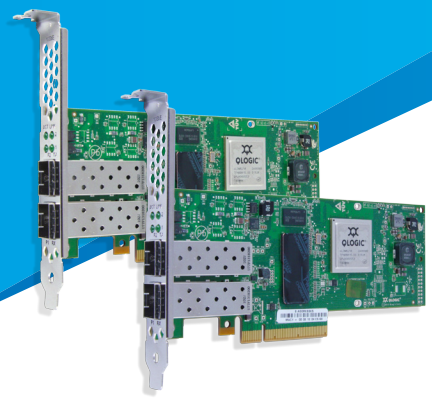
Introduction to Ethernet Latency

An Explanation of Latency and Latency Measurement



The primary difference in the various methods of latency measurement is the point in the software stack at which the latency is measured. The method that is appropriate for a given measurement is dependent on the component of overall latency that is in question.

INTRODUCTION

10Gigabit Ethernet (10GbE) is rapidly becoming the backbone of enterprise networks. The need for 10GbE is being driven by the deployment of servers with multi-core processors and applications with demanding requirements, such as virtualization, storage, backups, database clusters, and video-on-demand applications. 10GbE delivers increased performance, reliability, and ease of use in enterprise data centers, including virtualized environments. For networks, throughput is often the performance metric of choice.

However, there are many applications for which transaction latency is equally important. InfiniBand® and other proprietary I/O technologies are being used for High Performance Computing and clustering to deliver low-latency, system-level application performance. This technical brief defines Ethernet latency, describes test methods used to quantify latency, and identifies applications that can benefit from low latency.

MEASURING LATENCY

Ethernet latency can be measured with different tools and methods, such as specified by IEEE® RFC2544, netperf, or Ping Pong. Different methods measure latency at different points through the system software stack. Latency in a packet-switched network is stated as either one-way latency or round-trip time (RTT). One-way latency is the time required to send a packet from the source to the destination or the RTT divided by two (RTT/2)—the one-way latency from the source to the destination plus the one-way latency from the destination back to the source divided by two. RRT/2 is most often quoted, because it can be measured from a single clock. The technique for measuring round-trip latency through network devices and across networks:

- Uses a minimum of two network-connected systems
- Configures the systems to run collaborative software
- Sends packets between the systems
- Ensures the receiving system's collaborative software returns the packets to the sender
- Enables the sending system's measurement software to measure the time from when the packet was sent to the time it was returned

WHAT IS LATENCY?

The general definition of latency is the delay between a stimulus and a response. Latency in the context of networking is the time expended by propagation through the network medium and the adapter hardware, as well as the execution times of the software (OS and application). It impacts the time the application must wait for data to arrive at its destination.

Hardware latency contributors include the following:

- Traversing the network medium
- Traversing network switches or other network devices
- Propagation through the adapter silicon
- Propagation through the PCIe® bus
- Memory access times (both read and write)
- Lag between interrupt and ISR execution

Software latency contributors for networking include the following:

- Firmware running in an adapter
- The device driver controlling the adapter
- Operating system execution
- The portion of the network stack through which the data must flow
- The portion of the test application through which the data must flow

- One-way latency is measured by dividing the round-trip time in half
  - To accurately calculate one-way latency by dividing the round-trip time in half, the test systems' configuration must be perfectly symmetrical. This means that the systems must be configured identically with the same operating system (OS), OS settings, chip set architecture, PCIe slots used, and memory architecture
  - One-way latency is measured in this way due to the difficulty in synchronizing the system clocks. To calculate one-way latency between the systems, both the sending and the receiving system clocks would need to have exactly the same time. This would require the systems to exchange messages to synchronize their clock times, which is extremely difficult to achieve because the latency between the systems will skew the synchronizing messages by the transit time.

IEEE specification RFC2544 provides an industry-accepted method of measuring latency of store and forward devices. Store and forward devices receive an entire frame before evaluating the frame. They are the only devices considered for this paper.

In addition to testing with the RFC2544 method, latency can also be measured using benchmarks, such as netperf or Intel MPI Benchmarks' Ping Pong. All of these methods use the technique outlined above. However, they use different types of packets and different means of returning the packets. Netperf and Ping Pong are not covered by RFC2544.

Netperf can test latency with Request/Response tests (TCP_RR and UDP_RR). The Request/Response tests run at the application level of the network stack. This method of latency testing involves all of the layers of the stack. The stack's execution time contributes to the overall latency measurement. (See Figure 1.)

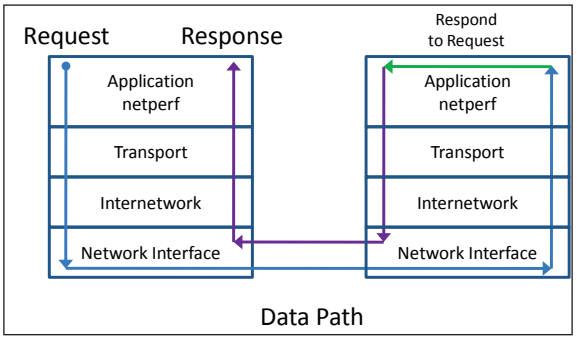


Figure 1. Netperf - Round-trip Data Path

- TCP Request/Response (TCP_RR) – A pair of connection-oriented sockets are established between a netperf client and server
- The netperf client and server exchange transactions as rapidly as possible during the test (a transaction being defined as the exchange of a single request and a single response)
- The round-trip and one-way trip latency times are inferred from the reported transaction rate:
  Transaction rate = transactions per second
  Round-trip latency (RTT) = 1 / transaction rate
- If the systems used in the test are symmetrical, meaning they are running the same software, using the same settings, and they have equal network and system performance, then it is feasible to calculate an accurate one-way latency:
  One-way latency (1/2 RTT) = (1 / transaction rate) / 2
- UDP Request/Response (UDP_RR) works the same as TCP_RR, with the exception that it uses connection-less sockets instead of connection-oriented sockets

Ping Pong is a method for measuring latency in a High Performance Computing cluster. Ping Pong is a test method which is included in the Pallas MPI Benchmarks (PMB) suite, now called Intel MPI Benchmarks (IMB) suite. (See Figure 2.)

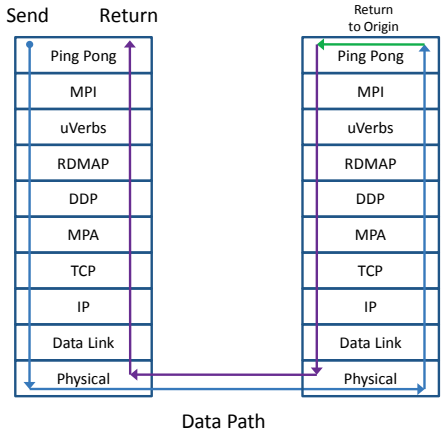


Figure 2. Ping Pong - Data Path

- Ping Pong tests even farther up the software stack than netperf
- Ping Pong tests latency from the perspective of a High Performance Computing application running over the Message Passing Interface (MPI) in the High Performance Computing stack
- Latency testing using Ping Pong measures the round-trip time of Remote Procedure Calls (RPCs) sent through the MPI
- The test divides the round-trip time by two to calculate the one-way latency of data from entry into the adapter to delivery to the application running on MPI

The RFC2544 latency test is the method for testing latency specified by IEEE RFC2544. While the specification doesn't explicitly require the use of IP forwarding, the test is universally conducted using IP forwarding. The RFC2544 latency test measures latency at layer 3 of the network stack, using IP forwarding to route packets from an input port to an output port. (See Figure 3.)

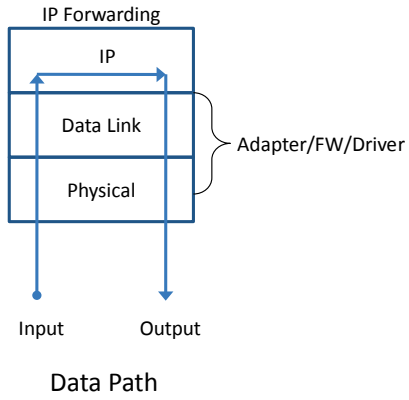


Figure 3. IP Forwarding Data Path Used with RFC2544 Latency Test

- Of the three methods for measuring latency (RFC2544, netperf, or Ping Pong), testing using RFC2544 with IP forwarding tests the least amount of software above the adapter and provides a view of the true capability of the adapter and its associated firmware

THE IMPACT OF LATENCY

Latency affects all network applications to some degree. The degree to which latency affects an application's performance depends on the application's programming model. Latency impacts an application's performance by forcing the application to stall while waiting for the arrival of a packet before it can continue to the next step in its processing.

Applications with programming models that are susceptible to performance degradation due to latency include the following:

- Applications that depend on the frequent delivery of one-at-a-time transactions, as opposed to the transfer of large quantities of data
- Applications that track or process real-time data, such as "low latency" applications

RFC2544 LATENCY TESTING

The RFC2544 specification states that the latency test's objective is to test latency as defined in RFC1242.

The RFC1242 latency definition for store and forward devices is:

- The time interval starting when the last bit of the input frame reaches the input port and ending when the first bit of the output frame is seen on the output port

RFC2544 stipulates that the latency test:

- Should be at least 120 seconds in duration
- Frame sizes to be used on Ethernet 64, 128, 256, 512, 1024, 1280, and 1518
- Should include an identifying tag in one frame after 60 seconds with the type of tag being implementation dependent
- Records the time at which the frame is fully transmitted (timestamp A)
- The receiver logic in the test equipment must recognize the tag information in the frame stream and record the time at which the tagged frame was received (timestamp B)
- This test should be performed with the test frame addressed to the same destination as the rest of the data stream, and also with each of the test frames addressed to a new destination network
- The test must be repeated at least 20 times with the reported value being the average of the recorded values
- The latency is timestamp B minus timestamp A, as per the relevant definition from RFC1242; namely, latency as defined for store and forward devices

Configuration and Data Path of the IEEE RFC2544 Latency Test

The diagram below (Figure 4) depicts an appropriate configuration for performing latency testing in accordance with RFC2544. The RFC2544 latency test requires that the DUT be configured to enable IP forwarding. In this configuration, the DUT acts like a router and forwards packets from one LAN to another through layer 3. The TX port of the tester is connected to the RX port on the DUT with a pair of IP addresses configured for LAN 0. The RX port of the tester is connected to the TX port of the DUT and this pair of IP addresses are configured for LAN 1. The DUT uses IP forwarding to route packets from its RX port on LAN 0 to its TX port on LAN 1.

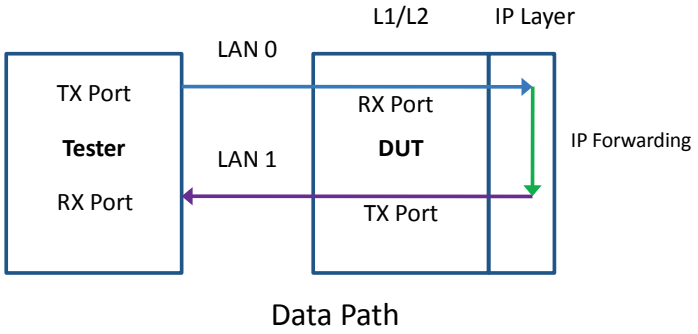


Figure 4. RFC2544 Latency Testing via IP Forwarding

The test system's implementation of RFC2544 latency testing uses a latency testing script. The test measures latency in conformance with RFC2544 for store and forward devices.

The RFC2544 test system's latency measurement script follows the RFC2544 specification to measure the time delta from the time when the last bit of the frame left the tester's transmit port to the time that the first bit of that frame arrived on the tester's receive port.

Latency Test Overhead Constants

The round-trip time measured in the test is actually more than the time it takes to transit the DUT's NIC and driver. The round-trip time also includes the propagation time of the network cables and the processing time for layer 3 to forward the packets from the RX port to the TX port. These times are presumably small in comparison to the overall round-trip time, and for a given test environment are also presumably constant. However, if tests are conducted in differing environments, then the skew introduced by the variation of cable length, CPU speed, interrupt type, or OS could affect the validity of relative measurements.

APPLICATION MAP FOR TECHNOLOGY

There are new technologies that are pushing latencies into the single-digit microsecond range when measured back-to-back in benchmark environments. Real-world applications and deployment environments requiring end-to-end data transfer can impact the overall latency when compared to specific benchmarks. End-to-end deployments may have significantly larger influence on overall latency and can exceed the adapter's latency.

Table 1. Applications and Associated Latency Requirements

Latency Range (µs)	Technology	Application
50 – 125	1Gb Ethernet (TCP/IP)	• Multi-tasking: multiple high-bandwidth applications running simultaneously • Bulk data transfer • Transactional database backup and applications • Web (front-end for data centers)
5 – 50	10Gb Ethernet (TCP/IP)	• Bulk data transfer • Real-time video streaming • Database backup and applications
3 – 5	RDMA, RoCEE, and iWARP	• High Performance Computing • High-Frequency Trading (HFT) • Inter-process communication (IPC) cluster • Low-latency applications
Sub-3	InfiniBand (QDR) and proprietary	• High Performance Computing • High Frequency Trading (HFT) • Ultra-low latency applications

CONCLUSION

Excessive latency limits the performance of network applications by delaying packet arrival. Network latency can be measured in several ways. The primary difference in the various methods of latency measurement is the point in the software stack at which the latency is measured. The method that is appropriate for a given measurement is dependent on the component of overall latency that is in question.

- **Latency measurements using netperf:** Measures latency of data moving between an application layer peer (netperf client) to another application layer peer (netperf server) on a remote system.

- **Latency measurements using Ping Pong:** Measures latency between peers existing above MPI that exchange RPCs.
- **Latency measurements using RFC2544 with IP forwarding**: Measures the latency contribution of the network hardware, firmware, driver, and IP layer.

When evaluating relative latency measurements, it is vital to understand the method used for the measurement and the environment in which the measurement was taken.

ABOUT CAVIUM

Cavium™, Inc. (NASDAQ: CAVM), offers a broad portfolio of infrastructure solutions for compute, security, storage, switching, connectivity and baseband processing. Cavium's highly integrated multi-core SoC products deliver software compatible solutions across low to high performance points enabling secure and intelligent functionality in Enterprise, Data Center and Service Provider Equipment. Cavium processors and solutions are supported by an extensive ecosystem of operating systems, tools, application stacks, hardware reference designs and other products. Cavium is headquartered in San Jose, CA with design centers in California, Massachusetts, India, Israel, China and Taiwan.

Follow us:

Corporate Headquarters Cavium, Inc. 2315 N. First Street San Jose, CA 95131 408-943-7100

International Offices UK | Ireland | Germany | France | India | Japan | China | Hong Kong | Singapore | Taiwan | Israel