Chapter 15 The Past, Present, and Future of Computing

15.1 Introduction

This chapter closes out our introduction to computing by stepping back and taking a look at its history, seeing how we arrived at where we are today, and then looking forward to what the future may hold.

Section 15.2 traces the evolving nature of human-computer interaction, from its early roots through present day interface technologies. Section 15.3 looks at the historical rate of progress in computing technology and considers possible future increases in computing capacity. In Section 15.4, we take a look at a number of promising technologies that could dramatically improve the capabilities of computing systems, beginning with improvements to today's semi-conductor based technology, then moving on to consider more advanced technologies, such as photonic computing, bio-computing, molecular-level computing, and quantum computing. Finally, Section 15.5 concludes this chapter, and thus the book, with a number of predictions based on the projected capabilities of future computing systems.

15.2 The evolving nature of human-computer interaction

When computers first began to gain wide spread acceptance in the late 1950's and early 1960's very few people interacted directly with them. Even the people who wrote computer programs did not always have physical access to the "machine room" where the computer was maintained by the technicians and operators.[1]

Programmers typed their programs and data on machines called keypunch machines. **Keypunches** were like typewriters, except that they punched holes in paper cards. Each **punch card** could hold up to 80 characters and represented one line of a program or data file. A punch card containing the FORTRAN statement:

```
DO 211  I=1,30
```

is illustrated in Figure 15.1.

Once a programmer had punched a program, he or she would give it to an operator. **Operators** were the people who actually worked directly with the machine. The operator would take the punched cards and place them in a card reader. The **card reader** would quickly process a deck of cards, sensing where the holes had been punched and thereby "reading" the characters that had been punched on them.

---

[1] Here is a video of Bell Labs Holmdel, New Jersey computer center from 1973 – a leading edge computing center at the time – http://www.youtube.com/watch?v=HMYiktO0D64

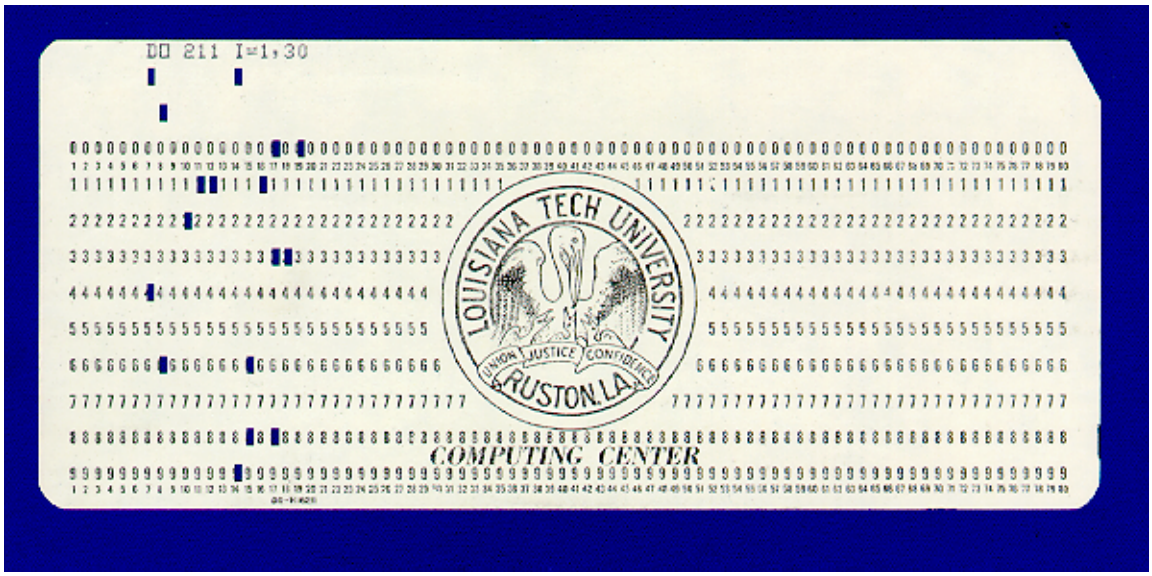

Figure 15.1: An 80 column punch card

After the program had been "read in," the computer (or Operator in the early days) would schedule the program to run when the needed resources became available. Following execution of the program, a listing of the program code together with its results would generally be printed on a line printer. **Line printers** were so named because of their ability to print an entire line (rather than a character) at a time.

As early as the 1960's a few visionaries imagined how computers might one day be used for personal information organization and communication. In 1968, Douglas Engelbart, the inventor of the mouse, gave a presentation to an audience of 1,000 computer professionals at the Fall Joint Computer Conference that illustrated a number of then revolutionary concepts that decades later would become commonplace. His demo included: word processing, video conferencing, hypertext links, and online collaboration tools enabling two people working remotely to share a desktop and mouse pointer. The entire presentation was displayed to the audience using an overhead projector. This demonstration was so far ahead of its time and so prescient in nature that it has come to be known as "The Mother of All Demos".[2] It would take decades for the kinds of applications Engelbart and his group envisioned to become practical as the cost of the hardware needed to support these apps was truly astronomical at the time.

As computers became more powerful they began to acquire the ability to handle the needs of multiple people, or **users**, at the same time. By the mid 1970's, most computers were large **mainframes** that filled entire rooms and often cost millions of dollars. People interacted with these computers predominately through character-based terminals. **Terminals** combined a keyboard with a video display device and looked much like a

[2] The complete demo – which runs 90 minutes in length is available online at: http://sloan.stanford.edu/MouseSite/1968Demo.html

personal computer does today. The difference was that the terminals of the mid 1970's lacked the ability to do any "computing" on their own. They simply acted as input/output devices for the mainframe computer – allowing people to enter their programs and data on a keyboard, rather than a keypunch, and to see the results of their computations on a display screen, rather than printed on paper. The vast majority of these terminals were limited to displaying text. They were unable to display graphical images.

A mainframe of the mid to late 1970's might have as many as 100 terminals attached to it. These terminals were often spread throughout an office, company, or college campus. Users communicated with the mainframe by issuing typed instructions called **commands**. Interfaces that supported the entry of typed commands were called **command line interfaces**. When first introduced, the command line interface was a major breakthrough. It allowed programmers to directly enter their programs and data into the computer, execute those programs, view the results, and make any needed corrections – all without the need for clunky keypunch machines, massive printers, and human operators. When personal computers began to emerge in the late 1970's and early 1980's, they too adopted the command line interface.

Despite its advantages, the command line was far from perfect. The names of the commands tended to be short and cryptic. For example, "ls" and "dir" are two common commands that request the computer to display the names of the user's files – "ls" stands for "list" and "dir" stands for "directory." To further complicate matters, most commands had a large number of options. These options were intended to increase the flexibility of the system. For example, the Unix command "ls -al" requests the computer to list all files in the current directory, including "hidden files," and to display the amount of disk space each file takes up, the last date it was modified, and which groups of users have access to it. Remembering the names of all of the commands and their options could be quite difficult, especially for users who did not interact with the system on a daily basis.

In addition to the commands themselves, the command line interface required users to remember the names of files and the directories in which they were stored. For example, a Unix command for copying a file called "exam.docx" that is stored in the "fall13" subdirectory of the "CSC100" directory, to a file called "old_exam.docx" in the "fall14" subdirectory of "CSC100", would look like the following:

```
cp   CSC100/fall13/exam.docx   CSC100/fall14/old_exam.docx
```

Entering such lengthy and cryptic commands was tedious and exacting. A single mistyped character could result in the computer doing something unexpected or generating an error – which at the very least would require the user to re-enter the (corrected) command.

While most everyone agreed that the command line interface had lots of issues, computer scientists were unsure how better interfaces could be developed. One popular approach was to have the computer system present a menu of choices. The user would select a

choice by entering a number associated with the item. While these menus worked well for systems that had only a small number of choices (say no more than eight), they became unwieldy for larger numbers of commands. The standard response to this was to group "similar" commands together in submenus. Systems with many commands often required multiple levels of menus.

Menu selection, by its very nature, was less flexible than the command line. Many people found this hierarchy of menus too restrictive in complex programs such as operating systems. Experienced users found menus to be slow. It often took much longer to walk through the menus, submenus, and sub-submenus to select a command, compared to just typing it in. This was especially true on early mainframes, which could take several seconds to respond to a user's request – due to the fact that they were serving several dozen users simultaneously. To make up for this lack of speed, the chief advantage of menus was supposed to be their ease of use. However, in large systems, even experienced users frequently found it difficult to quickly find the proper submenu that contained the command they wanted to execute.

Another approach taken by some computer scientists in the early 1970's was to try to make commands less cryptic. Their goal was to construct a natural language (such as English) interface. While most everyone agreed that understanding spoken English was out of reach of the technology of the day, many thought that computers might soon be able to handle some form of typed English. These systems never caught on for a number of reasons. First, they never really developed to the point where anything approaching a real human language could be entered. Second, the subset of English recognized often "devolved" into just another command line interface, albeit one with lengthy command names that required a lot of typing. Finally, these interface programs required a lot of resources which slowed down the response time of already slow computer systems.

The approach that eventually replaced the command line interface was developed at the Xerox **P**alo **A**lto **R**esearch **C**enter (PARC) in the 1970's as part of the Alto computer project. The genius of the PARC approach, which built off the work of Douglas Engelbart and others, stemmed from the researchers' recognition that in order for interfaces to become more useful they need to become simpler (from the user's point of view) not more complex. PARC researchers also recognized that producing an intuitive interface might require a substantial fraction of a computer's resources, but they were willing to pay this price in order for their Alto computers to be useful to non-computer specialists.

The researchers settled on the notion of representing the programs and data files that were stored in the computer as graphical images called icons. The computer screen was to be likened to a desktop on which these graphical icons could be placed. Programs could be run by selecting them with a pointing device, such as a mouse. Directories were to be represented by file folders that could be opened and closed by clicking on them. Files could be moved from one directory to another by "dragging" them from one open file folder to another.

In addition to the graphical user interface, the PARC researches also recognized that computers would be much more useful if networked together. Thus the Altos were connected together via an Ethernet, so that they could share documents and resources – including networked laser printers – and end-users could exchange email.

The interface paradigm developed by Xerox at PARC, which came to be known as the Graphical User Interface or GUI, eventually revolutionized computing, but it didn't make any money for Xerox. Although Xerox did create a "commercial" version of the Alto in 1981, called the Star (shown in Figure 15.2), few were sold. People debate the reasons for this market failure, but certainly the high cost (approximately $20,000 per workstation), slow speed, and limited capabilities contributed to the lack of market acceptance. Regardless of the lack of market success, the Xerox Star is the first commercial computer that would look "familiar" to a present-day computer user.

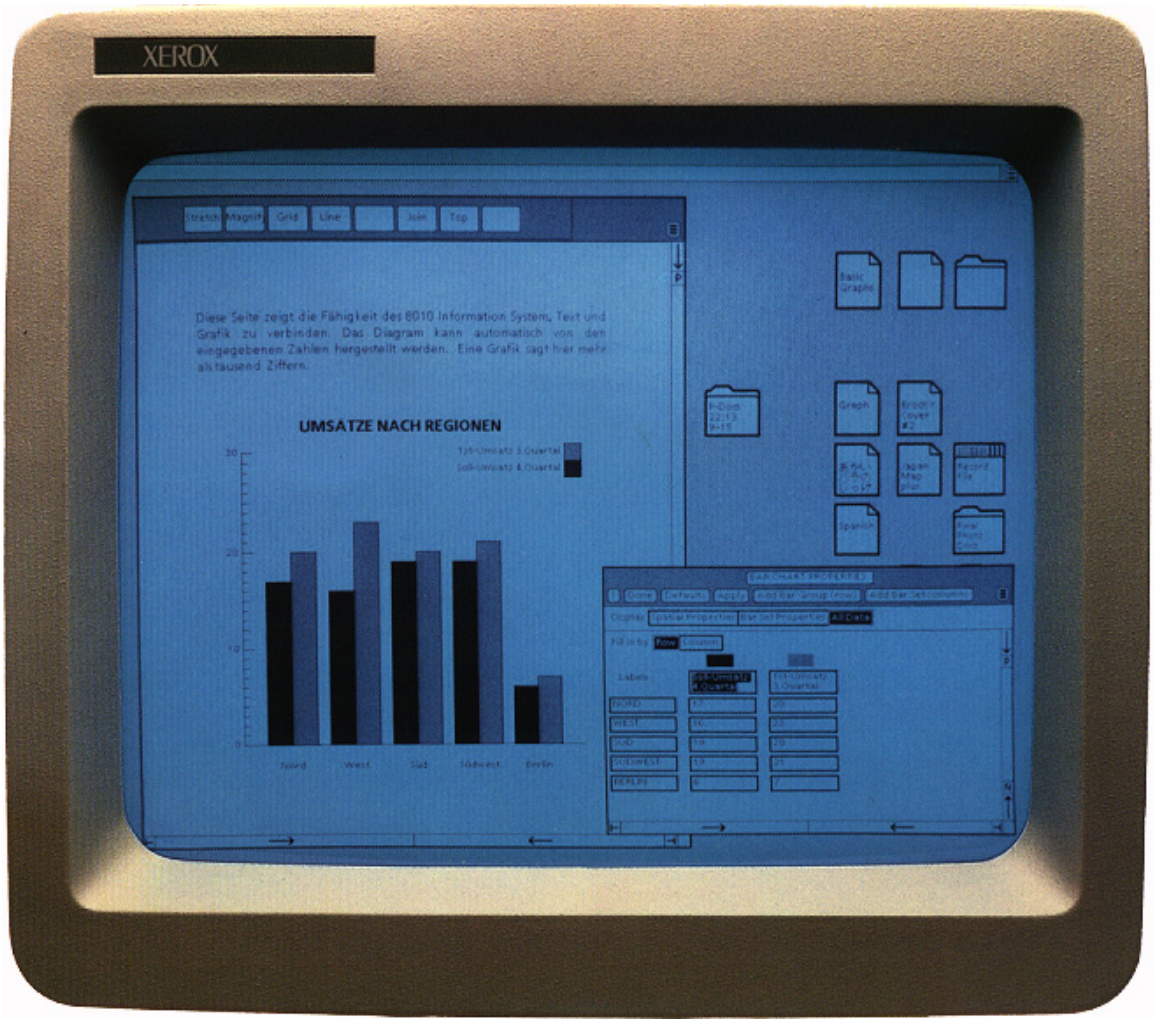

Figure 15.2: the Xerox Star – circa 1981

A turning point in the computer industry occurred when Steve Jobs, co-founder of a small "personal computer" company called Apple, was shown what the PARC team had created. Jobs immediately recognized the promise of mouse-based graphical user interfaces and put his engineers to work building an affordable personal computer based on the ideas developed at PARC.

Apple's first version of a computer based on PARC technology, Lisa, was a flop. However, Apple's second attempt, the Macintosh, first released in 1984, was a major success. Because of their intuitive nature, GUI's quickly became the human-computer interface of choice.

By the mid 1990's every major operating system, such as Windows95, MacOS, OS/2, and the various flavors of Unix, included a graphical user interface. This method of interacting with a computer proved so popular that 20 years later it is still the dominant form of human-computer interface in desktop and laptop environments.

An early version of the public Internet began to become popular in the late 1990's. During this period, end-users connected to the Internet and online services, such as AOL (America Online), using a modem and telephone. A modem was a device that allowed a computer to exchange digital information with another computer via audio tones transmitted over standard telephone lines. An end-user would dial the number of a computer system (or have the modem automatically do so). The computer being called would then 'answer the phone' and the two machines would exchange a series of tones to establish a connection. This process was called 'hand shaking' and was used to establish the speed with which data could be transferred between the two machines.

While access to the Internet was revolutionary for its time, the "Internet", as it existed in the late 1990's, would be barely recognizable compared to what we call the Internet today. There was no YouTube, Facebook, Google, Twitter, Netflix, Pandora, or Skype. In fact, many of these services couldn't have existed before the turn of the century as the data transmission speeds supported by dial up modems were *very* slow compared to today's broadband connections. How slow you ask? Well, a top-of-the-line dial up connection ran at around 56 Kbps (56,000 bits per second or about 7 Kbytes / sec) in the late 1990's. As of August 2013, www.netindex.com estimated US average download speeds of 18.7 Mbps (2.3 Mbytes / sec) – over 300 times faster than in the late 1990's. Because dial up connections were so slow, the web was mainly limited to text and lower resolution images. Streaming audio was of poor quality and considered 'cutting edge'. Steaming video was simply not practical.

This situation began to change around the turn of the century as broadband Internet connections started to become widespread. In the early days of broadband (circa 2001) there was stiff competition between phone companies and cable companies for who would provide high speed Internet access. Eventually, cable companies with their cable modems prevailed over telephone companies with their DSL (Digital Subscriber Line) technology.

By the midpoint of the first decade of the 21st century, many households had multiple computers and computer-like devices, such as game machines, so it was only natural to want to allow these devices to communicate with each other and with the wider Internet. Additionally, the growing popularity of laptops over desktops meant that people didn't want to have to plug their laptop into their cable/DSL modem in order to connect to the Internet. This situation led to the rapid adoption of wireless routers. Wireless routers establish LANs (Local Area Networks) which enable devices to connect to each other and the Internet using short range RF "radio frequency" transmitters and receivers.

Progressing in parallel with the rise of broadband Internet access and wireless Local Area Networks in the home, mobile communications were undergoing a sea change during the first decade of the 21st century. At the beginning of this period high-end professionals often carried around two separate devices: a PDA (Personal Digital Assistant) and a mobile phone. PDA's were devices that served as digital appointment calendars and note taking devices, with a modest number of additional features like calculators and clocks. They were generally devoid of any type of wireless access to the Internet and supported only low resolution black and white LCD (Liquid Crystal Displays) displays. Most mobile phones (called "cell phones" in the US) generally supported voice calls and little else.

By the midpoint of the first decade of the 21st century mobile phones began supplanting PDA's as the phones took on more and more of the features that had once been exclusive to PDA's – appointment calendars, address books, calculator functions, etc. The screens on mobile phones still tended to be small, and relatively low resolution, though color was being introduced and resolutions were improving.

With the June 2007 introduction of the iPhone and October 2008 release of the first Android phones, people's perceptions of mobile phones quickly morphed from primarily being devices for making telephone calls into "smart phones" – general purpose communication and computing devices that access the Internet over high speed mobile communication networks. Smart phones generally provide the functionality of: calendars, address books, clocks, calculators, portable music players, web browsers, digital cameras, video cameras, GPS navigation, email, photo albums, text messaging devices, and game machines. Oh, and they support making traditional phone calls too.

Apple's iOS and Google's Android smart phones also introduced the most substantial innovation in human computer interfaces since Xerox PARC developed the mouse driven graphical user interface in the 1970's. These phones support the direct manipulation of onscreen objects using multi-touch gestures such as: swipe, tap, pinch, and reverse pinch. For example, in order to reduce the size of an image or zoom out on a map, simply "pinch" your thumb and forefinger together. To expand an image or zoom in on a map, do the opposite, "reverse pinch" by moving your thumb and forefinger apart. These gestures provide a much more intuitive and expressive interface between humans and computers than the point and click graphical user interface.

By midpoint of the second decade of the 21st century smart phones were ubiquitous and high speed Internet access was available throughout most of the US. Phone numbers became predominately associated with individuals rather than with locations. Fixed 'land line' phones continued to exist, but primarily in business settings as more and more households discontinued traditional land line service – many viewing such service as an overpriced relic of a bygone era. The coverage area of high speed mobile networks continued to improve, as did the bandwidth available to smart devices with the move from 3G to 4G and LTE. (Mobile broadband speeds tended to generally reside in the range of 2 Mbps to 10 Mbps as of mid 2013.)

With the release of Amazon's Kindle in 2007 and Apple's iPad in 2010 the popularity of E-book readers and tablets began to soar. In May of 2013 International Data Corporation (IDC) projected that by the end of 2013 worldwide shipments of tablets would overtake portable computers (e.g., laptops) and by 2015 overtake the entire PC market, desktops and portables combined.[3]

As late 2013, it seemed the computer industry was poised for a number of rapid changes in the ways humans interact with computers. Google's Glass prototype, which implements an augmented reality interface, was already available to developers and a wide-scale consumer launch was promised for 2014. Voice based interfaces, for search and simple tasks like scheduling meetings were rapidly gaining popularity. Apple was rumored to be working on an 'iWatch" type device. And even virtual reality gaming, which has been promised for decades, appeared to finally be gaining traction with the Oculus Rift – which raised $2.4 million through a Kickstarter campaign and then secured an additional $16 million in venture capital in June 2013.

15.3 Moore's Law and its implications

15.3.1 What is Moore's Law?

In 1965, Gordon Moore, the co-founder of Intel, wrote a paper entitled "Cramming more components onto integrated circuits".[4] In this paper he attempted to predict the course of the microelectronics industry for the next ten years (through 1975). He noted that the number of components per integrated circuit had approximately doubled each year for the past five years and predicted that they would continue to do so for the next 10 years. In 1975, Moore noted that 9 doublings had taken place instead of the 10 he predicted and he modified his prediction of future progress to state that component densities would double every two years for the foreseeable future. According to Moore, a colleague, David House, noted that in addition to the number of transistors per integrated circuit doubling every two years the speed of those transistors was also improving – leading House to

[3] http://www.idc.com/getdoc.jsp?containerId=prUS24129713

[4] Electronics, Volume 38, Number 8, April 19, 1965. An archived copy is available online at http://download.intel.com/museum/Moores_Law/Articles-Press_Releases/Gordon_Moore_1965_Article.pdf

postulate that overall computing performance would double approximately every 18 months.[5]

Thus was born **Moore's Law** – the observation that computer performance for a fixed dollar cost roughly doubles every 18 months. Moore's Law is an observation about past progress in microelectronics – a "law" that has held true for over half a century. Computer engineers and chip designers use Moore's Law to extrapolate future increases in computer performance and to set research goals. In fact, it can be argued that this expectation of performance doubling every year and a half has, in a sense, become a self fulfilling prophesy – we expect computer performance to double every 18 months, so we work hard to ensure that such increases continue.

Moore's Law describes an exponentially increasing function. When an exponential function is graphed using a linear scale, as in Figure 15.3, it takes on the appearance of a hockey stick.

When one 'steps back' and looks at an exponential graphed over time it appears that for a long time very little progress occurs followed by an 'explosion' of progress at the end. However, this is an optical illusion of sorts, since a true exponential always has the characteristic hockey stick shape no matter which part of the curve you examine. Thus if you 'zoom in' on the 'relatively flat' section at the beginning of the graph you might be surprised to see the same characteristic hockey stick shape, just at a different scale.

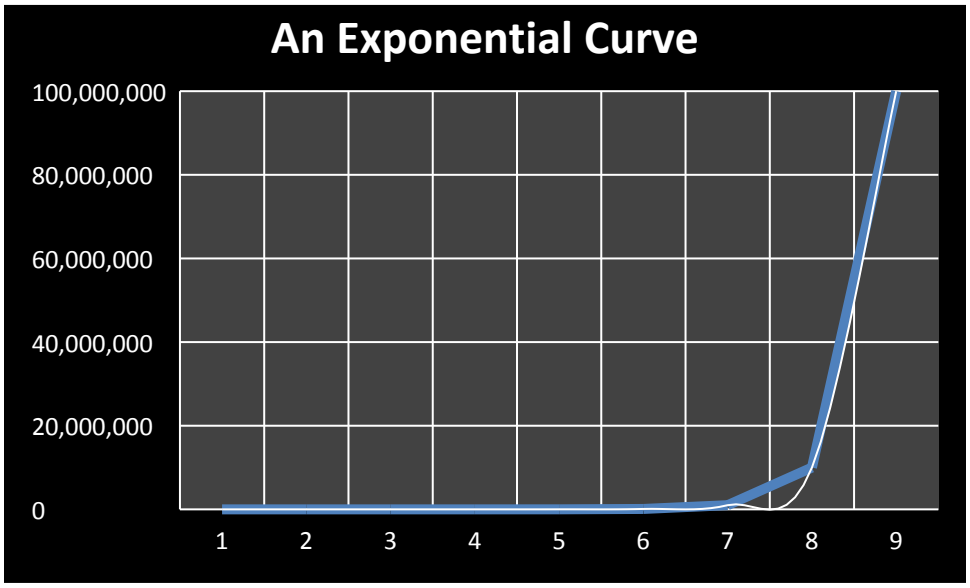

Figure 15.3: An exponential curve plotted using a linear scale

[5] "Excerpts from a conversation with Gordon Moore: Moore's Law", Intel, 2005. An archived copy is available online at ftp://download.intel.com/museum/Moores_Law/Video-Transcripts/Excepts_A_Conversation_with_Gordon_Moore.pdf

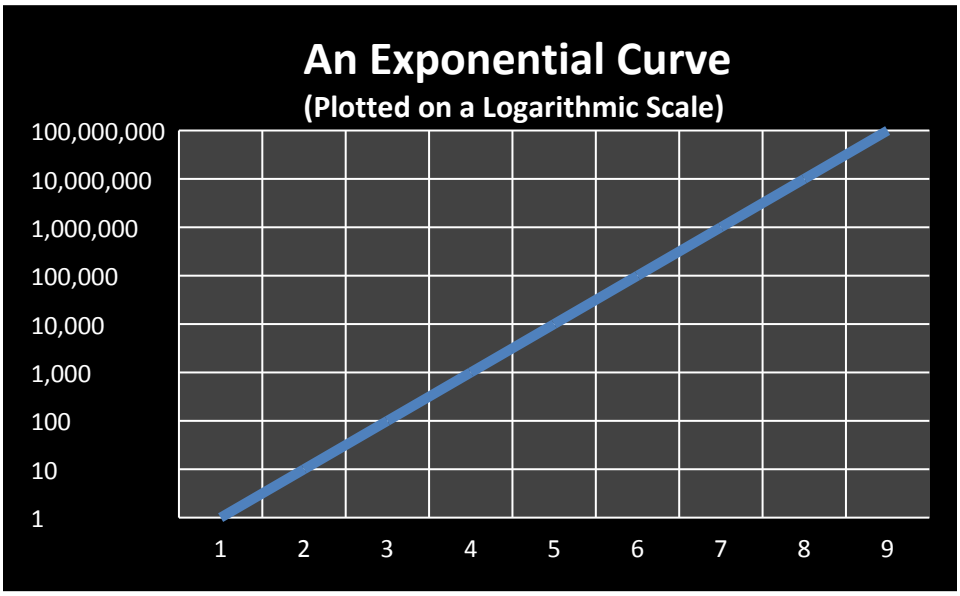

Figure 15.4: An exponential curve plotted using a logarithmic scale

In order to be able to see increases (or decreases) in the rate of progress more clearly, exponential processes are often graphed using a logarithmic scale. Instead of the vertical axis increasing in fixed sized steps, such as steps of 10,000,000 units in Figure 15.3, each vertical tick mark is X *times* the previous tick mark, where X is generally a small constant such as 2 or 10.

Figure 15.4 plots the same curve as that in 15.3 except that a logarithmic scale is used. Each horizontal line in Figure 15.4 represents ten times the previous horizontal line. Note that exponential functions appear as straight lines when plotted on a logarithmic scale.

Figure 15.5 displays the historical performance of Moore's Law from 1971 through 2011, graphed on a logarithmic scale with each vertical tick mark being 10 times the previous tick mark. The individual dots, or points, represent CPU's. Their location on the graph reflects their date of introduction (position along horizontal axis) and number of transistors (position along the vertical axis). The straight line represents a true exponential function that doubles every two years.[6]

[6] This figure comes from the Wikipedia article on Moore's Law. http://en.wikipedia.org/wiki/Moore's_law

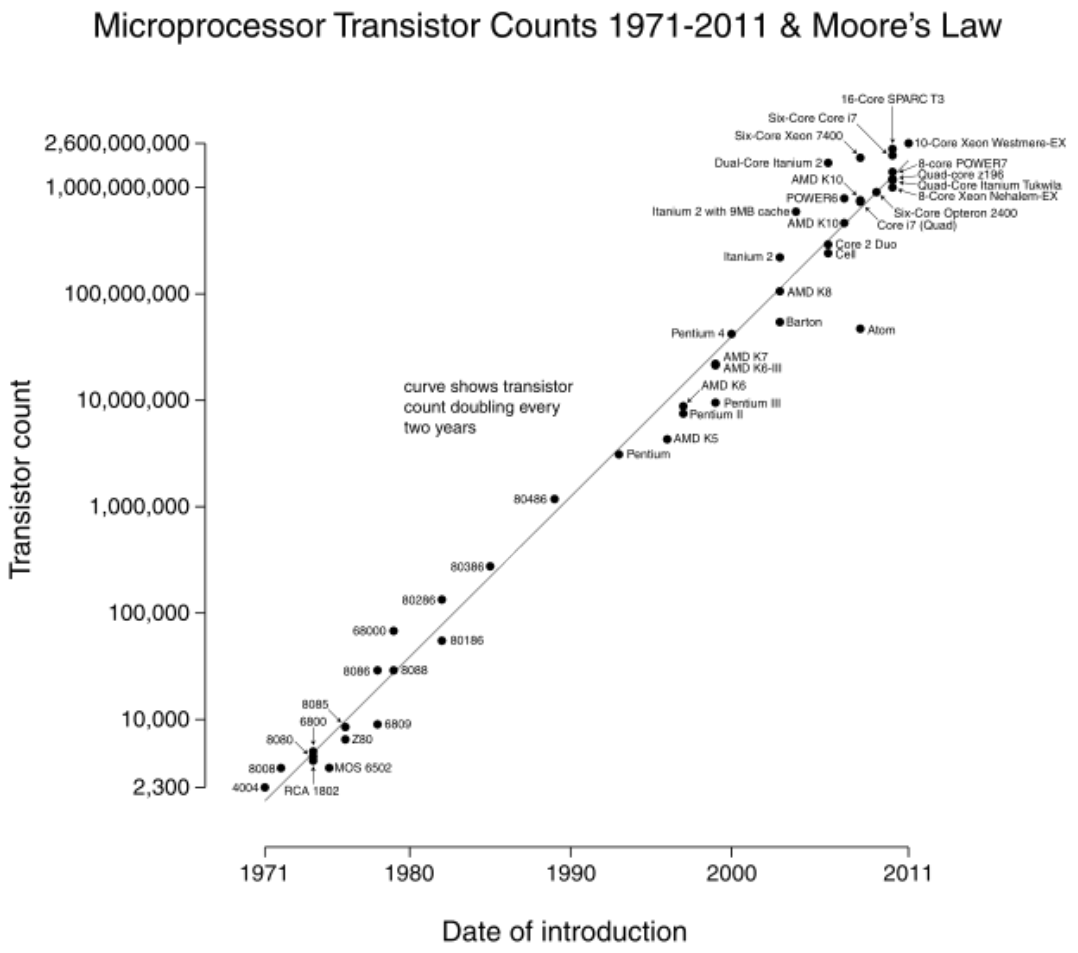

Figure 15.5: Forty years of Moore's Law – 1971 to 2011

As can be seen from Figure 15.5, transistor counts roughly doubled every two years from 1971 to 2011 in accordance with Moore's Law. Over this forty year period 20 doublings took place which translates to a *one million fold increase* in integrated circuit complexity.

Though Moore's Law specifically focuses on the rate of increase in the number of transistors in integrated circuits, this general trend of exponential growth has been observed in many other aspects of computing and information technology. Examples include: the exponentially decreasing cost per byte stored on hard drives, exponentially increasing network capacity and exponentially decreasing cost per byte transmitted, and exponentially decreasing cost of DNA sequencing. This more general observation concerning the exponential growth of computing and information technologies is sometimes called "Kurzweil's Law of Accelerating Returns" after Raymond "Ray" Kurzweil, the futurist who is the highest profile proponent of the school of thought that information and communication technologies are improving exponentially.

15.3.2 The exponential growth of supercomputer speed

Figure 15.6 presents a table of the world's fastest supercomputers over the past 20 years together with their speed in gigaFLOPs (billions of floating point operations per second) and the approximate date at which each came online. As can be seen from the table the speed of the world's fastest supercomputer increased more than 500,000 times over the period – from 60 gigaFLOPS (10^9 FLOPS, billons of FLOPS) to 33.86 petaFLOPS (10^{15} FLOPS, a million billion FLOPS). This difference represents just over 9 doublings in the 20 year period from 1993 to 2013, or a little less than once every 2 years.

DATE	GFLOP/S (R_MAX)	COMPUTER
June 1993	60	TMC CM-5/1024
November 1993	124	Fujitsu Numerical Wind Tunnel
June 1994	143	Intel XP / S140
November 1994	170	Fujitsu Numerical Wind Tunnel
June 1996	220	Hitachi SR 2201 / 1024
November 1996	368	Hitachi CP-PACS / 2048
June 1997	2,379	Intel ASCI Red
November 2000	7,226	IBM ASCI White, SP Power 3
June 2002	35,860	NEC Earth Simulator
November 2004	478,200	IBM BlueGene / L
August 2008	1,105,000	IBM Roadrunner
November 2009	1,759,000	Cray Jaguar – XT5-HE Opteron
November, 2010	2,566,000	NUDT Tianhe 1A
June 2011	10,510,000	Fujitsu K Computer, SPARC64
June 2012	16,324,751	IBM Sequoia BlueGen / Q
November 2012	17,600,000	Cray Titan – XK7 Opteron
June 2013	33,860,000	NUDT Tianhe 2

Figure 15.6: World's Fastest Supercomputers 1993 - 2013

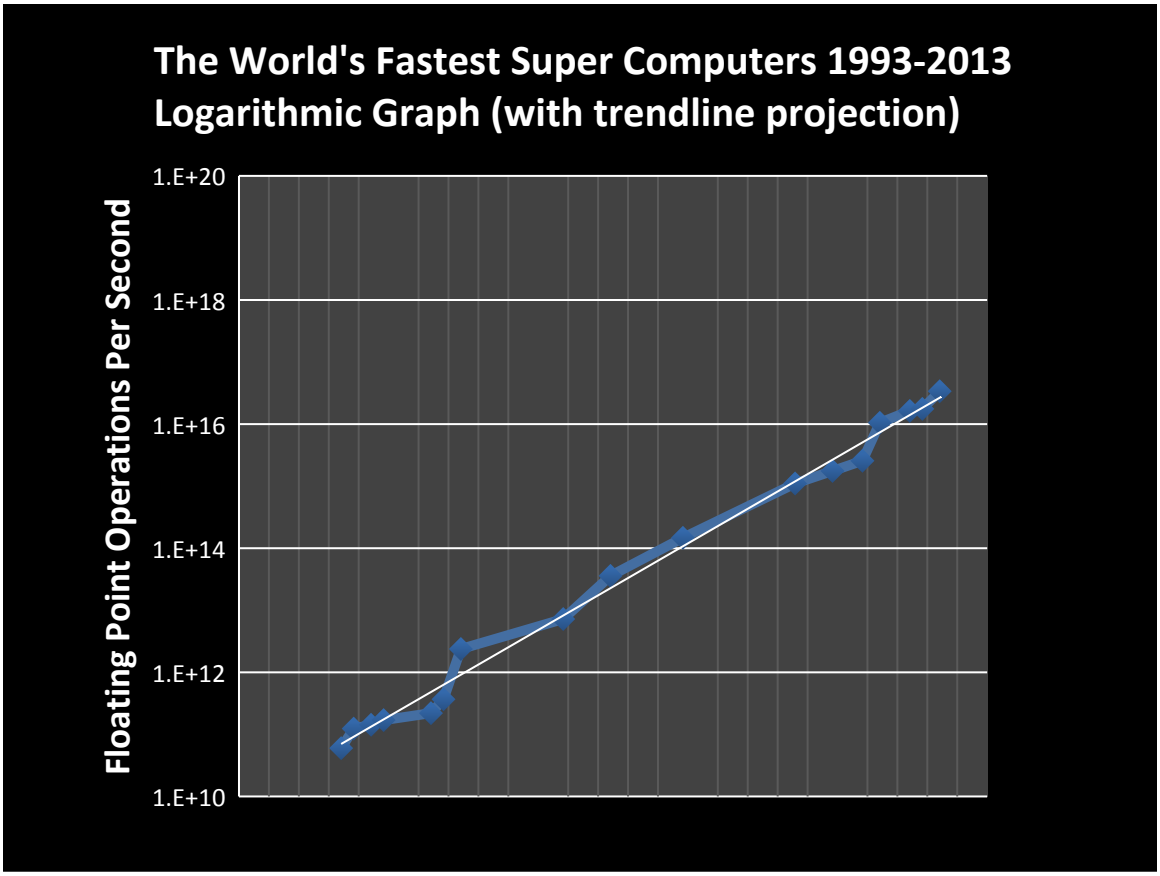

Figure 15.6: World's Fastest Supercomputers 1993 - 2013

Figure 15.6 plots the data on the world's fastest supercomputers from 1993 to 2013 which was presented in Figure 15.15. Figure 15.6 uses the logarithmic scale where exponential growth appears as a straight line. A trend line which projects historical growth through 2020 is included (shown as a yellow line in the figure). If the growth trend of the past 20 years continues to hold supercomputers should pass the one exaFLOPS (10^{18} FLOPS, a billion billon FLOPS) benchmark by 2020.

What could one do with an **exascale** machine – a machine that supports a billion billion floating point operations per second? Well, for one thing, current estimates are that 1 – 10 exaFLOPS should be sufficient for real-time human brain simulation and teams of scientists and engineers in the US and Europe are working diligently to reach this goal.[7]

15.3.3 But, surely this can't continue indefinitely?

Common sense would seem to dictate that the exponentially doubling that information technology has been undergoing cannot continue indefinitely. When it comes to the original formulation of Moore's Law for example there must be a limit to the size of

[7] "Why we need Exascale and why we won't get there by 2020", Horst Simon, Optical Interconnects Conference, May 6, 2013, page 49. Available at: http://www.slideshare.net/ultrafilter/exaflops

individual transistors, yes? As of 2013, "leading edge" fabrication technologies produced integrated circuits with minimum feature sizes in the 10 – 40 nanometer range. What this means is that some features, such as a wire, are no more than 40 nanometers wide. What is a nanometer? A **nanometer** is one billionth of a meter (10^{-9} meters). A strand of the DNA molecule is about 3 or 4 nanometers wide. Thus the minimum feature size of components in our most advanced integrated circuits is somewhere in the range of 3 to 10 times wider than the DNA double helix.

Our devices are nearing (some might argue already are at) molecular level scales. Molecules are, of course, composed of individual atoms. The diameter of a carbon atom is 0.15 nanometers – which is only about 100 times smaller than minimum feature size in some of today's (2013's) most advanced integrated circuits. Since minimum feature size cannot shrink below one atom wide, it looks like the reign of Moore's Law, at least in its 'standard' interpretation, is about over – a factor of 100 is less than 7 doublings (2^7 = 128) which would take 14 years to achieve if doublings took place at the historic rate of about once every 2 years.

This should not be surprising. Individual technologies tend to follow what are called "S" curves rather than true exponentials. An "s" curve, which is illustrated in Figure 15.7, is indistinguishable from an exponential in its early phases. Both begin by showing very little apparent progress over a long span of time, then suddenly progress appears to "take off". In an "s" curve, after this spurt of rapid progress, things settle down into a state where only minor, incremental progress is made. In a true exponential, things never settle down, but continue to accelerate without end.

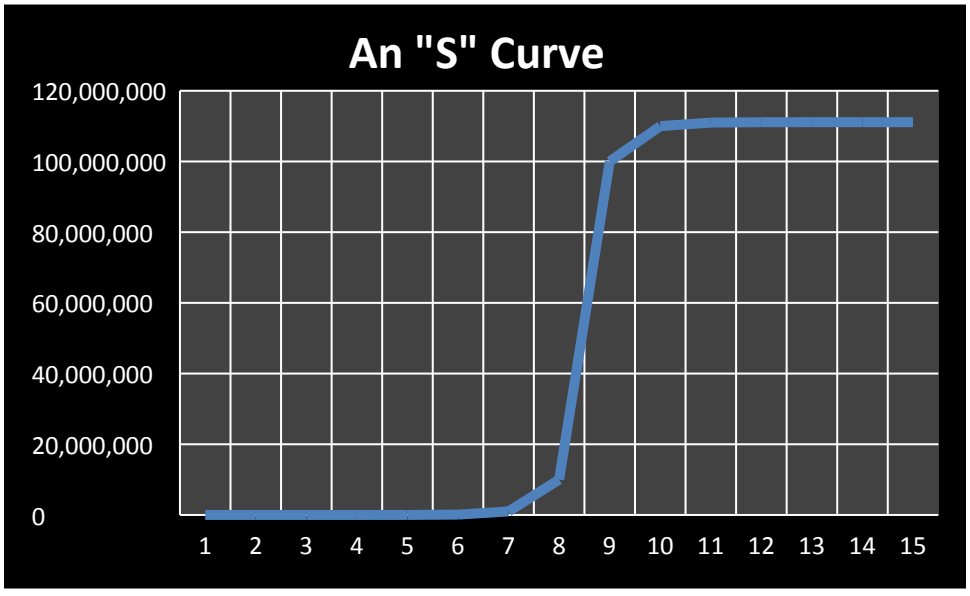

Figure 15.7: An "s" curve

Progress in aviation technology fits the "s" shape curve nicely. For thousands of years Man dreamed of being able to fly like a bird. For most of this time no real progress was made on the problem. The first manned hot air balloon flight didn't occur until November 1783 (about 230 years ago). It took another 120 years, until December 1903 (about 110 years ago) for the first powered airplane flight to occur. Just over half a century later, in October 1958, transatlantic passenger jet service was born with the introduction of the British Overseas Airways Corporation Comet 4. In the half century or so since the introduction of the passenger jet, transatlantic flight has certainly become much more common and much less expensive, but flight times have remained relatively constant – in fact with increased airport security, overtaxed air traffic control systems, and the retirement of the Concorde supersonic passenger plane, transatlantic travel time has actually increased in recent years.

Will advances in computing follow a trend similar to aviation (and many other technologies)? Will the exponential growth in computing technologies soon come to an end?

As mentioned earlier some technologists, such as Ray Kurzweil, argue that the exponential growth in computing extends back much farther than the integrated circuit era of Moore's Law and will continue far after the increasing transistor density of Moore's Law comes to an end.

Figure 15.8 contains a graph produced by Ray Kurzweil, and posted to Wikipedia as part of the "Accelerating Change" article, that captures Kurzweil's view that while individual computing technologies may follow "S" curves, overall progress in computing remains exponential by periodically switching to new underlying technologies.[8] According to this view, the electro-mechanical era of computing essentially began with Herman Hollerith's tabulating machine for the 1890 census, then moved on to relays in the 1930's, then vacuum tubes in the 1940's and 1950's, then transistors in the 1960's, and finally integrated circuits in the 1970's through today. While each of these underlying technologies eventually become mature, reaching their limits of speed and reliability, by leaping from one technology to another, gains in computing capabilities have continued to accelerate at an exponential rate for well over a century.

Given the chip designs that are already "in the pipeline", Moore's Law seems assured of continuing in the near term – say the next three to five years. In the longer term there are many exciting technologies on the horizon. The next section of this chapter presents a number of these technologies that show promise for delivering vastly more powerful computers than exist today – technologies that could continue generating exponential gains in computing for many decades to come.

[8] http://en.wikipedia.org/wiki/Accelerating_change

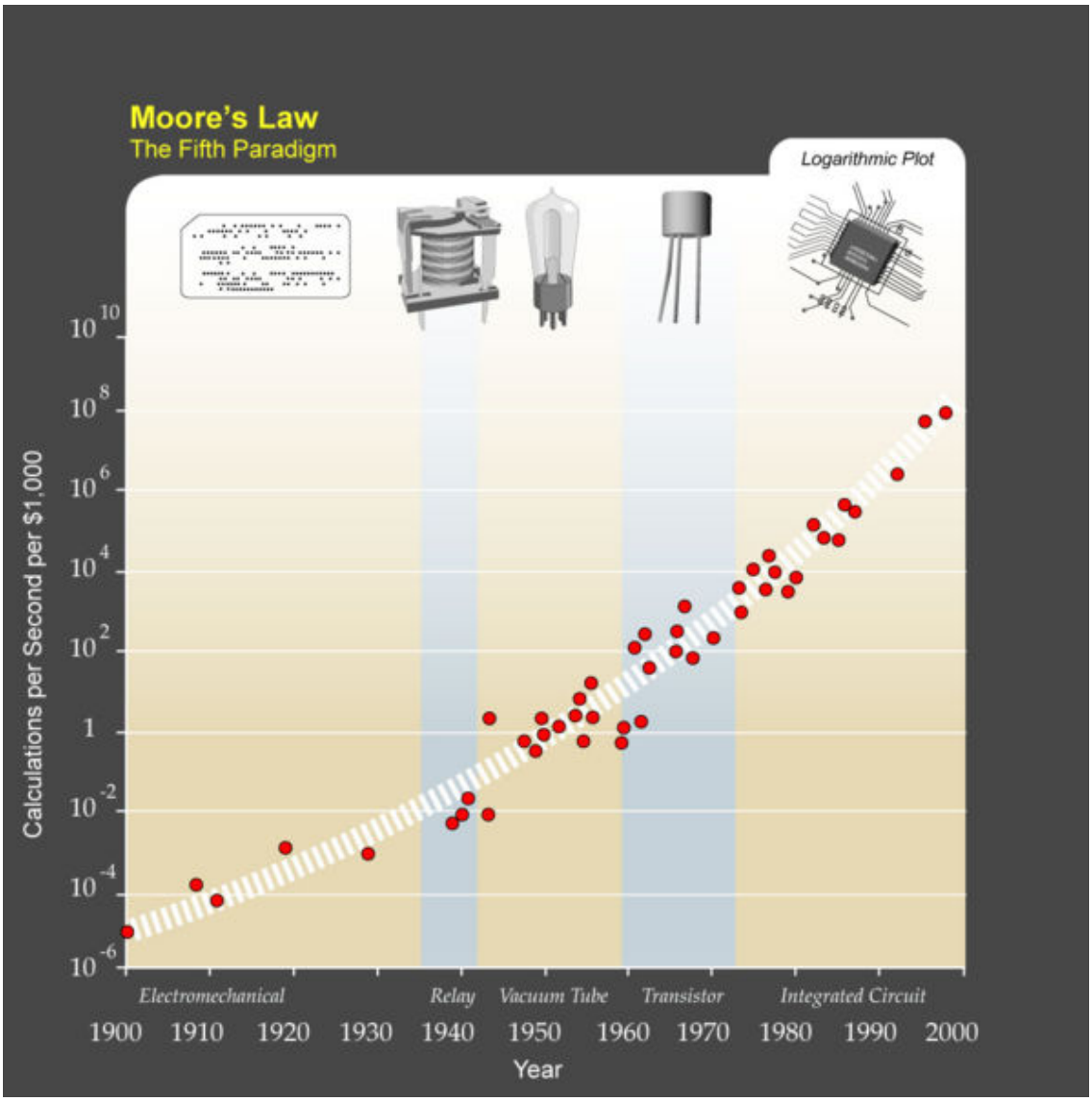

Figure 15.8: Kurzweil's Law of Accelerating Returns
(from Wikipedia "Accelerating Change" http://en.wikipedia.org/wiki/File:PPTMooresLawai.jpg)

15.4 Promising technologies for improving computing performance

In this section we examine a number of technologies that promise continued exponential increases in computing performance. We begin by looking at near-term improvements to existing semi-conductor based computing. We then turn our attention to four fundamentally different technologies that could revolutionize computing. These technologies are: optical computing (also known as photonic computing), bio-computing, molecular-scale nano-computing, and quantum computing.

Of course, there is no guarantee that all (or any) of these technologies will work out in practice. It could be that we will hit a roadblock when we need our next "technology fix"

to continue exponential progress. On the other hand, when one considers the number of promising lines of research continued exponential growth, while far from guaranteed, does seem likely – at least to this author.

15.4.1 Semi-conductor based computing

When it comes to integrated circuits constructed from silicon-based semi-conductors, to paraphrase Monty Python: They're "not dead yet."[9] Before we talk about the problems faced in continuing to improve integrated circuits, let's take a few moments to briefly review why this technology has been so successful for so long.

One of the primary reasons that smart phones, computers, and tablets are so inexpensive today and that progress in the field is so rapid is that integrated circuits (i.e., computer chips) are "manufactured" by a form of photography called **photolithography**.

In photolithography, the design of the integrated circuit is expressed as a very large image, called a **mask**. The mask is similar to a projector slide, in that light can be shown through the mask to produce an image of the circuit. However, in the chip manufacturing process instead of projecting a large image, the light is focused to project a very small image. Specifically, very short wavelength light, typically ultraviolet (UV) light with a wavelength of 193 nanometers (circa 2013), is shone through the mask and focused on a wafer of silicon that has been prepared with a **photoresist** – a chemical that changes properties when exposed to light. In the most common process, the photoresist hardens where exposed to light. During the "development" phase of the process the soft portions are washed, or etched, away by special chemical solutions, leaving a copy of the original mask image (much reduced in size) on the silicon wafer.

In general, the shorter the wavelength of light used in the photolithography process the smaller the image that can be projected onto the wafer of silicon, and thus the smaller the size of the individual features on the resulting chip. As mentioned earlier, as of 2013 cutting edge chips sport minimum feature sizes in the range of 10 to 40 nanometers. In order to generate such small features with UV light, multiple exposures using different masks is necessary. Much effort is being expended to make extreme ultraviolet (EUV) light, at 13.5 nanometers, economically practical by 2015.[10] EUV light will enable even smaller feature sizes and thus higher transistor density, resulting in faster processors.

As impressive as these numbers are, there are a variety of different technologies, such as electron beam lithography, X-ray lithography, and ion beam lithography, that can be used to create even smaller feature sizes and thus greater circuit densities. However, as feature sizes continue to shrink, we are beginning to approach fundamental barriers imposed by the laws of physics. For example, when wires are placed too close to one another, there

---

[9] Watch this clip from "Monty Python and the Holy Grail".

[10] http://arstechnica.com/information-technology/2013/08/moores-law-could-stay-on-track-with-extreme-uv-progress/

can be unintended "crosstalk" between them, where a signal jumps from one wire to another.

Heat dissipation and power consumption are two fundamental problems chip manufactures must battle. Up through the end of the 20th century and the beginning of the 21st, manufactures were able to make chips run faster by increasing the number of 'clock cycles' per second. While this approach enabled chips to do more work, to execute more instructions in a fixed amount of time, the faster a chip runs the more power it consumes and the more heat it produces.

One way around power consumption and heat dissipation issues is for computer chips to include multiple "cores" or CPU's. By having multiple CPU's working together in parallel more total work done can be done in a fixed amount of time, even if the individual CPUs run at somewhat slower speeds. As of 2013, most computers and smart devices contained multiple cores – commonly either two or four cores.

Since photolithography is basically a 'photographic' technique, the results are essentially two dimensional in nature – like a photograph on paper is two dimensional. However, just as individual photographs can be stacked together; three dimensional integrated circuits can be constructed by, among other methods, stacking multiple silicon wafers on top of each other and connecting the layers together using a technology such as "through-silicon via", or STV. Large scale manufacturing of 3D devices is projected for the 2014 / 2015 timeframe. One early benefit of this technology will most likely come in the form of improved DRAM (Dynamic Random Access Memory) devices. As of mid 2013, prototypes of the HMC (Hybrid Memory Cube) 3D design were claimed to be ten times faster than current high end (DDR3) memory chips.

Beyond multiple cores and three dimensional chip designs, another approach for extending the life of semi-conductor technology is to switch from silicon to other materials such as germanium or graphene or molybdenum disulfide. Materials scientists are studying these and many other materials as possible substitutes for silicon. Most conduct electricity better than silicon which could help with power consumption and heat dissipation problems, potentially allowing integrated circuits to run faster than today.

For many decades (at least since the 1980's) people have been predicting the end of silicon-based integrated circuits manufactured using photolithography. And these individuals have been proven consistently wrong. Despite this fact fundamental physics imposes limitations and we are starting to get close to these. Thus, while there is still life left in semi-conductor based integrated circuit technology, most computer engineers agree that the road ahead promises to become more and more difficult.

15.4.2 Optical computing

Optical computing, also known as **photonic computing**, focuses on constructing computing systems that use light (i.e., photons) instead of electricity (i.e., electrons) to carry out data transport and logic operations.

Our present computing and communications infrastructure is a hybrid of electronic and photonic components. Most high-speed, long-distance data and voice traffic is conducted over fiber optic cables, which transmit their data as pulses of light. Many high-speed local area networks use fiber optics as well – for example to move data between buildings on a college campus. Data manipulation, as opposed to data transport however, happens in the electronic domain. Thus, data must be continuously translated between the optical and electronic domains as it is shuffled from network to network. Unfortunately, electronics operate at speeds that are orders of magnitude slower than optics.

An analogy of the state of today's current data and communications networks that I find useful is to imagine the Interstate highway system which enables rapid and efficient vehicle transport between major cities – these Interstate highways are like the long distance fiber optic links in our networks. Now, imagine that instead of having freeway interchanges with their arching on and off ramps in major cities where Interstates meet, all our cities had only dirt paths for roads. If you were traveling across country by car travel time between cities wouldn't be a big deal, but passing through intermediate cities on your way to your final destination would be excruciating slow and painful. These dirt paths inside the cities are like the routers and switches, electronic devices, used to move data from one fiber optic cable to another

There is a lot of work going on in academia and corporate research labs to construct the basic building blocks of optical computers, such as photonic transistors. Optical computing promises many advantages over electronic computing, such as much faster speeds, smaller devices, and less power consumption. However, the obstacles to fully realizing optical computing are formidable.

Though all optical computing is probably several decades away – if it ever proves practical at all – hybrid electronic / photonic devices are a near term likelihood. IBM and Intel (among others) are working on **silicon photonics**, a technology that integrates both electronic and optical components onto a single silicon chip which can be manufactured using existing semi-conductor fabrication techniques. The underlying concept is that the individual components of a computer system, (e.g., the CPU and main memory) could exchange data between themselves optically rather than electronically. While the 'computing' would still be conducted in the electronic domain, data transport functions within the computer would move to the optical domain. The result would be systems that operate much faster than conventional electronics while using less power. It is anticipated that these hybrid systems will initially be used for data transport between individual integrated circuit chips, migrating over time to moving data between different regions of the same chip. In December 2012, IBM announced the first commercially viable silicon photonic chip.[11] Expect to see this technology roll out in the 2015 – 2020 timeframe, first in supercomputers and later in consumer devices.

[11] http://www-03.ibm.com/press/us/en/pressrelease/39641.wss

15.4.2 Bio-computing

One of the most important building blocks of life is the DNA (de-oxy-ribo-nucleic acid) molecule. Each individual cell of every living organism contains DNA. DNA is critical to all life because it stores the genetic information that describes how an organism works.

Our DNA represents all of the instructions, all of the data, that describe how to create a human and keep him or her functioning. In other words, human DNA contains the "program" or blueprint for building and "running" a human being. Subtle variations in DNA determine the individual physical characteristics that differ from person to person, such as facial features, and eye, hair, and skin color. DNA also determines a person's gender – either male or female.

DNA is a double helix shaped molecule composed of four subunits, or nucleotides, each consisting of about 25 atoms. The nucleotides are: adenine, thymine, guanine, and cytosine – which are usually abbreviated by their initials A, T, G, and C. Because of their chemical makeup, adenine always binds with thymine, and guanine with cytosine. For this reason A-T and G-C are referred to as the base pairs of DNA.

Human DNA consists of about 3.2 billion base pairs, each of which represents a single bit of data. A rather amazing consequence of this fact is that since 3.2 billion bits is only about 400 megabytes, a person's entire genome could be stored on a single audio CD, and more than a hundred human genomes on a 50 gigabyte blu-ray disc. Think about that the next time you pop a disc into your Xbox or PlayStation. That game disc holds far more information than your own biological blueprint.

On the other hand, the DNA molecule is a far more efficient data storage device than a blu-ray disc. This is apparent when we note that each and every one of the microscopic cells in our body contains a complete copy of our entire genome. Since each DNA base pair is composed of two 25 atom nucleotides, DNA uses about 50 atoms per bit. DNA thus expresses the entire human genome using a total of about 160 billion atoms (3.2 billion bits x 50 atoms per bit). While that may seem to be a large number of atoms, by the standards of our current computing and data storage technology, DNA is incredibly compact and efficient.

The field of **biological computing** (or **biocomputing**) attempts to harness the incredible information processing and storage capabilities of biology in order to construct computing systems. One of the most noteworthy early achievements in the field occurred in 1994 when Leonard Adleman showed how DNA could be used to solve an example of a very difficult optimization problem. In his experiment, Adleman used biological reactions to solve an instance of the Hamiltonian path problem.

The **Hamiltonian path** problem involves finding a path through a graph that visits every node in that graph exactly once. This problem can be envisioned by thinking of a map of towns connected by roadways. You want to find a route that visits each of the towns exactly once. You may start at any town and end at any other, but you must visit each

town on the map exactly once. While you might think solving such a problem is easy, the requirements that you can't skip any of the towns or visit any of them more than once make this problem quite tough.[12]

Adleman's approach exploited the massive parallelism found in conventional biochemistry. While biochemical reactions take place much more slowly than electronic computations, conventional biochemistry works on many billions of molecules at the same time. When each of these chemical reactions can be made to represent the search for a solution to a problem, the result is massive parallelism – searching many billions of options, all at the same time. Building on Adleman's basic approach, researchers have shown that a range of computing problems can be mapped to DNA and solved using biological reactions. However, this approach required that the problems be set up by hand, mapped to the appropriate biochemical regime, the required chemical reactions carried out in a test tube, and then the results interpreted.

A more recent approach to biocomputing focuses on building general purpose biological computers. Such computers would be able to store, transmit, and manipulate data. The fundamental building blocks of general purpose biological computers would be DNA and RNA, and as such these systems are often referred to as **DNA computers**. DNA computers would be constructed using the techniques of **synthetic biology** which is defined as the design and construction of biological devices and systems for useful purposes.[13]

On the data storage front, as of January 2013, a team of scientists from the European Bioinformatics Institute successfully encoded ¾ of a megabyte (750 kilobytes) of data in a strand of DNA – and later retrieved the data. The stored data included: a number of text files, a PDF, a JPEG color photograph, and an MP3 of a portion of Dr. Martin Luther King's famous "I Have a Dream" speech.[14] While significant cost and speed barriers remain, DNA possesses many positive characteristics for use as a long term storage medium, including the fact that it is highly compact and tends to be quite stable over time.

If the various obstacles to constructing DNA computers can be overcome, the practical benefits could be enormous. While DNA computers would be slow compared to today's electronic computers, they would be incredibly small, energy efficient, and highly compatible with existing biological systems. In fact, DNA computers can, in a very real

---

[12] Hamiltonian path, first posed by William Hamilton, is an example of an NP-Complete problem – a problem for which all known algorithms require exponential time to find a solution, but which once solved the solution can be verified correct quite rapidly (in polynomial time).

[13] "Adventures in Synthetic Biology" by Drew Endy, Isadora Deese, and Chuck Wadey provides an accessible introduction to synthetic biology in the form of a web comic.

[14] http://www.sciencenews.org/view/generic/id/347702/description/DNA_stores_poems_a_photo_and_a_speech

sense, be thought of as molecular level computers – since biological systems ultimately operate on molecules such as DNA and RNA.

Imagine a world where we could reprogram cellular biology. Such computers could potentially detect and then kill cancer cells, or be used to control insulin levels in diabetic patients. They have the potential to usher in a world of medicine at the cellular level.

Progress towards this dream is being made. In March 2013 a team of bioengineers at Stanford led by Drew Endy announced the creation of the first "biological transistor" from DNA and RNA called a **transcriptor**.[15] Endy's team illustrated how transcriptors could be used to implement *and*, *or*, *nand, nor* and other basic logic gates using a system they call Boolean Integrase Logic gates, or BIL gates.[16] Such transcriptor logic based BIL gates could be used to activate / deactivate the expression of particular genes under program control. While in its early days, this exciting field is well worth watching as its impact could be truly enormous.

15.4.3 Molecular-level computing.

In the previous section we looked at molecular level computing derived from biological systems. While biocomputing offers one path towards molecular level computing, in this section we explore another approach – that of directly engineering computing systems at the atomic level from the ground up without reference to biology.

The concept of atomic level computing traces its roots back to a speech given by Richard Feynman in 1959 to the American Physical Society at Cal Tech entitled "There's Plenty of Room at the Bottom"[17]. In his talk, Feynman anticipated the field of microelectronics and outlined some of the approaches, such as photolithography, that were later adopted for the production of integrated circuits. This was an amazing insight, given that computers filled entire rooms in his day and most people were talking about making them larger, not smaller. But Feynman didn't stop there; he went on to describe some of the advantages of, and approaches for, manipulating matter at the atomic scale. In other words, Feynman's speech anticipated the field of **nanotechnology**, which is concerned with the engineering of systems at the level of individual atoms.

Devices such as scanning tunneling microscopes (STMs) and atomic force microscopes (AFMs) can be used to image and even manipulate individual atoms. By 1990, this technology had developed to the point where researchers were able to spell out the letters "IBM" by positioning 35 xenon atoms on a nickel surface, and then create a picture of what they had done. We can easily imagine a memory system in which these xenon

---

[15] A YouTube video describing transcriptor logic is available here: http://www.youtube.com/watch?v=ahYZBeP_r5U

[16] Yes, this is a bad pun on Bill Gates name. Become a respected scientist and one day you too might be able to name your fundamental breakthrough something silly. ☺

[17] A published copy of this talk – well worth a read over half a century after it was given – can be found on the web at http://www.zyvex.com/nanotech/feynman.html. I encourage you to check it out.

atoms are laid out in a 2-D grid arrangement of rows and columns on the nickel surface. A "1" would be represented by the presence of a xenon atom at a particular row and column position, and a "0" by the absence of a xenon atom at row, column position. Both reading and writing of data could be accomplished by use of the STM or AFM.

Though the information density of such a system would be almost unimaginably high, such systems are impractical because the data read and write times would be very, very slow. Regardless of its impracticality, this example does make the point that atomic-scale memory densities are possible.

In addition to being able to store data, a computer system must also be able to manipulate that data under program control.

In the 1990's there was a lot of excitement surrounding the possibility of constructing molecular-level nanomechanical computer systems. These **nanomechanical computers**, first described in detail by K. Eric Drexler, would be mechanical computers built of rods and springs – but at the nanometer scale. Such nanotechnology promises far smaller and more energy efficient computers than exist today.

While the idea of directly engineering systems from the bottom up 'atom by atom' is intriguing, there has been little measurable progress towards the development of nanomechanical computers over the last few decades – and interest in this approach has waned. Some argue that Drexler's designs are simply impractical – that various laws of physics preclude the construction of reliable nanomechanical computers. On the other hand, the existence of naturally occurring biological (bio-molecular) systems that encode proteins and transcribe DNA prove that robust molecular-level computing systems can, and indeed already do, exist. The fact that we currently lack the ability to engineer nano-scale computers from the ground up does not mean that breakthroughs are not "just around the corner".

15.4.4 Quantum computing

The previous two sections looked at molecular-level computing and atomic-level information storage. The present section addresses the question of whether we can do even better. Are memory devices with storage densities greater than one bit per atom even fundamentally possible, let alone actually feasible? And what about computing power? Do the nanomechanical computers and/or DNA computers described in the previous sections represent the fundamental limits of computing density?

Quantum computers employ the laws of quantum physics (which reign in the subatomic world) to implement a form of massive parallelism. The computers we use today are based on classical physics – the physics of the "large" world around us. The fundamental unit of information in such systems is the bit, which, as we have learned, can be either "1" or "0", "on" or "off" – but never both. In quantum computing the fundamental unit of information is the quantum bit or **qubit**. A qubit may be "1", it may be "0", or it may simultaneously be both "1" and "0". This ability for quantum systems to be in multiple

states at the same time is called "superposition of states" – and though it may seem very odd to us, in physics this principle is well established and has been experimentally verified numerous times and in many different ways.

Quantum computing, if it proves to be practical, will be in many ways quite different from ordinary classical computing. Entirely new approaches to algorithm development must be devised to take advantage of the unique capabilities of quantum computers. An early achievement occurred in 1994 when Peter Shor developed an algorithm for quickly factoring large numbers on a quantum computer. Since the security of modern data encryption systems depend heavily on the difficulty of solving such problems, the existence of a working quantum computer running Shor's algorithm could have significant implications for cryptography.

While it is generally believed that quantum computers would be able to solve some problems far faster than they can be solved on traditional computers which employ classical physics, the exact range of problems that will prove amenable to a quantum approach and the speed increases that could be achieved are still open problems.

To date, no *general purpose* quantum computer has been constructed, but progress is being made.

The most visible, and arguably the most controversial, quantum computing project is the work being done by D-Wave Systems. In May 2011, D-Wave Systems announced the D-Wave One, a 128 qubit computer, which it followed up in 2012 with the D-Wave Two, a 512 qubit computer. These computers are not general purpose; instead they are built to perform a process known as "quantum annealing" which can be used to solve certain kinds of optimization problems.

When first announced there was significant skepticism in the scientific community as to whether the D-Wave computers were, in fact, employing quantum effects. More recently the tide has turned and most scientists working in the field accept that the D-Wave systems are quantum in nature. There is still controversy over whether, or how quickly, D-Wave's special purpose quantum computers will present a cost effective, practical alternative to classical computing for those problems that are suitable for solution using the quantum annealing approach.

15.5 Predictions

This chapter, and thus the text, concludes with some predictions as to what the future of computing holds. These predictions are organized into two categories: near term predictions, which the author thinks are likely to occur within the next 10 to 15 years – say by 2030, and longer term predictions which the author expects to take place by the midpoint of the 21st century (2050). Before diving into these predictions, we will first take a look at some past predictions that attempted to divine what life would be like in the first and second decades of the 21st century to see what past prognosticators got right and what they got wrong about our present.

15.5.1 Visions of future past

"It's tough to make predictions, especially about the future." – Yogi Berra[18]

Predicting the future is both difficult and in many ways thankless. In general, one often doesn't get much credit for successful predictions since they frequently appear 'obvious' in hindsight. A really accurate prediction sometimes isn't even recognized as a prediction – as can happen in television, movies, and commercials when you aren't really sure when the video was produced. If the prediction of present day technology is 'spot on' you just assume the movie was made at a later date than it actually was.

On the other hand one tends to get banged really hard for the predictions that do not come to pass. Incorrect predictions often appear so 'obviously' wrong in hindsight that our natural tendency is to laugh.

Despite these limitations, people tend to spend a considerable amount of time envisioning the future – be it something mundane like what they are planning to eat for dinner tonight or something more bold like designing a humans-to-Mars exploration strategy or a HyperLoop passenger transport system. Even when our predictions turn out to be wrong, I believe the act of prediction itself is valuable – envisioning the future often provides us with goals to work towards and the inspiration to carry on in the face of failures and disappointments.

In the chapter on artificial intelligence we talked extensively about the movie "2001: A Space Odyssey" which attempted to realistically depict advances in space travel and computing. Though most of the predictions depicted in the movie proved wildly optimistic, "2001" served as a catalyst for numerous individuals, such as this author, to pursue careers in science and technology. Many of these individuals went on to actually advance the state of the art in computing and robotics.

In 1967 (nearly 50 years ago) the Philco-Ford Company produced a short film entitled "1999 A.D." that attempted to predict what life would be like for the typical family at the turn of the century. One of the most interesting sequences in the film describes what computing will be like.[19] While the hairstyles, clothing, and background music, along with the somewhat offensive depiction of the roles of men and women, clearly date this video as being from the 1960's, if one looks beyond these shortcomings the video is really quite insightful. It predicts home computers and printers, Internet connections, email, online shopping, online banking, flat screen monitors, and web cams. The accuracy of these predictions is amazing when one considers that in 1967 computers of

---

[18] While this quote is generally attributed to Berra, a similar quote "Prediction is very difficult, especially if it's about the future." is attributed to physicist Niels Bohr. Some have suggested the joke first originated with Mark Twain (Samuel Clemens).

[19] Watch a portion of the film "1999 A.D." that depicts computing at the turn of the century as envisioned in 1967.

the day filled large rooms, took a small army of people to manage, and cost hundreds of thousands to millions of dollars - and there was, of course, no Internet, nor email, nor online banking and shopping. One interesting feature of the video that turned out to be wrong was the idea that home computer systems would need automated backup components that would kick in if a primary unit failed. While this idea seems antiquated today, back in the 1960s electronic components were unreliable and failed rather frequently, so this idea probably seemed quite reasonable at the time.

Twenty years later, in 1988 – just four years after the original version of the Mac was introduced – Apple produced a short film called "Knowledge Navigator" that aimed to show what computing would be like at the start of the second decade of the 21st century.[20]

The video got a surprising number of things right, including: the emergence of flat screen, high resolution, color displays; a portable, tablet-like form factor computer (though one that opened up like a thin book); support for touch and gesture-based input; wireless networking; built-in video camera, microphone, and speakers; video conferencing; a kind of Internet (though one seemingly focused on research and scholarly pursuits); and small high capacity external storage devices (similar to today's USB drives, though shown in the video as something more like the size and shape of a business card). This litany of successes is rather amazing – especially since some of the innovations, such as gesture-based interfaces have only become commonly available in the last 5 years or so.

The Knowledge Navigator video is, in fact, so accurate on so many fronts that it suffers from the "not enough credit" phenomenon mentioned above. It's too easy to forget – or be unaware of – how truly primitive by today's standards computers were when this video was made. We take flat screen, high resolution, color displays and wireless networks for granted now, but these things were essentially science fiction when Apple produced this video.

Despite its successes, the Knowledge Navigator video suffered from one major misstep. Apple envisioned that artificial intelligence would have progressed by now to the point where one would be able to converse with computers using human language and that operating systems would support an avatar-based interface – the Knowledge Navigator, a kind of 'butler in a box'. While it is true that Siri and Google Now enable people to have limited 'conversations' with their smart phones, the level of artificial intelligence needed to construct a Knowledge Navigator simply doesn't exist – at least not yet.

In 1993 AT&T produced a series of advertisements that attempted to envision life 15 to 20 years in the future.[21] These ads were known as the "You will" ads as they opened with a question of the form: "Have you ever …. " followed by some seemingly futuristic action, such as "crossed the country without stopping for directions", and then ended with the tag line "You will. And the company that will bring it to you: AT&T."

---

[20] Watch Apple's "Knowledge Navigator" video from 1988.

[21] Watch AT&T's "You Will" ad campaign which originally broadcast in 1993 and 1994.

While there are some misses in these ads, there are some surprising hits, including: in car GPS navigation with real time traffic updates and spoken directions, online education, books available online in electronic form, automated toll booths, online purchases of things like concert tickets, Skype-like video calls and video conferencing, NetFlix-like movies on demand (shown on large, rectangular, flat screen displays), wireless mobile computing, electronic medical records (which are just now becoming common). Misses include the idea that public phone booths would still exist, the expectation that voice recognition would be a common way of unlocking doors, and that AT&T would be a driving force in bringing many of the depicted innovations to market.

The author of this text has himself dabbled in making predictions concerning the future for quite some time. The text you are reading has existed in various forms and editions since it was first drafted in the late 1990's, and each edition has included predictions that were expected to come to true within 15 to 20 years. Thus, the earliest editions of the text included predictions that should have come to pass by now. So, before I make new predictions about the future, it is fair to ask how my past prognostications for today turned out.

One interesting prediction I made was the rise of e-books and "book readers" that sound a lot like today's tablets and smart phones. Here is the quote from the May 1997 edition of the text:

> *... computers also have many advantages over paper. Within a decade these advantages, such as multimedia, hypertext, and ease of information retrieval, will begin to overtake paper. ... These books will be viewed with a relatively inexpensive and portable "book reader." These "book readers" will really be more like a combination TV, telephone, and Web browser. Many publications, especially periodicals, will not be distributed in any physical form. Instead, your reader will be able to automatically download Time (or just the articles you are likely to find of interest) directly off the Net.*

Five years later, in the 2002 edition of the book I gave a more complete description of a smart phone. Here are some selected quotes from that edition:

> *With mobile devices, a mouse can be replaced with a touch sensitive display screen… The line between cell phones and PDAs is already starting to blur and this trend will accelerate…. In addition to calendar, address book, and scheduling applications, most phones will include web access capabilities over relatively high bandwidth data links…. Phones will also begin to include GPS (Global Positioning System) receivers. Incorporating GPS into a PDA Phone will allow the phone to always know where it is. That knowledge, combined with access to the Net, will enable all sorts of applications. Such a phone could give you turn-by-turn directions to the nearest In'N'Out Burger, determine when you have missed a turn and modify its instructions on how to get there based on this new information.*

It is important to remember that I wrote these words five years before the first iPhone was announced (2007) and six years before the Android (2008).

Another prediction, probably the one I'm most proud of because it deals with a social implication of then-future technology, was my suggestion that access to (what we now call) the Internet would forever change the nature of human conversation so that long drawn out arguments over a trivial questions of fact would come to an end. Here is the quote from the May 1997 edition when the Internet was in its infancy and people connected to it by low speed dial up modem:

> *... this technology will affect most aspects of our daily lives in ways both profound and mundane. Most questions of fact should be answerable anywhere and anytime by affordable access to the GII (Global Information Infrastructure). If nothing else this will be great for settling friendly disagreements quickly. Arguments of the type: "Star Trek originally premiered in September of 1967." "No, it premiered in September of 1968!" will be settled quickly and effortlessly by accessing the GII.*

Note that in 2013 simply asking Siri "When did the original Star Trek series premier?" brings up a results page that tells me it premiered on Thursday, September 8, 1966.

Of course, I'm picking and choosing quotes to highlight those predictions I got right. ☺ But it should be clear from the various examples presented in this section that by carefully observing existing and emerging trends it is sometimes possible to predict, at least in 'outline' form, certain future technological innovations.

15.5.2 Near-term predictions (next 10 to 15 years – before 2030)

In this section a number of technologies the author expects to become commonplace within the next 10 to 15 years – say by the year 2030 – are discussed. Before wowing you with the innovations we can expect to see, I'd like to take a moment to mention a few common sci-fi concepts that I don't believe will become reality by 2030.

First and foremost, while Santa may bring you more sophisticated, special purpose, Roomba-like robots over the next decade and a half, don't expect a general purpose household robot by 2030. By general purpose I mean a robot capable of doing typical household chores – cooking meals, doing the laundry, or loading the dishwasher and putting the dishes away afterward. While we are making progress in robotics, the problems are just far too deep for household robots to be anything more than expensive toys within the next 15 years. Sorry.

Next up, put away those notions of fully immersive virtual reality (VR) games that you 'plug in to' – PlayStation 5 and 6 won't come with that option. The kind of brain / computer interface necessary for that type of VR is simply too far away.

Finally, don't expect to be having meaningful unscripted conversations with NPC's (Non Player Characters) in your video games over the next 15 years. Artificial Intelligence systems capable of human-level dialog won't exist on your smart devices and game machines in this time period. As with robotics and VR the challenges are just too great to expect that degree of progress in only 15 years. That is not to say that NPC's won't become more sophisticated in their actions. I expect games will adopt advanced versions of systems like Skyrim's "Radiant AI" in order to enable NPC's to be somewhat more independent and lifelike in their actions.

15.5.2.1 Exponential doubling of computing speeds will continue through 2030

The age of Moore's Law – exponentially increasing transistor density in integrated circuits achieved by reductions in transistor size – will soon come to an end due to the underlying physics. However, based on the technologies described in Section 15.4, such as 3-D chip stacking and hybrid photonic/electronic components, I think it is a safe bet that exponential doubling of computing speeds and memory capacity for a fixed dollar will continue roughly at the rate of once every two years through at least 2030.

Thus, in 2030 your smart phone – or whatever smart device(s) are popular at the time – should be 100 to 200 times more powerful than the phone you carry in your pocket today. Some of the ways we will used this increased computing capacity are explored below.

15.5.2.2 Ubiquitous computing and personal networks will become common

Ubiquitous computing refers to the situation where highly-networked computing devices are embedded throughout our environment – in our homes, cars, classrooms, offices, and even our clothes. The trend towards ubiquitous computing, which has been going on for some time, will continue over the next 10 to 15 years.

Basic computer chips are already embedded throughout our environment – in our cars and household appliances (e.g., washing machines, dishwashers, and thermostats). These devices are beginning to incorporate wireless connectivity so that they can report their condition and be programmed remotely. More advanced systems reside in our consumer electronics (e.g., cable and satellite TV boxes, game machines, and televisions); and Internet connectivity is already standard in many of these devices.

Already most of us carry a smart phone wherever we go and are connected to the Internet 24-7. As of late 2013 we appear to be on the cusp of wearable technology taking off – Google Glass is in beta with an initial general release scheduled for 2014. There are strong rumors that Apple is looking at an iWatch concept. To be useful, these devices will need network access, but in the near term the power requirements for wi-fi will be out of reach. Thus, within the next three to five years smart phones will take on another important role – they will become the hub and Internet gateway for your personal network of devices. The devices themselves will communicate with your smart phone (over BlueTooth or something similar) and the smart phone will, in turn, provide the desired Internet access.

15.5.2.3 Voice-based interfaces will continue to gain acceptance

As personal networks, wearable smart devices, and computing systems embedded throughout our environment become commonplace; voice-based interfaces will become the predominant way of interacting with these technologies within 5 to 10 years. The reasoning behind this prediction is simple; a wristwatch size device only has enough surface space for a few buttons or small touch surface. In order to issue any kind of complex instruction – such as a search query or request to schedule a meeting – the flexibility of verbal interaction would seem to be required.

Those of us who grew up before voice interfaces were common will feel somewhat awkward at first talking to our wristwatches and other smart devices – but people can adapt quickly if they see an obvious benefit. Walk down any street or across any college campus these days and you will see that most people seem to be talking to themselves – not too many years ago that would have been a sign of mental instability, now days we recognize they are just talking on their phone. The same thing is happening with voice-based web searches and Siri-like commands. It takes some people a while to catch on, but once you try it, some things, like telling your phone to "wake me up in an hour" just seem quicker, easier, and more natural than manually setting an alarm.

Currently, as of 2013, we have to launch an app and/or press a button in order to for our smart devices to accept voice commands. This is beginning to change. First, our devices will listen just for keywords, like "Ok, Glass", "Ok, Google", or "Xbox…", and when the keyword is detected the system will "wake up" and accept a verbal command. As people become more comfortable with verbally interacting with computer systems, we will begin to allow our systems to 'listen in" continuously on our conversations. This will enable the systems to automatically provide helpful comments without being specifically asked to do so. For example, if you are in the middle of a conversation with a friend and the system hears you say something like "I think Nick's birthday is on Tuesday. We really need to figure out what we are going to do to help him celebrate." Depending on the circumstances, and your personal preferences, your intelligent assistant might offer the following: "Sorry to interrupt Mike, but Nick's birthday is next Monday not Tuesday."

15.5.2.4 Continuous health monitoring and personalized medicine will become routine

One benefit of wearable smart devices and computer systems embedded throughout our environment will be near continuous monitoring of our basic vital signs. The Kinect sensor on the Xbox One can detect a person's heart rate by observing the slight flushing our faces undergo with each heart beat – a change so subtle that humans are unable to detect it. If one is facing the sensor, Kinect can also detect whether a person's eyes are open or closed and by looking at their facial expressions get some idea of their mood. Even with this limited data, it might be possible for Xbox One to tell if someone is having a heart attack or in acute pain and, using its Internet connection, contact

emergency personnel. While Microsoft has no announced plans for fielding such an emergency medical application, it's clear that Xbox One has most of the needed hardware and software in place to perform such basic emergency medical monitoring and notification.

Wearable smart watches should be able to detect not only heart rate, but also skin temperature and (potentially) blood pressure. Being that the watch will most likely be worn 24-7 monitoring of these vital signs could be near continuous. Inexpensive disposable devices for performing more detailed medical tests – to monitor glucose levels for signs of diabetes, to check liver and kidney function, and screen for early warning signs of certain types of cancers – are all on the horizon. Early detection of potentially serious medical conditions could save untold lives and improve quality of life, while dramatically decreasing treatment costs.

Once a person is diagnosed with a condition that requires treatment, that treatment may be personalized to the individual. Such personalized medicine will require access to the patient's genome.

The cost of human genome sequencing is decreasing rapidly, from $100 million in 2001 to less than $10,000 per genome in 2013.[22] Within 10 to 15 years, sequencing will become standard medical practice. When costs drop below $1,000 per genome, infants will be sequenced immediately following birth (if they were not already sequenced prior to birth) and the rest of us will probably be sequenced at some point as part of a routine medial checkup. Having your complete genome on file in your medical records will enable doctors to warn you of conditions you are at risk of developing so that you can take preventive measures to avoid becoming ill. If you do become ill your genome will help doctors tailor any treatment program specifically to your body's particular needs.

The result of continuous health monitoring and personalized medicine is expected to be more effective and less costly medical care. By 2030 people should be living longer and healthier lives.

15.5.2.5 A limited form of virtual reality gaming will become popular

Virtual reality, or VR, is a human-computer interface technology in which a person is completely immersed in a simulated environment. While completely immersive virtual reality ala "The Matrix" won't happen by 2030, gamers shouldn't despair. Inexpensive head mounted displays capable of producing high resolution, high frame rate, stereoscopic color imagery, while accurately tracking head movement, will become a popular accessory to game machines over the next five to ten years. An early leader in this area is Oculus VR, makers of the Oculus Rift, which raised nearly $18.5 Million in 2013 – $2.4 million through a Kickstarter campaign and $16 million in venture capital.

[22] The Human Genome Project itself, the project to first sequence the human genome, cost on the order of $3 billon.

Head mounted displays may be paired up with omnidirectional treadmills, low friction concave surfaces (such as the Virtuix Omni), or some other device that allows one to 'walk in place' in any direction. Such a combined system would allow for far more immersive experiences than we have today. Just imagine the fun of exploring a Skyrim-like environment by walking through it.

Unfortunately, as good as these systems may become, the experience they provide will still be a far cry from Star Trek's holodeck.

15.5.2.6 Augmented reality will become commonplace

As we have seen, the aim of virtual reality is complete immersion in a virtual world. In **Augmented Reality**, or AR, the user can see the real world, but in addition computer technology is used to enhance or "augment" what the user sees. The most commonly discussed form of AR employs a mobile heads-up display together with position and orientation sensors capable of detecting where the user is and where he or she is looking. Given this interface, a computer will be able to superimpose computer generated virtual objects on top of the real-world scene.

The best known example of AR technology to date is Google Glass. As I write this chapter in late 2013, Google Glass is still in beta but is already generating lots of interest as well as lots of controversy. Some people think the technology looks geeky, and at this point they are right. But with improving technology, the sensors, display device, and computer hardware should soon fit into truly lightweight glasses. I believe that while there may be some initial resistance, within 10 to 15 years most people, including you, will probably wear some form of AR gear.[23]

Such technology has great potential for enhancing the world around us. For example, consider what happens when you have to make a connecting flight in a large and unfamiliar airport. Right now, we look up our flight number or destination city on airport monitors to find what gate our flight leaves from, then we check a map of the airport terminal, and finally make our way to the gate. With a properly programmed AR system, all you would have to do is follow the big green arrow that appears to be floating in mid air and it will lead you to your gate. Of course, only you would be able to see this arrow. Other travelers would see their own arrows leading them to their flights.

For another example, consider a chemistry class in which everyone is wearing AR gear. When the professor discusses sulfuric acid, the entire class would be able to see the H_2SO_4 molecule floating in mid air. Everyone could watch as simulations of chemical reactions took place all around them. The 3-D geometry of molecules and the role of reaction catalysts would be obvious.

[23]And don't worry about looking like a geek. AR glasses will come in designer styles and be thought of as very cool.

Despite the obvious benefits of AR technology, there is concern about its social implications – especially since Glass has a forward facing camera that can take pictures and record video. In fact, due to the potential for abuse, Google has announced that it will not develop face recognition applications for Glass.[24]

15.5.2.7 Intelligent assistants will become more intelligent

Over the next decade, most people in developed countries will begin interacting with intelligent assistants multiple times a day. These assistants will do more than simply manage your schedule and act as an interface to search engines – though those will still be important functions.

Deep Question Answering systems, of which IBM's Watson is a present-day example, will be integrated into these assistants. Classical 'search' will be replaced by 'conversation' where you ask a question, are given an answer, and can ask follow-up questions – just as you would do with a knowledgeable human. Note that these systems will be far more sophisticated than key word searches that return page after page of web links that may or may not contain the information you are looking for.

Currently, if you want an intelligent assistant to do something for you, you have to explicitly ask it to do so. I expect within the next few years our assistants will grow to the point where they will be able to offer up advice, suggestions, and answers to questions of interest without having to be asked. They will begin anticipating our needs, just as an excellent human assistant is able to do. In order for your assistant to offer this level of help, it will need to learn about you – your likes, dislikes, schedules, and routines. This is clearly the direction that Google is clearly headed with their Google Now product.

For example, when leaving on a recent trip I was surprised to find that Google Now knew my flight and hotel booking information, though I had not told it these things. Google Now apparently learned the details of my trip by scanning my email. Some people will find this creepy, and to be honest my first reaction was a kind of nervous shock – but once I started getting flight delay and gate change updates, and weather conditions for my destination city, I quickly got over those feelings and began to appreciate how Google Now could make travel a little less weary.

Google Now also does things like look at your movement patterns, figures out where you live and where you work and then feeds you route and travel time estimates – .updated with current traffic conditions.

The general idea is that the more these systems know about you the more helpful they can be – and not so coincidentally the more targeted the ads they can feed to you. Barring an

[24] This author hopes Google will change its stance on facial recognition as people become more familiar with the technology, since he is one of those individuals who can't remember people's names and would welcome a technology that would simply superimpose names directly beneath people's faces.

extreme backlash over privacy concerns, expect this trend of your intelligent assistant knowing more and more about you to continue over the next decade.

15.5.2.8 Self driving cars will start to appear on our roadways

Even though self driving cars have been 'just around the corner' since at least the 1950's, I'm reasonably confident that over the next 10 to 15 years auto-driving options will become available on higher end personal vehicles.[25] We probably won't call these vehicles "self-driving cars" because they will continue to support manual driving as the default option. Instead we will probably say something like "intelligent cruise control" is available on this model. In addition to self driving cars, self-driving taxis, 18-wheelers, and other service vehicles will be common by 2030.

The introduction of "intelligent cruise control" will be gradual and is, in fact, already happening. Systems that prevent you from following too closely to a vehicle in front of you – applying the brakes if necessary – are standard on some higher end models today. Systems to assist with steering in emergency situations are appearing. Some cars can now parallel park themselves. Systems that can handle driving on Interstates and freeways – including dealing with stop and go traffic – will probably be the next part of the driving experience to be automated. Systems capable of handling construction zones, understanding the hand signals of traffic cops, and coping with various kinds of inclement weather, such as ice and snow, will probably prove the most complex to develop.

15.5.2.9 Photorealistic virtual actors will start to appear.

Over the next 10 years or so I believe we will begin to see the use of computer generated, photo-realistic vactors (virtual actors) that are indistinguishable from real actors in movies and TV shows.[26] More specifically, I predict that within the next 15 years there will be a film featuring 1950's sex symbols Marilyn Monroe and James Dean – even though these two actors never worked together during their lives. Similarly, I wouldn't be surprised to see new episodes of the original Star Trek series, staring Shatner (Kirk), Nimoy (Spock), and Kelly (McCoy) in their mid-1960's prime. These characters will look, act, and sound (perhaps through motion capture and voice actors) very close, if not identical to the originals. If Hollywood chooses not to go in this direction, I believe fans of the original series will take it upon themselves to do so.[27]

---

[25] In fact as far back as the 1939 World's Fair, General Motors claimed in their Futurama exhibit that on the motorways of 1960 "safe distant between cars is maintained by automatic radio control'. http://www.youtube.com/watch?v=1cRoaPLvQx0 @ 14:30 to 14:50

[26] In the interest of full disclosure, I originally made this prediction in 2002 and am somewhat surprised it hasn't already come to pass. Of course, given the 2008 movie "The Curious Case of Benjamin Button" in which the face and head of the very old version of the character played by Brad Pitt is computer generated, some might argue the prediction has already become reality.

[27] Here is an example of the level of quality of some fan generated Star Trek inspired productions

Progress toward constructing vactors is ongoing. Hollywood is already very good at generating believable aliens (such as the Na'vi from Avatar) and human-like creatures (such as Gollum from The Hobbit and Dobby in the Harry Potter movies). Computer graphics have also been used quite effectively to generate much older versions of actors (such as Brad Pitt's character in "The Curious Case of Benjamin Button") and far less successfully to generate younger versions of actors (Patrick Stewart's Professor X and Ian McKellen's Magneto in the opening flashback scene of 2006's "X-Men: The Last Stand"; and the younger version of Jeff Bridges' character in Tron 2). Despite this progress, as of 2013, I am aware of no computer generated character that fits the definition of a completely photorealistic virtual actor.[28]

15.5.2.10 Artificial general intelligence will be achieved in supercomputers

I began this section on predictions for the next 15 years by stating a few things that would *not* happen. Among other things, I said: "Artificial Intelligence systems capable of human-level dialog won't exist on your smart devices and game machines in this time period." While I believe that statement to be true, I also believe that by 2030 we will have finally achieved artificial general intelligence through human brain simulation on supercomputers. In other words, an AI capable of passing the Turing Test will exist, but it will take the full resources of one of the world's fastest machines of the day to carry out the human brain simulation underlying that intelligence.

For nearly two decades prominent futurist Ray Kurzweil[29] has been predicting the emergence of exascale supercomputers capable of one billion billion computations per second (one exaflop) by 2020 running real time human brain simulations to enable full scale human AI by 2029. These predictions, which were once considered by most to be somewhere between fringe science and science fiction have now, at least in part, become mainstream science backed by government initiatives in the US and Europe with billions of research dollars promised over the next decade. Given this commitment, together with the progress being made understanding and simulating regions of the brain, and the long-term stability of Moore's Law that I have personally witnessed over my lifetime, I am willing to go on record saying that I believe Kurzweil is correct in his prediction that artificial general intelligence through human brain simulation will succeed by 2030.

The human brain contains about 100 billon (1×10^{11}) **neurons** or individual brain cells. Individual neurons can directly communicate with anywhere from hundreds to thousands other neurons by sending electro-chemical signals across the tiny gaps between neurons called synaptic gaps or **synapses**. Thus, the total number of synapses in the human brain is generally estimated to be somewhere between 100 trillion (1×10^{14}) to 1 quadrillion (1

[28] While no one would mistake the computer generated Kara for a human actress, the realism is rather amazing, especially when one considers this video was generated in real time on a PS3.

[29] And why should we listen to this guy? Ray Kurweil is considered the "father" of modern optical character recognition software (giving computers the ability to "read" printed text), text-to-speech synthesis technology (giving computers the ability to convert text into spoken words), and automated speech recognition technology (giving computers the ability to understand spoken languages).

X 10^{15}). These numbers together with results from early brain research projects, such as The Blue Brain Project which successfully simulated a small part of the rat brain, have led researchers to estimate that real time neural simulation of the entire human brain will require about 1 to 10 ExaFLOPS of computing power – in other words between 1 billion billion (1 X 10^{18}) and 10 billion billion (1 X 10^{19}) computations per second. (**FLOPS** stands for FLoating-point Operations Per Second and is a standard measure of computer speed based on the number of mathematical operations that can be processed each second.)

As mentioned in section 15.3.2, the fastest supercomputer in the world (as of June 2013) runs at 33.83 petaFLOPS, and the speed of such systems has historically doubled approximately every two years. Given the emphasis on supercomputing and the worldwide competition for 'bragging rights' to the country that has the fastest machine, supercomputing experts expect a computer capable of 1 or more exaFLOPS by 2020.

In addition to the progress being made in constructing the necessary hardware to support human brain simulation, work on understanding the brain itself is proceeding apace.[30] In Europe, Dr. Henry Markram's Human Brain Project, the successor to his highly successful Blue Brain Project, was funded by the European Union in January 2013 for up to $1.3 billion over a ten year period. In the United States, President Obama announced the BRAIN initiative in April 2013 with a projected budget of $3 billion over a ten year period. These two decade long projects focus on understanding the brain and ultimately building real time human brain simulations.

Thus, it looks increasingly likely that within 10 to 15 years we will have both the supercomputer hardware and the brain simulation software to implement a real time neural simulation of the human brain. Such a simulation should "appear" intelligent and be capable of passing the Turing Test – though it may take several years to do so as the simulation will probably need to be taught to use language in the same way that human children must be taught to speak.

When humans achieve the goal of building an artificial general intelligence we will have accomplished an amazing feat, at least as significant in human history as Man's first steps on the Moon. Even so, the first exascale computer running a human brain simulation will fill a large room and is projected to consume on the order of 20 to 30 megawatts of power. This compares to your brain which fits in your skull and consumes the equivalent of about 20 watts of power. In other words, even if these projects succeed our brains should still be about a million times more power efficient than a simulated brain running on a supercomputer.

---

[30] According to Horst Simon, the Deputy Director of the Lawrence Berkeley National Laboratory's National Energy Research Scientific Computing Center (NERSC) speaking in May 2013, our best brain simulations are at 4.5% of human scale, running at 1/83 real time speed. http://www.extremetech.com/computing/155941-supercomputing-director-bets-2000-that-we-wont-have-exascale-computing-by-2020

I will close this section by noting that human brain simulation is only one possible path to achieving artificial general intelligence and in some ways the least interesting. What I mean by this is that successfully copying a system that already produces intelligent behavior (the human brain) does not mean we will necessarily understand how that system produces intelligent behavior. It would be far more satisfying to 'engineer' intelligence directly. This is, in fact, the direction that AI research has been pursuing for well over half a century now.

Even though progress in AI has not been anywhere near as fast as researchers had originally hoped and expected, when one considers that today's fastest supercomputer is no more than 3% as powerful as the human brain, and our servers and mobile smart devices are just a tiny fraction of 1% as powerful as the human brain, the fact that we have made any measurable progress at all in artificial intelligence up to now is rather amazing. It seems only reasonable that problems, which are today totally intractable to engineered solutions, such as fluency in human languages and common sense reasoning, will surely fall when sufficient computational resources can be brought to bear.

15.5.3 Long-term predictions (2030 - 2050)

In this final section of the book, I'll go way out on a limb and make a few predictions beyond the 15 year horizon of the previous section – focusing on expectations for the 2030 to 2050 timeframe. Obviously, the further out one looks the more speculative the predictions become. Thus if you took the predictions of the previous section with a grain of salt, which you certainly should have, you should take these with at least three.

15.5.3.1 Human level artificial general intelligence will be common.

Assuming that computing power continues the exponential doubling it has historically enjoyed, we can expect the capacity of a 2030 supercomputer to be available "on the desktop" (or whatever the appropriate metaphor will be) by 2050. Under this assumption, I think it reasonable to conclude that software entities that appear to have human level intelligence will be commonplace by the midpoint of this century.

I believe that the "strong AI" verses "weak AI" debate, touched on in Chapter 14, will be a significant social issue by 2050. As you may recall, the difference between strong and weak AI is not based on behavior but on whether humans choose to view an AI as "really" intelligent or just "a good fake". If we believe our machines to be truly intelligent, then the issue of AI rights becomes very important. However, because the question of whether an AI can ever really be aware, in the same sense that humans are aware, has no objective answer, the issue will probably be approached on moral and religious grounds. Depending on whether the AIs derive from human brain simulations or whether they are engineered from first principles may tilt the debate one way or the other. Regardless, it seems clear that conflict is almost certain to arise.

15.5.3.2 Fully immersive virtual reality will appear.

Fully immersive virtual reality, indistinguishable from 'real' reality, will happen, but beyond the 2030 timeframe.

To date, military and civilian aircraft simulators used to train pilots are the closest we've come to implementing a true VR environment. These systems succeed because the pilot is sitting in a physical cockpit that is identical to the one in the actual aircraft. The computer system only has to simulate the view visible through the airplane's windows. Because the windows in the plane provide a relatively small viewing angle and the viewable objects are far away, the simulation can appear to be reasonably realistic. Advanced simulators are placed on gimbals that can change the orientation of the simulator under computer control to provide the physical sensation of nose up and nose down, or shake the simulator to mimic turbulence. Doing so further adds to the sense of immersion.

In section 15.5.2.5 we covered existing and near-term VR systems consisting of head mounted stereoscopic display devices and some kind of device to allow a person to 'walk in place'. These systems may be augmented with a data glove that provides haptic (tactile / touch) feedback. Regardless, these approaches to VR simply are not adequate to produce truly immersive experiences.

How does one reproduce the sensation of strolling about a forest? Say the user decides to reach down, pick up a stick, and throw it. What kind of I/O device could provide the realistic tactile feedback required for such an action? Surely, a data glove falls far short. What if the user decides to take a dip in the river that flows through the forest? Ultimately all of these questions reduce to "how does one build a Star Trek-type holodeck using real technology?"

I believe that truly immersive VR will require a brain computer interface of some sort. Specifically what I have in mind are five small devices – two attached to (or cutting across) the optic nerves, one for each ear, and one implanted at the base of the skull that attaches to (or cuts across) the spinal cord. Normally these devices would act in a passive manner, simply relaying signals from our eyes, ears, and limbs to our brain – for all intents and purposes if would be as if the devices weren't even there. In active mode however, these devices would take inputs from a simulated world and pass those inputs to the brain. They would also transmit the outputs from the brain to the simulation.

From the point of view of the outside world, the user would be immobile while in a simulation – probably lying safely in bed. From the user's point of view he or she would be completely immersed in a virtual world – it would look, sound, and feel completely real. You'd be able to run, jump, swim, etc. The only limitations would be that smell and taste would be missing and you wouldn't have sensation (from the simulated world) in your face, since those nerve signals don't pass through the optic nerves or spinal cord. These issues could be addressed with somewhat more invasive implants.

It is fun to speculate how this technology would be used. Personally, I think it will prove especially popular with the 80 year old plus crowd. Even as their physical bodies fail them, they could still be enjoying ski vacations in Aspen and scuba diving along the Great Barrier Reef. And what 80 year old wouldn't want to look, as well as feel, 30 again?

Various psychological conditions may arise where individuals become convinced that the "real" world is 'just another level' in some simulation. I can even foresee some rather unique legal defenses – perhaps where a man accused of violently murdering his boss pleads for manslaughter rather than first degree murder because he became confused as to reality verses simulation and he always murders his boss on the weekends in virtual reality as a harmless way of letting off steam – which is what he thought he was doing when he committed the actual murder.

While the technical and physical issues to implementing truly immersive VR are formidable, there is another set of problems that people have barely begun to even recognize, much less actually solve. A great deal of attention has been devoted to the concept of "playing" in VR environments, but much less has been paid to how people will "work" in such environments.

As a thought experiment, I can imagine writing a future version of this text using advanced VR to simulate a grand Victorian library where I sit at a massive oak desk in an oversized stuffed leather chair, typing away on a realistically rendered manual typewriter using paper that feels real to the touch, where mistakes must be erased by hand using an accurately simulated rubber eraser. While such a setting would be cool, obviously I'd be far less productive in such an environment than I am typing away on my laptop. What we'd like however is for VR to make us *more* productive, not less.

In order to work effectively, we will need to figure out how abstract concepts can be most effectively represented in virtual worlds. How will people retrieve, modify, and store information in a VR system? How will they solve problems? How will they conduct business transactions? In other words, what kind of simulation will provide an appropriate interface to a world of abstract concepts? Presently, no one knows. But this topic is sure to be fertile ground for research for years to come.

15.5.3.3 Ray Kurzweil's singularity will *not* take place by 2050, or even by 2100

As discussed above, scientists are beginning to converge on an estimate of 1 to 10 exaFLOPS as the amount of raw processing power needed to simulate a human brain in real time. Some, such as Ray Kurzweil, believe that human level cognition could eventually be accomplished on less powerful machines. Kurzweil estimates that 10 petaFLOPS, 1/100 of an exaFLOPS, should be sufficient.[31] Amazingly, since our fastest supercomputers now run at over 33 petaFLOPS (as of 2013), according to Kurzweil we

[31] The basic idea underlying this lower estimate is that your brain has lots of redundancy that we should be able to eventually eliminate. Once we figure out how various evolved systems function, we should be able to redesign them to function more efficiently.

should already possess the necessary supercomputer hardware to support an artificial general intelligence. If Kurzweil's estimate is correct, why haven't we already constructed an AI? The answer is simple, we haven't yet figured out how to generate the appropriate software.

For many years, Kurzweil has used the year 2029 as his target for when we will achieve human level intelligence in a machine – a machine capable of passing the Turing Test. When he first came out with these predictions, most 'reputable' scientists gently shook their heads and smiled. But that is changing rapidly.

Regardless of whether 10 petaFLOPS, or 10 exaFLOPS, or something in between turns out to be the magic number necessary to support human level artificial intelligence, we appear to be quite close to achieving the necessary hardware. Additionally, given recent progress in artificial vision (Kinect), speech recognition and generation (Siri, Google Now), deep question answering (IBM's Watson), and ongoing initiatives in the US and Europe to understand how the human brain works, the software component finally seems to be falling into place. Thus, Kurzweil's estimate for achieving AGI (artificial general intelligence) by 2029 now seems reasonable to many knowledgeable individuals.

Achieving artificial general intelligence, however, is not Kurzweil's most famous, nor most controversial, prediction. Essentially Kurzweil believes that the exponentially doubling trend in computer power will continue indefinitely – and this leads him to make an astonishing prediction: By 2045 he predicts a "technological singularity" will occur.

The **singularity** is a point in time at which the rate of technological change will become so rapid that unenhanced humans will no longer be able to comprehend what is happening. Kurzweil believes that humans (or at least a significant portion of the human population) will choose to merge with super intelligent AIs leading to a post-singularity world of near immortal, near God-like beings.[32] [33]

Obviously, this is a controversial idea that many dismiss as a form of religion – "rapture of the nerds" is a phrase that is sometimes heard.

But, let's look at this prediction in a bit more detail, taking for the sake of argument that we will have achieve one human AI equivalent in a supercomputer by 2030 – something that many now view as reasonable. **IF** the exponential increase in computing speeds continues at the rate of one doubling approximately every 2 years, in 20 years computing should have increased 1,000 fold. Thus, by 2050 a "laptop equivalent" (which runs about 1,000 times slower than a supercomputer) should possess one human level equivalent and a supercomputer should possess 1,000 human level equivalents. In another 20 years, by 2070, the laptop will support 1,000 human level equivalents and the supercomputer one

[32] Here is a trailer for the film "Transcendent Man" an excellent documentary about the life of Ray Kurzweil and his ideas concerning the Singularity.

[33] Here is a very brief summary of Kurzweil's "Six Epochs of Evolution" by Jason Silva.

million human level equivalents. By 2090, one million human level equivalents in the average laptop and a billion human level equivalents in the supercomputer.

Under such circumstances how long do **you** think humans would remain the dominant intellectual force on the planet? How long would we remain in control? Wouldn't the only option be to "join" them?

While I do think many of the predictions that Kurzweil makes are reasonable and will come to pass over the course of the next 50 years – such as artificial general intelligence, greatly expanded human lifespans, and nanotech assemblers able to construct most physical goods (including food, clothing, and shelter) directly from raw materials at almost zero cost – I do **not** believe the singularity will occur during the 21st century.

My reason for this view is simple. Humans are not ready to 'step aside' as the premier species on the planet. Moore's Law (and any successor that continues the exponential doubling of computing power) is not a physical "law". This progress happens only because we humans work hard to make it happen. In my opinion no sane human would want his laptop to be a thousand times smarter than he is. So, we won't build such things. We will begin to limit the advance of future computer systems – either their intelligence or their 'desire for independence' so that we humans remain firmly in control for the foreseeable future.

Eventually, at some point in the distant future, humans may decide that merging with intelligence systems to create super powerful merged human / AI hybrids is a desirable thing. These hybrids would presumably blend the creativity and insight of human beings with the exponentially increasing power of computing hardware to become a fundamentally different type of "post-human". Thus, I'm not saying the singularity will *never* happen, just that it is not something that will occur any time "soon".