Topic for today:

The datapath

CPU Basics

- The computer's CPU fetches, decodes, and executes program instructions.
- The two principal parts of the CPU are the *datapath* and the *control unit*.
  - The datapath consists of an arithmetic-logic unit and storage units (registers) that are interconnected by a data bus that is also connected to main memory.
  - Various CPU components perform sequenced operations according to signals provided by its control unit.

2

Recall

An Arithmetic Logic Unit (ALU) is a combinational circuit which performs arithmetic and logic operations on a sequence of input bits to produce a sequence of output bits. The operation to be performed is determined by a control signal.

Datapath

The *datapath* consists of the ALU, one or more registers, and one or more internal buses connecting them.

Controlling the datapath

The flow of data on the datapath is determined by control signals, which select ALU functions and control reading/writing of the registers.

General purpose registers

Many registers have dedicated purposes. Registers which may be accessed by the programmer to store arbitrary data are general purpose registers. If there is only one general purpose register, it is called the accumulator. (Compare this to the memory of a typical low-end hand-held calculator.)

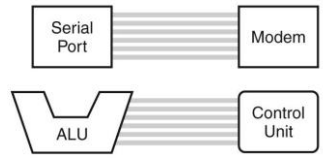
Registers

- Registers hold data that can be readily accessed by the CPU.
- They can be implemented using D flip-flops.
  - A 32-bit register requires 32 D flip-flops.
- The arithmetic-logic unit (ALU) carries out logical and arithmetic operations as directed by the control unit.
- The control unit determines which actions to carry out according to the values in a program counter register and a status register.

The Bus

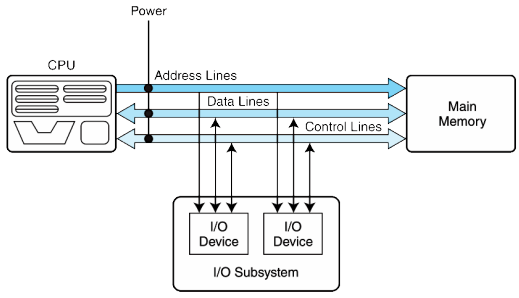
- The CPU shares data with other system components by way of a data bus.
  - A bus is a set of wires that simultaneously convey a single bit along each line.
- Two types of buses are commonly found in computer systems: *point-to-point*, and *multipoint* buses.

These are point-to-point buses:



8

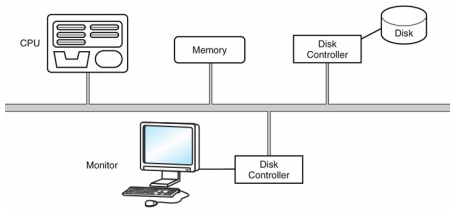
The Bus



9

The Bus

- A multipoint bus is shown below.
- Because a multipoint bus is a shared resource, access to it is controlled through protocols, which are built into the hardware.



10

The Bus

In a master-slave configuration, where more than one device can be the bus master, concurrent bus master requests must be arbitrated.

Four categories of bus arbitration are:

- **Daisy chain:** Permissions are passed from the highest-priority device to the lowest.
- **Centralized parallel:** Each device is directly connected to an arbitration circuit.
- **Distributed using self-detection:** Devices decide which gets the bus among themselves.
- **Distributed using collision-detection:** Any device can try to use the bus. If its data collides with the data of another device, it tries again.

11

Clocks

- Every computer contains at least one clock that synchronizes the activities of its components.
- A fixed number of clock cycles are required to carry out each data movement or computational operation.
- The clock frequency, measured in megahertz or gigahertz, determines the speed with which all operations are carried out.
- Clock cycle time is the reciprocal of clock frequency.
  - An 800 MHz clock has a cycle time of 1.25 ns.

12

Clocks

- Clock speed should not be confused with CPU performance.
- The CPU time required to run a program is given by the general performance equation:

$$\text{CPU Time} = \frac{\text{seconds}}{\text{program}} = \frac{\text{instructions}}{\text{program}} \times \frac{\text{avg. cycles}}{\text{instruction}} \times \frac{\text{seconds}}{\text{cycle}}$$

  - We see that we can improve CPU throughput when we reduce the number of instructions in a program, reduce the number of cycles per instruction, or reduce the number of nanoseconds per clock cycle.

We will return to this important equation in later chapters.

13

The Input/Output Subsystem

- A computer communicates with the outside world through its input/output (I/O) subsystem.
- I/O devices connect to the CPU through various interfaces.
- I/O can be memory-mapped-- where the I/O device behaves like main memory from the CPU's point of view.
- Or I/O can be instruction-based, where the CPU has a specialized I/O instruction set.

14

Memory Organization

- Computer memory consists of a linear array of addressable storage cells that are similar to registers.
- Memory can be **byte-addressable**, or **word-addressable**, where a word typically consists of two or more bytes.
- Memory is constructed of RAM chips, often referred to in terms of length × width.
- If the memory word size of the machine is 16 bits, then a 4M × 16 RAM chip gives us 4 megabytes of 16-bit memory locations.

15

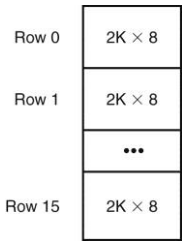
Memory Organization

- How does the computer access a memory location corresponds to a particular address?
- We observe that 4M can be expressed as $2^2 \times 2^{20} = 2^{22}$ words.
- The memory locations for this memory are numbered 0 through $2^{22}-1$.
- Thus, the memory bus of this system requires at least 22 address lines.
  - The address lines "count" from 0 to $2^{22} - 1$ in binary. Each line is either "on" or "off" indicating the location of the desired memory element.

16

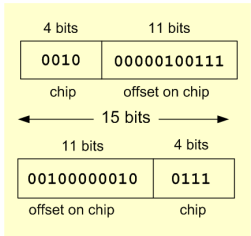
Memory Organization

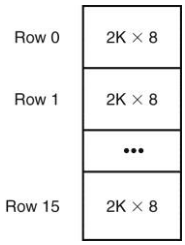
- Example: Suppose we have a memory consisting of 16 2K x 8 bit chips.
  - Memory is 32K = $2^5 \times 2^{10} = 2^{15}$
  - 15 bits are needed for each address.
  - We need 4 bits to select the chip, and 11 bits for the offset into the chip that selects the byte.



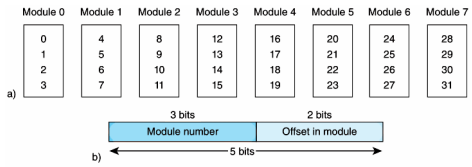
Memory Organization

- In high-order interleaving the high-order 4 bits select the chip.
- In low-order interleaving the low-order 4 bits select the chip.





Memory Organization

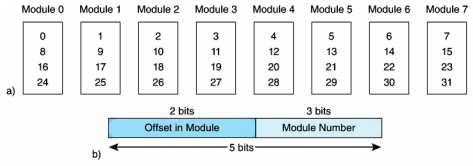


Module	Decimal Word Address	Binary Address	Address Split per Given Structure	Module Number	Offset in Module
Module 0	0	00000	000 00	0	0
	1	00001	000 01	0	1
	2	00010	000 10	0	2
	3	00011	000 11	0	3
Module 1	4	00100	001 00	1	0
	5	00101	001 01	1	1
	6	00110	001 10	1	2
	7	00111	001 11	1	3

c)

a) High-Order Memory Interleaving b) Address Structure c) First Two Modules

Memory Organization



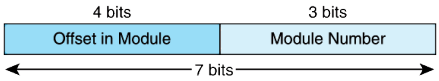
Module	Decimal Word Address	Binary Address	Address Split per Given Structure	Offset in Module	Module Number
Module 0	0	00000	00 000	0	0
	8	01000	01 000	1	0
	16	10000	10 000	2	0
	24	11000	11 000	3	0
Module 1	1	00001	00 001	0	1
	9	01001	01 001	1	1
	17	10001	10 001	2	1
	25	11001	11 001	3	1

c)

a) Low-Order Memory Interleaving b) Address Structure c) First Two Modules

Memory Organization

- EXAMPLE 4.1 Suppose we have a 128-word memory that is 8-way low-order interleaved
  - which means it uses 8 memory banks; $8 = 2^3$
- So we use the low-order 3 bits to identify the bank.
- Because we have 128 words, we need 7 bits for each address ($128 = 2^7$).



21

Interrupts

- The normal execution of a program is altered when an event of higher-priority occurs. The CPU is alerted to such an event through an interrupt.
- Interrupts can be triggered by I/O requests, arithmetic errors (such as division by zero), or when an invalid instruction is encountered.
- Each interrupt is associated with a procedure that directs the actions of the CPU when an interrupt occurs.
  - Nonmaskable interrupts are high-priority interrupts that cannot be ignored.

22