

Why Latency Matters

How Bandwidth and Latency Affect Application Performance

Edwin Hoffman CDO

OVERVIEW

Much has been written about network bandwidth and the ever-increasing need for more of it. Bandwidth is essentially the measurement of the capacity of a network connection and is typically expressed in terms of bits per second. However *bandwidth*, while important, is only one component of the perceived performance of a network. The other essential ingredient is *latency*.

LATENCY VS. BANDWIDTH

Latency is a measure of the time delay in processing network traffic. It is the total time for a network packet to travel from the application on one server, through the network adapter, over the wire, through the second adapter and into an application on another server. Latency is affected by distance, operating system and protocol overhead, the number and characteristics of the devices the data must pass through (including network adapters, switches, and so on), and additional load or congestion on the network.

Figure 1 shows the simplest one-way representation of a latency measurement, although round-trip measurements can be considered as well.

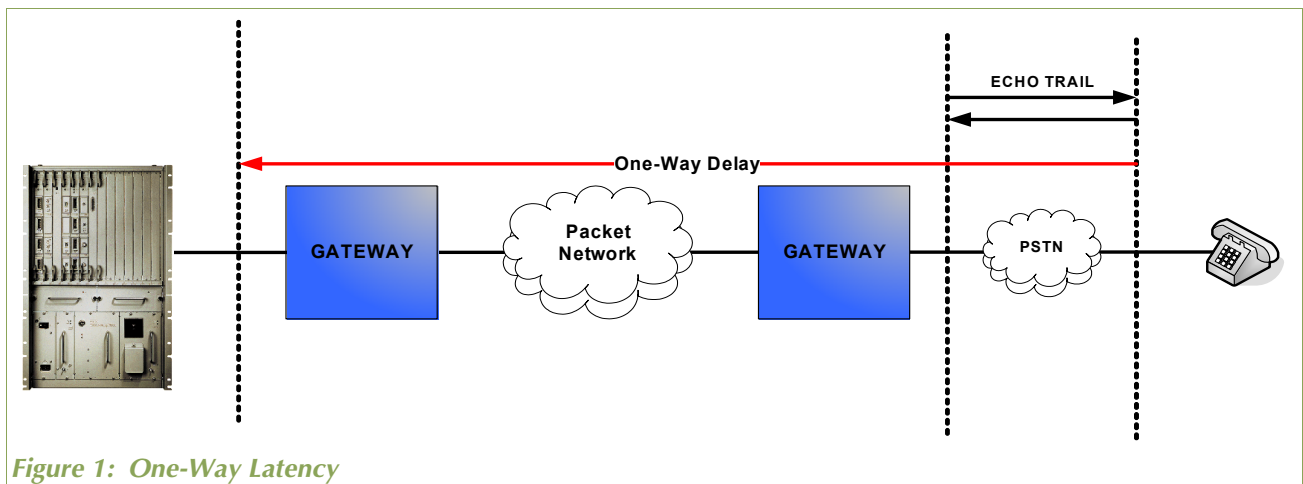


Figure 1: One-Way Latency

How are bandwidth and latency related? High latencies can negatively impact the effective (versus theoretical) bandwidth that can be achieved. When latencies are high, the protocol overhead can overwhelm the work needed to deliver the data. High latencies can create a bottleneck, which prevents full utilization of the network, thus decreasing network bandwidth.

Latency reduces overall performance by limiting how fast an application can get the data it needs, and limits the overall size and scalability of a cluster by limiting the number of messages that can be effectively put on the wire.

For example, a VoIP connection will suffer degradation if the overall latency exceeds 150 milliseconds (round-trip time). Most of VoIP latency is caused by the codec (audio encoder/decoder), and packetization at each end of the connection. Typically, as little as 100 microseconds can be left for the LAN/WAN connection to use in delivery of VoIP data packets.

Latency variation (Jitter) will cause the “effective latency” of a conversation to increase, so systems with significant jitter also cause problems. The amount of bandwidth can reduce congestion-induced latency, but this simply delays the inevitable need to seek a low-latency network solution.

OTHER LATENCY-SENSITIVE CONSIDERATIONS

A further reason latency is an important consideration is that bandwidth can always be increased by adding more channels and running these in parallel. Nevertheless, the latency of a device is a true limiting factor. Even if you need to transfer only a very tiny piece of information such as a processing lock or acknowledgement, there is a minimum time (the latency) that must never be exceeded.

The effective impact of high latency on short-message traffic is much greater than on larger data transfers because the latency can represent a much higher percentage of the total transmission time. At a typical 1-GbE connection with 100-microsecond latency, only large messages greater than 12.5 kilobytes approach wire speed. But, if the latency drops to 10 microseconds, then messages as small as 125 bytes can move at wire speed, thus increasing bandwidth utilization and CPU efficiency.

Clustered databases are one example of particularly latency-sensitive applications. They generate large numbers of short messages, which are typically sent for synchronization between nodes in the cluster. Hence, high latencies place a limitation on the effective number of nodes in a cluster. Cluster scalability is also related to the time it takes to access storage devices. Databases must have exclusive access to data (called locking), and the longer it takes to acquire locks from other nodes (due to high latency), the worse the performance and hence the scalability.

Another example of a latency-sensitive application is Online Transaction Processing (OLTP), where lower latency means more transactions can be processed.

REDUCING LATENCY

Low-latency (10-microsecond) interconnect solutions have been available for years in proprietary implementations from various vendors, and more recently from others implementing InfiniBand and alternative architectures. They dramatically improve cluster performance, but at the cost of requiring installation of a completely new and different network fabric.

More recently, with the completion of standard specifications for Remote Direct Memory Access (RDMA) over TCP/IP, new standard network adapters are becoming available, offering low latency using standard Ethernet networks. Whereas, much has been made of Ethernet Host Bus Adaptors (HBA) with less than 10-microsecond latencies, the Ethernet network must be able to allow data to flow with low enough latency to sustain the 10-microsecond HBA.

Recently, Foundry® Networks claimed the crown for the lowest latency multi-layer switching of less than 10 microseconds, which would seem to enable RDMA transport over TCP/IP, but two factors still exist. A practical latency for the Foundry *Mucho Grande* switch from Gigabit port to Gigabit port across the backplane turned out to be as high as 120 microseconds, or 12 times the HBA speed. In addition, TCP is a notoriously slow protocol that has been dropped in favor of UDP by most of the networking world that needs low latency and high performance. Examples include Storage/IP, VoIP using SIP, Video/IP all using User Datagram Protocol (UDP) as the transport protocol. The true network problem is that UDP is poorly served in layer 2/3/4 switches and can add latency as traffic levels rise.

The Ether-Raptor switches of Raptor Networks Technology are tested at 4.58 microseconds for Gigabit-to-Gigabit transport, and the switches can maintain these latencies on a hop-to-hop basis over long distances with the only latency addition caused by the physics related to the speed of light. For example, a Raptor Adaptive Switch Technology (RAST™) link over 10 km generates a total one-way latency of less than 10 microseconds, or the same delay as the HBA.

Up until now, the need to create clusters with disaster recovery (avoidance) has been prevented by the latency induced by intersite communications. Ether-Raptor and RAST have removed that barrier.

SUMMARY

Although bandwidth will continue to play a major role in optimizing network performance, practical network design demonstrates that latency matters as much as, if not more than, bandwidth. As distributed n-tiered architectures,

Corporate Headquarters: 1241 E. Dyer Road, Suite 150 Santa Ana, CA 92705

Phone: 949-623-9300 / Fax: 949-623-9400 / Web: www.raptor-networks.com / E-mail: info@raptor-networks.com

Raptor Networks Technology, Inc. reserves the right to make changes without further notice to any products or data herein to improve reliability, function, or design. Information furnished by Raptor Networks Technology, Inc. is believed to be accurate and reliable. However, Raptor Networks Technology, Inc. does not assume any liability arising out of the application or use of this information, nor the application or use of any product or circuit described herein, neither does it convey any license under its patent rights nor the rights of others.

Raptor Networks Technology, Inc. is a registered trademark and RAST is a trademark of Raptor Networks Technology, Inc. All other trademarks are the property of their respective owners.

CD-WP1000 06/14/05