Q1. What are the trends in HPC and parallel computer architectures?

**Performance Trends in HPC:**

The number of floating-point operations executed per second can be used as an indication for analyzing the performance trends of computers. Figure 1.1 below shows a steady increase in the number of floating-point operations/second.

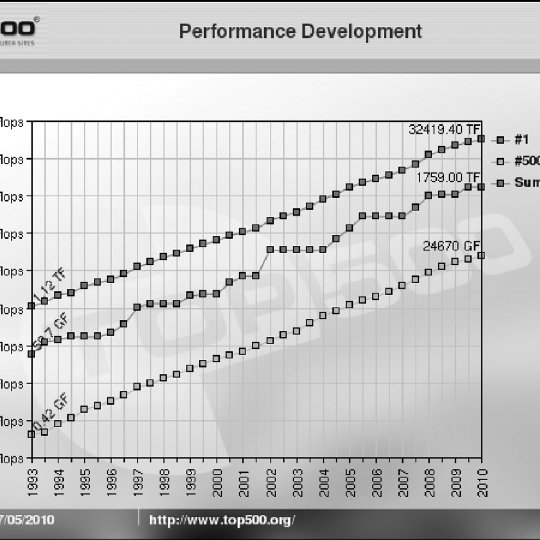


Figure .1

**Technology trends in HPC:**

According to Moore’s law, the number of transistors that can be incorporated into a single chip will roughly double every 1.5 years (See yellow line in fig 1.2). However, we notice that the same trend does not hold for other factors such as clock frequency and power consumption. In recent years, it has been observed that the clock frequency has flatlined due to power dissipation limits, so how is there still a performance gain? The answer lies in the shift to parallel computing, where the number of cores/processors has increased to compensate for the limitations to improving single-processor speeds. (See black dots, fig 1.2). By increasing the number of cores/processors, tasks can be distributed among the different cores/processors and executed in-parallel for better performance.

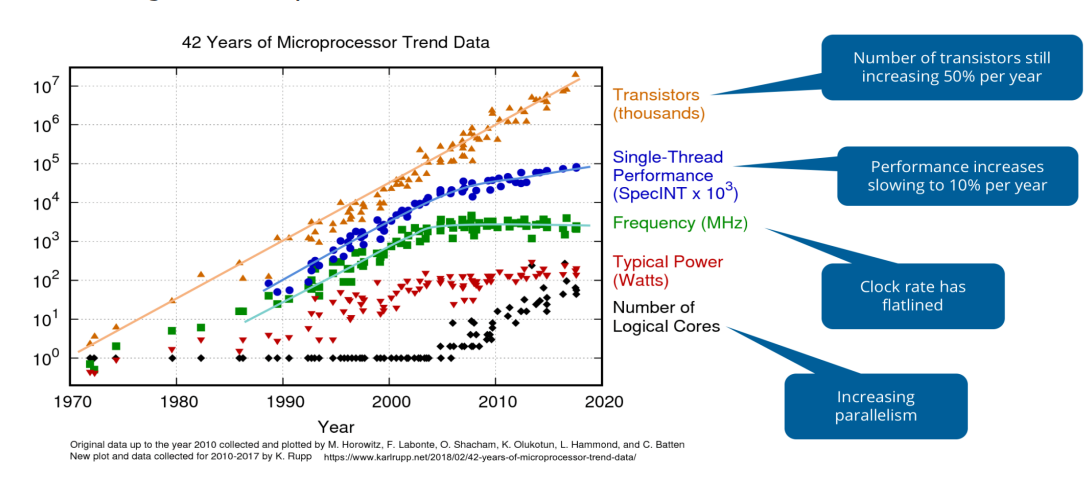


Figure 1.2

**Shift to heterogenous architecture**

Nowadays, it is popular to have a heterogenous architecture which includes both a CPU and GPUs (Graphics Processing Unit), for achieving high performance. GPUs have many more computing units when compared to CPUs, so can be used for achieving high computation throughput. However, each GPU core is not powerful as a single CPU, so applications generally use a combination of both CPUs and GPUs for the best performance.

**Cluster computing**

With the improvement of network connectivity, tasks can be distributed among multiple inter-connected computers working in parallel for performance improvement.

Q2. What is cache memory? What is the reason of introducing cache memories?

Cache memory is a type of fast, relatively small memory that is stored on computer hardware. The purpose of cache memory is to store program instructions and data that are frequently used by the computer during its general operations. Cache makes use of SRAM (faster), whereas main memory uses DRAM (slower).

Cache has generally smaller memory size when compared to the main memory because it is more expensive. It is also placed closer to the CPU compared to main memory and disk drives for faster access (See fig 2.1)

Cache is also usually present at multiple levels (Fig 1.3). Each level has a different memory size and access time.

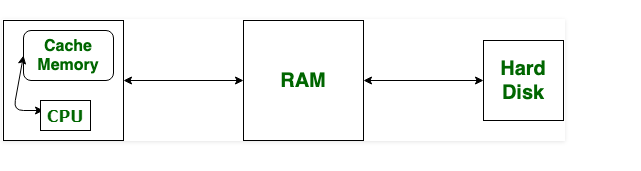


Figure .1

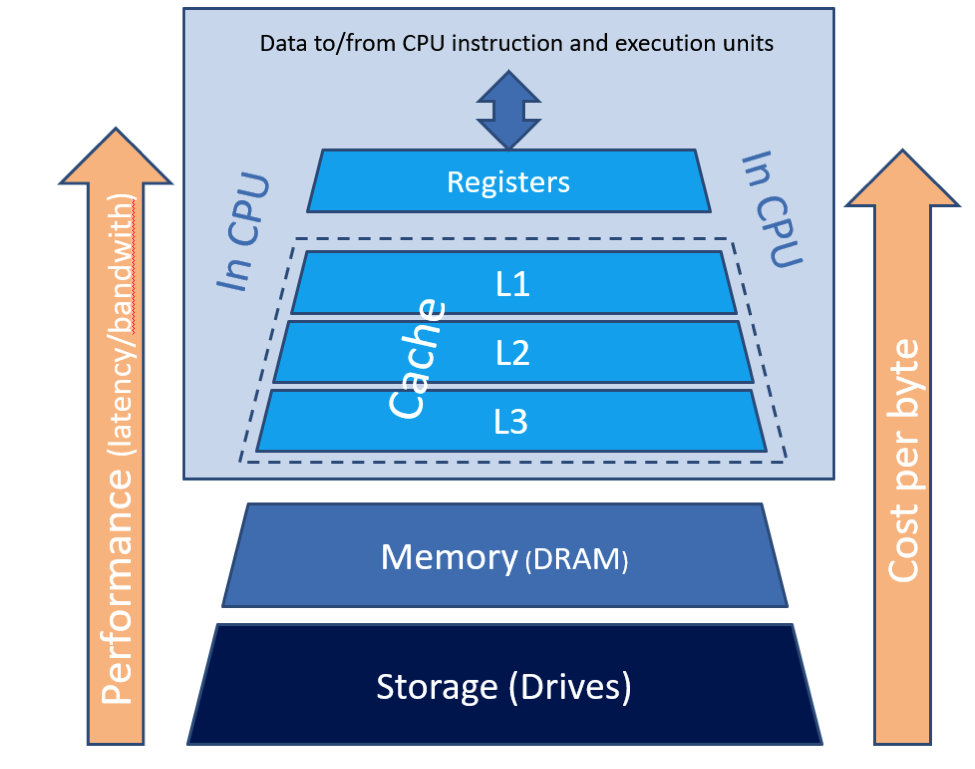


Figure 2.2

Advantages of using cache:

1. Fast access of frequently used data and instructions.
2. Can reduce the load on the data/instruction bus, allowing more processors to be connected to the bus.