Tutorial-2: Real-Time Computer Systems and Architecture

A 1. **Processor-memory:** Data may be transferred from processor to memory or from memory to processor.

Processor-I/O: Data may be transferred to or from a peripheral device by transferring between the processor and an I/O module.

Data processing: The processor may perform some arithmetic or logic operation on data.

Control: An instruction may specify that the sequence of execution be altered.

A 2. **Instruction address calculation (iac):** Determine the address of the next instruction to be executed.

Instruction fetch (if): Read instruction from its memory location into the processor.

I**nstruction operation decoding (iod):** Analyze instruction to determine type of operation to be performed and operand(s) to be used.

Operand address calculation (oac): If the operation involves reference to an operand in memory or available via I/O, then determine the address of the operand.

Operand fetch (of): Fetch the operand from memory or read it in from I/O.

Data operation (do): Perform the operation indicated in the instruction.

Operand store (os): Write the result into memory or out to I/O.

A 3. **Physical layer:** Consists of the actual wires carrying the signals, as well as circuitry and logic to support ancillary features required in the transmission and receipt of the 1s and 0s. The unit of transfer at the Physical layer is 20 bits, which is called a **Phit** (physical unit).

Link layer: Responsible for reliable transmission and flow control. The Link layer's unit of transfer is an 80-bit Flit (flow control unit).

Routing layer: Provides the framework for directing packets through the fabric.

Protocol layer: The high-level set of rules for exchanging packets of data between devices. A packet is comprised of an integral number of Flits.

A 4. **Physical layer:** Consists of the actual wires carrying the signals, as well as circuitry and logic to support ancillary features required in the transmission and receipt of the 1s and 0s.

Data link layer: Is responsible for reliable transmission and flow control. Data packets generated and consumed by the DLL are called Data Link Layer Packets (DLLPs).

Transaction layer: Generates and consumes data packets used to implement load/store data transfer mechanisms and also manages the flow control of those packets between the two components on a link. Data packets generated and consumed by the TL are called Transaction Layer Packets (TLPs).

A 5. a) 2^{24} = 16 MBytes

b) (1) If the local address bus is 32 bits, the whole address can be transferred at once and decoded in memory. However, because the data bus is only 16 bits, it will require 2 cycles to fetch a 32-bit instruction or operand.
(2) The 16 bits of the address placed on the address bus can't access the whole memory. Thus a more complex memory interface control is needed to latch the first part of the address and then the second part (because the microprocessor will end in two steps). For a 32-bit address, one may assume the first half will decode to access a "row" in memory, while the second half is sent later to access a "column" in memory. In addition to the two-step address operation, the microprocessor will need 2 cycles to fetch the 32-bit instruction/operand.

c) The program counter must be at least 24 bits. Typically, a 32-bit microprocessor will have a 32-bit external address bus and a 32-bit program counter, unless on-chip segment registers are used that may work with a smaller program counter. If the instruction register is to contain the whole instruction, it will have to be 32-bits long; if it will contain only the op code (called the op code register) then it will have to be 8 bits long.

A 6. Clock cycle =1/8MHz= 125 ns

Bus cycle = 4 x 125 ns = 500 ns
2 bytes transferred every 500 ns; thus transfer rate = 4 MBytes/sec

Doubling the frequency may mean adopting a new chip manufacturing technology (assuming each instructions will have the same number of clock cycles); doubling the external data bus means wider (maybe newer) on-chip data bus drivers/latches and modifications to the bus control logic. In the first case, the speed of the memory chips will also need to double (roughly) not to slow down the microprocessor; in the second case, the "wordlength" of the memory will have to double to be able to send/receive 32-bit quantities.

A 7. a) Input from the Teletype is stored in INPR. The INPR will only accept data from the Teletype when FGI=0. When data arrives, it is stored in INPR, and FGI is set to 1. The CPU periodically checks FGI. If FGI =1, the CPU transfers the contents of INPR to the AC and sets FGI to 0.
When the CPU has data to send to the Teletype, it checks FGO. If FGO = 0, the CPU must wait. If FGO = 1, the CPU transfers the contents of the AC to OUTR and sets FGO to 0. The Teletype sets FGI to 1 after the word is printed.

b) The process described in (a) is very wasteful. The CPU, which is much faster than the Teletype, must repeatedly check FGI and FGO. If interrupts are used, the Teletype can issue an interrupt to the CPU whenever it is ready to accept or send data. The IEN register can be set by the CPU (under programmer control)

A 8. a) During a single bus cycle, the 8-bit microprocessor transfers one byte while the 16-bit microprocessor transfers two bytes. The 16-bit microprocessor has twice the data transfer rate.

b) Suppose we do 100 transfers of operands and instructions, of which 50 are one byte long and 50 are two bytes long. The 8-bit microprocessor takes 50 + (2 x 50) = 150 bus cycles for the transfer. The 16-bit microprocessor requires 50 + 50 = 100 bus cycles. Thus, the data transfer rates differ by a factor of 1.5.

A 9. **Sequential access:** Memory is organized into units of data, called records. Access must be made in a specific linear sequence.

Direct access: Individual blocks or records have a unique address based on physical location. Access is accomplished by direct access to reach a general vicinity plus sequential searching, counting, or waiting to reach the final location.

Random access: Each addressable location in memory has a unique, physically wired-in addressing mechanism. The time to access a given location is independent of the sequence of prior accesses and is constant.

A 10. In a cache system, **direct mapping** maps each block of main memory into only one possible cache line.

Associative mapping permits each main memory block to be loaded into any line of the cache.

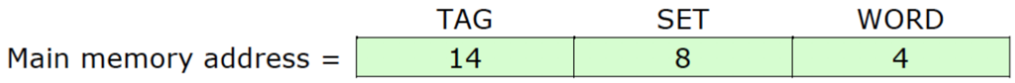
In set-associative mapping, the cache is divided into a number of sets of cache lines; each main memory block can be mapped into any line in a particular set.

A 11. One field identifies a unique word or byte within a block of main memory. The remaining two fields specify one of the blocks of main memory. These two fields are a line field, which identifies one of the lines of the cache, and a tag field, which identifies one of the blocks that can fit into that line.

A 12. A tag field uniquely identifies a block of main memory. A word field identifies a unique word or byte within a block of main memory.

A 13. One field identifies a unique word or byte within a block of main memory. The remaining two fields specify one of the blocks of main memory. These two fields are a set field, which identifies one of the sets of the cache, and a tag field, which identifies one of the blocks that can fit into that set.

A 14. There are a total of 8 kbytes/16 bytes = 512 lines in the cache. Thus the cache consists of 256 sets of 2 lines each. Therefore 8 bits are needed to identify the set number. For the 64-Mbyte main memory, a 26-bit address is needed. Main memory consists of 64-Mbyte/16 bytes = 2^{22} blocks. Therefore, the set plus tag lengths must be 22 bits, so the tag length is 14 bits and the word field length is 4 bits.



	TAG	SET	WORD
Main memory address =	14	8	4

A 15. a) 8 leftmost bits = tag; 5 middle bits = line number; 3 rightmost bits = byte number

b) slot 3; slot 6; slot 3; slot 21

c) Bytes with addresses 0001 1010 0001 1000 through 0001 1010 0001 1111 are stored in the cache

d) 256 bytes

e) Because two items with two different memory addresses can be stored in the same place in the cache. The tag is used to distinguish between them.

A 16. **Human readable:** Suitable for communicating with the computer user.
Machine readable: Suitable for communicating with equipment.
Communication: Suitable for communicating with remote devices.

A 17. **Programmed I/O:** The processor issues an I/O command, on behalf of a process, to an I/O module; that process then busy-waits for the operation to be completed before proceeding.

Interrupt-driven I/O: The processor issues an I/O command on behalf of a process, continues to execute subsequent instructions, and is interrupted by the I/O module when the latter has completed its work. The subsequent instructions may be in the same process, if it is not necessary for that process to wait for the completion of the I/O. Otherwise, the process is suspended pending the interrupt and other work is performed.

Direct memory access (DMA): A DMA module controls the exchange of data between main memory and an I/O module. The processor sends a request for the transfer of a block of data to the DMA module and is interrupted only after the entire block has been transferred.

A 18. With **memory-mapped I/O**, there is a single address space for memory locations and I/O devices. The processor treats the status and data registers of I/O modules as memory locations and uses the same machine instructions to access both memory and I/O devices.

With **isolated I/O**, a command specifies whether the address refers to a memory location or an I/O device. The full range of addresses may be available for both.

A 19. Four general categories of techniques are in common use: multiple interrupt lines; software poll; daisy chain (hardware poll, vectored); bus arbitration (vectored).

A 20. In the first addressing mode, 2^8 = 256 ports can be addressed. Typically, this would allow 128 devices to be addressed. However, an opcode specifies either an input or output operation, so it is possible to reuse the addresses, so that there are 256 input port addresses and 256 output port addresses. In the second addressing mode, 2^{16} = 64K port addresses are possible.

A 21. In direct addressing mode, an instruction can address up to 2^{16} = 64K ports. In indirect addressing mode, the port address resides in a 16-bit registers, so again, the instruction can address up to 2^{16} = 64K ports.

A 22. Using non-block I/O instructions, the transfer takes 20 x 128 = 2560 clock cycles. With block I/O, the transfer takes 5 x 128 = 640 clock cycles (ignoring the one-time fetching of the iterative instruction and its operands). The speedup is (2560 – 640)/2560 = 0.75, or 75%.

A 23.

a) Each I/O device requires one output (from the point of view of the processor) port for commands and one input port for status.

b) The first device requires only one port for data, while the second devices requires and input data port and an output data port. Because each device requires one command and one status port, the total number of ports is seven.

c) seven.

A 24. The operating system (OS) is the software that controls the execution of programs on a processor and that manages the processor's resources

A 25. **Program creation:** The operating system provides a variety of facilities and services, such as editors and debuggers, to assist the programmer in creating programs.

Program execution: A number of tasks need to be performed to execute a program. Instructions and data must be loaded into main memory, I/O devices and files must be initialized, and other resources must be prepared.

Access to I/O devices: Each I/O device requires its own peculiar set of instructions or control signals for operation.

Controlled access to files: In the case of files, control must include an understanding of not only the nature of the I/O device (disk drive, tape drive) but also the file format on the storage medium.

System access: In the case of a shared or public system, the operating system controls access to the system as a whole and to specific system resources.

Error detection and response: A variety of errors can occur while a computer system is running.

Accounting: A good operating system will collect usage statistics for various resources and monitor performance parameters such as response time.

A 26. A process is a program in execution, together with all the state information required for execution.

A 27. I/O-bound programs use relatively little processor time and are therefore favored by the algorithm. However, if a processor-bound process is denied processor time for a sufficiently long period of time, the same algorithm will grant the processor to that process because it has not used the processor at all in the recent past. Therefore, a processor-bound process will not be permanently denied access