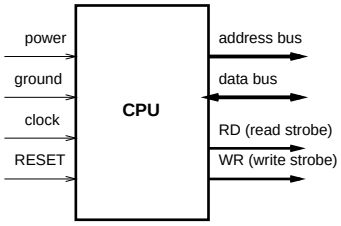
Microcomputer Architecture

This lecture gives an overview of the architecture of a simple microcomputer. It describes the operation of the CPU address and data buses during read and write cycles.
After this lecture you should be able to: (1) show how the following buses and signals are connected in a microcomputer system: power and ground, address and data buses, read and write strobes, and chip enables; (2) give the sequence of signals that must appear on the address, data, and control lines of a memory or I/O chip in order to read or write a particular data value to/from a particular address; (3) explain the purpose of these lines; and (4) compute the number of address lines required for a given memory size or vice-versa.

CPU

A microcomputer is usually assembled using a microprocessor chip (CPU[1]) connected to memory chips and I/O (input/output) chips. This section decribes the functions of the signals appearing on the pins on the CPU chip.

The following diagram shows the input and output signals on a typical CPU. The thick lines shows buses and the arrows indicate whether the signal is an input, output or, as in the case of the data bus, if it can alternate between the two directions.



As with any other digital logic chip, the CPU chip needs pins to supply power and ground. The CPU also needs a clock signal (a signal that periodically switches from high to low). This clock input is used by the processor to synchronize its internal operations. The processor is a (very complicated) state machine and the clock is used to sequence between the processor's states. The clock speed typically ranges from 32 kHz to several hundred MHz.

The chip has two buses: the address bus and the data bus. These two buses are used by the processor to communicate with memory and I/O chips.

An address bus of N bits can be used to select (address) one of 2^N bytes in the microcomputer's memory. Typical address bus sizes ("widths") are 16 bits (most 8-bit microprocessors), 20 bits (the 8088 CPU used in the original IBM PC), 24 bits (the 68000 CPU used in the lab computer) and 32 bits (the chips used in modern microprocessors). The individual signals in the bus are usually given the labels $A_0, A_1, A_2, \ldots, A_{N-1}$.

Exercise: Approximately how many bytes can each of the above processors address? *Hint: Use the approximation that* 2^{10} *is about 1000.*

The data bus is used to transfer data between memory or I/O peripherals and the CPU. The data bus width in modern computers is always an even multiple of 8 bits. For simplicity we will only consider CPUs with an 8-bit data bus width. The individual signals in the data bus are usually given the labels $D_0, D_1, \ldots, D_7$.

The address and data buses connect the CPU to the memory and I/O chips. Multiple memory devices can be connected in parallel to these buses but additional logic circuitry ("address decoders") are required to ensure that only one device is enabled at a time.

Finally, there are a number of control signals. We will only consider three of these: the read and write strobe outputs (RD and WR), and the RESET input.

The RD and WR outputs are used to control memory devices. If the RD output is high, the CPU is reading from memory (a "read" cycle); if the WR output is high the CPU is writing to memory (a "write" cycle).

The RESET input pin resets the processor to a known initial state. This is usually done when power is first applied or if the processor gets stuck while executing a buggy program. When the RESET pin is brought high the processor stops executing the current instruction sequence and restarts execution at an address that contains a program to restart the computer.

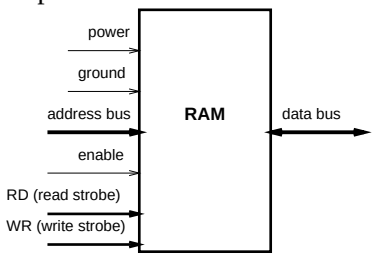
[1] Central Processing Unit

Memory

Memory chips are digital logic devices used in microcomputer systems to store programs and data.

These chips are of two main types: RAM and ROM. The contents of RAM (random access memory) chips can be both read and re-written by the CPU while the contents of ROM (read-only memory) chips can only be read. The advantage of a ROM is that, unlike a RAM, it's contents are retained when power is removed.

The diagram below shows the signal pins on a typical RAM chip:



As with any other logic device, it has power and ground pins. In a RAM the address bus is an input that is used to tell the memory chip which address is being accessed. The data bus is used to convey the data between the CPU and the RAM. It is a "bi-directional" bus since it is an input during a write operation and an output during a read operation. There is also an "enable" input that must be high for the RAM to do anything at all.

Exercise: How many address inputs are required on a 64 kByte (2^{16} byte) byte-wide RAM?

The following operations take place during a read operation ("cycle"):

- the CPU puts the address of the desired memory location on the address bus
- the CPU turns its data bus into an input
- the CPU asserts (brings high) the RD signal line
- the RAM detects the high signal on the RD line and turns its data bus into an output
- the RAM looks up the value current value stored for that memory location and, after a short delay (the access time), outputs it on the data bus
- the CPU read the value from the memory

The following table shows the values of the different signals over time during a read cycle where the CPU reads the value 0x32 from address 0x105:

address bus	data bus	RD	WR
0x105	-	0	0
0x105	-	1	0
0x105	0x32	1	0

The following operations take place during a write cycle:

- the CPU puts the address of the desired memory location on the address bus
- the CPU turns its data bus into an output
- the CPU puts the value to be stored on the data bus
- the CPU asserts (brings high) the WR signal line
- the RAM detects the high signal on the WR line and turns its data bus into an input
- the RAM stores the value currrently on the data bus into the desired memory location inside the RAM

Exercise: Write out a table similar to that above showing the values appearing on the two strobes and the address and data buses when the value 0x33 is written to address 0x1200.

It may be helpful to consider that the memory chip operates like an array. The input to the chip is like the array index and the data value read or written by the memory chip corresponds to the current content of that element of the array.

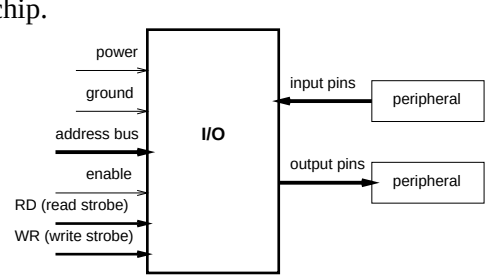
There are many types of ROM chips. The most common are EPROMs (erasable programmable read-only memory) which can be erased by exposing them to strong ultra-violet light for several minutes and then re-written using a device called a device programmer. EEPROMs (electrically erasable programmable read-only memory) are similar to EPROMs, but they can be quickly erased by the device programmer before being re-programmed.

The interface between a CPU and a ROM is similar to that for a RAM except that there is no write cycle.

Exercise: Would a ROM have a write strobe (WR) input?

I/O Chips

I/O chips are used to allow the CPU to interface with peripherals (keyboards, printers, etc). An I/O chip is similar to a memory chip but during a write cycle the chip "stores" the value on the output pins of the chip instead of storing the value internally. These output pins in turn are connected to the peripheral. A read of the I/O chip causes the value currently on the input pins to be transferred to the CPU. This allows the CPU to monitor the state of the input pins on the I/O chip.



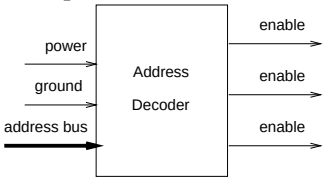
There are many different I/O chips available. They usually include additional logic circuits to make it easier for the CPU to deal with specific peripherals such as modems or hard disks. Later in the course we will look at a few common interface chips.

Address Decoders

A microcomputer often uses several memory and I/O chips, each of which is smaller than the total amount of memory that the microprocessor (CPU) can address. For example, a CPU with with a 16-bit address bus can address a 64 kBytes of memory but may be used in a system with a 16 kByte RAM, a 32 kByte EPROM and a 4-byte I/O chip.

Exercise: How many address lines are required by each of the above chips?

The purpose of the address decoder is to look at the address output by the CPU and enable individual memory or I/O chips. The following diagram shows the inputs and outputs of an address decoder:



Exercise: Draw a diagram showing how a CPU, RAM, EPROM, I/O chip and address decoder would be connected to build a microcomputer. Show the connections of the data and address buses and the read and write strobes. Use arrows at each chip to indicate whether a particular signal is an input or an output.

Microcontrollers

A microcontroller is a single-chip microcomputer. One chip includes the CPU, a RAM, an I/O chip and an EPROM or EEPROM. This allows all of the pins on the chip to be used for I/O.