Analyzing the Network Response Time and Load Balancing

Mădalina MLAK, Bucharest, Romania, Madalina.Mlak@ie.ase.ro

This paper describes the network response time which represent a significant indicator for network performance and how load balancing can improve the network performance.
***Keywords**: response time, load balancing, throughput, network performance*

Response time is the total time it takes from when a request will wait to receiving a response.
For network, the number of bits counted divided by the media speed determines the network response time.
A basic advice regarding response time has been about the same for 40 years ago [1].

2. Types of response time

The main categories of response time are the following:

- CPU or processing response time;
In real-time systems the response time of process or thread is the time elapsed between when task is ready to execute and the time when is finishes the job.
In data processing, the response time perceived by the end user is the interval between the instant at which a user at a terminal enters a request for a response from a computer and the instant at which the first character of the response is received at the terminal.

- **disk response time**;

Disk response time is the total time it takes to serve a request on disk.

- **monitor response time**;

Response time for monitors is the amount of time a pixel in a monitor takes to go from active to inactive and back to active again. Long response time creates blur pattern around moving objects.

- server response time;
Server response time is the total time it takes to serve a request on server.
The response time of a server is the sun of the weight of the processes assigned to it.

- **proxy response time**;

Proxy response time is the total time it takes to serve a client request.

- **message response time**;

For communications the response time is the interval between receipt the end of transmission of an inquiry a message and the beginning of transmissions of a response message to the source station.

- **network response time**.

Network response time refers to the amount of time it takes to send a request and wait for a response to come back using a network. Network response time involve bi-directional traffic from source to destination (only in the client-to-server direction).
In the following, is referring network response time.

3. Factors which affect the network response time

The following situations can affect the network response time:

- **a congestion path or limited path can indicate a slower response time**;

If the response time for each successive user will be longer that the previous one, that result that the average response time will grow without limits and the network will have congestion. A solve of this problem are limiting the number of users and this implies that the average response time has a stabile value.

- if the connection is broken, the response time is absent;
A solve of this physic problem is to use ping message to localize the place.
Use ping to find response time from several hosts of the network. If the response time from one host is absent or is not in a good field, you can analyze the host. Physical problems are caused by the following reasons: missing terminations, loose cables, loose connectors, improper grounding or reflection of signal.

- when the network response time become greater and the operating of networks are timing out, the reason may be broadcast storm (burst).
As conscience the user cannot log to servers

or access services. In the end, the network becomes unusable. A solve of this problem is growing of throughput.

- a growing in using of network can determine a poor response time and a sloe network performance.

A solve of this problem is segmentation of network and determine the users' domains.

4. Measuring and calculating response time

Response time measures the total time that system takes to serve a request. This time includes latency. Do not confuse response time with latency.

Latency represent the delay in communicating a message "spend" in the network. Latency is the amount of time takes for a packet to reach its destination after it is transmit and it increases when the network has a bottleneck.

Latency is affected by physical component and utilization of network.

Response time can affect by changes in the processing time of your system and by changes in latency, which occur due the change in hardware resource or utilization.

Response time measurement may be made by:

- **operating system commands**;
- **operating system performance monitor**;
- **applications**.

For example, UNIX command time output the amount of elapsed time (real), the user CPU time and the system CPU time and is good for CPU response time. You can use this command to invoke an instance of your application or perform a data base operation.

In Windows, you can use Performance Monitor to measure user time, processor time and elapsed time.

The output of ping (one per second) gives you the round trip response time (RTT).

Ping doesn't use TCP or UDP to measure response time. It uses ICMP (Internet Control Management Protocol), so the test can lead to inaccurate results. ICMP packets are smaller than TCP or UDP packets and ICMP may be handled differently by connecting devices such as routers, because has a diminished priority from busy routers and can produce unreliable results. Putting ping test near the devices for analyze you can reduce ping overhead and can produce more reliable results. For small network the difference between ping and other TCP applications for test response time may be negligible. I-t is much better to use for measure response time a separate monitoring application to collect highly reliable measurements for analyze.

Note: The response time of a single request isn't representative in all cases for the typical response time of the system. For have a good measure of response time will usually calculate the average response time of many requests.

To measure response time choose a traffic type for which you like to investigate response time. For example: HTTP traffic it will be analyzes response time on the Web.

Response time is usually measured in seconds/request.

To calculate response time use formula:

response time = latency + processing time

On systems where the CPU is not necessarily busy, such as a client system, the benefit of offloading checksum may be seen in better network response time, rather than in noticeably improved throughput.

Parallel database systems and multi-user mod is required to achieve high throughput and for reducing response time.

Throughput and response time are linearly related and the influence is reciprocal of one another. Throughput and response time are inversely proportional.

Throughput is the rate at which a computer or network sends or receives data.

To calculate throughput use formula:

throughput = 1/ response time

Response time is a performance indicator for a network, like the throughput.

If response time increases when the load is increased, the degradation of system performance (implicit network) is greater. A poorly scaling application is indicated by response time that degrades gradually as more users are added.

We can reduce the response time for each user by reducing wait times or increase throughput. This throughput will be main-

tained.
And the driver must periodically adjust the hardware and software to maximize the throughput and response time.

5. Response time for Web applications

End-to-end Web performance is influent by many factors including Web server platform, protocol and network characteristics.
Exists the following alternatives Web serving architectures [2] on client-perceived response time and the information may be place like following:

- single server location;

The classical approach of storing information in a single data center that uses a single network provider.

- single server location with multiple ISPs;

The data center provides direct connectivity to many ISPs to provide performance and availability usually with additional control over BGP routing.

- multiple server locations;

The information is placed in few regional data center to improve performance and availability.

- widely distributed servers.

The information is served from a distributed network which has a set of small servers spread across many service providers.
Each of these steps in communication introducing its own delay in getting a Web page from the serve to the user and those delays are cumulative:

- the throughput of the server;
- server's connection to Internet;
- Internet itself (has bottleneck especially for cross-continent connections and for use at peak hours);
- the user's connection to Internet;
- speed of the user's browser application;
- speed of the user's computer;
- upgrade of the software and for the driver of hardware.

6. Managing network response time

Performance management can manage response time. Network performance is measured in throughput and response time. For performance of system and for the network performance it is desirable that all processes get a minimum response time (waiting time) and a maximum CPU utilization and throughput.
Response time varies with the amount of bandwidth available for applications and for communications.
To improve the traffic it is recommended load balancing and traffic analyze.

7. Load balancing

Load balancing is the process of improving the performance of system through a distribution of loads among the processors.
In network refers to balancing a workload amongst multiple computer devices.
The load balancing analysis is based on load balancing algorithms.
Load balancing and capacity planning improve the network performance and has effect on network response time.
Load balancing options [3] for network connections:

- least numbers of open connections to the server;

At Web sites where are servers with difference between performance, this option performs very well.

- weight to each server with percent;

In this option, servers with a higher activity will receive a larger percentage of connections.

- Round Robin policy;

This option directs the network connection to the next server and treats all servers as equal numbers of connections or response time. Round Robin is superior because not exists propagation delay. When a server is not responding we can determinate that.

- fastest connections to the server;

This option directs the network connections to the server with the fastest response time. Web-server performance does not follow a linear progression of response time to number of connections. In this situation, this option will tend to overload a particular server before moving on to another.

- setting maximum number of connections.

Specifying the maximum number if connections for each server and setting this limit which a server can accept, you can avoid exceeding the capacity limit of the server.

Bibliography

[1] Miller, R.B. – Response time in man-computer conversational transaction – Vol. 33 Proc. AFIPS Fall Join Computer Conference, 1968

[2] IBM research

[3] www.cisco.com - Load Balancing Options

[4] Kai Shen, Tao Yang, and Lingkun Chu - Cluster Load Balancing for Fine-grain Network Services - Department of Computer Science - University of California at Santa Barbara

[5] Erhard Rahm, Robert Marek - Analysis of Dynamic Load Balancing Strategies for Parallel Shared Nothing Database Systems - University of Kaiserslautem , Germany

[6] Sameer Bataineh, Jamal Al-Karaki - Closed form solution for response time of fault tolerant network of processors - Computers & Electrical Engineering, Volume 30, Issue 4 , June 2004, Pages 291-308

[7] Robin Schumacher - Response-Time Analysis Made Easy in Oracle Database 10*g* (www.oracle.com)