Chapter 0
Where We Are – How We Got Here

"…a device in which an individual stores all his books, records, and communications, and which…may be consulted with exceeding speed and flexibility."

—Vannevar Bush

This book will be in the hands of students in the third decade of the 21st century. Although there were early attempts at mechanical computing machines, electronic, programmable, digital computers have been around for only about 80 years.

Figure 0-1
Vannevar Bush
U.S. Library of Congress

In the twenty-first century, digital computers are everywhere. You almost certainly have one with you now, although you may call it a "phone." You are more likely to read this book on a digital device than on paper.

An electronic digital computer is a machine. It has mechanisms for receiving input of information, for processing and storing information, and for producing and communicating results. You will find computers in all kinds of specialized devices, but the computer itself is one of the most general-purpose machines ever devised.

The more general a technology, the more and more varied are the uses to which it will be put. We use computers for everything from preparing tax returns and processing them to landing airplanes, generating medical images, and displaying pictures of cats.

0.1 Computing in the 21st Century

As we begin the 21st century, those ubiquitous "phones" are the most numerous computers. We use them to communicate with both voice and text, but each one also has more computing power than all of NASA during the first manned moon landing in the summer of 1969. Although phones are the most numerous computers, they aren't the only ones. Many individuals and nearly all businesses also have larger personal-use computers, from tablets like the iPad to laptop computers and desktop computers. People use more powerful personal-use computers called workstations for graphic arts and scientific computing.

Huge data centers house thousands of pizza-box size computers and power enterprises like Google, Facebook, and Amazon along with many others. Similar data centers offer "cloud" computing and storage to individuals and to businesses of all sizes from tiny to Fortune 10.

Mainframe computers, machines with very large memories and enormous input and output capabilities, once called "dinosaurs," still power about two thirds of the Fortune 100 businesses. The principal reasons are outstanding ability to perform input and output, which we will cover in more detail in Chapter 4, and the ability to handle hundreds or thousands of transactions per second while maintaining database consistency. Database technology is covered in Chapter 6.

Supercomputers, computing machines vastly more powerful than those used by most people, perform such jobs as weather forecasting and complex simulations. We now call the use of supercomputers **high performance computing**.

Tiny computers embedded in machinery from microwave ovens to automobiles, ships, and airplanes perform control functions formerly handled by special-purpose mechanical or electronic devices. Using programmable controllers in this way makes the functions they perform both more flexible and less expensive than the custom components they replaced. They may, however, make the devices in which they are embedded more difficult to repair because repairs might require software changes. Making those changes may be complex, and manufacturers often consider their embedded software to be proprietary and a trade secret.

Not only are computers everywhere, nearly all are also interconnected. Most of the students reading this book cannot remember a time when Internet access was not available to nearly everyone. Today Internet access is available with our phones, in our homes and places of work, and in our cars in most parts of the United States. The 5G cellular technology being widely deployed as this is written and the 6G technology coming next promise to make speeds in the billions of bits per second available nearly everywhere.

We use computers primarily for communication, from text messages and phone conversations to surfing web sites and streaming sound and video. We also use computers for storage of very rapidly growing amounts of data, for control functions, and for computational tasks like simulations. Making computing machines "intelligent," for some definition of intelligence, has been a goal of computer scientists for nearly as long as there have been computers. The 21st century has seen remarkable progress in machine learning and artificial intelligence. As storage, processing speed, and input and output capabilities have improved, it has become possible to analyze truly huge quantities of data. The ability to process "big data" has spawned the new discipline of **data science and analytics**. Those increases in storage capacity and processing speed have also enabled what is called **generative artificial intelligence**, the technology of programs like ChatGPT.

None of this will seem new to most readers of this book, but the way we got here is a fascinating story. It probably began in pre-history when people first began to count their possessions and to trade with one another. Advances in mathematics from the 17th century onward laid some of the foundations. Although there have been mechanical counting and computing devices since ancient times, the foundation for today's electronic computing machines was laid much later.

0.2 Early Ideas

Computers operate on numbers, and the development of systems of numbers suitable for use as the internal representation of data in a computer is a story of its own. That story appears in the next chapter.

A fundamental characteristic of computers is that they are **programmable**. Different programs allow the computer to do different things; computers are general purpose machines. Possibly the first programmable machine was invented in the years just after 1801 by French weaver Joseph Marie Jacquard. Jacquard invented a weaving loom that could be programmed by punching holes in cards. Once a set of cards was prepared, the same fabric design could be woven over and over, error-free. Different cards wove a different design. The Jacquard loom was programmable. In the 1960s and 1970s, punched cards similar to Jacquard's were used as input for electronic computers. Jacquard looms are still used in the 21st century, but the programs are stored in electronic memory, not on punched cards.

Celestial navigation – navigation using the positions of the sun, moon, planets, and stars – was important to the British Navy, merchant shipping, and military forces everywhere until the development of radio navigation systems in the 20th century. Celestial navigation used a sighting instrument called a sextant to measure angles to celestial objects and a book of tables called an ephemeris to determine the position of that celestial object at that time. Preparation of the tables was time-consuming, and calculation, transcription, or typesetting errors could result in lost ships.

About 1822 Charles Babbage, a British mathematician and philosopher, proposed to automate the production of such tables, including the automation of printing plates, to eliminate sources of error. With funding by the British government, Babbage designed and began construction of a mechanical computing machine called the Difference Engine. He and the craftsman hired to build the Difference Engine disagreed over costs and the project was never completed. Babbage designed an improved version but was not able to get funding to complete it.

The Difference Engine was a special-purpose mechanical computer. Confident that his design was correct,[1] Babbage went on to design a general-purpose mechanical computing machine, the Analytical Engine, that was remarkably similar to a modern computer. The part Babbage called the mill corresponds to the central processing unit of a modern computer, and the part Babbage called the store corresponds to the memory.

Lord Byron's[2] daughter, Augusta Ada King, Countess of Lovelace, had corresponded with Babbage because of her interest in mathematics. She was fascinated by the Analytical En-

1 It was. Near the end of the 20th century, the Science Museum in London built two Difference Engines from Babbage's plans and to the manufacturing tolerances of the 19th century. The Difference Engines worked as intended.

2 The same Lord Byron you met or will meet in English classes.

gine, and in her *Notes* included programs to be run on the Analytical Engine, possibly making Lady Ada the first computer programmer.

In 1854, George Boole, another British mathematician and philosopher, published *An Investigation of the Laws of Thought* in which he expanded on his earlier idea of an algebra over finite sets of discrete values. We now call that Boolean algebra. It was not until 1937 that Claude Shannon, an American mathematician and electrical engineer, wrote in his master's thesis that, if the discrete values of Boolean algebra were zero and one, or on and off, they could describe switching systems, and that Boolean algebra could be used to optimize switching systems. We'll meet Boole and Shannon again in Chapter Three.

0.3 The World War II Era

During the World War II era, several important inventions were made at about the same time by different people or groups in different places. Simultaneous invention occurred at other times in history, too. Some historians say these inventions were possible "because it's time." That is, the necessary knowledge and technology had been developed to allow invention to occur. By the 1930s electricity was widely available in cities and towns, electrical engineering was an established discipline, and electronic components like vacuum tubes were being mass-produced. Babbage's idea of a programmable computing device was over a hundred years old, as was George Boole's algebra. The Boolean algebra was applied to switching by Shannon in 1937. Alan Turing's 1936 doctoral dissertation described what Alonzo Church later called a Turing machine.

The World War itself provided incentive to complete quickly anything related to the war effort and provided liberal funding for research and development.

Howard Aiken of Harvard conceived the Harvard Mark I computing machine, which IBM called the Automatic Sequence Controlled Calculator. Aiken presented the concept to IBM in 1937 and IBM funded and built it. The Mark I included and improved upon many of the features of Babbage's Analytical Engine. One of the uses of the Mark I was the production of mathematical tables, Babbage's original intention over a hundred years earlier.

Also in 1937, John Vincent Atanasoff of Iowa State College and graduate student Clifford Berry conceived and began work on the Atanasoff-Berry Computer, or ABC. The ABC used binary numbers internally and had an electronic arithmetic and logic unit. World War II interrupted Atanasoff's work on the ABC and it fell into obscurity.

Artillery in World War II was aimed using firing tables much like the tables of an ephemeris. Each kind of artillery piece needed its own firing tables, and the tables included such factors as kind of shell, kind of propellant, distance to the target, and wind direction and velocity. They were computed by hand by women whose job title was "computer." The process was slow, and new artillery pieces were being developed faster than firing tables could be prepared for them. Army Lieutenant Herman Goldstine convinced his superiors that electronic computation of firing tables would get them into the field faster. That was the beginning of the ENIAC[3] project. J. Presper Eckert, John Mauchly, and many others were involved in the design and construction of ENIAC at the University of Pennsylvania. ENIAC began solving operational problems in December 1945. The first real problem was a calculation for the Los Alamos National Laboratory related to the development of the hydrogen bomb, World War II having ended in September.

ENIAC was entirely electronic, having no mechanical parts. However, it did not yet have the capability to store programs. A mechanism to allow for stored programs, and eventually an electronic memory, were added later. ENIAC used vacuum tubes as its computing elements. ENIAC remained in operation until 1956.

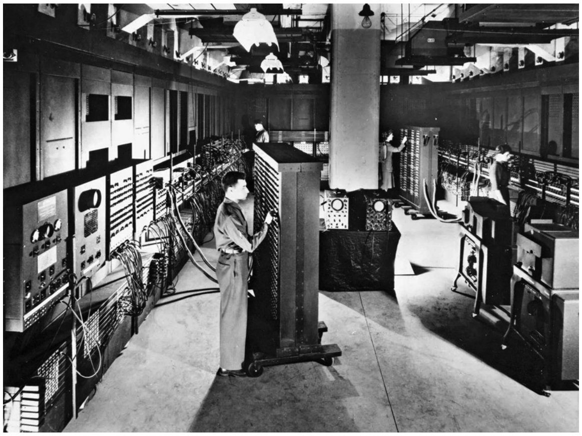

Figure 0-2
ENIAC

U.S. Army

Those building the ENIAC learned quite a lot about the design of electronic computers, much of it too late to be implemented in ENIAC itself. Eckert and Mauchly proposed to the Army's Ballistic Research Laboratory another computer to be called EDVAC.[4] John von Neumann,[5] then at the Los Alamos National Laboratory, was serving as a consultant on the ENIAC project. In June, 1945 while returning by train to Los Alamos from the Moore School of Electrical Engineering in Philadelphia, von Neumann compiled a set of design notes which he called *First Draft of a Report on the EDVAC*.[6] Von Neumann mailed the handwritten notes to Herman Goldstine, who was the Army's project manager for the construction of computers at the Moore School. Although von Neumann's notes were a distillation of the work of many people, Goldstine had the notes typed and circulated listing von Neumann as the sole author.

Von Neumann's *First Draft* described what came to be known as the **von Neumann architecture** and is a description of how computers are still designed. He described a computer with an arithmetic and logic unit, a control unit, a memory, input and output, and storage. Programs and data were to be stored in the same memory, instructions to be executed in

3 Electronic Numerical Integrator and Computer

4 Electronic Discrete Variable Automatic Computer

5 Pronounced "von NOY-man."

6 If you are curious about what a description of how computers are still designed looks like, you can find copies of the *First Draft* on line or buy a typeset copy of all 47 pages!

sequence, and data represented as binary numbers.

There were significant developments in computing in the United Kingdom during World War II as well, but, because of the British Official Secrets Act, most were not revealed until the 1970s. A key figure was Alan Turing, who designed a code-breaking machine called a Bombe based on an earlier design from Poland. Turing also contributed to the design of the Colossus electronic computers. Probably Turing's most important contribution was the description, in his 1936 doctoral dissertation of what we now call a Turing machine. The **Turing machine** is a mathematical model of computation that can be used to prove that there exist problems that "look computational" but cannot be solved.

Eckert and Mauchly, having built ENIAC and EDVAC, formed the Eckert–Mauchly Computer Corporation to design, build, and sell electronic computers. Their first product was called UNIVAC.[7] It is important as the first computer design of which multiple units were sold commercially.

0.4 Vannevar Bush and the Memex

> *…a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. (Bush, 1945).*

Apart from the somewhat outdated language and the citation date at the end, which is a dead giveaway, you might think this was a description of a 21st century smartphone. This remarkable description is from Vannevar[8] Bush's essay, "As We May Think," published in the *Atlantic Monthly* in July, 1945.

Dr. Bush was describing, not a smartphone, but a hypothetical device he called a **memex**. The size of an office desk, the memex was for information storage and retrieval. Bush was not trying to predict future technology. The memex was based on technology available at the time or likely to be available in the five- to ten-year future. For example, Bush discussed a camera that could take pictures that would be instantly available without development needing chemicals and a darkroom. The first Polaroid instant camera was sold in 1948; the first digital camera was not made until 1975, thirty years after Bush's article.

Bush's memex addressed the information storage and retrieval needs of that time, 1945. He was concerned by what we now call information overload. An important part of the memex was enough storage to hold all the documents an individual might need and the ability to connect documents with associative links, which we now call hyperlinks. Bush's description included the ability to respond to voice commands and even to take dictation automatically. He envisioned a head-mounted camera that would allow the user to photograph whatever

7 Universal Automatic Computer

8 Pronounced "va-NEE-var," to rhyme with achiever.

was being seen. That was realized in 2013 as Google Glass.[9] In short, Bush was describing using technology to manage information in ways that were impossible for an unaided human. In doing so, he foresaw much of the technology which we take for granted in the 21st century.

0.5 Transistors, Chips and Moore's Law

The ENIAC was an important advance in electronic computing, but it had an important physical limitation: its computational elements were vacuum tubes. Vacuum tubes are made of glass, and so relatively delicate to begin with. If the vacuum is lost, the tube will not function. Vacuum tubes work on the principle of thermionic emission. The cathode (negative element) was heated by a heater similar to the filament in an incandescent light bulb. Like the filaments in incandescent light bulbs, those heaters eventually failed; the tube burned out. At first several tubes a day failed. By running the tubes at a slightly lower voltage, they reduced the failure rate to one every two or so days (Randall, 2006). Doubling the computing power by doubling the number of tubes would lead to a crash a day.

Figure 0-3
A vacuum tube.
Anucha Parbrohm/123RF

An other problem was heat. The ENIAC consumed about 150 kW of power, the equivalent of 1,500 100-watt incandescent bulbs and produced as much heat as 1,500 incandescent bulbs. It required two twelve-horsepower blowers to conduct that heat away (Williams, 1985).

The transistor, invented at Bell Laboratories in 1947 and independently by Herbert Mataré in 1948, could perform the same switching functions as vacuum tubes. That solved both problems. Transistors are solid-state devices, so do not need a vacuum or glass envelope. There is no heater, so transistors consume much less power than vacuum tubes and produce much less heat. A transistor as a discrete electronic component is about the size of a pencil eraser, while a vacuum tube is about the size of a human thumb.

The next step was the invention in 1958 of the integrated circuit "chip" nearly simultaneously by Robert Noyce, Jack Kilby, and Kurt Lehovec. An integrated circuit packages many transistors and other electronic devices on a single semiconductor chip. The first commercial chips had tens of transistors. Modern chips can have billions.

In 1965, Gordon Moore, then director of research and development at Fairchild Semiconductor, wrote in the *Electronics* magazine that the density of devices on integrated circuits doubles about every year, and that he expected this trend to continue at least until 1975

9 For those who may not remember, Google Glass was mounted on a frame like eyeglasses. It could take pictures or video, respond to voice commands, and display information in the wearer's field of vision. Google Glass Enterprise Edition 2 was intended for workers in industrial, medical, or similar settings. Google discontinued Google Glass Enterprise Edition in March, 2023.

(Moore, 1965). Moore later revised his estimate to doubling about every two years. This came to be known as Moore's Law, although it is an empirical observation, not a physical law. Moore died in March, 2023 at the age of 94.

The significance of Moore's Law is this: the cost of a semiconductor chip is roughly proportional to the area and the computational power is roughly proportional to the number of devices. If the density doubles every 12 to 24 months, in two years or less, we can get the same computational power at half the cost or twice the power at the same cost.

People have predicted the end of Moore's Law several times in the last decade. In Chapter 3 you will meet Chenming Hu, whose invention of the FinFET transistor probably extends the life of Moore's Law to 2030 or later. (Perry, 2020) Following the FinFET will be Gate All Around and 3D complementary transistors that will further extend Moore's Law. (Radosavljevic & Kavalieros, 2022)

0.6 Computing Grows Up

The field of computing did not leap directly from 1950 to 2023. There was a lot of growing up, and it was not without growing pains.

Computers probably first came to the attention of the general public when CBS television and Remington Univac arranged to use a UNIVAC computer to attempt to predict the outcome of the U.S. presidential election of 1952 from early returns. News staff, including Walter Cronkite and Charles Collingsworth, described the computer as a "giant electronic brain" in television news programs. Programming correctly predicted the outcome of the election very early in the evening, but CBS held the prediction for several hours because they didn't trust it.

In 1957 the Soviet Union launched Sputnik, the first artificial Earth satellite. That feat convinced the American public and policy-makers that the Soviet Union was far ahead of the United States scientifically and began an effort to increase science education in the U.S.

As you saw in Section 0.5, the integrated circuit "chip" was invented in 1958. It is impossible to overstate the importance of this invention. Without integrated circuits, the other advances in computing would have been impossible.

Beginning about 1959, IBM and other computer manufacturers introduced computers that mid-sized companies could afford. In the same year, Digital Equipment Corporation released the first minicomputer, the PDP-1.

In 1969, ARPA, the U.S. Advanced Research Projects Agency began work on what would become the Internet.

Also in 1969, the Nippon Calculating Machine Corporation approached Intel with a request to design a chip set for an electronic calculator. The result was the Intel 4004, a complete four-bit CPU on a single chip, introduced in 1971. It was quickly followed by the Intel 8008, an eight-bit CPU. (That pocket calculator cost nearly $400, the equivalent of almost $3,000 in 2023 dollars.)

The first microcomputer, the Datapoint 2200, was introduced in 1971 at a price of $7,800 and equipped with 8 K bytes of memory. The first microcomputer that an individual could reasonably afford was the MITS Altair in 1975. It came as a $200 kit and included 256 bytes of memory. Paul Allen and Bill Gates wrote a version of the Basic programming language for the Altair and that was the beginning of Microsoft. Apple Computer didn't come along until 1976 and the IBM PC wasn't released until 1981.

Figure 0-4
Datapoint 2200.
Ken Shirriff

Commercial access to the Internet was available in 1989 and that same year Sir Tim Berners-Lee invented the World Wide Web. Bill Gates and Microsoft thought the Internet would be a passing fad, and so were very late to adapt to the idea of ubiquitous connectivity. Microsoft did eventually adjust to the fact that the Internet is here to stay.

Vannevar Bush's memex was realized as a pocket device with the introduction in the mid-1990s of the Palm Pilot, a personal digital assistant small enough to be carried in a pocket. Cellular phones were available then, too, but they only made phone calls. The Blackberry could send two-way text messages.

From about 1995, everyone was concerned with the so-called Y2K problem. Programmers had represented dates using only two digits for the year, and suddenly four would be needed. That caused high anxiety and the expenditure of hundreds of millions of dollars to remediate millions of computer programs amid dire predictions of disaster. About the worst that happened was that, in the early morning of January 1, 2000, the U.S. Naval Observatory briefly reported the date as "January 1, 19100." (Oops!)

Smart Phones combine the functions of that personal digital assistant and the cellular telephone. The Handspring Treo 180, in 2002, was the first fully-integrated smart phone. Now we have smart phone operating systems from Apple and Google, among others.

Remote job entry facilities allowed the use of distant computers in the 1960s, and timesharing systems were available before 1970. Cloud computing was first visualized by J.C.R. Licklider as "the Intergalactic Network" in about 1969. What we call cloud computing – self-provisioned and scalable computing – didn't arrive until the first decade of the 21st century.

We continue to make huge advances in both hardware and software. The "big thing" as this is being written is generative artificial intelligence, such as ChatGPT. Other advances, like fully autonomous automobiles, seem relatively far away despite decades of work. And, we still have to live with Wirth's Law: "software is getting slower more rapidly than hardware is becoming faster."[10]

0.7 The Internet and the World Wide Web

Most of the people reading this book are accustomed to sending a message to a friend or acquaintance anywhere in the world and having it delivered in seconds or less. An observation called Metcalfe's Law states that the value of a network to a user is proportional to the number of people one can contact. The *total* value is $n \times (n-1) / 2$, the number of connections. (Shapiro & Varian, 1998)[11] According to the Pew Research Center (2021), as of February, 2021, 77% of U.S. households had broadband Internet access at home and another 15% did not have broadband access but use smartphones as their primary means of Internet access at home. An astonishing 92% of Americans have Internet access at home or via a personal device. Of those who don't, poverty or geography explain some. There are even some Americans who choose not to have Internet access.

The Internet is an *Inter*net to emphasize that it is not a simple network, it is a network of networks. Your home or apartment router acts as a gateway to a connection to any other Internet-connected network in the world. It is *The* Internet because there's only one, connecting nearly all the world's networks. Metcalf's Law shows the enormous utility of having a single internetwork to connect all the world's networks.

The Internet had its beginnings in 1966 at ARPA, the U.S. Defense Advanced Research Projects Agency. A design based on the work of J. C. R. Licklider resulted, by 1969, in a packet-switched network of four nodes. Packet switching and the Transmission Control Protocol, TCP, both of which are integral to the modern Internet, were developed at ARPA. We will visit those fundamental concepts in detail in Chapter 5.

A very large number of standards, called RFCs, came out of the research funded by ARPA. A few of those are still in use, and many are direct ancestors of today's standards.

ARPANet was primarily a military network with universities and other sites connected under contract to ARPA. Military sites were later segregated onto their own network, with work completed in 1984. In 1986 the U.S. National Science Foundation founded a network using packet switching and the TCP protocol. The NSFNet was eventually opened to commercial access and finally operated entirely by commercial organizations since 1995.

The Internet of the 1980s was limited to electronic mail and file transfers. The University of Minnesota's Gopher[12] protocol provided a mechanism for indexing and searching Internet resources. There are still Gopher servers running, and Gopher clients are available for Win-

10 You will meet Dr. Wirth, whose name is pronounced "veert," in Chapter 6.

11 According to Shapiro and Varian, this observation was formulated by George Gilder and attributed to Metcalfe.

12 The University of Minnesota's athletic teams are the Golden Gophers.