Computer components and performance

Recall

Response time and throughput

If you were running a program on two different desktop computers, you can say that the faster one is the desktop computer that gets the job done first. If you were running a datacenter have several servers (server can be understood as a computer) running jobs submitted by many users, you can say that the faster computer was the one that completed the most jobs during a day. As an individual computer user, you are interested in reducing **response time** (the time between the start and completion of a task, or *execution time*). On the other hand, datacenter managers are often interested in increasing **throughput** (somewhere called *bandwidth*, the total amount of work done in a given time, in physics we called frequency but we changed the number of motions with the number of jobs).

Time

Time is the measure of computer performance: the computer that performs the same amount of work in the least time is the fastest.

We called the total time to complete a task, including disk accesses, memory accesses, input/output (I/O) activities, operating system overhead - everything (the time between the start and completion of a task) is **response time**, or **elapsed time** and we have formula

$$t_{resp} = t_{end} - t_{start} = \text{CPU time} + \text{IO wait time}$$

In **elapsed time**, we have a smaller time which is **CPU time** (or *CPU execution time*). **CPU time** is just only time spent processing a given job (Discounts I/O time, other jobs shares, it's actual time the CPU spends computing for a specific task). **CPU time** comprises user CPU time and system CPU time.

Improve performance

We can improve performance by improve response time (t_{resp}) or throughput

But, if you use **faster CPU** $\implies t_{resp}$ reduced $\implies$ throughput increased because time for jobs decreased. Hence, in this case, both of response time and throughput are improve

Another case, we can add more CPUs. In this case, throughput improved because in one time more jobs are done at the same time, so response time ***may be*** improve

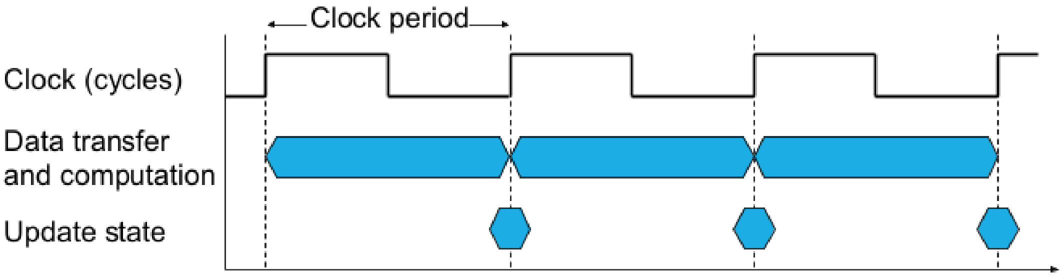
Example: In computer, if you want to do some jobs (Like 5 jobs), CPU maybe process each a job. For example, if CPU process job 1, 4 jobs left in queue, and if we have more CPU, we can decrease jobs in queue. So, we can decrease IO time because we do not need more time for transferring jobs from queue to CPU

Key term

- Response time (execution time): The total time required for the computer to complete a task, including disk accesses, memory accesses, I/O activities, operating system overhead, CPU execution time, and so on.
- Throughput (bandwidth): Another measure of performance, it is the number of tasks completed per unit time.
- CPU time: It is actual time the CPU spends computing for a specific task (Discounts I/O time, other jobs shares)

CPU Clocking

As mentioned above, computer users we need to care about time (t_{resp}) but we need new unit convenient to think about performance in other metrics. Almost all computers are constructed using a clock that determines when events take place in the hardware. These discrete time intervals are called **clock cycles** (or *ticks, clock ticks, clock periods, clocks, cycles*). (Designers refer to the length of a clock period both as the time for a complete clock cycle (e.g., 250ps) and as the clock rate (e.g., 4 GHz), which is the inverse of the clock period)



CPU Time

Instruction count and CPI

Iron Law

Other metrics

MIPS

MFLOPS

Amdahl's Law

Power