Ch.1 Computer System Overview

- What is the main function of the computer?

  to execute programs

- What are the 4 structural elements of a computer?

  1. Processor
     a. it controls the operation of the computer and performs its data processing functions
     b. it exchanges data with memory
  2. Main Memory
     a. it stores data and programs
     b. this memory is typically volatile (when the computer is shut down, the contents of the memory are lost)
     c. it is referred to as real or primary memory
  3. I/O modules
     a. they move data between the computer and its external environment (devices, disks aka secondary memory devices, communications equipment, and terminals)
  4. System bus
     a. it provides for communication among processors, main memory and I/O modules

- Where there is only one processor, what is it called?

  the CPU or central processing unit

- Are the contents of disk memory saved even when the computer is shut down?

yes

▼ To exchange data with memory, the processor makes use of two internal registers. What are they?

- a memory address register (MAR) specifies the address in memory for the next read or write
- a memory buffer register (MBR) contains the data to be written into memory, or receives the data read from memory

▼ What is a register?

- high-speed memory internal to the CPU.
- some registers are available to the programmer via the machine instruction set; others are only used by the CPU

▼ What is an I/O address register (I/OAR)?

it specifies a particular I/O device

▼ What is an I/O buffer register (I/OBR)?

it used for the exchange of data between an I/O module and the processor

▼ A memory module consists of

a set of locations (that can be interpreted as instructions or data) defined by sequentially numbered addresses

▼ An I/O module contains internal buffers for what?

temporarily storing data

▼ What do multiprocessors contain?

each chip called a socket contains multiple processors called cores, each with multiple levels of large memory caches, and multiple logical processors sharing the execution units of each core

▼ As of 2010, what do laptops contain?

- 2-4 cores

- each core with 2 hardware threads
- total: 4 or 8 logical processors

▼ What do GPUs or Graphical Processing Units do?

- they provide efficient computation on arrays of data using Single-Instruction Multiple Data (SIMD) techniques pioneered in supercomputers
- they are used for general numerical processing

▼ CPUs are gaining the capability of operating on arrays of data...how?

- with powerful vector units integrated into the processor architecture of the x86 and AMD64 families

▼ What are Digital Signal Processors (DSPs)?

they are embedded in I/O devices to deal with streaming signals like audio or video

▼ What do fixed function units do?

they support computations like encoding/decoding speech and video, encryption and security

▼ What is System on a Chip (SoC)?

putting the components of a system on the same chip to satisfy the requirements of handheld devices

▼ What are the 2 steps of instruction processing?

1. The processor reads (fetches) instructions from memory one at a time
2. The processor executes each instruction

▼ What does program execution consist of?

Repeating the process of instruction fetch and instruction execution

▼ What is an instruction cycle?

the processing required for a single instruction (fetch stage then execute stage)

▼ What is the fetched instruction loaded into?

the instruction register

▼ What are the 4 types of actions a processor can take given an instruction?

1. Processor-memory: data transferred from processor to memory, or from memory to processor
2. Processor-I/O: data transferred to or from a peripheral device by transferring between the processor and an I/O module
3. Data processing: the processor may perform some arithmetic or logic operation on data
4. Control: An instruction may specify that the sequence of execution be altered.

▼ What is an Interrupt?

- a mechanism by which other modules (I/O, memory) may interrupt the normal sequencing of the processor.

▼ What are the 4 classes of Interrupts?

1. Program - generated by some condition that occurs as the result of an instruction execution
2. Timer - generated by a timer within the processor
3. I/O - generated by an I/O controller to signal normal completion of an operation or to signal a variety of error conditions
4. Hardware failure - generated by a failure such as power failure or memory parity error

▼ Why are interrupts used?

With interrupts, the processor can be engaged in executing other instructions while an I/O operation is in progress.

▼ What does an interrupt do for the user program?

The processor and the OS suspend the normal sequence of execution.

When the interrupt processing is completed, execution of the user program resumes.

▼ What is the interrupt stage in the instruction cycle?

the cycle checks for interrupts after the execute stage.

if an interrupt is pending, the processor suspends execution of the current program and executes an interrupt-handler routine.

▼ What is the sequence of events that an interrupt triggers in both the processor hardware and in software?

Hardware

1. Device controller or other system hardware issues an interrupt
2. Processor finishes execution of current instruction
3. Processor signals acknowledgement of interrupt
4. Processor pushes PSW (program status word) and PC (the location of the next instruction to be executed which is contained in the program counter) onto control stack
5. Processor loads new PC value based on interrupt

Then: Software

6. Save reminder of process state information
7. Process interrupt
8. Restore process state information
9. Restore old PSW and PC

▼ What is a block?

a collection of contiguous records that are recorded as a unit

▼ What are 2 ways to deal with multiple interrupts?

1. A disabled interrupt means the processor ignores any new interrupt request signal
2. Define priorities for interrupts. This allows an interrupt of higher priority to cause a lower-priority interrupt handler to be interrupted

▼ What 3 questions sum up design constraints on computer's memory?

- How much?
- How fast?
- How expensive?

▼ What are the trade-offs of memory?

- faster access time, greater cost per bit
- greater capacity, smaller cost per bit
- greater capacity, slower access speed

▼ How does the memory hierarchy address tradeoffs? (top-to-bottom of the pyramid)

1. Decreasing cost per bit
2. Increasing capacity
3. Increasing access time
4. Decreasing frequency of access to the memory by the processor

▼ What is hit ratio?

in two-level memory, the fraction of all memory accesses that are found in faster memory (e.g. the cache)

▼ What is locality of reference?

the tendency of a processor to access the same set of memory locations repetitively over a short period of time

▼ What is secondary or auxiliary memory?

external, nonvolatile memory used to store program and data files

▼ What is a hard disk used for?

to provide an extension to main memory known as virtual memory

▼ What is a disk cache?

a buffer usually kept in main memory that functions as a cache of disk blocks between disk memory and the rest of main memory

▼ What is cache memory?

- a memory that is smaller and faster than main memory and it is interposed between the processor and main memory
- the cache acts as a buffer for recently used memory locations

▼ What is the memory cycle time?

the rate at which the processor can execute instructions

▼ What does the cache contain a copy of?

a portion of main memory

▼ When the processor attempts to read a byte or word of memory, a check is made where?

the cache

▼ If something is not in the cache what happens?

a block of main memory consisting of a fixed number of bytes is read into the cache

▼ What is the intention of cache memory?

to provide memory access time approaching that of the fastest memories available AND support a large memory size that has the price of less expensive types of semiconductor memories

▼ Cache access order?

Level 1 or L1 cache

L2

L3

▼ L2 vs L1 cache

L2 cache is slower and typically larger than L1

▼ L3 vs L2 cache

L3 cache is slower and typically larger than L2

▼ What is block size?

the unit of data exchanged between cache and main memory

▼ What does the mapping function do?

it determines which cache location the block will occupy

▼ The more flexible the mapping function...

- the more scope we have to design a replacement algorithm to maximize the hit ratio
- the more complex the circuitry is required to search the cache to determine if a given block is the in cache

▼ What is the replacement algorithm?

it chooses within the constraints of the mapping function which block to replace when a new block is loaded into the cache and the cache already has all slots filled with other blocks

▼ What is the LRU least-recently-used algorithm?

a strategy to replace the block that has been in the cache longest with no reference to it

▼ What is the write policy?

it dictates when to write back an update to the contents of a block in the cache to the main memory

▼ What are 3 techniques possible for I/O operations?

- programmed IO
- interrupt-driven IO
- direct memory access DMA

▼ What is programmed IO?

IO in which the CPU issues an IO command to an IO module and must then wait for the operation to complete before proceeding

▼ What is interrupt-driven IO?

the processor issues an IO command to a module then goes on to do some other useful work

▼ What is direct memory access?

- it is used when large volumes of data are to be moved
- the DMA function can be performed by a separate module on the system bus or incorporated into an IO module

▼ How does DMA work?

The processor issues a command to the DMA module by sending:

- whether a read or write is requested
- the address of the IO device involved
- the starting location in memory to read data from or write data to
- the number of words to be read or written

▼ What is parallelism?

it refers to techniques to make programs faster by performing several computations at the same time

▼ What are 3 approaches providing parallelism by replicating processors?

- symmetric multiprocessors SMPs
- multicore computers
- clusters

▼ What is an SMP or symmetric multiprocessor?

a stand-alone computer system with:

- 2 or more similar processors of comparable capability
- these processors share the same main memory and IO facilities
- all processors share access to IO devices
- all processors can perform the same functions aka they are symmetric

- the system is controlled by an integrated operating system that provides interaction between processors and their programs at the job, task, file and data element levels

▼ What are the potential advantages of an SMP organization over a uniprocessor?

- performance
  - portions of the work can be done in parallel
- availability
  - a failure of a single processor does not halt the machine
- incremental growth
  - can enhance the system performance by adding an additional processor
- scaling
  - vendors can offer a range of products with different price and performance characteristics based on the number of processors configured in the system

▼ What is synchronization?

2 or more processors coordinate their activities based on a condition

▼ What is a multicore computer or chip multiprocessor?

it combines 2 or more processors (called cores) on a single piece of silicon (called a die)

▼ What does a core consist of?

all the components of an independent processor, such as registers, ALU, pipeline hardware, and control unit, plus L1 instruction and data caches.

in contemporary multicore chips, they include L2 cache and in some cases L3 cache