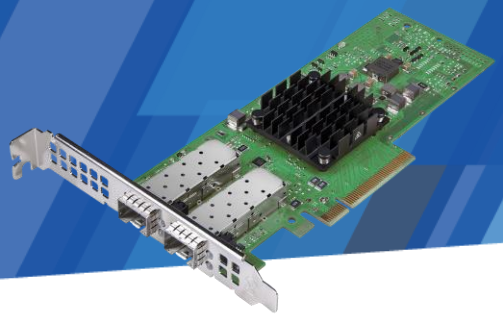
Demartek

Evaluation of Broadcom NetXtreme 25Gb Ethernet Adapters

Broadcom demonstrates the industry's most deterministic latency – the critical element to providing consistent, scalable performance for enterprise and cloud environments.



Executive Summary

As customers of all sizes begin to adopt 25Gb Ethernet (25GbE) technology, enterprises and cloud data centers need to consider the scalability and consistency of performance of the Ethernet adapters, also known as network interface cards (NICs).

In particular, cloud-scale data centers, where hundreds to thousands of NICs are deployed, need to consider the range of latencies produced by NICs. Tail latency matters more than minimum or average latency at large scale, because the tail latency experienced by several NICs at any given moment can cause unexpected and disruptive application delays, slowing down the overall network performance. Predictable and low-tail latency is critical to high-performance cloud-scale data centers.

Broadcom commissioned Demartek to analyze the performance of the Broadcom® NetXtreme® 25GbE NIC in terms of performance consistency, comparing these results to the equivalent Mellanox® ConnectX®-4 25GbE NIC.

Key Findings

> Broadcom's 25GbE adapter has very small latency variability, while the competitive adapter has very high long-tail latency.

> Broadcom's 25GbE adapter shows not only lower average latency but considerably lower and more consistent maximum latency.

> Broadcom's 25GbE adapter provides higher lossless frame rates than the competitive adapter.

> In RoCEv2 tests, Broadcom's 25GbE adapter shows higher port throughput and more consistent latency than the competitive adapter under network congestion.

Evaluation of Broadcom NetXtreme 25Gb Ethernet Adapters

Demartek
January 2018

Latency and Throughput Metrics in Real-World Environments

In this paper, Demartek measures the Broadcom NetXtreme NIC's performance and compares it against the competition in cloud computing environments. Cloud computing environments are characterized by their use of commodity hardware where tasks are distributed to a large number of servers to achieve higher performance.

In these cloud environments, the maximum latency (often referred to as the "long-tail latency") is particularly important because application performance depends on the worst-case response time, where the last server to respond determines the application performance. So, if Server X takes five seconds to reply, the application performance must wait for Server X, even if all other servers have replied. This dynamic is highlighted in the paper *The Tail at Scale* by Dean and Barroso from Google[1].

With this in mind, latency and throughput are measured, two of the most common performance metrics.

However, for latency, most traditional benchmark tests measure performance under little to no traffic. This is unrealistic and generally not helpful in measuring real-world performance.

In the testing performed, we focused on the following:

- Realistic network traffic scenarios
  - Latency measured with loaded traffic
  - For RoCEv2 tests, congestion was injected to mimic a realistic network scenario
- Latency that matters
  - Tail latency
  - Latency distribution

[1] https://research.google.com/pubs/pub40801.html

Evaluation of Broadcom NetXtreme 25Gb Ethernet Adapters

Demartek
January 2018

RFC2544 – L2 Performance Tests

RFC2544 tests are commonly used to evaluate Ethernet adapters for throughput and latency in transmit and receive tests. For these tests, dual-port 25GbE adapters were tested using Ixia test equipment and the Ixia IxNetwork tools. IPv4 traffic was transmitted into one port of the adapter and returned via the second port.

Throughput tests determine the maximum rate at which no offered frames are dropped.

Packet sizes are 64, 128, 256, 512, 1024, 1280 and 1518 bytes.

Ixia Test Equipment

- XGS2 Chassis
- Novus-R100GE8Q28 + FAN + 10G + 25G + 40G + 50G blade

Server

- 2x Intel Xeon E5-2690 v4, 2.6 GHz 14/28t
- 64 GB RAM
- RHEL 7.3, Kernel: 3.10.0-514.el7.x86_64

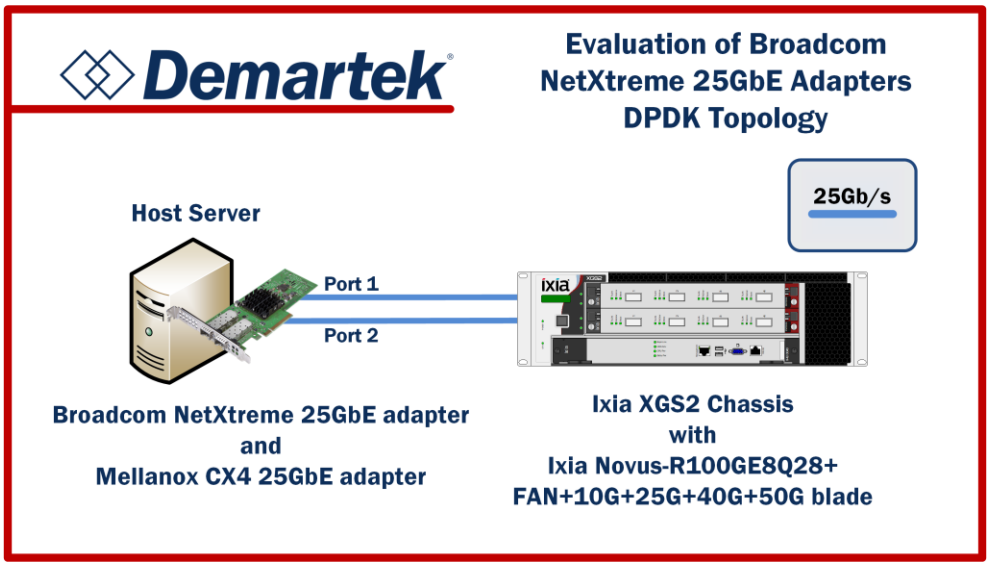
Broadcom NetXtreme 25GbE Adapter

- Broadcom P225p (BCM957414A4142CC)
- DPDK 17.08
- Software:
  - bnxt_en: 1.8.29
  - bnxt_re: 20.8.0.16
  - rocelib: 20.8.0.7
  - firmware: 20.8.61.0

Mellanox CX-4 25Gbe Adapter

- Mellanox CX4-L (MCX4121A-ACAT)
- DPDK MLNX_DPDK_16.11_2.3
- OFED: 4.1.1
- Software:
  - driver: mlx5_core
  - OFED version: 4.1-1.0.0.x
  - firmware: 14.20.1010

Mapping for the Mellanox adapter was performed using procedures described at http://www.mellanox.com/page/products_dyn?product_family=209&mtag=pmd_for_dpdk



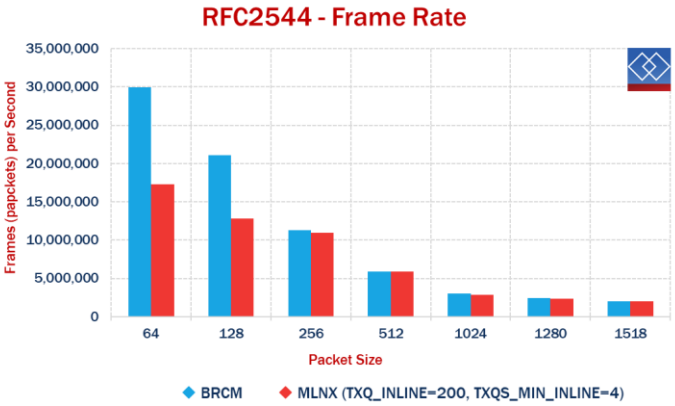
Evaluation of Broadcom NetXtreme 25Gb Ethernet Adapters

Performance Results – RFC2544 (L2 Tests)

These tests handle packet routing between the test hardware and the adapter installed in the host server, focusing on small packets and utilization the adapter, driver and IP portion of the TCP/IP stack.

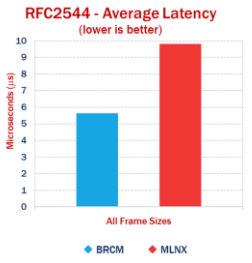
Lossless Frame Rate

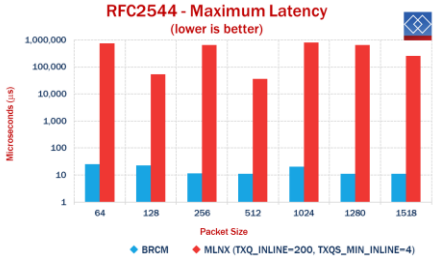
In RFC2544 testing, Ixia starts at full line rate and reduces the offered load until no packets drop. Broadcom's adapter performs better than the competition for smaller frames.



Average and Maximum Latency

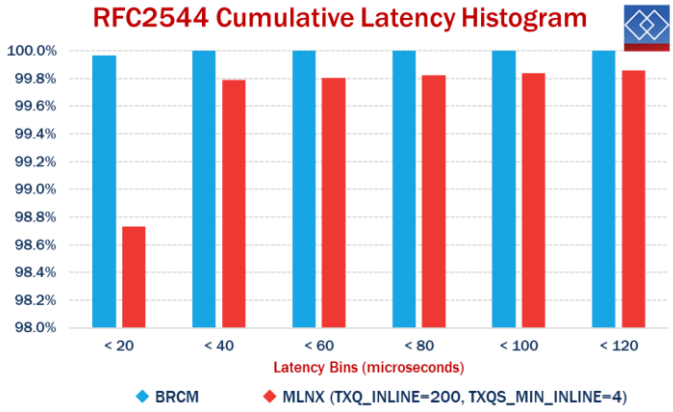
The Broadcom adapter shows lower average latency and considerably lower and more consistent maximum latency. The competition's maximum latency spikes up to a few 100 milliseconds.

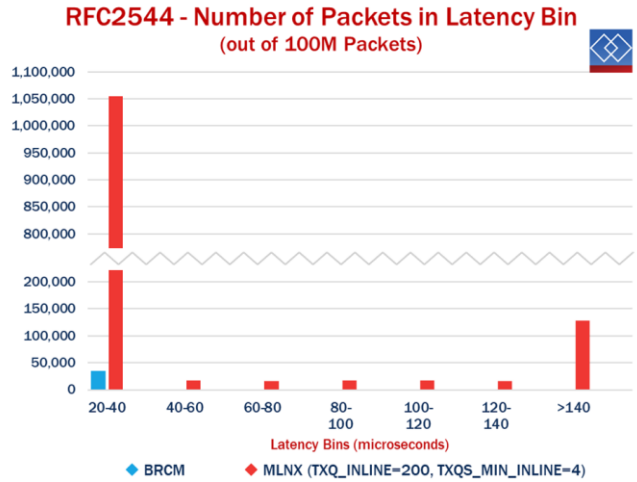




Long-tail Latency

The Broadcom adapter shows only 35K packets (out of 100M – approximately three seconds of line rate operation) outside 20 µsec and all packets within 40 µsec, while the competition shows 1.3M packets outside 10 µsec and 145K packets still not received after 120 µsec.





Evaluation of Broadcom NetXtreme 25Gb Ethernet Adapters

Demartek
January 2018

RoCEv2 Performance Tests

For the next two tests, two clients (servers #1 and #2) generated RoCEv2 network traffic targeted at server #3, creating network congestion with 2:1 incast flows. Throughput and latency were measured.

These RoCE features were enabled:

- Priority Flow Control (PFC)
- Explicit Congestion Notification (ECN)

The workload was generated by Perftest-3.4-09. Message sizes of 1K, 2K, 4K, 8K, 16K, 64K, 128K, 256K and 512K bytes were sent using QP levels ranging from 5 to 100.

Servers (3x)

- 1x Intel Xeon E5-2667 v4, 3.2 GHz 8c/16t
- 64 GB RAM
- RHEL 7.3, Kernel: 3.10.0-514.el7.x86_64

Broadcom NetXtreme 25GbE Adapter (3x)

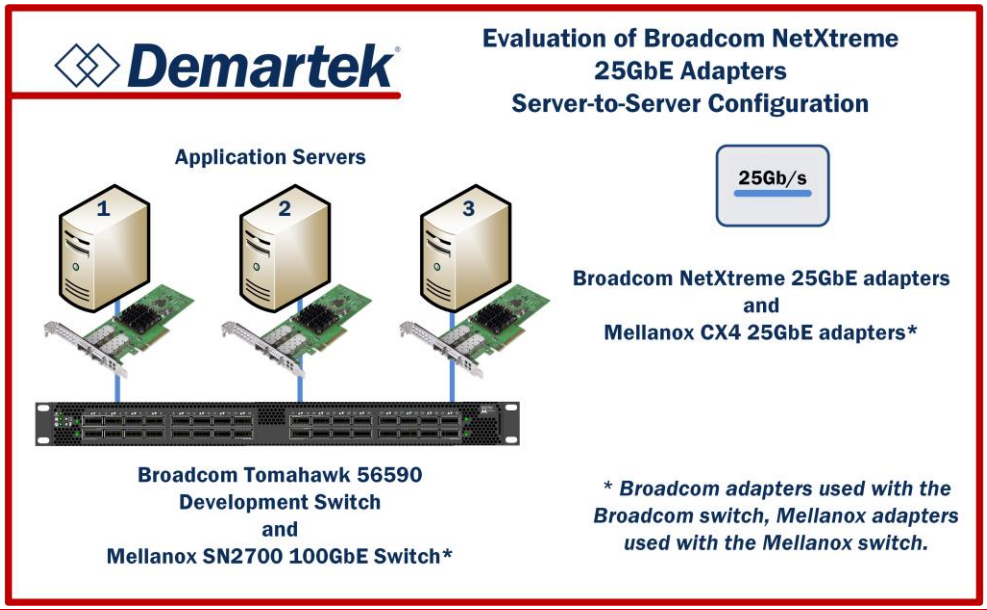
- Broadcom P225p (BCM957414A4142CC)
- OFED: 4.8 GA
- Software:
  - bnxt_en: 1.8.29
  - bnxt_re: 20.8.0.16
  - rocelib: 20.8.0.7
  - firmware: 20.8.61.0
  - roce firmware: 20.8.29.0

Mellanox CX-4 25Gbe Adapter (3x)

- Mellanox CX4-L (MCX4121A-ACAT)
- OFED: 4.1.1
- Software:
  - driver: 4.1-1.0.2
  - firmware: 14.20.1010

Ethernet Switch (1x)

- Mellanox SN2700-1, MNLNX-OS 3.6.4112
- Broadcom Tomahawk 56960 Development system, 14x40G + 14x100G + 4x100G (32 QSFP), 7.8.20.18, Linux 3.10.59-6795c564



Evaluation of Broadcom NetXtreme 25Gb Ethernet Adapters

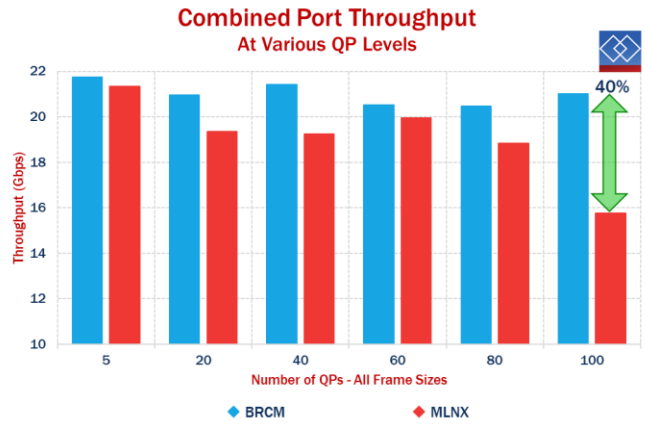
Demartek
January 2018

Performance Results – RoCEv2 Tests

These tests measure RoCE efficiency for two servers simultaneously communicating with a third server.

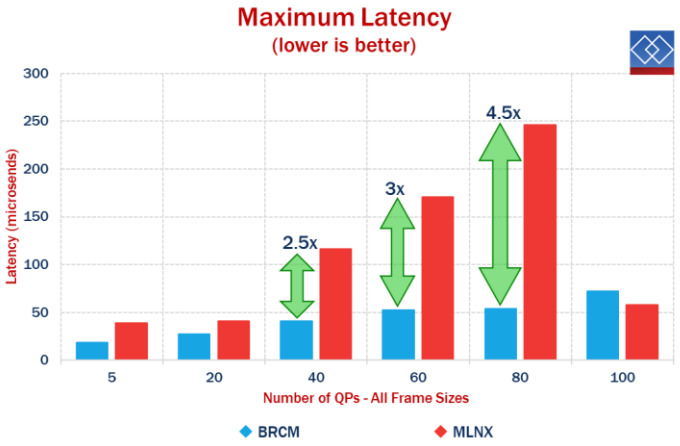
Port Throughput

The Broadcom NetXtreme 25GbE adapter performs with higher and more consistent combined-port throughput for all QP levels. At high QP levels, the Broadcom adapter sustains 40% higher (21Gb vs 15Gb) throughput than the competition.



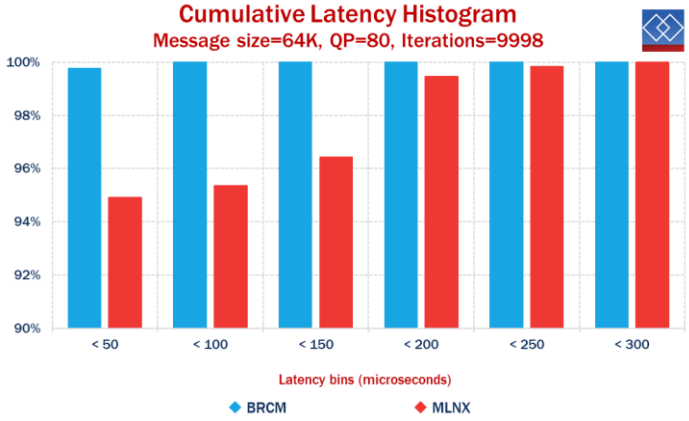
Maximum Latency

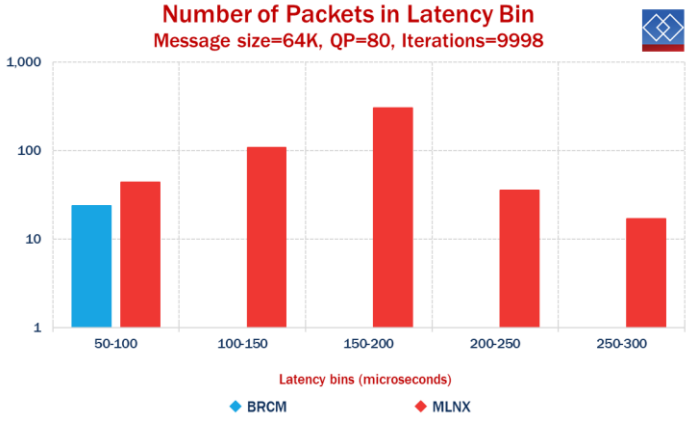
The Broadcom adapter demonstrates considerably lower and more consistent maximum latency than the competition. For some QP levels, the maximum latency of the competitive adapter spikes over 300 µsec.



Long-tail Latency

The Broadcom adapter showed only 24 (out of 9998) packets outside 50 µsec and all packets within 100 µsec, while the competition showed 509 packets outside 50 µsec and 17 packets still not received after 250 µsec.





Evaluation of Broadcom NetXtreme 25Gb Ethernet Adapters

Demartek
January 2018

Summary and Conclusion

For large-scale installations, it is important to have consistent performance across all of the adapters so that applications provide predictable performance across hundreds or thousands of nodes. This consistency applies to frames per second, throughput and latency.

With a large number of network adapters deployed in an enterprise or cloud data center, the probability that one or more of these adapters is experiencing its maximum latency at any given moment is high. Wildly fluctuating latencies can lead to, at best, uncertainty for application performance and, at worst, application failures.

In the tests conducted for this report, the Broadcom NetXtreme 25GbE network adapter provided a much tighter set of latency results from minimum to maximum latency than the competitive adapter. Moreover, Broadcom's maximum latency under heavy load was up to five times lower than the competitive adapters.

Large-scale cloud environments would do well to use the Broadcom NetXtreme 25GbE adapter for their deployments.

The most current version of this report is available at http://www.demartek.com/Demartek_Broadcom_NetXtreme_25G_Adapters_2018-01.html on the Demartek website.