

# Computer Performance Latency and Throughput

Posted on October 7, 2019

Computer performance refers to the amount of work that a computer system can accomplish. Based on different contexts, when evaluating high computer performance, one should consider high throughput. Throughput refers to the rate at which a computer processes its instructions. Other things to consider include short response times in processing certain types of work, minimum usage of computer resources, high bandwidth, and a faster rate of compressing and decompressing data (Thomas 11). In addition, the computer system should guarantee high availability while also providing a shorter data transmission time. Computer software performance determines the quality of the software, which defines the interaction between humans and computers.

In understanding the aspect of computer performance, there is a need to examine what is meant by performance engineering (Thomas 13). This term applies in the field of software engineering to refer to all sets of roles, practices, skills implemented, tools used, activities occurring, and deliverables arising at every stage of the system development life cycle (SDLC). All those elements ensure that system design and analysis manage to achieve the performance requirements outlined in the solution. Besides, performance engineering encompasses all trade-offs that define different types of performances. Therefore, enhancing the working of the CPU will require improving all performance aspects. Besides, application performance engineering (APE) within performance engineering serves the purpose of overcoming any challenges relating to application performance within environments such as cloud computing, mobile development, and terrestrial IT (Thomas 17). Here, elements such as skills, practices, and deliverables within every stage of the application lifecycle ensure that system development manages to achieve all the [non-functional](#) performance requirements.

## Aspects of Performance

Aspects of performance refer to things that one can measure when evaluating a system's performance. These aspects refer to elements such as response time, latency, availability, scalability, and efficiency, among others.

## Availability

A system's availability is measured by examining its reliability; an increase in reliability translates to improvements in the system's availability as it results in less downtime. Another strategy for improving the availability of a system includes increasing its maintainability and testability. Enhancement of maintainability becomes easier in comparison to improvements in reliability

(Osterhage 13). Maintainability estimates entail the repair rates and related processes being adequately managed to guarantee accuracy. However, reliability estimates provide uncertainties as they are substantial and often result in availability problems even with extremely high maintainability levels.

## Response Time

Response time refers to the time taken by a computer system to generate the expected output based on some given input data. In addition, the time that a system takes to process requests for services helps define its response time. Thus, the field of computing derives three elements that sum up the response time. First, there is service time, which refers to the time taken to work on a requested task (Osterhage 18). Second, there is the wait time, which is the duration taken by a process in a queue before it commences execution. Lastly, there is the transmission of time. This refers to the time taken to capture input data, execute, and generate feedback for the requestor.

## Processing Speed

The choice of computer architecture depends on the speed which the system can manage to run the pre-existing and pre-compiled computer software. Therefore, the operating frequency determines the speed of the computer system. This means that the higher the processing speed, the more powerful a computer system is.

## Channel Capacity

Channel capacity refers to the maximum bound that a communication channel can accommodate while transmitting information at a specific rate. Based on the noisy-channel coding theorem, the limiting information rate defines a channel's capacity by comparing the units of information transmitted over a time unit and with the least probability of errors (Osterhage 25). Therefore, the maximum rate of information traversing between the input and the output of a channel defines its capacity. The input distribution determines the level of maximization.

## Latency

Latency refers to the time delay within a computer system, especially when executing different processes. Limited velocity, which determines the rate of any physical interaction, defines a system's latency. This velocity is usually a figure equal to the speed of light or lower. Therefore, any physical system with spatial dimensions that do not equate to zero has to experience some level of latency. Determining the lower limit of latency in communication depends on the medium of communication used. Moreover, latency can also refer to the delay between inputting user commands to the system

and the time the computer will generate the results (Osterhage 29). Therefore, latency has a stronger impact on user satisfaction and usability with regard to human-computer interaction. The development of real-time computing systems should achieve the worst-case response, requiring the CPU to be set at low interrupt latency, especially when there is a deterministic response.

## Bandwidth

The consumed or available data, which forms the communication resources measured in bit-rate, generates the bandwidth in computer networking. The standard unit of measurement of bandwidth is bits per second, expressed as bits/s (Osterhage 31). Also, bandwidth can define channel capacity, net bit rate, or the maximum throughput of a communication path within a computer system.

## Throughput

Throughput refers to the rate of executing a process. Measuring throughput units usually takes into account the reciprocal of the outcome of the propagation delay expressed as seconds per output. Therefore, one can equate throughput to a computational device, such as an embedded processor, performing a dedicated function, such as simplification of system analysis. Other aspects of performance include scalability, which refers to the ability to enlarge the network or computer system to accommodate more processes (Osterhage 37). Also, there is relative efficiency and power consumption, which entails the amount of energy a computer system utilizes when executing different operations. Besides these aspects, there are several other factors that can negatively impact the performance of a computer.

## Factors Affecting Computer Performance

At times a computer may be slow or fast when executing different processes. The speed usually depends on multiple factors such as the memory space on the hard disk, the CPU speed, the size of the random access memory (RAM), the ability of the computer to multitask, the type of graphics card installed, and the process of defragmenting the files.

## The CPU Speed

CPU speed, as well as the clock speed, refers to the frequency of executing instructions within the processor. The speed also refers to the frequency of processing data within the CPU (Jouppi et al. 20). The unit of measuring the CPU speed is the number of cycles per second, also known as megahertz (MHz). The higher the clock speed, the higher the performance of the CPU. Therefore, the execution of computer processes will be faster with high clock speed.

## Size of the RAM

Random Access Memory (RAM) is a vital component of every computer. The RAM is used to store information and data about current processes running on the computer. Therefore, retrieving the data saved in RAM should be fast, simple, and straightforward (Jouppi et al. 25). A computer's performance is directly proportional to the size of its RAM. Therefore, the larger the computer's RAM, the higher the performance. Limited RAM results in the computer slowing down.

## The speed of the Hard Disk

Hard disk speed refers to the rate of writing and reading files in the memory storage capacity. However, disk speed varies significantly due to factors such as the type of the drive, the manufacturer, as well as its usage (Jouppi et al. 29). Therefore, a computer system with a hard disk that maintains high access speeds offers better performance.

## Hard Disk Space

A hard disk with bigger storage space helps ensure that the computer performs well. Therefore, a smaller disk space causes a computer to slow down. Storing more data in the hard disk means there will be further utilization of the memory, hence reducing the amount of memory that a processor requires to run its operations.

## Number of Processes Running on the Computer

The multitasking technique usually results in slightly slower computer performance. Multitasking utilizes more physical memory while running different applications, as each process has to compete for available resources. On the other hand, if a computer is running only one application, the single process makes use of all memory, and hence execution will be faster (Jouppi et al. 33). Therefore, an increase in the number of applications running at the same time leads to reduced computer performances. Usually, when such applications compete for the available resources, it may result in a deadlock. A deadlock interrupts the computer's working causing significantly slower process execution.

## Type of Graphic Card

A graphics card usually affects image rendering speeds. Powerful graphics cards are essential in fields that require image and video rendering, such as computer design and architecture fields (Jouppi et al. 36). The quality of pictures, and videos, and the ability to animate things depends on a computer's

performance. Graphics design tools usually encompass most processes running simultaneously. Hence the choice of the graphics card will determine how fast or slow the computer system will execute its processes. Therefore, integrating a powerful graphics card will enhance the computer's performance, especially when dealing with graphical images and videos.

## Defragmenting files

Defragmenting files is necessary as it enhances easier readability and execution of files. However, defragmenting files forces the computer to implement multitasking techniques that may reduce the computer's performance. Therefore, the process of defragmenting files can only be possible and faster while utilizing computer systems with higher CPU speeds. They also should have significantly large RAM (Jouppi et al. 41). A computer system with a high computation speed will easily achieve process multithreading.

## Conclusion

The performance of a computer system is essential when executing different processes, especially if they involve the input and output of data. Therefore, various computer performance aspects determine how it manages to respond to different processes. The determinant factors include availability, processing speed, channel capacity, latency, and the power that the computer consumes. Also, it is evident that several factors affect a computer's performance. They include the CPU speed, the size of the RAM, disk space, and the graphics card installed in the system. Therefore, it will be easier to develop a computer system that will meet all user requirements at a faster rate only if all the factors that affect the system's performance are considered.

## Works Cited

Jouppi, Norman P., et al. "In-Data Center Performance Analysis of a Tensor Processing Unit." *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ACM, 2017.

Osterhage, Wolfgang. *Computer Performance Optimization: Systems – Applications – Processes*. Berlin: Springer Berlin Publishing House, 2015.

Thomas, Nigel. *Computer Performance Engineering: Proceedings*. Berlin: Springer Berlin Publishing House, 2011.