CHAPTER 1

Introduction

1.1 Overview 1
1.2 The Main Components of a Computer 3
1.3 An Example System: Wading through the Jargon 4
1.4 Standards Organizations 18
1.5 Historical Development 19
   1.5.1 Generation Zero: Mechanical Calculating Machines (1642–1945) 20
   1.5.2 The First Generation: Vacuum Tube Computers (1945–1953) 22
   1.5.3 The Second Generation: Transistorized Computers (1954–1965) 27
   1.5.4 The Third Generation: Integrated Circuit Computers (1965–1980) 29
   1.5.5 The Fourth Generation: VLSI Computers (1980–???) 30
   1.5.6 Moore’s Law 33
1.6 The Computer Level Hierarchy 34
1.7 Cloud Computing: Computing as a Service 37
1.8 The von Neumann Model 40
1.9 Non-von Neumann Models 43
1.10 Parallel Processors and Parallel Computing 44
1.11 Parallelism: Enabler of Machine Intelligence—Deep Blue and Watson 47

Chapter Summary 49
1.1 Overview 1

- Why study computer organization and architecture?
  - Design better programs, including system software such as compilers, operating systems, and device drivers.
  - Optimize program behavior.
  - Evaluate (benchmark) computer system performance.
  - Understand time, space, and price tradeoffs.

- Computer Organization
  - We must become familiar with how various circuits and components fit together to create working computer system.
  - How does a computer work?

- Computer Architecture:
  - It focuses on the structure and behavior of the computer and refers to the logical aspects of system implementation as seen by the programmer.
  - Computer architecture includes many elements such as instruction sets and formats, operation code, data types, the number and types of registers, addressing modes, main memory access methods, and various I/O mechanisms.
  - How do I design a computer?

- The computer architecture for a given machine is the combination of its hardware components plus its instruction set architecture (ISA).
- The ISA is the agreed-upon interface between all the software that runs on the machine and the hardware that executes it. The ISA allows you to talk to the machine.

1.2 The Main Components of a Computer 3

- There is no clear distinction between matters related to computer organization and matters relevant to computer architecture.
- Principle of Equivalence of Hardware and Software:
  - Anything that can be done with software can also be done with hardware, and anything that can be done with hardware can also be done with software.
- At the most basic level, a computer is a device consisting of three pieces:
  - A processor to interpret and execute programs
  - A memory to store both data and programs
  - A mechanism for transferring data to and from the outside world.
1.3 An Example System: Wading through the Jargon

![A Typical Computer Advertisement](image)

**FIGURE 1.1** A Typical Computer Advertisement

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kilo</td>
<td>K</td>
<td>1 thousand  = 10³</td>
<td>2¹⁰ = 1024</td>
<td>Milli</td>
<td>m</td>
<td>1 thousandth = 10⁻³</td>
<td>2⁻¹⁰</td>
</tr>
<tr>
<td>Mega</td>
<td>M</td>
<td>1 million  = 10⁶</td>
<td>2²⁰</td>
<td>Micro</td>
<td>μ</td>
<td>1 millionth = 10⁻⁶</td>
<td>2⁻²⁰</td>
</tr>
<tr>
<td>Giga</td>
<td>G</td>
<td>1 billion  = 10⁹</td>
<td>2³⁰</td>
<td>Nano</td>
<td>n</td>
<td>1 billionth = 10⁻⁹</td>
<td>2⁻³⁰</td>
</tr>
<tr>
<td>Tera</td>
<td>T</td>
<td>1 trillion = 10¹²</td>
<td>2⁴⁰</td>
<td>Pico</td>
<td>p</td>
<td>1 trillionth = 10⁻¹²</td>
<td>2⁻⁴⁰</td>
</tr>
<tr>
<td>Peta</td>
<td>P</td>
<td>1 quadrillion = 10¹⁵</td>
<td>2⁵⁰</td>
<td>Femto</td>
<td>f</td>
<td>1 quadrillionth = 10⁻¹⁵</td>
<td>2⁻⁵⁰</td>
</tr>
<tr>
<td>Exa</td>
<td>E</td>
<td>1 quintillion = 10¹⁸</td>
<td>2⁶⁰</td>
<td>Atto</td>
<td>a</td>
<td>1 quintillionth = 10⁻¹⁸</td>
<td>2⁻⁶⁰</td>
</tr>
<tr>
<td>Zetta</td>
<td>Z</td>
<td>1 sextillion = 10²¹</td>
<td>2⁷⁰</td>
<td>Zepto</td>
<td>z</td>
<td>1 sextillionth = 10⁻²¹</td>
<td>2⁻⁷⁰</td>
</tr>
<tr>
<td>Yotta</td>
<td>Y</td>
<td>1 septillion = 10²⁴</td>
<td>2⁸⁰</td>
<td>Yocto</td>
<td>y</td>
<td>1 septillionth = 10⁻²⁴</td>
<td>2⁻⁸⁰</td>
</tr>
</tbody>
</table>

**TABLE 1.1** Common Prefixes Associated with Computer Organization and Architecture

- Clock frequencies are measured in cycles per second, or Hertz
  - Hertz = clock cycles per second (frequency)
  - 1MHz = 1,000,000Hz
  - Processor speeds are measured in MHz or GHz.
- Byte = a unit of storage
  - 1KB = 2¹⁰ = 1,024 Bytes
  - 1MB = 2²⁰ = 1,048,576 Bytes
  - 1GB = 2³⁰ = 1,099,511,627,776 Bytes
  - Main memory (RAM) is measured in GB
  - Disk storage is measured in GB for small systems, TB (2⁴⁰) for large systems.
• Measures of time and space:
  o Millisecond = 1 thousandth \((10^{-3})\) of a second
    ▪ Hard disk drive access times are often 10 to 20 milliseconds.
  o Nanosecond = 1 billionth \((10^{-9})\) of a second
    ▪ Main memory access times are often 50 to 70 nanoseconds.
  o Micron (micrometer) = 1 millionth \((10^{-6})\) of a meter
    ▪ Circuits on computer chips are measured in microns.
• A bus operating at 133MHz has a cycle time of 7.52 nanoseconds:
  o \(133,000,000\) cycles/second = 7.52ns/cycle
  o Bus: a group of wires that moves data and instruction to various places within the computer
• Computers with large main memory capacity can run larger programs with greater speed than computers having small memories.
• SDRAM: Synchronous Dynamic Random Access Memory
  o RAM is an acronym for random access memory.
  o Random access means that memory contents can be accessed directly if you know its location.
• Cache is a type of temporary memory that can be accessed faster than RAM.
  o The cache in our system has a capacity of kilobytes (KB), which is much smaller than main memory.
  o Level 1 cache (L1): a small, fast memory cache that is built into the microprocessor chip and helps speed up access to frequently used data
  o Level 2 cache (L2): a collection of fast, built-in memory chips situated between the microprocessor and main memory
  o In Chapter 6 you will learn how cache works, and that a bigger cache isn’t always better.
• Hard Drive:
  o SATA: Serial Advanced Technology Attachment
  o EIDE: Enhanced Integrated Drive Electronics
• USB:
  o USB (Universal Serial Bus) is a popular external bus that supports Plug-and-Play (the ability to configure devices automatically) as well as hot plugging (the ability to add and remove devices while the computer is running).
• Ports:
  o Whereas the system bus is responsible for all data movement internal to the computer, ports allow movement of data to and from devices external to the computer.
• Serial ports vs. Parallel ports:
  o Serial ports transfer data by sending a series of electrical pulses across one or two data lines. Parallel ports use at least eight data lines, which are energized simultaneously to transmit data.
• Peripheral Component Interconnect (PCI) is one such I/O bus that supports the connection of multiple peripheral devices. PCI, developed by the Intel Corporation, operates at high speeds and also supports Plug-and-Play such as PCI modem and sound card.
• AGP (Accelerated Graphical Port) graphics card
1.4 Standards Organizations

- The Institute of Electrical and Electronic Engineers (IEEE)
  - Promotes the interests of the worldwide electrical engineering community.
  - Establishes standards for computer components, data representation, and signaling protocols, among many other things.
- The International Telecommunications Union (ITU)
  - Concerns itself with the interoperability of telecommunications systems, including data communications and telephony.
- National groups establish standards within their respective countries:
  - The American National Standards Institute (ANSI)
  - The British Standards Institution (BSI)
- The International Organization for Standardization (ISO)
  - Establishes worldwide standards for everything from screw threads to photographic film.
  - Is influential in formulating standards for computer hardware and software, including their methods of manufacture.
1.5 Historical Development

In modern times computer evolution is usually classified into four generations according to the salient technology of the era.

1.5.1 Generation Zero: Mechanical Calculating Machines (1642–1945)

- Calculating Clock - Wilhelm Schickard (1592 - 1635).
- Pascaline - Blaise Pascal (1623 - 1662).
- Difference Engine - Charles Babbage (1791 - 1871), also designed but never built the Analytical Engine.
- Punched card tabulating machines - Herman Hollerith (1860 - 1929).

1.5.2 The First Generation: Vacuum Tube Computers (1945–1953)

- Electronic Numerical Integrator and Computer (ENIAC)
  - John Mauchly and J. Presper Eckert, University of Pennsylvania, introduced to the public in 1946
  - The first all-electronic, general-purpose digital computer.
  - This machine used 17,468 vacuum tubes, occupied 1,800 square feet of floor space, weighted 30 tons, and consumed 174 kilowatts of power.
- Vacuum tubes are still used in audio amplifiers.

1.5.3 The Second Generation: Transistorized Computers (1954–1965)

- In 1948, three researchers with Bell Laboratories – John Bardeen, Walter Brattain, and William Shockley – invented the transistor.
- Transistors consume less power than vacuum tubes, are smaller, and work more reliably.
  - Control Data Corporation (CDC) under the Seymour Cray, built CDC 6600, the world’s first supercomputer. The $10 million CDC 6600 could perform 10 million instructions per second, used 60-bit words, and had an astounding 128 kilowords of main memory.
1.5.4 The Third Generation: Integrated Circuit Computers (1965–1980)

- Jack Kilby invented the integrated circuit (IC) or microchip.
- Integrated Circuit: Multiple transistor were integrated onto on chip
- IBM 360
- DEC PDP-8 and PDP-11
- The Cray-1, in stark contrast to the CDC 6600, could execute over 160 million instructions per second and could support 8 megabytes of memory.

1.5.5 The Fourth Generation: VLSI Computers (1980–????)

- VLSI (Very Large Scale Integration): more than 10,000 components per chip.
- ENIAC-on-a-chip project, 1997
- VLSI allowed Intel, in 1971, to create the world’s first microprocessor, the 4004, which was a fully functional, 4-bit system that ran at 108KHz.
- Intel also introduced the random access memory (RAM) chip, accommodating 4 kilobits of memory on a single chip.

1.5.6 Moore’s Law

- Visit
  - http://en.wikipedia.org/wiki/Moore’s_law
- In 1965, Intel founder Gordon Moore stated, “The density of transistors in an integrated circuit will double every year.”
- The current version of this prediction is usually conveyed as “the density of silicon chips doubles very 18 months.”
1.6 The Computer Level Hierarchy

We call the hypothetical computer at each level a virtual machine. Each level’s virtual machine executes its own particular set of instructions, calling upon machines at lower levels to carry out the tasks when necessary.

Level 6, the User Level, is composed of applications such as world processors, graphics packages, or games.

Level 5, the High-Level Language Level, consists of languages such as C, C++, FORTRAN, Lisp, Pascal, and Prolog.
- These languages must be translated (using either a compiler or an interpreter) to a language the machine can understand.
- Compiled languages are translated into assembly language and then assembled into machine code (They are translated to the next lower level).
- Even though a programmer must know about data types and the instructions available for those types, she need not know about how those types are actually implemented.

Level 4, the Assembly Language Level, encompasses some type of assembly language.
- One-to-one translation: One assembly language instruction is translated to exactly one machine language.

- Level 3, the System Software Level, deals with operating system instructions.
  - This level is responsible for multiprogramming, protecting memory, synchronizing processes, and various other important functions.
  - Often, instructions translated from assembly language to machine language are passed through this level unmodified.

- Level 2, the Instruction Set Architecture (ISA) or Machine Level, consists of the machine language recognized by the particular architecture of the computer system. We will study ISA in Chapter 4 and 5.

- Level 1, the Control Level, is where a control unit makes sure that instructions are decoded and executed properly and that data is moved where and when it should be.
  - Control units can be designed in one of two ways: They can be hardwired or they can be microprogrammed.
  - In hardwired control units, control signals emanated from blocks of digital logic components: fast, very difficult to modify.
  - A microprogram is a program written in a low-level language that is implemented directly by the hardware: slow, easily to modify.

- Level 0, the Digital Logic Level, is where we find the physical components to the computer system: the gates and wires. Chapter 3 presents the Digital Logic Level.
1.7 Cloud Computing: Computing as a Service 37

- Computer users typically do not care about terabytes of storage and gigahertz of processor speed.
- Many companies outsource their data centers to third-party specialists, who agree to provide computing services for a fee. These arrangements are managed through service-level agreements (SLAs).
- Rather than pay a third party to run a company-owned data center, another approach is to buy computing services from someone else’s data center and connect to it via the Internet.
- A Cloud computing platform is defined in terms of the services that it provides rather than its physical configuration.
- Cloud computing models:
  - Software as a Service (SaaS):
    - A Cloud provider might offer an entire application over the Internet, with no components installed locally.
    - The consumer of this service buy application services. The consumer of this service does not maintain the application or need to be at all concerned with the infrastructure in any way.
- Well-known examples include Gmail, Dropbox, GoToMeeting, and Netflix.
  
  **Platform as a Service (PaaS):**
  - PaaS provides server hardware, operating systems, database services, security components, and backup and recovery services. The PasS provider manages performance and availability of the environment,
  - The customer manages the *applications* hosted in the PassS Cloud. The customer is typically billed monthly per megabytes of storage, processor utilization and megabyte of data transferred.
  - Well-known PaaS providers include Google App Engine and Microsoft Windows Azure Cloud Services.
  
  **Infrastructure as a Service (IaaS):**
  - IaaS provides only server hardware, secure network access to the servers, and backup and recovery services.
  - The customer is responsible for all system software including the operating system and databases. IassS is typically billed by the number of *virtual machines* used, megabytes of storage, and megabytes of data transferred but at a lower rate than PassS
  - Well-known IaaS platforms include Amazon EC2, Google Compute Engine, Microsoft Azure Services Platform, Rackspace, and HP Cloud.
  
  Cloud storage is a limited type of IaaS that includes services such as Dropbox, Google Drive, and Amazon.com’s Cloud Drive.
1.8 The von Neumann Model

![Diagram of the von Neumann Architecture]

Today’s stored-program computers have the following characteristics:
- Three hardware systems:
  - A central processing unit (CPU)
  - A main memory system
  - An I/O system
- The capacity to carry out sequential instruction processing.
- A single data path between the CPU and main memory.
  - This single path is known as the von Neumann bottleneck.
- This architecture runs programs in what is known as the **von Neumann execution cycle** (also called the fetch-decode-execute cycles), which describes how the machine works. One iteration of the cycle is as follows:
  - The control unit **fetches** the next instruction from memory using the program counter to determine where the instruction is located.
  - The instruction is **decoded** into a language that the ALU can understand.
  - Any data operands required to execute the instruction are fetched from memory and placed into registers within the CPU.
  - The ALU **executes** the instruction and places results in registers or memory.
FIGURE 1.6 The Modified von Neumann Architecture, Adding a System Bus
1.9 Non-von Neumann Models 43

- von Neumann computer execute instructions sequentially and are therefore extremely well suited to **sequential** processing.
- **Harvard architecture**: Computer systems have **separate** buses for data and instructions.
- Many non-von Neumann systems provide **special-purpose** processors to offload work from the main CPU.
1.10 Parallel Processors and Parallel Computing

- Parallel processors are technically not classified as von Neumann machines because they do not process instructions sequentially.
- Parallel processing allows a computer to *simultaneously* work on subparts of a problem.
- Parallel computing
  - In the late 1960s, high-performance computer systems were equipped with dual processors to increase computational throughput.
  - In the 1970s supercomputer systems were introduced with 32 processors.
  - Supercomputers with 1,000 processors were built in the 1980s.
  - In 1999, IBM announced its Blue Gene system containing over *1 million* processors, each with its own dedicated memory.
- Multicore architectures are parallel processing machines that allow for multiple processing units (often called cores) on a *single* chip.
- Each core has its own ALU and set of registers, but all processors **share memory** and other resources.
- “Dual core” differs from “Dual processor.”
  - Dual-processor machines, for example, have two processors, but each processor plugs into the motherboard *separately*.
  - All cores in multicore machines are integrated into the *same* chip.
- Multi-core systems provide the ability to multitask
  - For example, browse the Web while burning a CD
- Multithreaded applications spread mini-processes, threads, across one or more processors for increased throughput.
  - Programs are divided up into thread, which can be thought of as mini-processes.
  - For example, a web browser is multithreaded; one thread can download text, which each image is controlled and downloaded by a separated thread.
- Examples of non-von Neumann languages including:
  - Lucid: for dataflow
  - QCL: Quantum Computation Language for quantum computer
  - **VHDL** and Verilog: Languages used to program FPGAs
1.11 Parallelism: Enabler of Machine Intelligence — Deep Blue and Watson

- The quest for machine intelligence has been ongoing for over 300 years.
- The 20th Century witnessed the first machines that could be human grandmasters at chess when Deep Blue beat Garry Kasparov in 1997.
- But the machine and the algorithm relied on a brute force solution, although impressive, hardly “intelligent” by any measure.
- Any definition of true machine “intelligence” would have to include the ability to acquire new knowledge independent of direct human intervention, and the ability to solve problems using incomplete and perhaps contradictory information.
- This is precisely what IBM achieved when it is built the machine named Watson.
- Watson proved this when it beats two human Jeopardy! champions on February 16, 2011.
- Watson had a massively parallel architecture dubbed DeepQA (Deep Question and Answer).
- The system relied on 90 IBM POWER 750 servers.
- Each server was equipped with four POWER7 processors, and each POWER7 processor had eight cores, giving a total of 2880 processor cores.
- While playing Jeopardy!, each core had access to 16TB of main memory and 4TB of storage.
- Watson's technology has been put to work in treating cancer.
  - Commercial products based on Watson technology, including “Interactive Care Insights for Oncology” and “Interactive Care Reviewer,” are now available.
- Watson is also becoming more compact: Watson can now be run on a single POWER 750 server.
- Watson has surely given us a glimpse into the future of computing.
Chapter Summary 49

- A brief overview of computer organization and computer architecture.
- Principle of Equivalence of Hardware and Software
- Moore’s Law
- The von Neumann architecture is predominant in today’s general-purpose computers.