**Difference between Response time & Throughput**  
  
**Response time**: This is the time difference between submission of a request until the response begins to be received. The response time should be as low as possible so that a large number of interactive users receive an acceptable response time.  
  
**Throughput**: The number of processes that are completed per unit time is called the throughput.  
  
It is desirable to maximize CPU utilization and throughput and to minimize turnaround time and response time.  
  
**An example to further clarify the** **difference between response time and throughput**  
  
Take an example of disk performance, a common measure of disk performance is the response time. This measures the elapsed time from when a command is issued to the disk drive to when data transfer is completed and the disk drive signals completion of the command. Response time is basically a measure of how fast a drive is in servicing a request.  
  
Another measurement of drive performance is the throughput or its capacity to handle work. Throughput is commonly measured in one of the two ways: the number of I/Os per second or the amount of data transferred per second.

Response time and throughput are related. The response time for an average transaction tends to decrease as you increase overall throughput.

However, you can decrease the response time for a specific query, at the expense of overall throughput, by allocating a disproportionate amount of resources to that query. Conversely, you can maintain overall throughput by restricting the resources that the database allocates to a large query.

The trade-off between throughput and response time becomes evident when you try to balance the ongoing need for high transaction throughput with an immediate need to perform a large decision-support query. The more resources that you apply to the query, the fewer you have available to process transactions, and the larger the impact your query can have on transaction throughput. Conversely, the fewer resources you allow the query, the longer the query takes.

The Throughput graph shows the amount of data in bytes that the Vusers received from the server in a second. When we compare this with the transaction response time, we will notice that as throughput decreased, the response time also decreased. Similarly, the peak throughput and highest response time would occur approximately at the same time.