From Silicon to CPUs !

One of the most fundamental components in the manufacture of electronic devices, such as a CPU or memory, is a **switch**. Computers are constructed from thousands to millions of switches connected together. In modern computers, components called **transistors** act as electronic switches. A brief look at the history of computing reveals a movement from mechanical to electromechanical to electronic to solid state electronic components being used as switches to construct more and more powerful computers as illustrated in Figure 7.9:

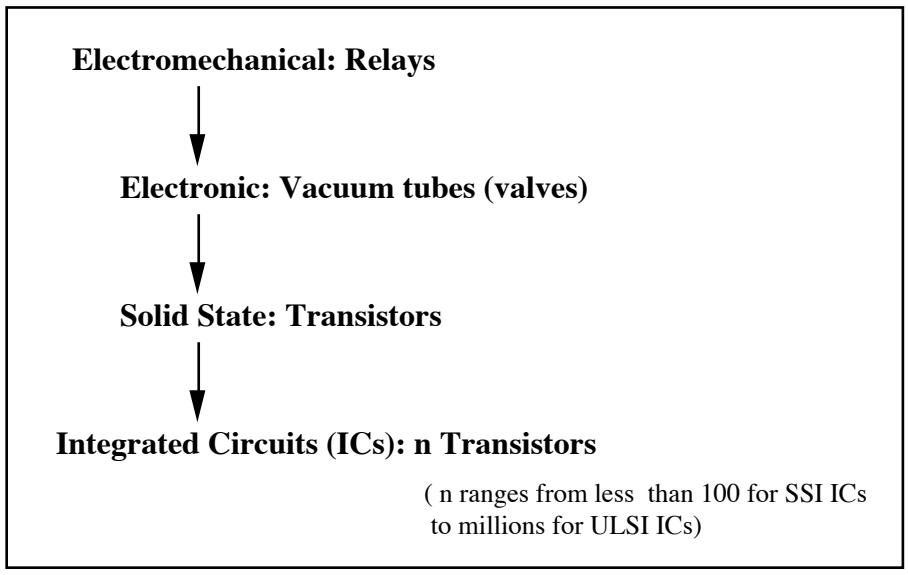

Figure 7.9: Evolution of switching technology

Transistors are fundamental components in the construction of computers. In crude terms, they act as electronic switches, i.e. they allow information to pass or not to pass under certain conditions. The invention of the transistor, at Bell Labs in 1948, revolutionised the development of computers. Transistors were much smaller, more rugged, cheaper to make and far more reliable than the valves which they replaced. The development of **integrated circuits** (ICs) allowed the construction of a number of transistors on a single piece of silicon (the material out of which IC's are made). IC's are also called **silicon chips** or simply **chips**. The number of transistors on a chip is determined by its level of

integration. Early chips had only a few transistors and used small-scale integration (**SSI**). As technology developed, medium-scale integration (**MSI**) allowed hundreds of transistors per chip, large-scale integration (**LSI**) allowed thousands of transistors per chip, very large scale integration (**VLSI**) allowed hundreds of thousands of transistors per chip. Today, millions of transistors are available per chip, e.g. Intel's Pentium microprocessor consists of an IC with in excess of 3 million transistors. This level of integration is called ultra large scale integration (**ULSI**). In 1965, Gordon Moore (chairman of Intel at the time) predicted that the complexity of integrated circuits would double every two years (this has become known as **Moore's Law**). This prediction has proved very reliable to date and it seems likely that it will remain so over the next ten years.

7.7.1 Logic Circuits

Transistors are the fundamental components from which are constructed **logic circuits**. These logic circuits are in turn the basic building blocks of the CPU. They include devices called **gates** and **flip-flops**. (A single gate may use up to 6 transistors.) Gates and flip-flops are used to construct more complex circuits such as **adders**, **decoders**, **registers** and **counters**. These circuits in turn are used to build **ALUs** and **control units**, i.e. CPUs as illustrated in Figure 7.10. By way of analogy, houses are made up of rooms, rooms are made up of walls, walls are made up of bricks and bricks are made up of sand and cement. If by analogy, houses correspond to CPUs, then the sand and cement corresponds to transistors. In one sense this analogy is quite appropriate since transistors are developed on silicon chips and silicon is the essential element of sand!

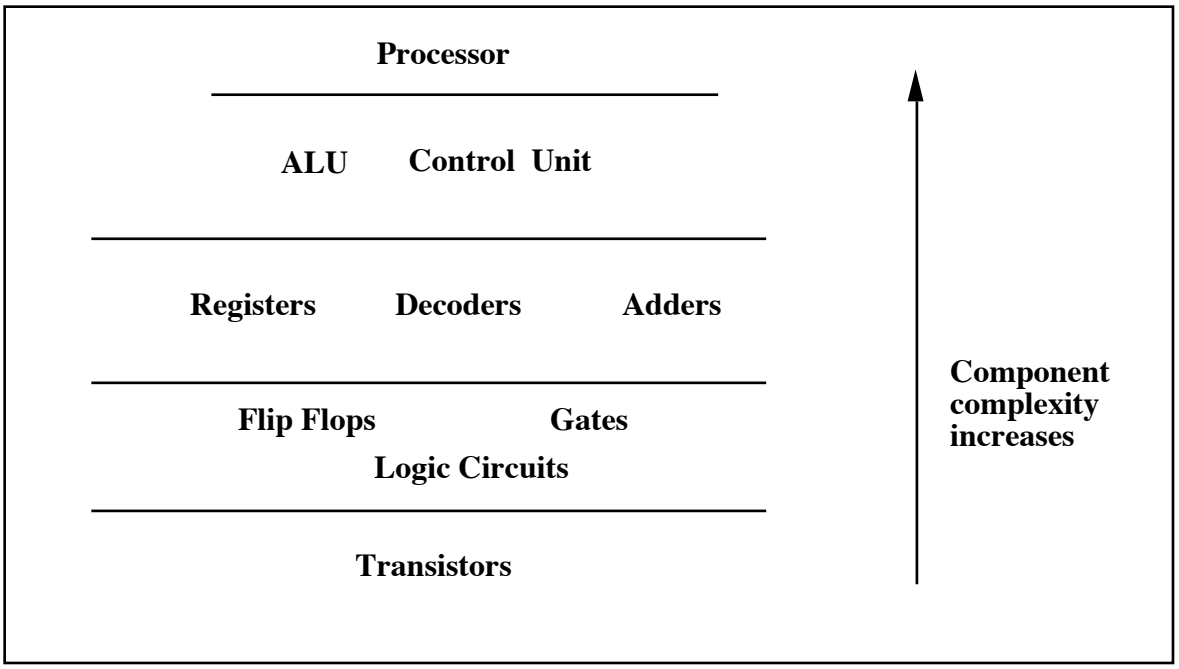

Figure 7.10: From Transistors to Processors

Different names are used for the basic logic circuits such as **binary** circuits, **digital** circuits, **Boolean** circuits and **gates**. They are called logic circuits because they perform the logic operations (e.g. AND, OR etc.). There are a number of types of logic gates such as and, or, xor, nand and nor gates. The term **gate** refers to the fact that they act like gates, letting some signals through and blocking others, depending on their inputs. The term **Boolean** comes from George Boole, the originator of Boolean Algebra.

Logic circuits fall into two classes: sequential logic circuits and combinatorial logic circuits. Combinatorial circuits are those where the output is at all times a function of the current inputs to the circuits (**no** feedback is allowed from the outputs to the inputs). **Decoders** and **adders** are important examples of such circuits, used in the construction of digital systems such as CPUs. Sequential circuits are those where the outputs depend on **past** inputs as well as **current** inputs (they allow feedback). They are sequential in that the output depends on the sequence leading to the present situation. As a consequence, such circuits exhibit **memory**, i.e.

they can retain information. The **flip-flop** is one of the best known sequential circuits. There are a number of types of flip-flop such as the R-S flip-flop, J-K flip-flop and **D-type** flip-flop. **Registers** can be constructed from D-type flip-flops and so D-type flip-flops are commonly used in computers. Memory may also be implemented using flip-flops, as may be **shift registers** and **counters**, which are also important computer components.

In summary, gates and flip-flops, which are constructed out of transistors, are the basic building blocks of computers.

7.7.2 Chip Fabrication

Silicon chips have a surface area of similar dimensions to a thumb nail (or smaller) and are three dimensional structures composed of microscopically thin layers (perhaps as many as 20) of insulating and conducting material on top of the silicon. The manufacturing process is extremely complex and expensive. Silicon is a **semiconductor** which means that it can be altered to act as either a conductor allowing electricity to flow or as an insulator preventing the flow of electricity. Silicon is first processed into circular wafers and these are then used in the fabrication of chips. The silicon wafer goes through a long and complex process which results in the circuitry for a semiconductor device such as a microprocessor or RAM being developed on the wafer. It should be noted that each wafer contains from several to hundreds of the particular device being produced. Figure 7.11 illustrates an 8-inch silicon wafer containing microprocessor chips.

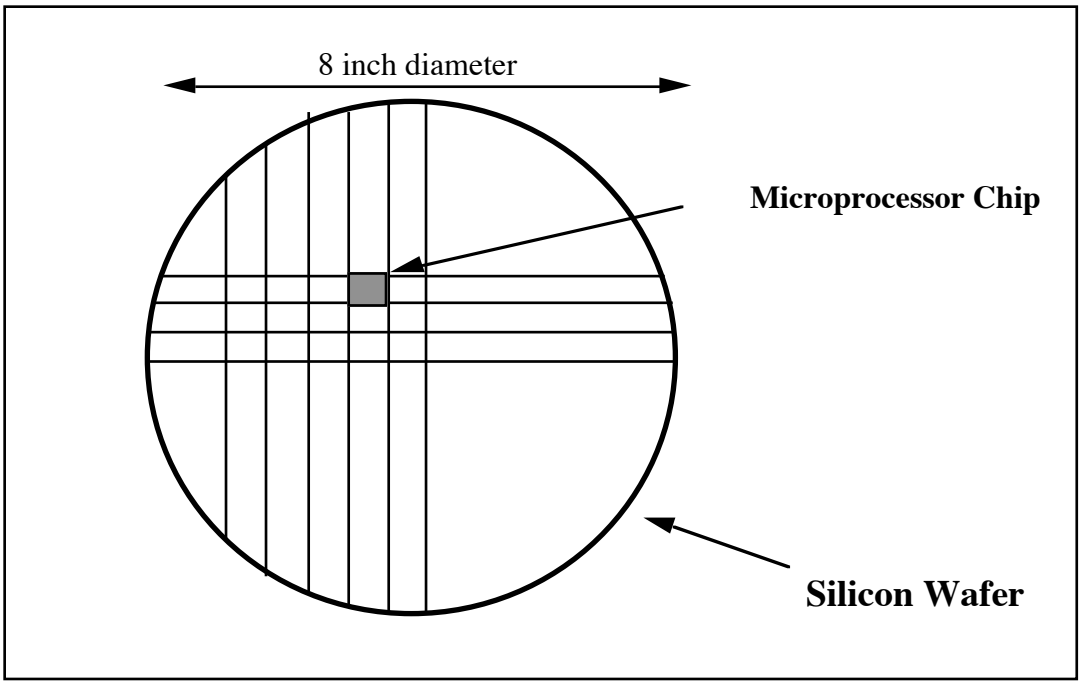

Figure 7.11: A single silicon wafer can contain a large number of microprocessors

The individual chips are obtained by cutting them out of the wafer with a high precision diamond saw. Before this process, the wafer goes through a "wafer sort" facility which is an electrical test of each chip on the wafer to ensure they are working correctly. Malfunctioning chips are marked with ink so they may be discarded. The percentage of functioning chips is referred to as the **yield** of the wafer. Yields vary substantially depending on the complexity of the device being produced, the feature size used and other factors. While manufacturers are slow to release actual figures, yields as low as 50% are reported and it is accepted that 80-90% yields are very good. A high chip failure rate should not be surprising, given the complexity of the production task and the sub-micron feature size. A single short circuit, caused by two wires touching in a million plus transistor chip, is enough to cause chip failure!

The **feature size** refers, in simple terms, to the size of a transistor or to the width of the wires connecting transistors on the chip. One micron (one thousandth of a millimetre) is a common feature size and state of the art

chips are using **sub-micron** feature sizes from 0.8 to 0.3 microns. Intel are producing 80486 and Pentium chips using 0.8-micron process technology at the time of writing while 0.6-micron technology will be in use for these microprocessors, before this text is published. The smaller the feature size, the more transistors there are available on a given chip area. This allows more microprocessors for example to be obtained from a single silicon wafer. It also means that a given microprocessor will be smaller, runs faster and uses less power than its predecessor using a larger feature size. Since more of these smaller chips can be obtained from a single wafer, each chip will cost less which is one of the reasons for cheaper processor chips. (Market forces are another powerful reason!) In addition, reduced feature size it makes it possible to make more complex microprocessors, such as the Pentium, which uses in excess of 3 million transistors and its planned successor (code-named the P6) will use around 5.5 million transistors. An obvious way to increase the number of transistors on a chip is to increase the area of silicon used for each chip (the **die size**), however, this can lead to problems. Assume that a fixed number faults occur randomly on the silicon wafer illustrated in Figure 7.11. A single fault will render an individual chip useless. The larger the die size for the individual chip, the greater the waste in terms of area of silicon, when a fault arises on a chip. For example, if a wafer were to contain 40 chips and ten faults occur randomly, then up to 10 of the 40 chips may be useless giving up to 25% wastage. On the other hand, if there are 200 chips on the wafer, we would only have 5% wastage with 10 faults. Hence, there is a trade-off between die size and yield, i.e. a larger die size leads to a decrease in yield.

When the chips have been extracted from the silicon wafer (they are now called **dies**) they are **packaged**. They are first connected to a set of leads that enable the chip to communicate with other devices. Then the chip is encased in plastic or ceramic material which protects it from contamination. Three common forms of packaging are used. DIP (Dual In-Line Packaging) is a familiar form where up to 64 metal leads or pins protrude from each side of the chip. These pins allow the chip to be inserted into holes on a printed circuit board (PCB). Because of the

increasing complexity of chips, more pins are required than DIP can easily support. Another form of packaging is PGA (Pin Grid Array) packaging. Here, the metal pins (up to several hundred) protrude from layers inside the packaging and plug into sockets on PCBs. Another form of packaging is called SMT (Surface Mount Technology) where the metal leads are on all four sides of the chip and are soldered directly to the conductors on the surface of a PCB. These chips can be mounted on both sides of the same PCB and can accommodate chips requiring more than 200 leads.

Different types of chip technologies such as **MOS** (e.g. CMOS, nMOS) and **bipolar** technologies (e.g. TTL, ECL)[1], have been developed that govern such things as the speed of operation and power consumption of a chip. MOS technology allows a high level of integration and but not as high a speed of operation, as bipolar technology which provides for very high operating speeds. On the other hand bipolar chips are very expensive and consume larger amounts of power than say CMOS chips. ECL chips tend to be used in supercomputers, where speed of operation is a critical factor. In modern microprocessor designs, both technologies may in fact be used on the same chip, giving rise to BiCMOS chips. The speed critical parts of the microprocessor can be implemented in bipolar technology, while the less critical parts can be implemented in the slower CMOS technology.

One way that the performance of an architecture can be improved, without any changes to the instruction set architecture or organisation, is, to use a faster chip technology.

The advances in chip design are to a large extent due to the development of CAD (Computer Aided Design) tools that allow designers create new chip designs and simulate how they will work as well as test them for design errors. It would be impossible to design state of the art microprocessors without such software.

---

[1]MOS (Metal Oxide Semiconductor), CMOS (Complimentary MOS), nMOS (Negative channel MOS), TTL (Transistor Transistor Logic), ECL (Emitter Coupled Logic).

7.8 Summary

A number of techniques for improving processor performance were described in this chapter such as **pipelining**, **superscalar** architectures and the use of **cache** memory. This was followed by a discussion of **RISC** and **CISC** architectures and their relative merits. The complex area of computer performance was then tackled, outlining such metrics as MIPS, MFLOPS, and SPECmarks and the problems associated with them. Finally, a brief introduction to the world of semiconductor technology and chip fabrication, was presented.

Exercises

7.1 Explain the concept of pipelining and define flowthrough time and throughput.

7.2 What would the throughput and flowthrough time of a pipeline, made up of 5 stages where the stages take the following units of time to complete: 1, 2, 1, 3, 2 ?

7.3 Why are branch instructions a problem for pipelining? What other problems arise with pipelining?

7.4 What prevents the parallel execution of every pair of instructions in a superscalar machine with multiple functional units? Explain how instruction re-ordering can alleviate this problem.

7.5 What is cache memory and why is it so important? What is the average memory access time for a machine with a cache hit rate of 95%, where the cache access time is 10ns and the memory access time is 80ns?

7.6 Compare a typical RISC processor with a CISC processor using a table with headings such as: instruction length, addressing modes, instruction set size, clock cycles per second.

7.7 Why is measuring computer performance so difficult ? What are the problems associated with using MIPS, FLOPS, benchmarks and SPECmarks as performance metrics?