University of Linkoping - Department of Computer Science
Course Leader: Erik Larsson

1. The von Neumann Architecture

The principles:

(a) Data and instructions are both stored in main memory (stored program concept).
(b) The content of the memory is addressable by its location.
(c) Instructions are executed sequentially, i.e. from one instruction to the next, unless the order is explicitly modified.
(d) The organization (architecture) consists of:
(e) The central processing unit (CPU) containing: the control unit (CU) which coordinates the execution of instructions, and the arithmetic/logic unit (ALU) which performs all arithmetic and logical operations.
(f) The main memory (RAM).

All von Neumann designs are general-purpose computers. They can solve different problems depending on the program they execute.

2. The central processing unit (CPU)

The primary function of the CPU is to execute the instructions fetched from the main memory (RAM). An instruction tells the CPU to perform basic operations, i.e. an arithmetic or logic operation, or to transfer data from/to main memory. The CPU interprets (decodes) the instruction to be executed which "tells" the other components what to do. The CPU also includes a set of registers which are temporary storage devices that store data and intermediate results.

3. The Instruction Cycle

Each instruction is performed in steps. The steps corresponding to one instruction are referred together as the instruction cycle.

(a) Fetch Instruction.
(b) Decode Content.
(c) Fetch Operand.
(d) Execute Instruction.

4. Memory Organization

(a) The main memory (RAM) is used to store the program and data which are currently manipulated by the CPU.
(b) The secondary memories, e.g. disks, flash, cards, and other storage media, provide the long-term storage of large amounts of data and programs.
(c) Before data and programs in the secondary memory can be manipulated by the CPU, they must first be loaded into the main memory.
(d) The most important characteristics of a memory is its speed, size, and cost, which are mainly constrained by the technology used for its implementation.
(e) Typically, the main memory is fast and of limited size, and the secondary memory is relatively slow and of very large size.

The most widely used technology for main memories is semiconductor chips. The most common memory type is the random access memory (RAM). The information stored in semiconductor memories will be lost when electrical power is removed.

5. Problems with the Memory System

Main problem: We need a memory system to fit many programs and to work at a speed comparable to that of the processor. However, processors work at a very high rate and need large memories. Memories are much slower than processors.
Facts: The larger a memory, the slower it is. The faster the memory, the greater the cost/bit.
Solution: The designer aims to make a composite memory system that combines a small fast memory as well as a large slow main memory. This system behaves (most of the time) like a large fast memory. This two level principle above can be extended into a hierarchy of many levels including cache memories and secondary memories (disks).

6. Cache Memory

The cache is a small, expensive, and very fast memory that retains copies of recently used information stored in main memory (RAM). It is used as a bridge between CPU and RAM in order to speed up the memory retrieval process. Cache operates transparently to the programmer automatically deciding which values to keep or overwrite. Cache has its own logic which is entirely based on hardware. This arrangement ensures that decisions are taken very fast, typically within a few nanoseconds.

(a) The processor operates effiently when the memory items it requires are held in the cache rather than in the RAM. Items are instructions and/or data.
(b) The overall system performance depends strongly on the proportion of the memory accesses satisfied by the cache. Therefore:

(b1) An access to an item found in the cache: “hit”.
(b2) An access to an item not found in the cache: “miss”.
(c) The proportion of all memory accesses satisfied by the cache: “hit rate”.
(d) The proportion of all memory accesses not satisfied by the cache: “miss rate”.
(e) The miss rate of a well-designed cache is only a few (%), e.g. 10-20 (%).

Contemporary designs feature both internal (on-chip) as well external (off-chip) caches. Such designs ensure that the “hit rate” is high most of the times. Thus, a series of caches, numbered as L1, L2, L3, etc. are possible. Level 1 cache is the fastest and smallest of all; Level 2 is larger and slightly slower than Level 1; Level 3 is again larger and slightly slower than Level 2.

The processor (CPU) checks the entire cache system using the following sequence:
L1 => L2 => L3. In the event of a cache miss, it fetches items from the RAM.

7. Virtual Memory

The virtual program space is much larger than the physical memory (RAM). It is divided into equal, fixed-size segments called pages. Also, the RAM is organized as a sequence of frames and a page is assigned to an available frame for storage (page size = frame size). The page is the basic unit of information moved between main memory and disk by the virtual memory system. Common page sizes are: 2 - 16 Kbytes.

When a page is not found in the main memory, a page fault is produced, and the OS loads the missing page into the RAM. Then, a physical address is generated, and then the CPU resumes its activities.

8. Input / Output Devices

Input and output devices provide the means for people to make use of computer systems. Some I/O devices also function as interfaces between a computers and other systems. Such interfaces include analog-to-digital (A/D) and digital-to-analog (D/A) converters.