence of an optical space-to-ground network architecture is that the network of ground stations must be geographically dispersed so that there will always be a high probability of finding a ground station that has a sky clear of clouds. Studies have shown that only a few such ground stations would be required.

Optical computing may be a means of building small, dedicated computers for special algorithms requiring very high performance, especially for real-time optimization problems, such as simulated annealing algorithms executed on neural networks. While it is not clear that such algorithms have direct applicability to concerns of NASA, miniaturization of computers with supercomputer-class performance is of direct concern because of the leverage gained from processing onboard spacecraft. It would be very advantageous to have ultra-miniaturized computers onboard spacecraft capable of performing image compression in real time. Photonic computers may not be competitive with electronic ones in the next 10 to 20 years for general-purpose

computing, but the comparison is very algorithm-dependent. Algorithms for which massive connectivity and parallelism are appropriate may give the edge to photonics.

Holographic storage will be a method well-matched to the requirements of onboard mass storage, since besides the density and access-time advantages, there are no moving parts, giving high reliability. Very high storage capabilities onboard spacecraft open up new options, not only for supporting onboard processing at very high throughput rates, but also for store-and-forward and burst-transmission communications architectures. A more exotic use of holographic storage is in the form of associative memory as a way to attack the problem of high-speed database search. High-speed database search and retrieval is one important aspect of the implementation and successful exploitation of the data from the Earth Observing System. Holographic storage also may have significant impacts on spacecraft design. Spacecraft may no longer be tethered to relay satellites that are needed to provide continuous coverage, since buffer times could be longer and readout rates faster than real time. At the least, it will open up new options for spacecraft weight reduction and for store-and-forward and burst-transmission network architectures.

There are additional benefits to photonic devices that are of use to NASA for space applications that have less value to the mass-market, such as non-mechanical interconnects (for enhanced reliability) and immunity to EMI and charged particles. According to some military project managers, these may turn out to be as important as processing power [22].

3.2 TECHNOLOGY ROADMAP

The following is a technology roadmap, or hypothetical timeline, that summarizes the major developments in photonics and how they will be applied over the next 25 years.

1995–2000: Optical computing will remain in the research stage, with a relatively small number of optical computers deployed for specialized applications. Higher capacity optical switches for telecommunications will become available. Photon-tunneling scanning probe microscopy may be applied to commercial storage products while research and development of holographic storage continues. Spatial light modulators with high performance and low cost may be available for use in specialized optical computers.

2000–2010: Optical computing units will be applied to hardware neural network computers. There may be an attempt to market general purpose optical computers, although this is very uncertain. High-performance spatial light modulators will become widely available. Holographic storage products will be introduced while other optical storage technologies mature.

2010–2020: Photonics will be successfully applied to molecular computing and storage, bringing about orders-of-magnitude higher performance.

3.3 FEASIBILITY AND RISK

Practical holographic storage faced difficult implementation problems during the 1970s in both the materials science aspect and the spatial light modulator aspect. The materials problem seems to have been overcome in the 1980s, while the spatial light modulator remains the weak link. A business-as-usual attitude seems to be standing in the way of aggressive commercial development of this technology. Industrial research consortia, such as the Microelectronics and Computer Technology Corporation, that have been working on the technology have had problems migrating their efforts to sponsor companies for commercial development. Still, the technical merits of holographic storage are convincing. Potential competition is looming on the horizon from advanced solid-state concepts, such as single-electron transistors, quantum-well devices, and other quantum-effect techniques, as well as from molecular-scale storage concepts. Optical storage is well ahead of these in terms of maturity, though. And, optics is seen as the preferred method of interconnection with molecular-scale devices. A major commercial application for very high-density storage will be digitized HDTV over fiber-optic cable. Huge amounts of high-speed storage are required for this

application. This application should be watched for opportunities for risk reduction through the use of commercially developed hardware.

While all-optical computers will be developed for specialized applications, it is not clear that general purpose optical computers will become commercially viable. It is more likely that hybrid optical/electronic processors will compete with traditional electronic and newer molecular computing devices.

References

1. Rhea, J., "Rome's Three Keys to the Future", *Air Force Magazine*, November 1991, p. 36.
2. Neff, J., et al., "Two-Dimensional Spatial Light Modulators: A Tutorial", *Proc. IEEE*, Vol. 78, No. 5, May 1990, p. 826.
3. Berra, P. B., et al., "The Impact of Optics on Data and Knowledge-Base Processing Systems", *IEEE Trans. on Knowledge and Data Eng.*, Vol. 1, March 1989, p. 111.
4. Huang, A., "Architectural Considerations Involved in the Design of an Optical Computer", *Proc. IEEE*, Vol. 72, No. 7, July 1984, p. 780.
5. Prise, M. E., "Optical Computing Using Self-Electro-Optic Effect Devices", *Proc. SPIE*, Vol. 1214, January 1990, p. 3.
6. Brown, C., "A Photonic Future," *OEM*, April 1, 1994, page 82.
7. Derman, G., "Large Systems Forego Electrical Connection for Opticals," *EE Times*, October 9, 1995, page 55.
8. Fisher, A. D., and J.N. Lee, "Development Issues for MCP-based Spatial Light Modulators", in *Proc. OSA Top. Meet. Spatial Light Modulators Appl.*, Vol. 8, S. Lake Tahoe, NV, June 1988, pp. 60–63.
9. Clarke, P., "AT&T Plan to Aid Optical Computing," *EE Times*, August 29, 1994, page 4.
10. Shimazu, M., "Optical Computing Comes of Age", *Photonics Spectra*, November 1992, p. 66.
11. Berra, P. B., et al., "Optical Database/Knowledge Base Machines", *Applied Optics*, Vol. 29, January 10, 1990, p. 195.
12. Neff, J., "Photonics Inside the Computing Mainframe", *Photonics Spectra*, June 1992, p. 145.
13. Hinton, H. S., "Switching to Photonics", *IEEE Spectrum*, February 1992, p. 42.
14. Mahowald, M. A., and Carver Mead, "The Silicon Retina", *Scientific American*, May 1991, p.76.
15. Berra, P. B., et al., "Optics and Supercomputing", *Proc. IEEE*, Vol. 77, December 1989, p. 1797.
16. Farhat, N. H., et al., "Optical Implementation of the Hopfield Model", *Applied Optics*, Vol. 24, No. 10, May 15, 1985, p. 1469.
17. Abu-Mostafa, Y. B., and D. Psaltis, "Optical Neural Computers", *Scientific American*, March 1987, p. 88.
18. Kelly, C., "Trends in Computing Architectures", internal SAIC memorandum, 1992.
19. McAulay, A. D., "Optical Computer Architectures for Supervised Learning Systems", *IEEE Computer*, May 1992, p. 72.
20. Clarkson, M. A., "The Quest for the Molecular Computer", *BYTE Magazine*, May 1989, p. 268.

21. Stein, R. M., "Terabyte Memories with the Speed of Light", *BYTE Magazine*, March 1992, p. 168.

22. Bell, T. E., "Optical Computing: A Field in Flux", *IEEE Spectrum*, August 1986, p. 34.

ANNEX 1. ACRONYMS AND ABBREVIATIONS

ANSI	American National Standards Institute
AOTF	Acousto-optic tunable filter
ARPA	Advanced Research Projects Agency
ATM	Asynchronous transfer mode
BMDO	Ballistic Missile Defense Organization
CCSDS	Consultative Committee on Space Data Systems
CISC	Complex instruction set computer
CMOS	Complementary metal-oxide semiconductor
CNRI	Corporation for National Research Initiatives
CPM	Colliding-pulse, mode-locked (laser)
CPU	Central processing unit
dB	Decibel
DEC	Digital Equipment Corporation
DOD	Department of Defense
DOE	Department of Energy
DOS	Digital operating system
DRAM	Dynamic random access memory
ECL	Emitter-coupled logic
EMI	Electromagnetic interference
E/O	Electro-optic, electro-optical
EOS	Earth Observing System
FLOPS	Floating point operations per second
FM	Frequency modulation
FT	Fourier transform
Gbps	Gigabits per second (billion bits per second)
giga-FLOPS	Billion floating point operations per second
GHz	Gigahertz (billion cycles per second)
HDTV	High-definition television
HIRIS	High resolution imaging spectrometer
HTML	Hyper text markup language
IEEE	Institute of Electrical and Electronic Engineers
I/O	Input/output
IR	Infrared
JPL	Jet Propulsion Laboratory
kbps	Kilobits per second
km	Kilometer
LED	Light-emmitting diode
LEO	Low earth orbit
Mbps	Megabits per second
Mbytes	Megabytes per second
MCC	Microelectronics and Computer Technology Corporation
MCM	Multi-chip module
MHz	Megahertz (million cycles per second)
MIMD	Multiple-instruction, multiple-data
MIPS	Million instructions per second
MODIS	Moderate resolution imaging spectrometer
MPP	Massively parallel processor
NASA	National Aeronautics and Space Administration
NCSA	National Center for Supercomputer Applications
nm	Nanometer (one billionth of a meter)
NREN	National Education and Research Network

NSA	National Security Agency
NSF	National Science Foundation
ORAM	Optical random access memory
OSC	Office of Space Communications
PC	Personal computer
RAID	Redundant array of inexpensive disks
RAM	Random access memory
RF	Radio frequency
RISC	Reduced instruction set computing
SAIC	Science Applications International Corporation
SAR	Synthetic aperture radar
SDI	Strategic Defense Initiative
SDIO	Strategic Defense Initiative Office
SEED	Self-electro-optic effect device
SIMD	Single-instruction, multiple data
SMP	Symmetric multi-processing
SRAM	Static random access memory
Tbps	Terabits per second (thousand billion bits per second)
TDRSS	Tracking and Data Relay Satellite System
tera-FLOPS	Thousand billion floating point operations per second
THz	Terahertz (thousand billion cycles per second)
UNIX	(de-facto standard multi-tasking operating system for workstations)
VLIW	Very long instruction word
VLSI	Very large scale integration
WORM	Write-once read-many

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1996	Final Contractor Report

4. TITLE AND SUBTITLE

Technology Directions for the 21st Century
Volume I

5. FUNDING NUMBERS

WU–315–90–81
DAEA32–96–D–0001

6. AUTHOR(S)

Giles F. Crimi, Henry Verheggen, William McIntosh, and Robert Botta

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Science Applications International Corporation
1710 Goodridge Drive
McLean, Virginia 22102

8. PERFORMING ORGANIZATION REPORT NUMBER

E–10240

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135–3191

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA CR–198478

11. SUPPLEMENTARY NOTES

Project Manager, Denise S. Ponchak, Space Electronics Division, NASA Lewis Research Center, organization code 5610, (216) 433–3465.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Categories 32, 17, 33, 60, and 76

This publication is available from the NASA Center for AeroSpace Information, (301) 621–0390.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

For several decades, semiconductor device density and performance have been doubling about every 18 months (Moore's Law). With present photolithography techniques, this rate can continue for only about another 10 years. Continued improvement will need to rely on newer technologies. Transition from the current micron range for transistor size to the nanometer range will permit Moore's Law to operate well beyond 10 years. The technologies that will enable this extension include: single-electron transistors; quantum well devices; spin transistors; and nanotechnology and molecular engineering. Continuation of Moore's Law will rely on huge capital investments for manufacture as well as on new technologies. Much will depend on the fortunes of Intel, the premier chip manufacturer, which, in turn, depend on the development of mass-market applications and volume sales for chips of higher and higher density. The technology drivers are seen by different forecasters to include video/multimedia applications, digital signal processing, and business automation. Moore's Law will affect NASA in the areas of communications and space technology by reducing size and power requirements for data processing and data fusion functions to be performed onboard spacecraft. In addition, NASA will have the opportunity to be a pioneering contributor to nanotechnology research without incurring huge expenses.

14. SUBJECT TERMS

Moore's Law; Semiconductor technology; Computing technology; Storage technology; Photonics technology

15. NUMBER OF PAGES

55

16. PRICE CODE

A04

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

National Aeronautics and
Space Administration

Lewis Research Center
21000 Brookpark Rd.
Cleveland, OH 44135-3191