COMPUTER ORGANIZATION & DESIGN (ARM ED.) ONLY CHAPTERS 1,2,4,5,6 PG. 77

PERSONAL COMPUTER (PC)- computer designed for use by an individual, w/ a graphics display, a keyboard, and a mouse.

SERVER- Computer used for running larger programs for multiple users, often simultaneously, and typically accessed only via a network.
-emphasis on dependability

SUPERCOMPUTER- Class of computers w/ highest performance & cost; Configured as servers & cost tens-hundreds of millions of dollars.

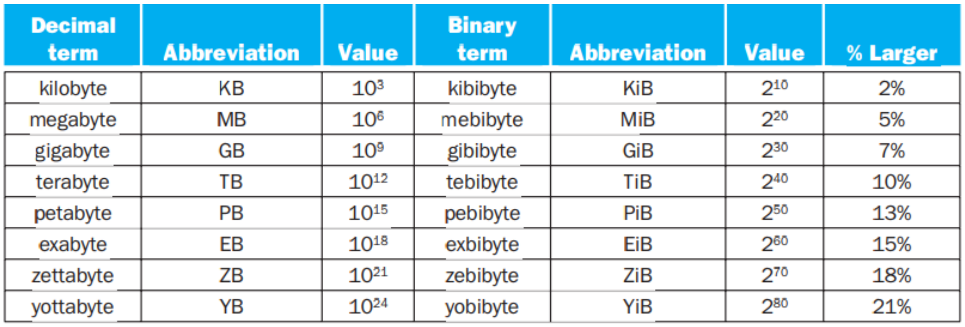
TERABYTE (TB) Originally 1,099,511,627,776 (240) bytes, although communications and secondary storage systems developers started using the term to mean 1,000,000,000,000 (1012) bytes. To reduce confusion, we now use the term

TEBIBYTE (TIB) for 240 bytes, defining terabyte (TB) to mean 1012 bytes. Figure 1.1 shows full range of decimal / binary values / names

EMBEDDED COMPUTER- computer inside another device. Runs one predetermined application or collection of software.

-lower tolerance for failure -dependability is achieved primarily through simplicity

-Many embedded processors are designed using **processor cores**, version of a processor written in **hardware description language** -such as Verilog or VHDL



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%

WELCOME TO THE POST-PC ERA

-Tablets are the fastest growing category, nearly doubling between 2011 and 2012. Recent PCs and traditional cell phone categories are relatively flat or declining.

-PERSONAL MOBILE DEVICES (PMDs)- small wireless devices to connect to the Internet; Rely on batteries for power & software is installed by downloading apps. Conventional examples are smart phones and tablets.

-**CLOUD COMPUTING**- a large collections of servers to provide services over the Internet; some providers rent dynamically varying numbers of servers as a utility.

-Giant datacenters are known as **Warehouse Scale Computers (WSCs)**

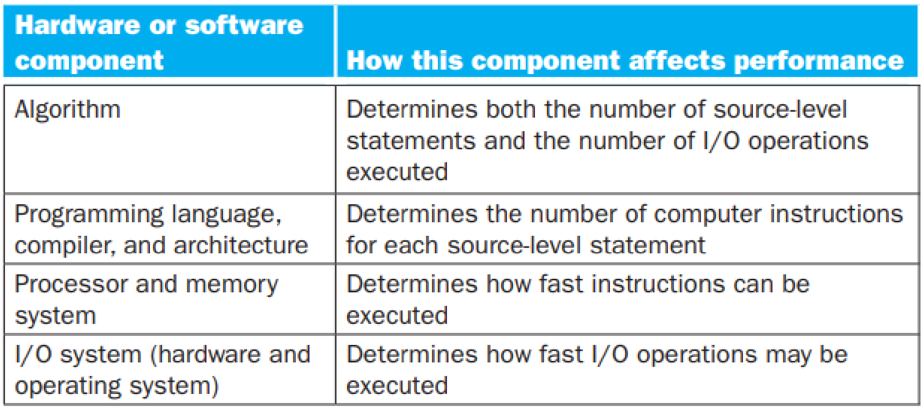
-Software as a Service (SaaS)- Delivers software & data as a service over the Internet, usually via a thin program such as a browser that runs on local client devices, instead of binary code that must be installed, and runs wholly on that device.

-Examples include web search and social networking.

-Today's programmers need to worry about *energy efficiency* of their programs running either on the PMD or in the Cloud

-**Multicore Microprocessor**- A microprocessor containing multiple processors ("cores") in a single integrated circuit. Parallelism.

Understanding Program Performance



Hardware or software component	How this component affects performance
Algorithm	Determines both the number of source-level statements and the number of I/O operations executed
Programming language, compiler, and architecture	Determines the number of computer instructions for each source-level statement
Processor and memory system	Determines how fast instructions can be executed
I/O system (hardware and operating system)	Determines how fast I/O operations may be executed

-liquid crystal display (LCD) A display technology using a thin layer of liquid polymers that can be used to transmit or block light according to whether a charge is applied.

-active matrix display A liquid crystal display using a transistor to control the transmission of light at each individual pixel.

-pixel The smallest individual picture element. Screens are composed of hundreds of thousands to millions of pixels, organized in a matrix.

-computer hardware support for graphics consists mainly of a raster refresh buffer, or frame buffer, to store the bit map.

-integrated circuit Also called a chip. A device combining dozens to millions of transistors.

-central processor unit (CPU) Also called processor. The active part of the computer, which contains the datapath and control and which adds numbers, tests numbers, signals I/O devices to activate, and so on.

-datapath The component of the processor that performs arithmetic operations.

-control The component of the processor that commands the datapath, memory, and I/O devices according to the instructions of the program.

-memory The storage area in which programs are kept when they are running and that contains the data needed by the running programs.

-dynamic random access memory (DRAM) Memory built as an integrated circuit; it provides random access to any location. Access times are 50 nanoseconds and cost per gigabyte in 2012 was $5 to $10.

-cache memory A small, fast memory that acts as a buffer for a slower, larger memory.

-static random access memory (SRAM) Also memory built as an integrated circuit, but faster and less dense than DRAM.

-Most important abstractions is the interface between the hardware and the lowest-level software. Because of its importance, it is given a special name: the instruction set architecture, or simply architecture

-instruction set architecture encompasses all the information necessary to write a machine language program that will run correctly, including instructions, registers, memory access, I/O, and so on.

-application binary interface (ABI) The user portion of the instruction set plus the operating system interfaces used by application programmers. It defines a standard for binary portability across computers.

-volatile memory Storage, such as DRAM, that retains data only if it is receiving power.

-nonvolatile memory A form of memory that retains data even in the absence of a power source and that is used to store programs between runs. A DVD disk is nonvolatile.

-flash memory is the standard secondary memory for PMDs

-main memory Also called primary memory. Memory used to hold programs while they are running; typically consists of DRAM in today's computers.

-secondary memory- Nonvolatile memory used to store programs and data between runs; typically consists of flash memory in PMDs and magnetic disks in servers.

-**magnetic disk** Also called hard disk. A form of nonvolatile secondary memory composed of rotating platters coated with a magnetic recording material. Because they are rotating mechanical devices, access times are about 5 to 20 milliseconds and cost per gigabyte in 2012 was $0.05 to $0.10.

-**flash memory** A nonvolatile semiconductor memory. It is cheaper / slower than DRAM but more expensive per bit and faster than magnetic disks. Access times are about 5 to 50 microseconds and cost per gigabyte in 2012 was $0.75 to $1.00.

-Networked computers have several major advantages: ■ **Communication**: Information is exchanged between computers at high speeds. ■ **Resource sharing**: Rather than each computer having its own I/O devices, computers on the network can share I/O devices. ■ **Nonlocal access**: By connecting computers over long distances, users need not be near the computer they are using.

-local area network (**LAN**) A network designed to carry data within a geographically confined area, typically within a single building.

-wide area network (**WAN**) A network extended over hundreds of kilometers that can span a continent.

-transistor An on/off switch controlled by an electric signal.

-very large-scale integrated (VLSI) circuit- A device containing hundreds of thousands to millions of transistors.

-silicon A natural element that is a semiconductor.

-semiconductor A substance that does not conduct electricity well.

-add materials to silicon that allow tiny areas to transform into one of three devices: ■ Excellent conductors of electricity (using either microscopic copper or aluminum wire) ■ Excellent insulators from electricity (like plastic sheathing or glass) ■ Areas that can conduct or insulate under specific conditions (as a switch)

-silicon crystal ingot- A rod composed of a silicon crystal that is between 8 and 12 inches in diameter and about 12 to 24 inches long.

-wafer- A slice from a silicon ingot no more than 0.1 inches thick, used to create chips.

-defect A microscopic flaw in a wafer or in patterning steps that can result in the failure of the die containing that defect.

-die The individual rectangular sections that are cut from a wafer, more informally known as chips.

-yield The percentage of good dies from the total number of dies on the wafer.

-A key factor in determining the cost of an integrated circuit is volume. Which of the following are reasons why a chip made in high volume should cost less? 1. With high volumes, the manufacturing process can be tuned to a particular design, increasing the yield. 2. It is less work to design a high-volume part than a low-volume part. 3. The masks used to make the chip are expensive, so the cost per chip is lower for higher volumes. 4. Engineering development costs are high and largely independent of volume; thus, the development cost per die is lower with high-volume parts. 5. High-volume parts usually have smaller die sizes than low-volume parts and therefore, have higher yield per wafer.

1.6 Performance

-response time Also called execution time. The total time required for the computer to complete a task, including disk accesses, memory accesses, I/O activities, operating system overhead, CPU execution time, and so on.

-throughput Also called bandwidth. Another measure of performance, it is the number of tasks completed per unit time.

-CPU execution time Also called CPU time. The actual time the CPU spends computing for a specific task.

-user CPU time The CPU time spent in a program itself.

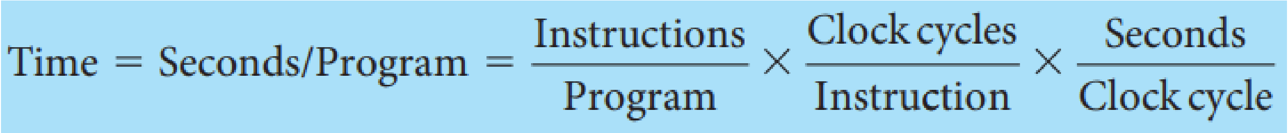
-system CPU time The CPU time spent in the operating system performing tasks on behalf of the program.

-clock cycle Also called tick, clock tick, clock period, clock, or cycle. The time for one clock period, usually of the processor clock, which runs at a constant rate.

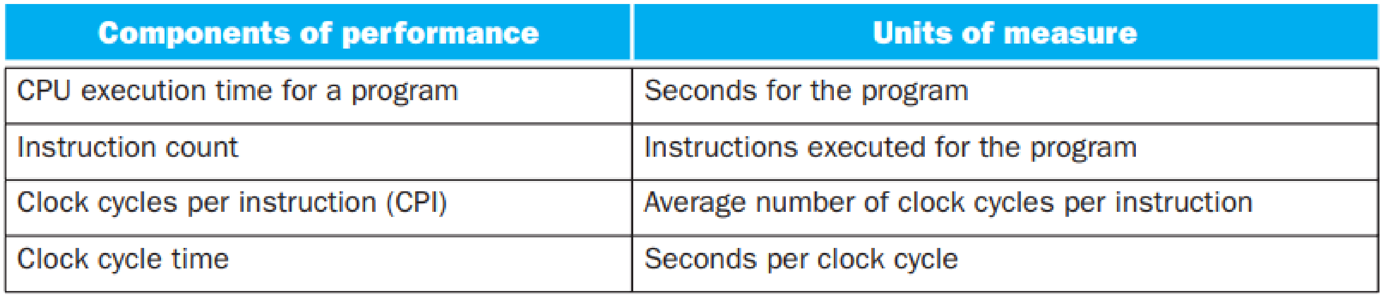
-clock period The length of each clock cycle

-clock cycles per instruction (CPI) Average number of clock cycles per instruction for a program or program fragment.

-instruction count The number of instructions executed by the program.



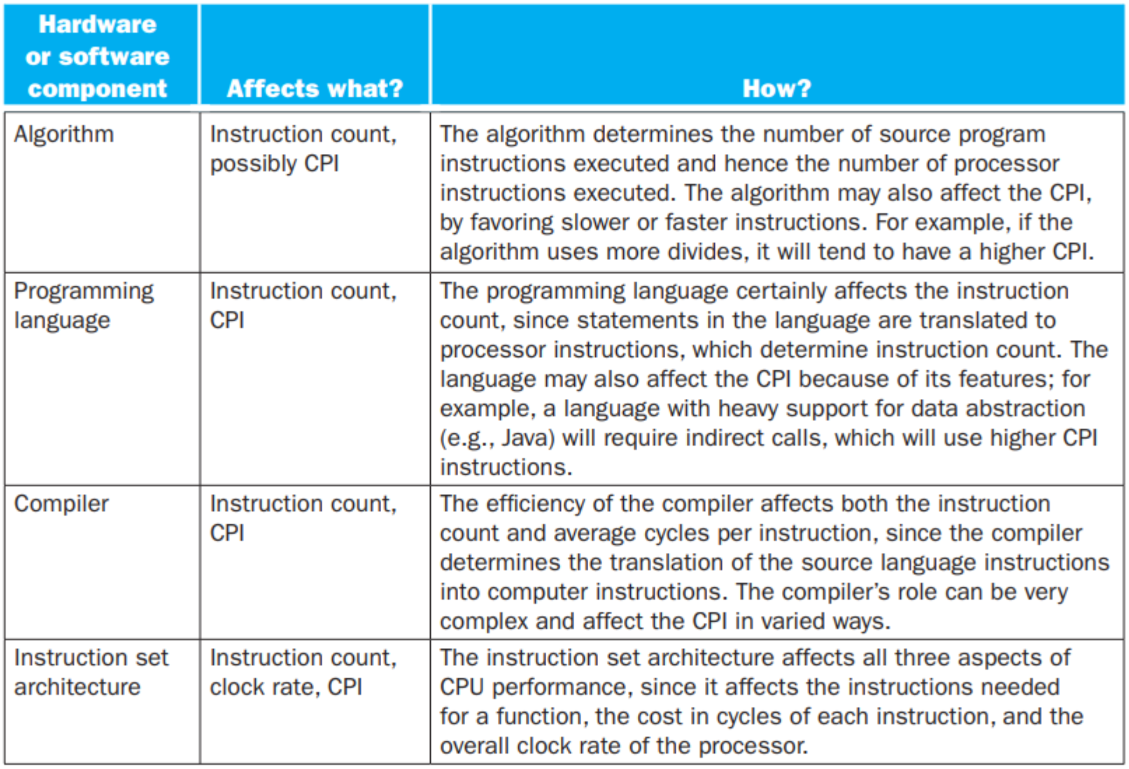
$$\text{Time} = \text{Seconds/Program} = \frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Clock cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Clock cycle}}$$



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.

---instruction mix A measure of the dynamic frequency of instructions across one or many programs.

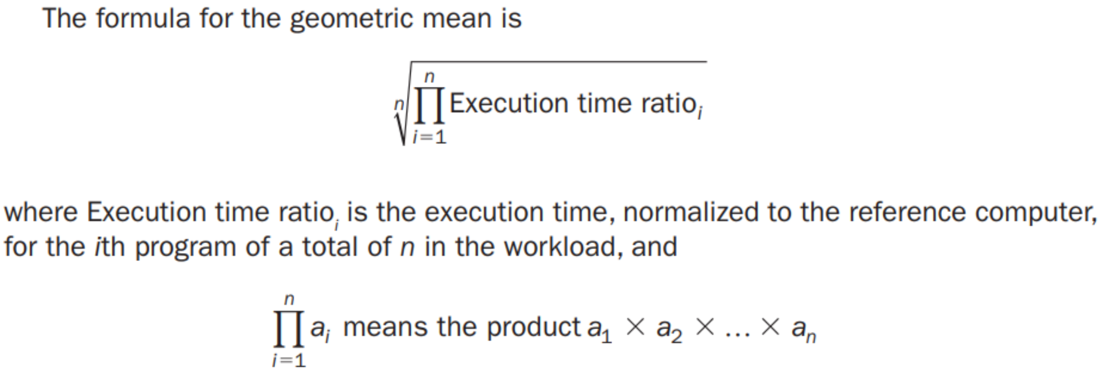


Hardware or software component	Affects what?	How?
Algorithm	Instruction count, possibly CPI	The algorithm determines the number of source program instructions executed and hence the number of processor instructions executed. The algorithm may also affect the CPI, by favoring slower or faster instructions. For example, if the algorithm uses more divides, it will tend to have a higher CPI.
Programming language	Instruction count, CPI	The programming language certainly affects the instruction count, since statements in the language are translated to processor instructions, which determine instruction count. The language may also affect the CPI because of its features; for example, a language with heavy support for data abstraction (e.g., Java) will require indirect calls, which will use higher CPI instructions.
Compiler	Instruction count, CPI	The efficiency of the compiler affects both the instruction count and average cycles per instruction, since the compiler determines the translation of the source language instructions into computer instructions. The compiler's role can be very complex and affect the CPI in varied ways.
Instruction set architecture	Instruction count, clock rate, CPI	The instruction set architecture affects all three aspects of CPU performance, since it affects the instructions needed for a function, the cost in cycles of each instruction, and the overall clock rate of the processor.

-**workload** A set of programs run on a computer that is either the actual collection of applications run by a user or constructed from real programs to approximate such a mix. A typical workload specifies both the programs and the relative frequencies.

-**benchmark** A program selected for use in comparing computer performance.

-



The formula for the geometric mean is

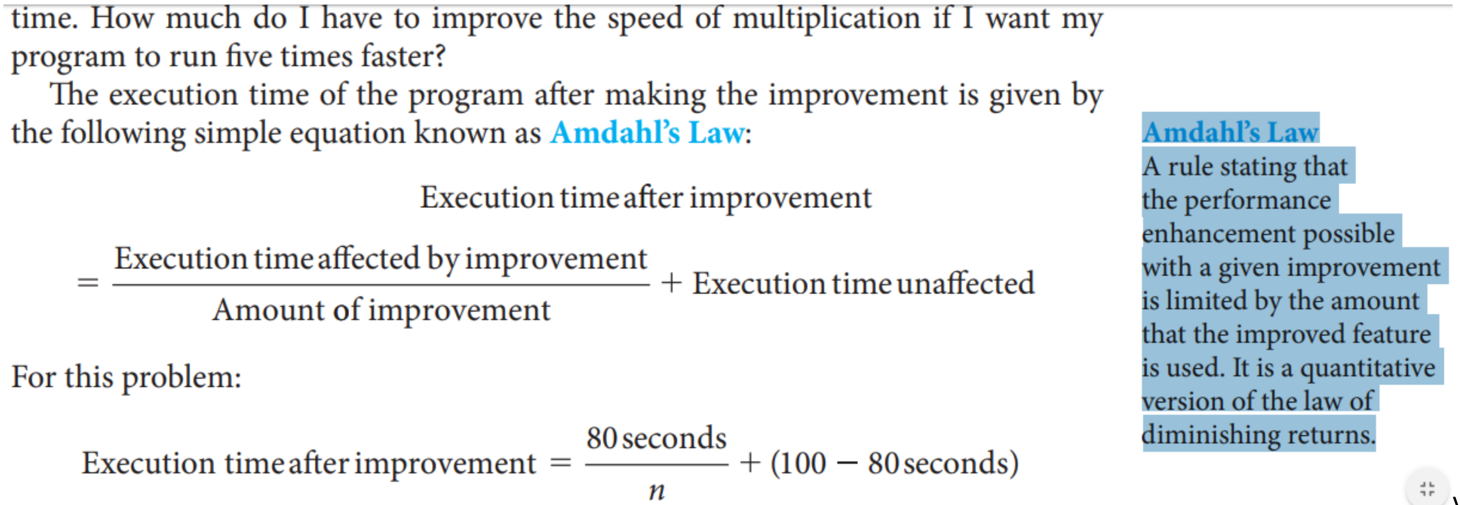
$$\sqrt[n]{\prod_{i=1}^{n} \text{Execution time ratio}_i}$$

where Execution time ratio$_i$ is the execution time, normalized to the reference computer, for the *i*th program of a total of *n* in the workload, and

$$\prod_{i=1}^{n} a_i \text{ means the product } a_1 \times a_2 \times \ldots \times a_n$$

a = the range of i->n

a_i -> a_n



time. How much do I have to improve the speed of multiplication if I want my program to run five times faster?

The execution time of the program after making the improvement is given by the following simple equation known as **Amdahl's Law**:

$$\text{Execution time after improvement} = \frac{\text{Execution time affected by improvement}}{\text{Amount of improvement}} + \text{Execution time unaffected}$$

For this problem:

$$\text{Execution time after improvement} = \frac{80\text{ seconds}}{n} + (100 - 80\text{ seconds})$$

Amdahl's Law
A rule stating that the performance enhancement possible with a given improvement is limited by the amount that the improved feature is used. It is a quantitative version of the law of diminishing returns.

vc

-million instructions per second (MIPS) A measurement of program execution speed based on the number of millions of instructions. MIPS is computed as the instruction count divided by the product of the execution time and 10^6 .

-