Decimal term	Abbreviation	Value	Binary term	Abbreviation	Value	% Larger
kilobyte	KB	10^{3}	kibibyte	KiB	2^{10}	2%
megabyte	MB	10^{6}	mebibyte	MiB	2^{20}	5%
gigabyte	GB	10^{9}	gibibyte	GiB	2^{30}	7%
terabyte	TB	10^{12}	tebibyte	TiB	2^{40}	10%
petabyte	PB	10^{15}	pebibyte	PiB	2^{50}	13%
exabyte	EB	10^{18}	exbibyte	EiB	2^{60}	15%
zettabyte	ZB	10^{21}	zebibyte	ZiB	2^{70}	18%
yottabyte	YB	10^{24}	yobibyte	YiB	2^{80}	21%
ronnabyte	RB	10^{27}	robibyte	RiB	2^{90}	24%
queccabyte	QB	10^{30}	quebibyte	QiB	2^{100}	27%

Figure 1.1 The 2^X vs. 10^Y 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.

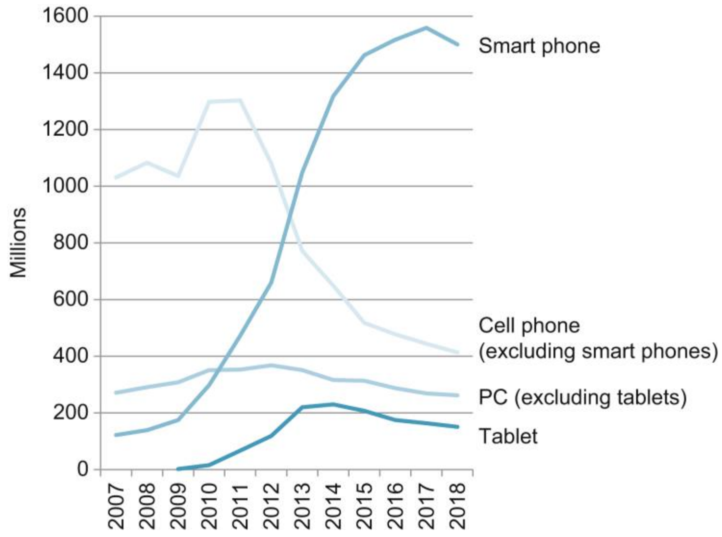

Figure 1.2 The number manufactured per year of tablets and smart phones, which reflect the post-PC era, versus personal computers and traditional cell phones. Smart phones represent the recent growth in the cell phone industry, and they passed PCs in 2011. PCs, tablets, and traditional cell phone categories are declining. The peak volume years 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.

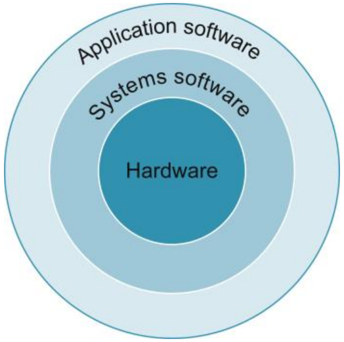

Figure 1.3 A simplified view of hardware and software as hierarchical layers, shown as concentric circles with hardware in the center and application software outermost. In complex applications, there are often multiple layers of application software as well. For example, a database system may run on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

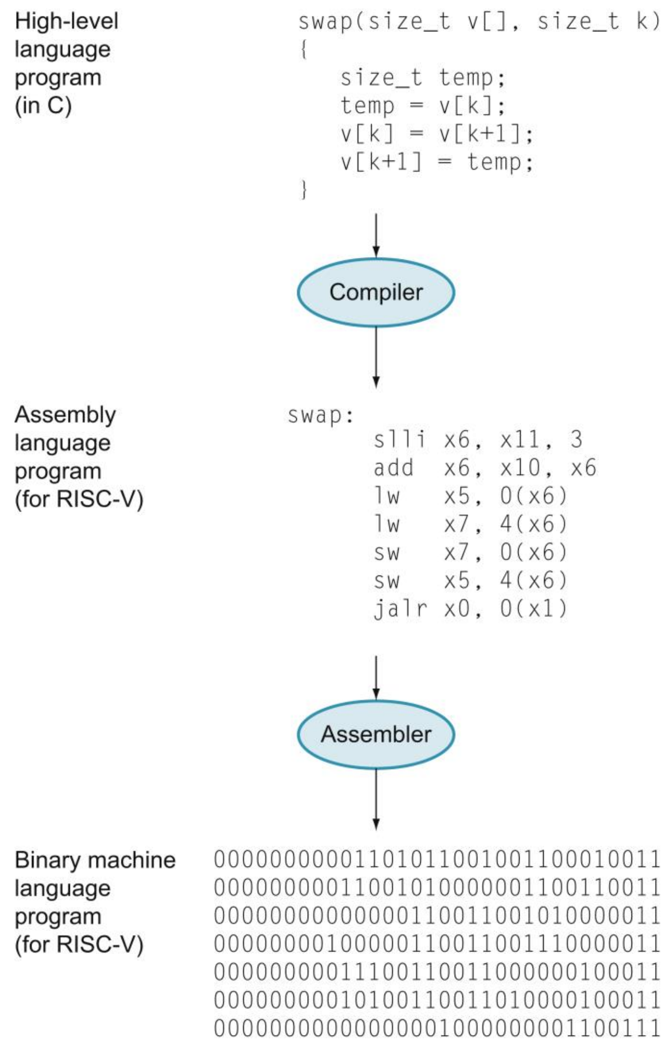

Figure 1.4 C program compiled into assembly language and then assembled into binary machine language. Although the translation from high-level language to binary machine language is shown in two steps, some compilers cut out the middleman and produce binary machine language directly. These languages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

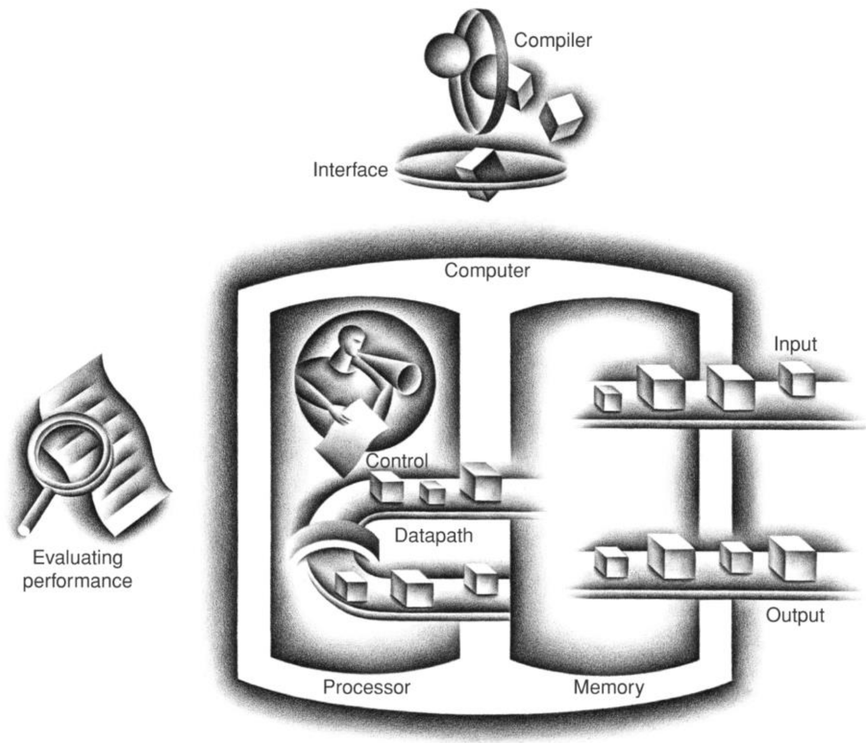

Figure 1.5 The organization of a computer, showing the five classic components. The processor gets instructions and data from memory. Input writes data to memory, and output reads data from memory. Control sends the signals that determine the operations of the datapath, memory, input, and output.

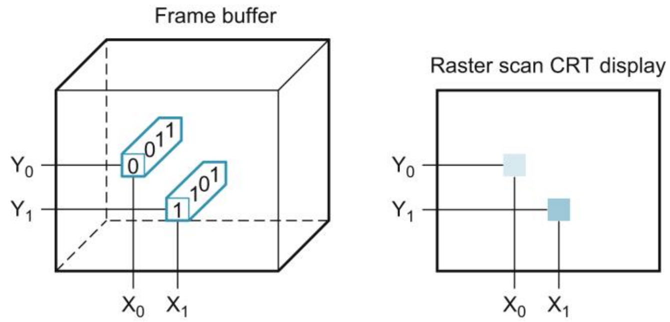

Figure 1.6 Each coordinate in the frame buffer on the left determines the shade of the corresponding coordinate for the raster scan CRT display on the right. Pixel (X_0, Y_0) contains the bit pattern 0011, which is a lighter shade on the screen than the bit pattern 1101 in pixel (X_1, Y_1).

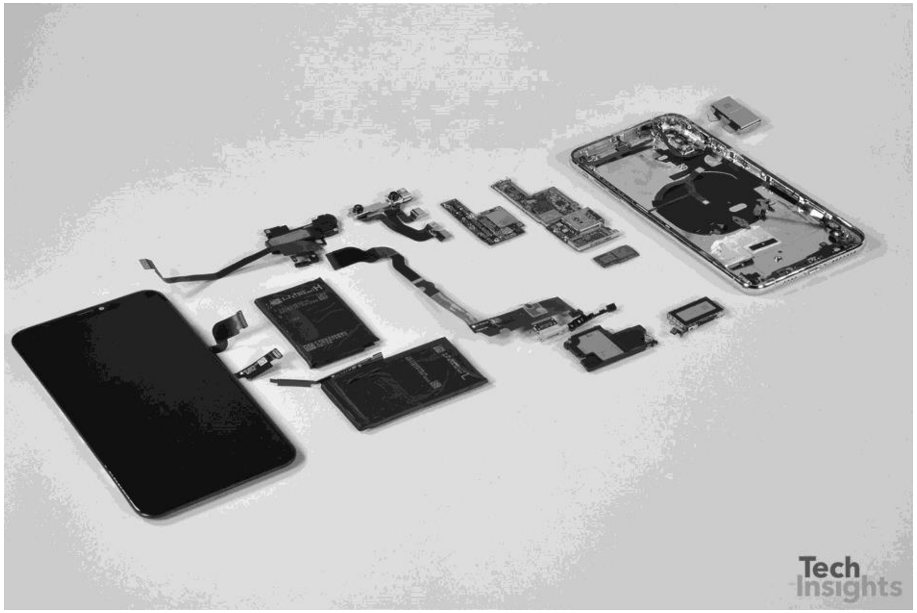

FIGURE 1.7 Components of the Apple iPhone XS Max cell phone. At the left is the capacitive multitouch screen and LCD display. Next to it is the battery. To the far right is the metal frame that attaches the LCD to the back of the iPhone. The small components in the center are what we think of as the 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 memory. (Courtesy TechIngishts, www.techIngishts.com)

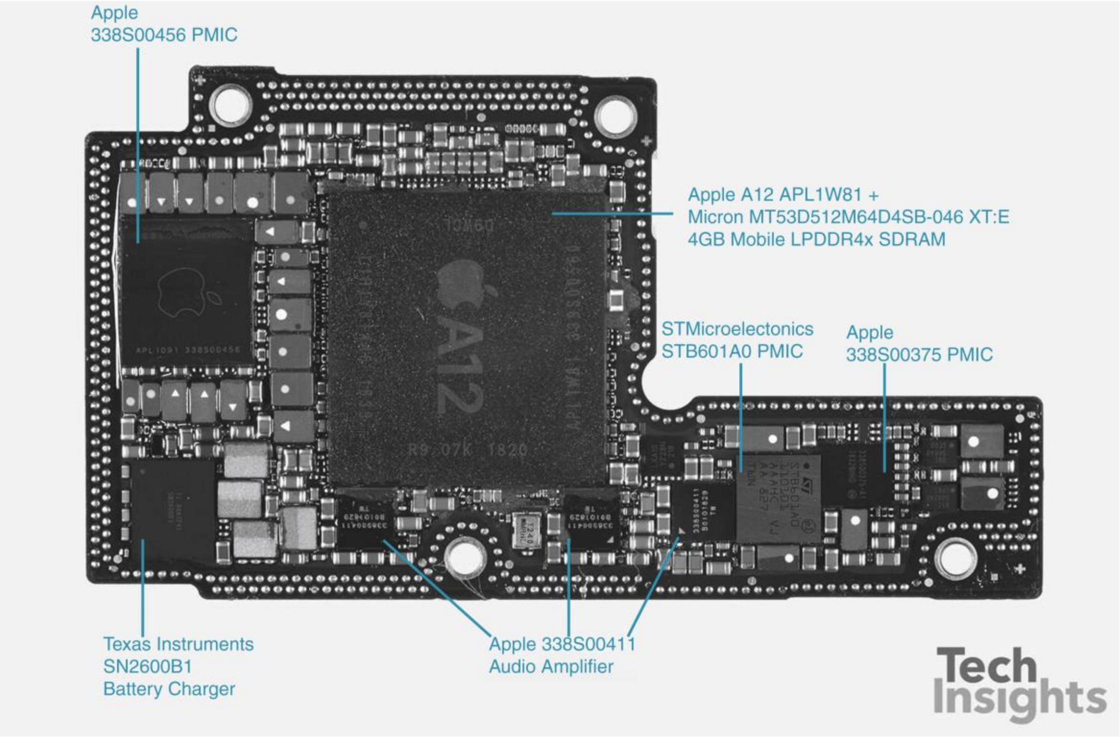

FIGURE 1.8 The logic board of Apple iPhone XS Max in Figure 1.7. The large integrated circuit in the middle is the Apple A12 chip, which contains two large and four small ARM processor cores that run at 2.5 GHz, as well as 2 GiB of main memory inside the package. Figure 1.9 shows a photograph of the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

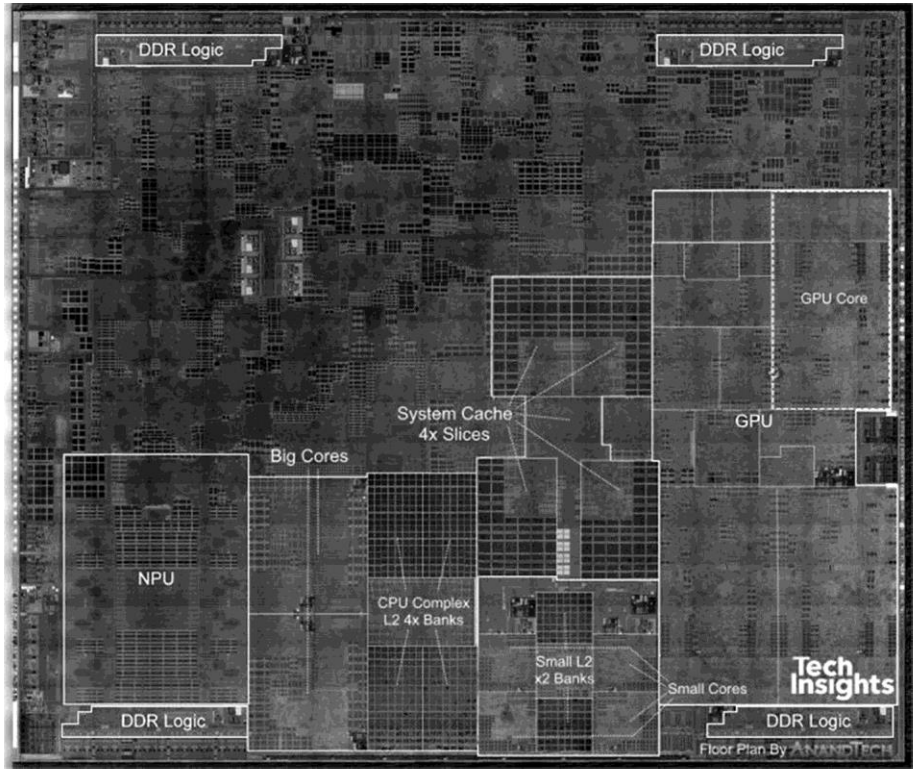

FIGURE 1.9 The processor integrated circuit inside the A12 package. The size of chip is 8.4 by 9.91 mm, and it was manufactured originally in a 7-nm process (see Section 1.5). It has two identical ARM processors or cores in the lower middle of the chip, four small cores on the lower right of the chip, a graphics processing 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 memory (L2) banks for the big and small cores (see Chapter 5). At the top and bottom of the chip are interfaces to the main memory (DDR DRAM). (Courtesy TechInsights, www.techinsights.com)

Year	Technology used in computers	Relative performance/unit cost
1951	Vacuum tube	1
1965	Transistor	35
1975	Integrated circuit	900
1995	Very large-scale integrated circuit	2,400,000
2020	Ultra large-scale integrated circuit	500,000,000,000

Figure 1.10 Relative performance per unit cost of technologies used in computers over time. Source: Computer Museum, Boston, with 2013 extrapolated by the authors.

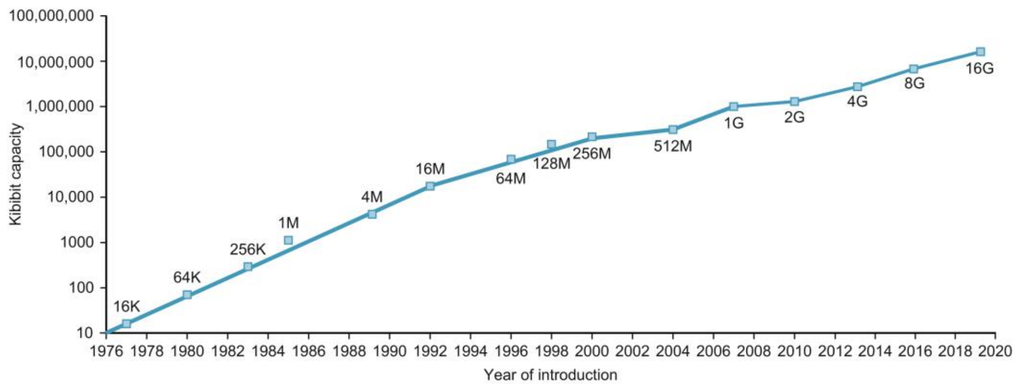

Figure 1.11 Growth of capacity per DRAM chip over time. The *y*-axis is measured in kibibits (210 bits). The DRAM industry quadrupled capacity almost every three years, a 60% increase per year, for 20 years. In recent years, the rate has slowed down and is somewhat closer to doubling every three years. 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.

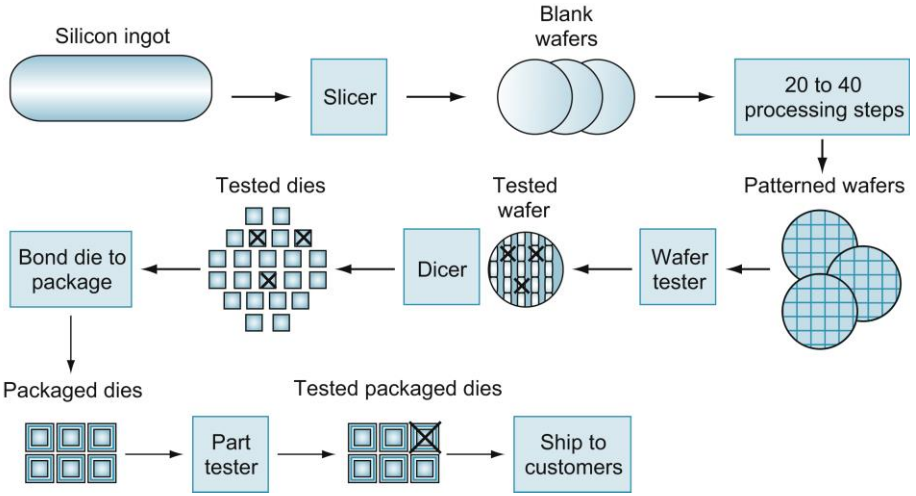

Figure 1.12 The chip manufacturing process. After being sliced from the silicon ingot, blank wafers are put through 20 to 40 steps to create patterned wafers (see Figure 1.13). These patterned wafers are then tested with a wafer tester, and a map of the good parts is made. Next, the wafers are diced 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.

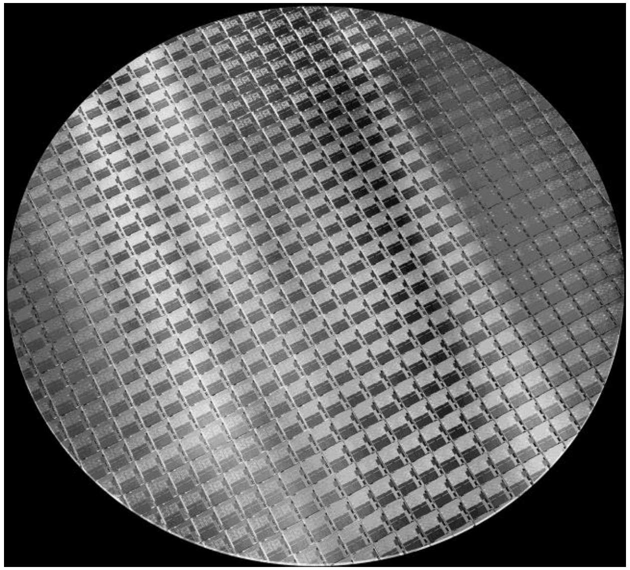

Figure 1.13 A 12-inch (300mm) wafer this 10nm wafer contains 10th Gen Intel® Core™ processors, code-named "Ice Lake" (Courtesy Intel). The number of dies on this 300 mm (12 inch) wafer at 100% yield is 506. According to AnandTech1, each Ice Lake die is 11.4 by 10.7 mm. The several dozen partially rounded chips at the boundaries of the wafer are useless; they are included because it's easier to create the masks used to pattern the silicon. This die uses a 10-nanometer 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.

Airplane	Passenger capacity	Cruising range (miles)	Cruising speed (m.p.h.)	Passenger throughput (passengers × m.p.h.)
Boeing 737	240	3000	564	135,360
BAC/Sud Concorde	132	4000	1350	178,200
Boeing 777-200LR	301	9395	554	166,761
Airbus A380-800	853	8477	587	500,711

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).

Components of performance	Units of measure
CPU execution time for a program	Seconds for the program
Instruction count	Instructions executed for the program
Clock cycles per instruction (CPI)	Average number of clock cycles per instruction
Clock cycle time	Seconds per clock cycle

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

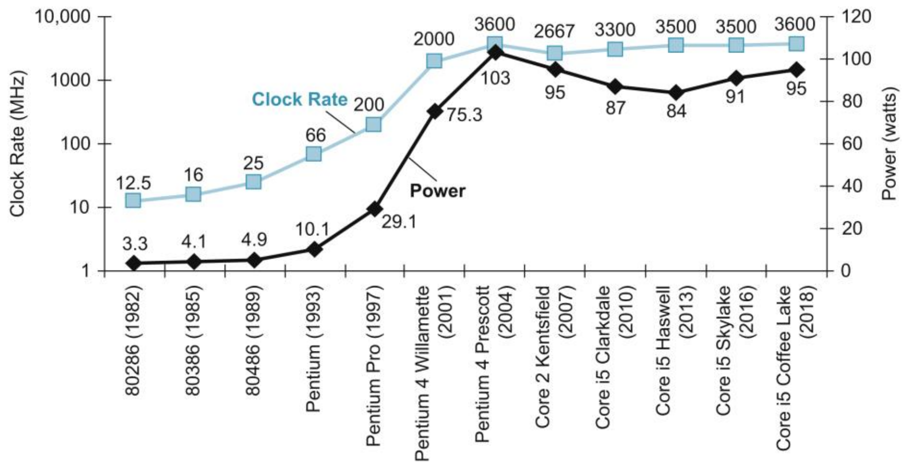

Figure 1.16 Clock rate and power for Intel x86 microprocessors over nine generations and 36 years. The Pentium 4 made a dramatic jump in clock rate and power but less so in performance. The Prescott thermal problems led to the abandonment of the Pentium 4 line. The Core 2 line reverts to a simpler pipeline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

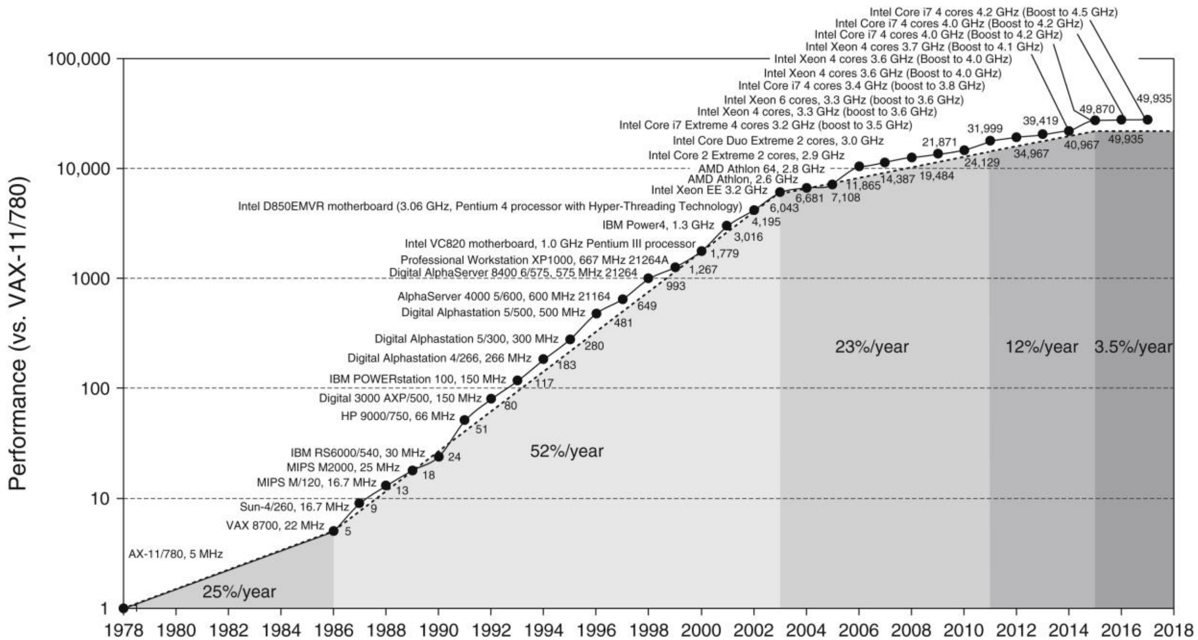

Figure 1.17 Growth in processor performance since the mid-1980s. This chart plots performance relative to the VAX 11/780 as measured by the SPECint benchmarks (see Section 1.11). Prior to the mid-1980s, processor performance growth was largely technologydriven and averaged about 25% per year. The increase 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 larger 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.

Description	Name	Instruction Count x 10^9	CPI	Clock cycle time (seconds x 10^{-9})	Execution Time (seconds)	Reference Time (seconds)	SPECratio
Perl interpreter	perlbench	2684	0.42	0.556	627	1774	2.83
GNU C compiler	gcc	2322	0.67	0.556	863	3976	4.61
Route planning	mcf	1786	1.22	0.556	1215	4721	3.89
Discrete Event simulation - computer network	omnetpp	1107	0.82	0.556	507	1630	3.21
XML to HTML conversion via XSLT	xalancbmk	1314	0.75	0.556	549	1417	2.58
Video compression	x264	4488	0.32	0.556	813	1763	2.17
Artificial Intelligence: alpha-beta tree search (Chess)	deepsjeng	2216	0.57	0.556	698	1432	2.05
Artificial Intelligence: Monte Carlo tree search (Go)	leela	2236	0.79	0.556	987	1703	1.73
Artificial Intelligence: recursive solution generator (Sudoku)	exchange2	6683	0.46	0.556	1718	2939	1.71
General data compression	xz	8533	1.32	0.556	6290	6182	0.98
Geometric mean	–	–	–	–	–	–	2.36

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.

Target Load %	Performance (ssj_ops)	Average Power (watts)
100%	4,864,136	347
90%	4,389,196	312
80%	3,905,724	278
70%	3,418,737	241
60%	2,925,811	212
50%	2,439,017	183
40%	1,951,394	160
30%	1,461,411	141
20%	974,045	128
10%	485,973	115
0%	0	48
Overall Sum	26,815,444	2,165
$\sum$ssj_ops / $\sum$power =		12,385

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.

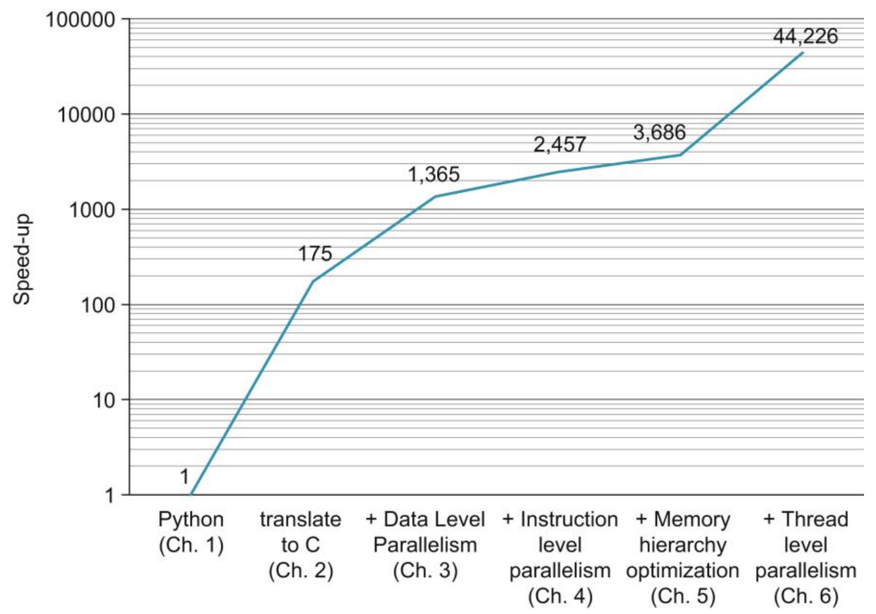

Figure 1.20 Optimizations of matrix multiply program in Python in the next five chapters of this book.

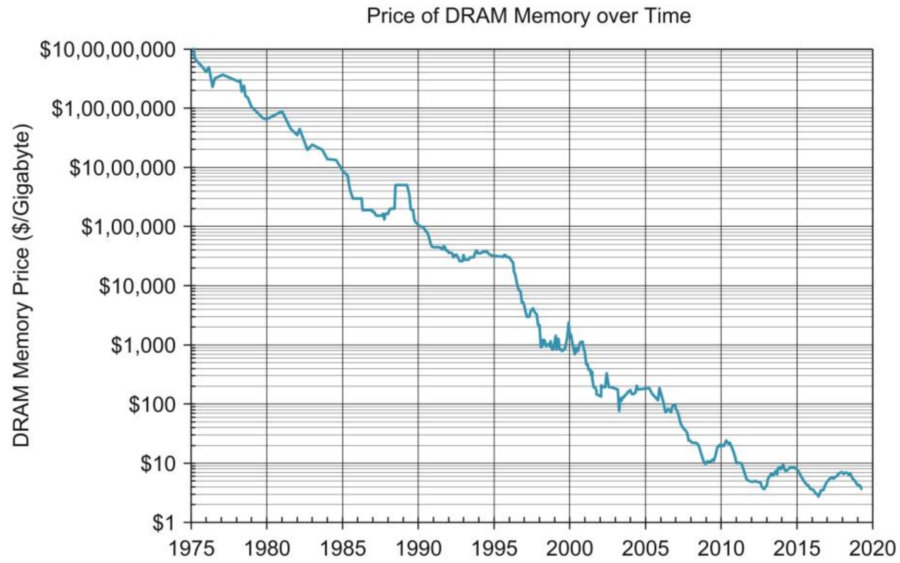

Figure 1.21 Price of memory per gigabyte between 1975 and 2020. (Source: https://jcmit.net/ memoryprice.htm)

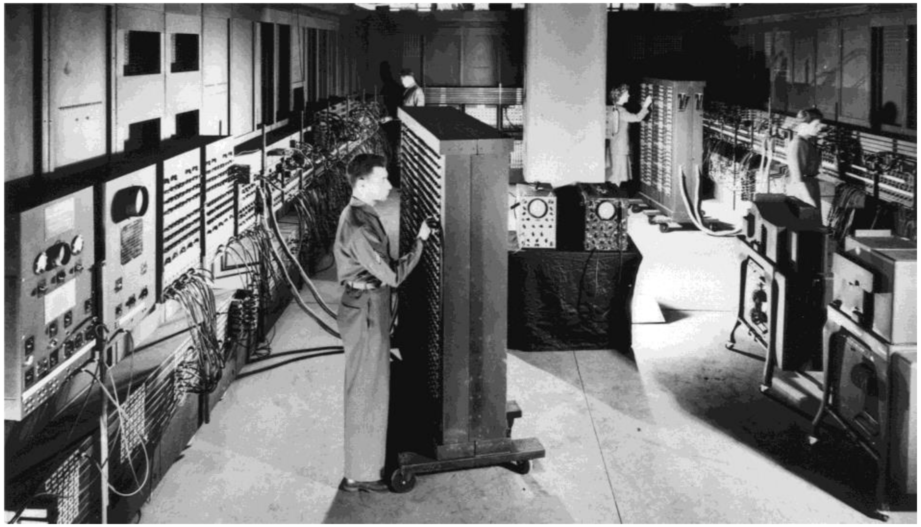

FIGURE e1.13.1 ENIAC, the world’s first general-purpose electronic computer.

FIGURE e1.13.2 UNIVAC I, the first commercial computer in the United States. It correctly predicted the outcome of the 1952 presidential election, but its initial forecast was withheld from broadcast because experts doubted the use of such early results.

FIGURE e1.13.3 IBM System/360 computers: models 40, 50, 65, and 75 were all introduced in 1964. These four models varied in cost and performance by a factor of almost 10; it grows to 25 if we include models 20 and 30 (not shown). The clock rate, range of memory sizes, and approximate price for only the processor and memory of average size: (a) model 40, 1.6 MHz, 32 KB–256 KB, $225,000; (b) model 50, 2.0 MHz, 128 KB–256 KB, $550,000; (c) model 65, 5.0 MHz, 256 KB–1 MB, $1,200,000; and (d) model 75, 5.1 MHz, 256 KB–1 MB, $1,900,000. Adding I/O devices typically increased the price by factors of 1.8 to 3.5, with higher factors for cheaper models.

FIGURE e1.13.4 Cray-1, the first commercial vector supercomputer, announced in 1976. This machine had the unusual distinction of being both the fastest computer for scientific applications and the computer with the best price/performance for those applications. Viewed from the top, the computer looks like the letter *C*. Seymour Cray passed away in 1996 because of injuries sustained in an automobile accident. At the time of his death, this 70-year-old computer pioneer was working on his vision of the next generation of supercomputers. *(See www.cray.com for more details.)*

FIGURE e1.13.5 The Apple IIc Plus. Designed by Steve Wozniak, the Apple IIc set standards of cost and reliability for the industry.

FIGURE e1.12.6 The Xerox Alto was the primary inspiration for the modern desktop computer. It included a mouse, a bit-mapped scheme, a Windows-based user interface, and a local network connection.

Year	Name	Size (cu. ft.)	Power (watts)	Performance (adds/sec)	Memory (KB)	Price	Price/ performance vs. UNIVAC	Adjusted price (2007 $)	Adjusted price/ performance vs. UNIVAC
1951	UNIVAC I	1000	125,000	2000	48	$1,000,000	1	$7,670,724	1
1964	IBM S/360 model 50	60	10,000	500,000	64	$1,000,000	263	$6,018,798	319
1965	PDP-8	8	500	330,000	4	$16,000	10,855	$94,685	13,367
1976	Cray-1	58	60,000	166,000,000	32,000	$4,000,000	21,842	$13,509,798	47,127
1981	IBM PC	1	150	240,000	256	$3000	42,105	$6859	134,208
1991	HP 9000/ model 750	2	500	50,000,000	16,384	$7400	3,556,188	$11,807	16,241,889
1996	Intel PPro PC (200 MHz)	2	500	400,000,000	16,384	$4400	47,846,890	$6211	247,021,234
2003	Intel Pentium 4 PC (3.0 GHz)	2	500	6,000,000,000	262,144	$1600	1,875,000,000	$2009	11,451,750,000
2007	AMD Barcelona PC (2.5 GHz)	2	250	20,000,000,000	2,097,152	$800	12,500,000,000	$800	95,884,051,042

FIGURE e1.13.7 Characteristics of key commercial computers since 1950, in actual dollars and in 2007 dollars adjusted for inflation. The last row assumes we can fully utilize the potential performance of the four cores in Barcelona. In contrast to Figure e1.13.3, here the price of the IBM S/360 model 50 includes I/O devices. *(Source: The Computer History Museum and Producer Price Index for Industrial Commodities.)*

p	# arith inst.	# L/S inst.	# branch inst.	cycles	ex. time	speedup
1	2.56E9	1.28E9	2.56E8	1.92E10	9.60	1.00
2	1.83E9	9.14E8	2.56E8	1.41E10	7.04	1.36
4	9.14E8	4.57E8	2.56E8	7.68E9	3.84	2.50
8	4.57E8	2.29E8	2.56E8	4.48E9	2.24	4.29

p	ex. time
1	41.0
2	29.3
4	14.6
8	7.33

processors	exec. time/ processor	time w/overhead	speedup	actual speedup/ideal speedup
1	100			
2	50	54	100/54 = 1.85	1.85/2 = .93
4	25	29	100/29 = 3.44	3.44/4 = 0.86
8	12.5	16.5	100/16.5 = 6.06	6.06/8 = 0.75
16	6.25	10.25	100/10.25 = 9.76	9.76/16 = 0.61

Desktop Processor	Year	Tech	Max. Clock Speed (GHz)	Integer IPC/core	Cores	Max. DRAM Bandwidth (GB/s)	SP Floating Point (Gflop/s)	L3 cache (MiB)
Westmere i7-620	2100	32	3.33	4	2	17.1	107	4
Ivy Bridge i7-3770K	2013	22	3.90	6	4	25.6	250	8
Broadwell i7-6700K	2015	14	4.20	8	4	34.1	269	8
Kaby Lake i7-7700K	2017	14	4.50	8	4	38.4	288	8
Coffee Lake i7-9700K	2019	14	4.90	8	8	42.7	627	12
Imp./year		20%	4%	7%	15%	10%	19%	12%
Doubles every		4 years	18 years	10 years	5 years	7 years	4 years	6 years