Figure 1.1 The 2X vs. 10Y bytes ambiguity was resolved by adding a binary notation for all the common size terms. In the last column we note how much larger the binary term is than its corresponding decimal term, which is compounded as we head down the chart. These prefixes work for bits as well as bytes, so gigabit (Gb) is 109 bits while gibibits (Gib) is 230 bits. The society that runs the metric system created the decimal prefixes, with the last two proposed only in 2019 in anticipation of the global capacity of storage systems. All the names are derived from the entymology in Latin of the powers of 1000 that they represent.

Figure 1.2 The number manufactured per year of tablets and smart phones, which reflect the PostPC era, versus personal computers and traditional cell phones. Smart phones represent the recent growth in the cell phone industry, and they passed PCs in 2011. PCs, tablets, and traditional cell phone categories are declining. The peak volume years text are 2011 for cell phones, 2013 for PCs, and 2014 for tablets. PCs fell from 20% of total units shipped in 2007 to 10% in 2018.

Figure 1.3 A simplified view of hardware and software as hierarchical layers, shown as concentric circles with hardware in the center and applications software outermost. In complex applications, there are often multiple layers of application software as well. For example, a database system may run on top of the systems software hosting an application, which in turn runs on top of the database.

Figure 1.4 C program compiled into assembly language and then assembled into binary machine language. Although the translation from high-level language to binary machine language is shown in two steps, some compilers cut out the middleman and produce binary machine language directly. These languages and this program are examined in more detail in Chapter 2.

Figure 1.5 The organization of a computer, showing the five classic components. The processor gets instructions and data from memory. Input writes data to memory, and output reads data from memory. Control sends the signals that determine the operations of the datapath, memory, input, and output.

Figure 1.6 Each coordinate in the frame buffer on the left determines the shade of the corresponding coordinate for the raster scan CRT display on the right. Pixel (X0, Y0) contains the bit pattern 0011, which is a lighter shade on the screen than the bit pattern 1101 in pixel (X1, Y1).

Figure 1.7 Components of the Apple iPhone Xs Max cell phone. At the left is the capacitive multitouch screen and LCD display. Next to it is the battery. To the far right is the metal frame that attaches the LCD to the back of the iPhone. The small components surrounding in the center are what we think of as the computer; they are not simple rectangles to fit compactly inside the case next to the battery. Figure 1.8 shows a close-up of the board to the left of the metal case, which is the logic printed circuit board that contains the processor and the memory (Courtesy TechInsights, www.techIngishts.com).

Figure 1.8 The logic board of Apple iPhone Xs Max in Figure 1.7. The large integrated circuit in the middle is the Apple A12 chip, which contains two large ARM processor cores and four little ARM processor cores that run at 2.5 GHz, as well as 2 GiB of main memory inside the package. Figure 1.9 shows a photograph of the processor chip inside the A12 package. A similar-sized chip on a symmetric board attached to the back is the 64 GiB flash memory chip for nonvolatile storage. The other chips on the board include power management integrated controller and audio amplifier chips (Courtesy TechInsights, www.techIngishts.com).

Figure 1.9 The processor integrated circuit inside the A12 package. The size of chip is 8.4 by 9.91 mm, and it was manufactured originally in a 7-nm process (see Section 1.5). It has two identical ARM big processors or cores in the lower middle of the chip, four small cores on the lower right of the chip, a graphical processor unit (GPU) on the far right (see Section 6.6), and a domain-specific accelerator for neural networks (see Section 6.7), called the NPU, on the far left. In the middle are second-level cache memories (L2) for the big and small cores (see Chapter 5). At the top and bottom of the chip are interfaces to main memory (DDR DRAM) (Courtesy TechInsights, www.techinsights.com, and AnandTech, www.anandtech.com).

Figure 1.10 Relative performance per unit cost of technologies used in computers over time. Source: Computer Museum, Boston, with 2020 extrapolated by the authors. See  **Section 1.13**.

Figure 1.11 Growth of capacity per DRAM chip over time. The y-axis is measured in kibibits (210 bits). The DRAM industry quadrupled capacity almost every three years, a 60% increase per year, for 20 years. In recent years, the rate has slowed down and is somewhat closer to doubling every three years. With the slowing of Moore’s Law and difficulties in reliable manufacturing of smaller DRAM cells given the challenging aspect ratios of their three-dimensional structure.

Figure 1.12 The chip manufacturing process. After being sliced from the silicon ingot, blank wafers are put through 20 to 40 steps to create patterned wafers (see Figure 1.13). These patterned wafers are then tested with a wafer tester, and a map of the good parts is made. Then, the wafers are diced into dies (see Figure 1.9). In this figure, one wafer produced 20 dies, of which 17 passed testing. (X means the die is bad.) The yield of good dies in this case was 17/20, or 85%. These good dies are then bonded into packages and tested one more time before shipping the packaged parts to customers. One bad packaged part was found in this final test.

Figure 1.13 A 12-inch (300-mm) wafer this 10nm wafer contains 10th Gen Intel® Core™ processors, code-named “Ice Lake” (Courtesy Intel). The number of dies on this 300-mm (12-inch) wafer at 100% yield is 506. According to AnandTech,1 each Ice Lake die is 11.4 by 10.7 mm. The several dozen partially rounded chips at the boundaries of the wafer are useless; they are included because it is easier to create the masks used to pattern the silicon. This die uses a 10-nm technology, which means that the smallest features are approximately 10 nm in size, although they are typically somewhat smaller than the actual feature size, which refers to the size of the transistors as “drawn” versus the final manufactured size.

Figure 1.14 The capacity, range, and speed for a number of commercial airplanes. The last column shows the rate at which the airplane transports passengers, which is the capacity times the cruising speed (ignoring range and takeoff and landing times).

Figure 1.15 The basic components of performance and how each is measured.

Figure 1.16 Clock rate and Power for Intel x86 microprocessors over nine generations and 36 years. The Pentium 4 made a dramatic jump in clock rate and power but less so in performance. The Prescott thermal problems led to the abandonment of the Pentium 4 line. The Core 2 line reverts to a simpler pipeline with lower clock rates and multiple processors per chip. The Core i5 pipelines follow in its footsteps.

Figure 1.17 Growth in processor performance since the mid-1980s. This chart plots performance relative to the VAX 11/780 as measured by the SPECint benchmarks (see Section 1.11). Prior to the mid-1980s, processor performance growth was largely technology-driven and averaged about 25% per year. The increase in growth to about 52% since then is attributable to more advanced architectural and organizational ideas. The higher annual performance improvement of 52% since the mid-1980s meant performance was about a factor of seven higher in 2002 than it would have been had it stayed at 25%. Since 2002, the limits of power, available instruction-level parallelism, and long memory latency have slowed uniprocessor performance recently, to about 3.5% per year. (From Hennssey JL, Patterson DA. Computer Architecture: A Qualitative Approach, ed 6. Waltham, MA: Elsevier, 2017.)

Figure 1.18 SPECspeed 2017 Integer benchmarks running on a 1.8 GHz Intel Xeon E5-2650L. As the equation on page 35 explains, execution time is the product of the three factors in this table: instruction count in billions, clocks per instruction (CPI), and clock cycle time in nanoseconds. SPECratio is simply the reference time, which is supplied by SPEC, divided by the measured execution time. The single number quoted as SPECspeed 2017 Integer is the geometric mean of the SPECratios. SPECspeed 2017 has multiple input files for perlbench, gcc, x264, and xz. For this figure, execution time and total clock cycles are the sum running times of these programs for all inputs.

Figure 1.19 SPECpower\_ssj2008 running on a dual socket 2.2 GHz Intel Xeon Platinum 8276L with 192 GiB of DRAM and one 80 GB SSD disk.

Figure 1.20 Optimizations of matrix multiply program in Python in the next five chapters of this book.

Figure 1.21 Price of memory per gigabyte between 1975 and 2020. (Source: <https://jcmit.net/memoryprice.htm>)