networktest

A Whole Lot of Ports:

Juniper Networks QFabric

System Assessment

March 2012

Executive Summary

Juniper Networks commissioned Network Test to assess the performance, interoperability, and usability of its **QFabric System**, a converged switch fabric for cloud and large data center applications tested with **1,536 10-Gbit/s Ethernet ports.**

Even at this unprecedented scale – by far the largest ever in a public switch test – this project loaded the QFabric System to only one-quarter of its maximum capacity of **6,144 10-Gbit/s Ethernet ports.**

Using industry-standard RFC benchmarks representing the most rigorous possible test cases, engineers stress-tested QFabric System performance in terms of unicast and multicast throughput and latency with separate events for Layer 2 and Layer 3 traffic. Engineers also assessed interoperability, a key consideration when adding QFabric technology incrementally into existing data center networks, and evaluated device management.

Key results from QFabric System testing include the following:

- ✓ All QFabric System components operated as a single device, simplifying network management and reducing operational complexity
- ✓ Throughput for Layer 2 traffic was virtually identical in store-and-forward and cut-through modes, with rates approaching the maximum channel capacity for most frame sizes
- ✓ QFabric forwarding delay is low and consistent across all tests (less than 5 microseconds for all frames sizes up to 512 bytes) when offered loads below the throughput rate
- ✓ Average latency for Layer 3 traffic is 10 microseconds or less for most frame sizes tested
- ✓ Multicast throughput was close to line rate in all tests, regardless of frame length, with the system moving traffic at speeds of up to 15.3 terabits per second
- ✓ Multicast average latency was low and consistent in all tests, never exceeding 4 microseconds
- ✓ The QFabric System successfully interoperated with Cisco Nexus 7010 and Cisco Catalyst 6506-E switch/routers when using common data center protocols such as link aggregation, OSPF equal cost multipath, and BGP

The remainder of this document discusses the test results in more detail. Besides presenting the test results, each section describes the test objective and procedure, as well as its meaning for network architects and network managers.

The QFabric System Test Bed

Figure 1 shows the test bed used in this project, encompassing 1,536 10-Gbit/s Ethernet edge ports; 128 redundant 40-Gbit/s fabric uplinks; and an out-of-band gigabit Ethernet management network. **From a network management perspective, all the various QFabric System components operated as a single device.** Engineers used one Junos configuration file to define all interfaces and protocols. The configuration syntax is identical to Junos on other Juniper platforms. **Even in very large cloud and data**

center applications, the QFabric System allows the entire data center network infrastructure to be managed as a single entity.

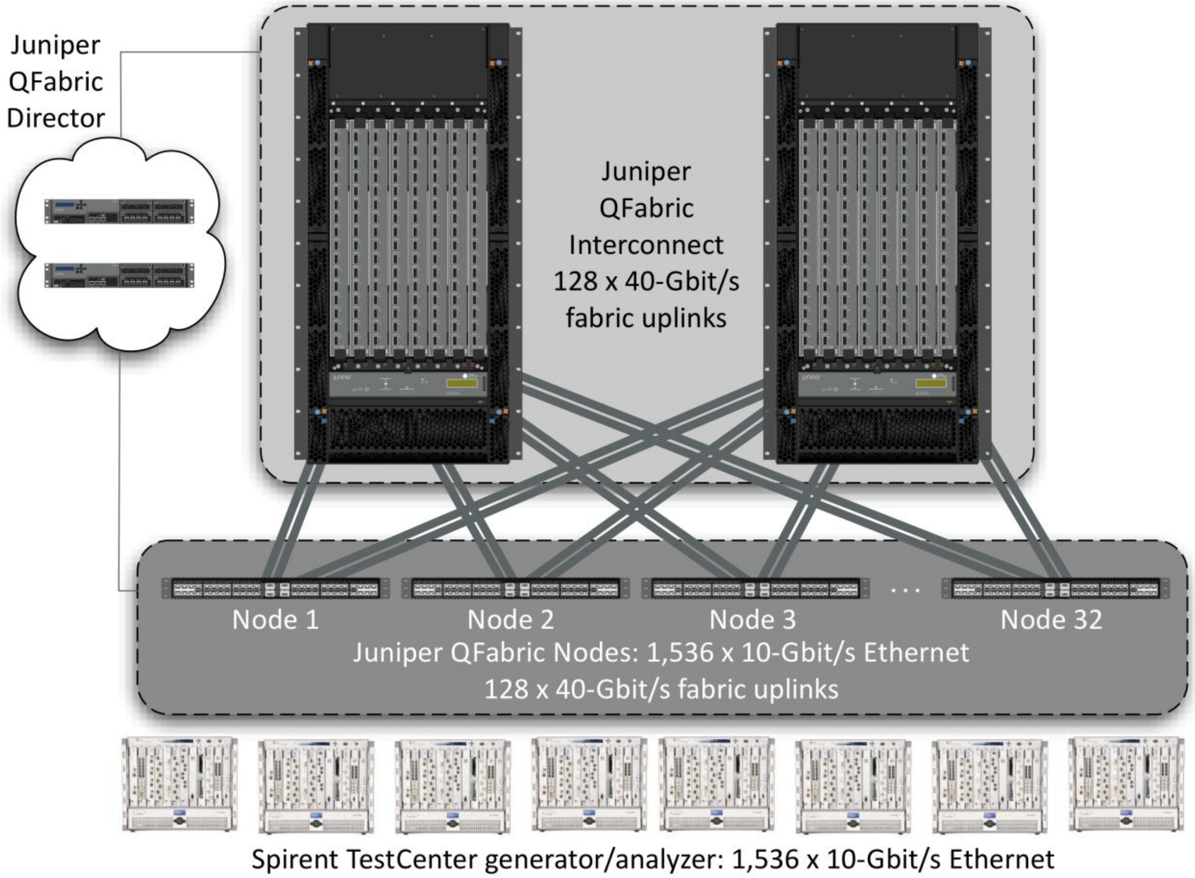

Juniper QFabric Director
Juniper QFabric Interconnect 128 x 40-Gbit/s fabric uplinks
Node 1
Node 2
Node 3
...
Node 32
Juniper QFabric Nodes: 1,536 x 10-Gbit/s Ethernet
128 x 40-Gbit/s fabric uplinks
Spirent TestCenter generator/analyzer: 1,536 x 10-Gbit/s Ethernet

Figure 1: The QFabric System Test Bed

The test bed illustrates how the QFabric System takes the key pieces of a conventional modular switch and separates them into individual physical components, while still managing the entire system as one entity. The QFabric System's three components are:

- the QFabric Director (QFX3100), which performs control-plane functions analogous to the Routing Engine module in Juniper switches and routers
- the QFabric Interconnect (QFX3008), which ties together switch ports in the same way as a modular switch's backplane
- the QFabric Nodes (QFX3500), which are analogous to line cards in a modular switch

In this test, Juniper used 32 QFabric Nodes, each with 48 10-Gbit/s Ethernet ports and four 40-Gbit/s fabric uplink ports.

Building a 1,536-port test bed is a massive undertaking. The test bed fully occupied four standard 42U racks – two apiece for the Juniper QFabric System components and the Spirent TestCenter traffic generator/analyzers, which used Spirent's 32-port HyperMetrics dX modules. **Because of QFabric's modular design, the QFabric Nodes could have been physically dispersed – for example, with one in each of 32 racks scattered throughout a data center.**

The QFabric System used standard structured cabling, with 3-meter direct-attached copper (DAC) cables between Spirent test ports and the Juniper QFabric System and MPO multimode OM3 fiber cables between the QFabric Nodes and the QFabric Interconnect devices. The out-of-band management network used copper gigabit Ethernet ports, tied together with a pair of Juniper EX4200 switches using Virtual Chassis technology (not shown in figure).

As large as this test bed was, it is nowhere near the limit of QFabric scalability. A pair of QFX3008 Interconnects can accommodate up to 128 top-of-rack switches, for a total of 6,144 10-Gbit/s Ethernet ports in a single system.

The high-level goals for this project were to assess QFabric System performance, interoperability, and usability. Performance tests covered throughput and delay for unicast and multicast traffic, using the Spirent test tool to run RFC benchmarks across 1,536 10-Gbit/s Ethernet ports. Interoperability tests sought to verify that the QFabric System would work with switches/routers from Cisco Systems running common data center protocols such as link aggregation, OSPF equal cost multipath, and BGP. There were no separate usability tests, but engineers verified during all other tests that all QFabric System components were capable of being managed as a single system, reducing operational complexity.

Unicast Throughput

The primary goal of the unicast throughput tests was to determine how fast the QFabric System moved traffic with zero frame loss.

QFabric allows network architects to build blocking or nonblocking switch fabrics. In this case, the overall system was 3:1 oversubscribed for unicast traffic (with 48 10-Gbit/s Ethernet ports on each QFabric Node [480 Gbit/s of capacity in each direction] attached to the interconnect nodes via four 40-Gbit/s fabric uplink ports [160 Gbit/s of capacity in each direction]).

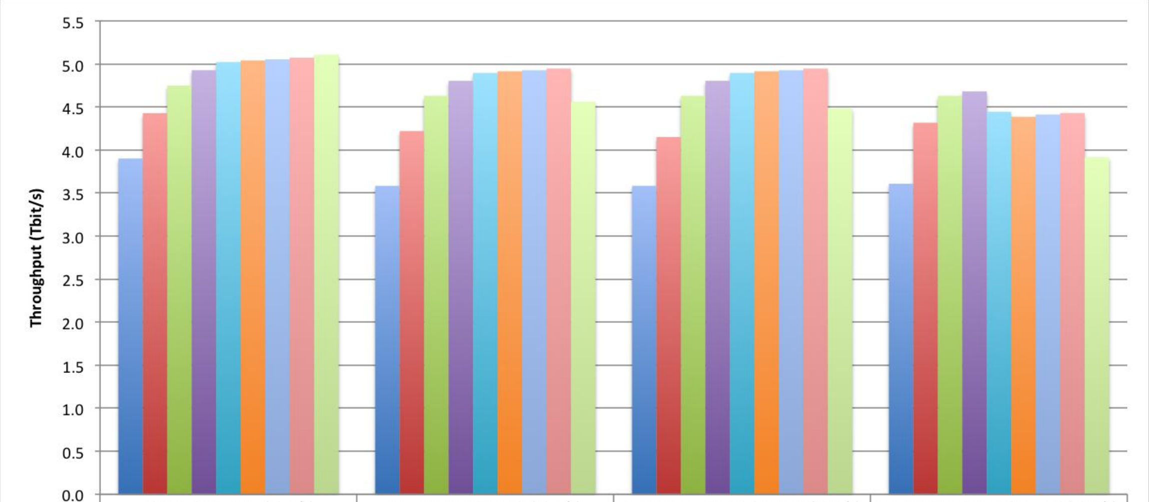
As defined in RFCs 1242 and 2544, throughput is the maximum rate at which a system can forward traffic with zero loss. With a 3:1 oversubscription, this meant the QFabric System's channel capacity was essentially one-third that of the QFabric Nodes. (This document uses the term "channel capacity" to describe the maximum transmission rate a system will support; RFC 4689 also refers to the same concept as "forwarding capacity.") Engineers repeated these tests with both Layer 2 Ethernet and Layer 3 IP traffic, in all cases using a fully meshed traffic pattern, the most stressful possible test case.

With a fully meshed pattern, the Spirent test instruments attached to all ports offered traffic destined to all other ports. This is far more stressful on the switching fabric than port-pair tests and better describes the limits of system performance. Network architects and managers have a reasonable expectation of being able to send traffic between any arbitrary sets of ports in a system, not just selected port pairs. With results from a fully meshed traffic test, users can be confident they are truly seeing the limits of system performance.

In the Layer 2 tests, engineers configured the Spirent TestCenter traffic generator to offer raw Ethernet frames with pseudorandom MAC addresses as described in RFC 4814. These frames had no IP or other upper-layer headers, forcing the QFabric System's hashing algorithms to make path selection decisions based solely on destination and source MAC addresses.

This Ethernet-only configuration is meaningful for data centers with large, flat Layer 2 domains, such as those carrying non-IP storage traffic. Other common use cases for large, flat Layer 2 domains in the data center include virtualization, where large broadcast domains ensure seamless migration of virtual machines between physical hosts; and converged data centers, where data and non-IP storage traffic shares a single high-speed fabric for transport.

The QFabric System has the ability to forward traffic in both store-and-forward and cut-through modes, with the latter sometimes used in data centers with latency-sensitive applications. Test engineers conducted separate throughput tests in each mode. **Throughput for Layer 2 traffic was virtually identical in store-and-forward and cut-through modes, with rates approaching channel capacity for most frame sizes.** Rates for 64- and 9,216-byte jumbo frames were slightly lower than the channel capacity rate. Figure 2 presents the results of throughput testing.



	Channel capacity (Tbit/s)	Layer-2 cut-through (Tbit/s)	Layer-2 store-and-forward (Tbit/s)	Layer-3 store-and-forward (Tbit/s)
64	3.901	3.584	3.584	3.608
128	4.428	4.220	4.151	4.317
256	4.749	4.630	4.630	4.630
512	4.928	4.804	4.804	4.681
1,024	5.022	4.896	4.896	4.444
1,280	5.041	4.915	4.915	4.386
1,518	5.053	4.927	4.927	4.412
2,176	5.073	4.947	4.947	4.430
9,216	5.109	4.561	4.483	3.918

Figure 2: Unicast Throughput

Engineers then moved on to throughput tests involving IP traffic. Here, engineers configured each edge port on the QFabric Nodes to use a different IP subnet, with IP networks assigned in sequential order (e.g., port 1 used 11.1.1.0/24, port 2 used 11.1.2.0/24, and so on). Test traffic contained IP and UDP headers, with the latter containing random source and destination port numbers. In the Layer 3 tests, the QFabric System's hashing algorithms used a combination of IP and UDP data to select interconnect forwarding paths. In these Layer 3 tests, engineers configured the QFabric Nodes to use their default store-and-forward mode.

When handling IP traffic, the QFabric System exhibited higher throughput with smaller (64- and 128-byte) frames and equivalent throughput for medium-sized 256-byte frames compared with Layer 2 results.

Layer 3 throughput rates for frame lengths of 512 bytes and above were lower than those in the Layer 2 tests. During these tests, engineers observed some unevenness in traffic distribution on the links between QFabric Nodes and QFabric Interconnect devices.

Given different traffic content, IP throughput may well change. More randomness in IP addressing might lead to higher Layer 3 throughput; conversely, less randomness in UDP port numbers might reduce throughput. Defining a one-size-fits-all model of IP test traffic was definitely not a goal of this project. The traffic parameters described here represent *one* model of network traffic; other models may result in different throughput rates.

Unicast Forwarding Delay and Latency

Delay is a critical consideration in the data center. With traffic increasingly moving between servers a few meters apart rather than across the global Internet, every microsecond can have an impact on application performance. For some applications such as video, voice, and high-frequency trading, delay is even more important than throughput.

In the context of network device benchmarking, the term *latency* differs somewhat from its colloquial definition. As discussed in RFC 2544, latency describes delay at, and only at, the throughput rate. As such, it's really a description of buffering capacity. RFC 4689 introduces a complementary metric, *forwarding delay,* that describes delay at any load, not just the throughput rate.

Network Test used both metrics in this project – latency at the throughput rate, and forwarding delay for loads below the throughput rate. Because no production network operates at 100 percent utilization for any significant duration, forwarding delay is useful in understanding delays under common operating conditions. Latency is important in describing system limits, in this extreme case when all 1,536 ports are loaded to the maximum zero-drop rate.

To get a more complete picture of the QFabric System's load vs. delay curve, Network Test conducted measurements of forwarding delay by offering fully meshed traffic at 5, 10, 15, and 20 percent of theoretical line rate, as well as at the throughput rate. (Remember that with the 3:1 oversubscription of the QFabric Nodes, channel capacity is equivalent to 33.3 percent of the theoretical maximum.)

Figure 3 presents measurements from the Layer 2 store-and-forward tests. **QFabric forwarding delay was much lower in step tests than delays observed at the throughput rate.** Forwarding delay is less than 5 microseconds for frame sizes up to 512 bytes with loads of up to 20 percent of theoretical line rate. At the throughput rate – the maximum stress point where all 1,536 ports are fully loaded and switch buffers are filled to capacity – latency is still less than 10 microseconds for 64- and 128-byte frames, and never exceeds 40 microseconds for any frame length.

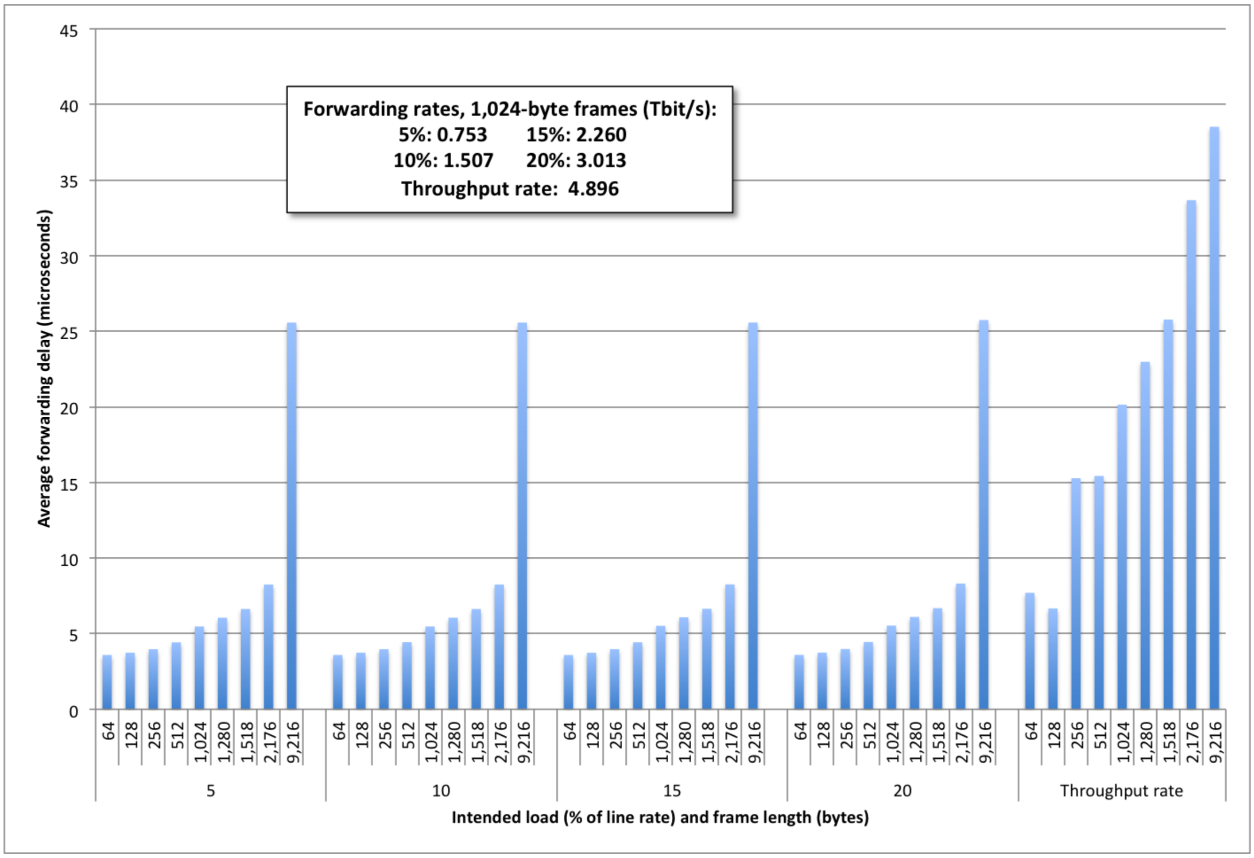


Figure 3: Forwarding Delay in Layer 2 Store-and-Forward Mode

To get a sense of how heavy a load the QFabric System handled, the inset box in Figure 3 shows throughput rates for 1,024-byte frames, the midpoint of all sizes tested. **Forwarding delay was essentially identical for fully meshed, 1536-port loads of up to 3.0 Tbit/s, and increased only at the throughput rate approaching 5.0 Tbit/s.**

Network Test also measured forwarding delay and latency with the QFabric System configured in Layer 3 store-and-forward mode. **As in the Layer 2 tests, forwarding delay was less than 5 microseconds for**

frame sizes up to 512 bytes at intended loads of up to 20 percent of line rate, and delays were consistent across various intended loads. Figure 4 presents the Layer 3 forwarding delay measurements with the QFabric System configured in store-and-forward mode.

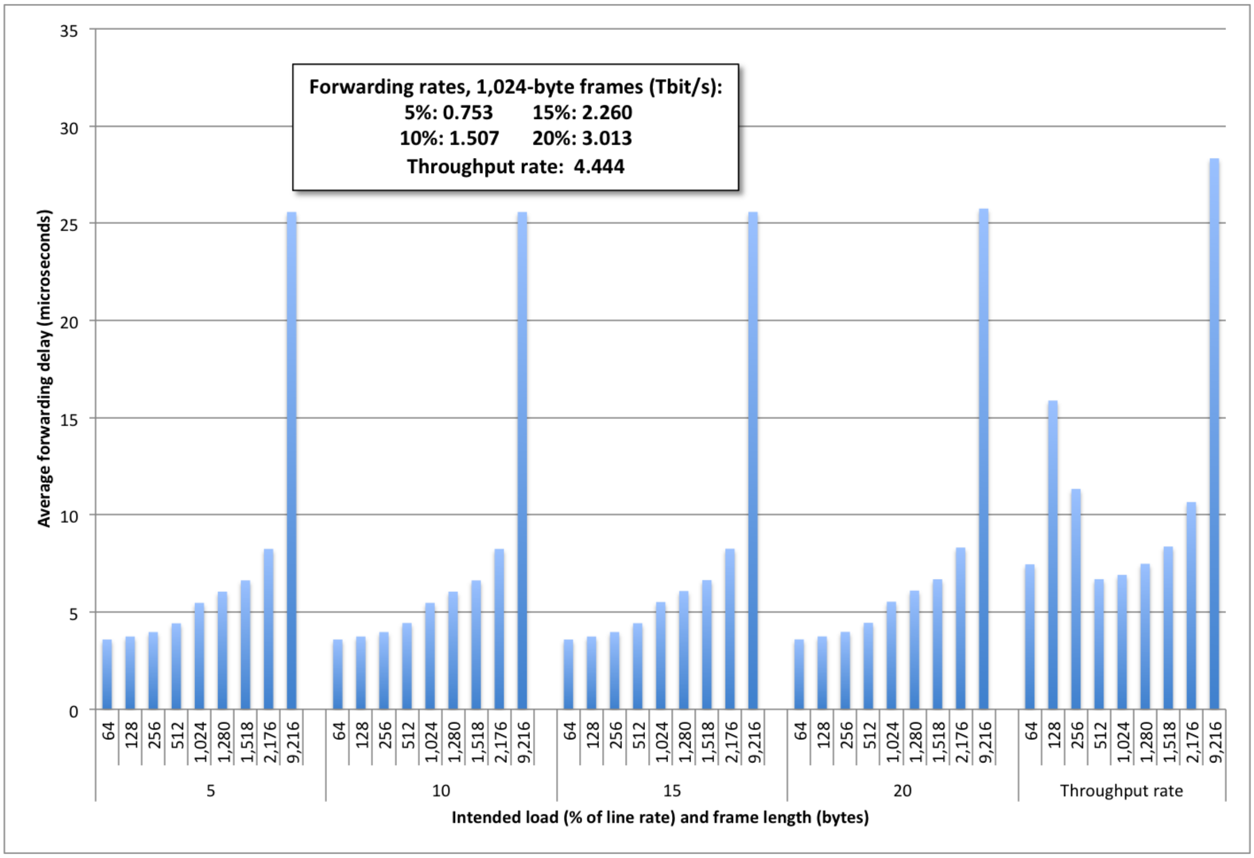


Figure 4: Forwarding Delay in Layer 3 Store-and-Forward Mode

The inset box in Figure 4 presents the rates used to obtain these measurements. **Here again, forwarding delay was essentially identical for fully meshed, 1536-port loads of up to 3.0 Tbit/s, and increased only at the throughput rate of more than 4.4 Tbit/s.**

As noted, Network Test also measured latency with the QFabric System configured in cut-through mode. RFC 1242 requires that cut-through latency use a different measurement method than with store-and-forward devices.[1] It's improper to do direct comparisons between the two, so the results for cut-through latency are presented separately here, in Figure 5**.**

[1] With cut-through devices, RFC 1242 requires measurement using a first-in, first-out (FIFO) method; in contrast, the RFC requires a last-in, first out (LIFO) method when measuring the latency of store-and-forward devices. Measurements from these two methods will differ by at least the time it takes to put each frame on the wire (sometimes known as *serialization delay* or *frame insertion time*).

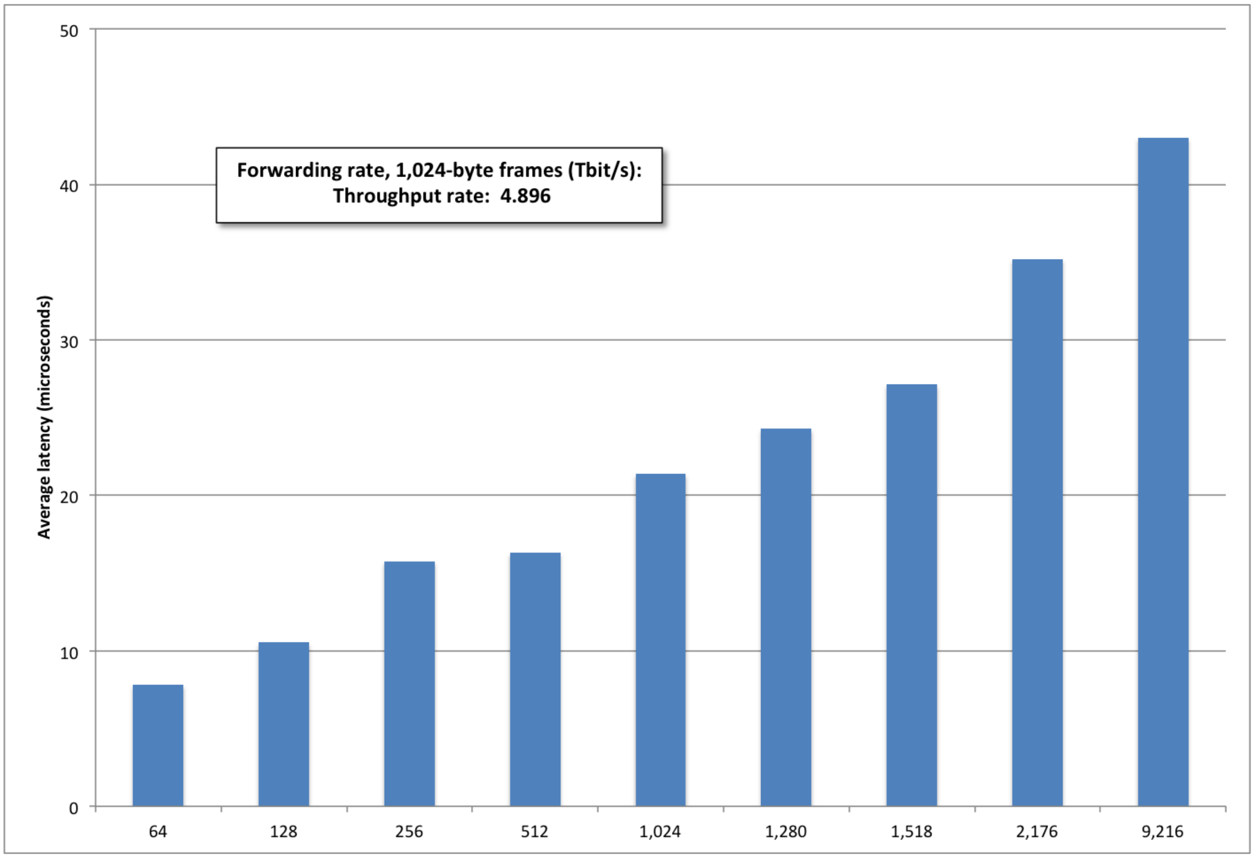


Figure 5: Layer 2 Cut-Through Latency

In the cut-through tests, average latency at the throughput rate appears slightly higher than in the store-and-forward tests, but this is due mainly to differences in the measurement methods involved. Network Test was unable to complete load-vs.-delay measurements in Layer 2 cut-through mode, as in the store-and-forward tests, so results shown here are only for the throughput (maximum zero-drop) rate. The inset box shows the load used in these tests for 1,024-byte frames, the midpoint of frame sizes tested.

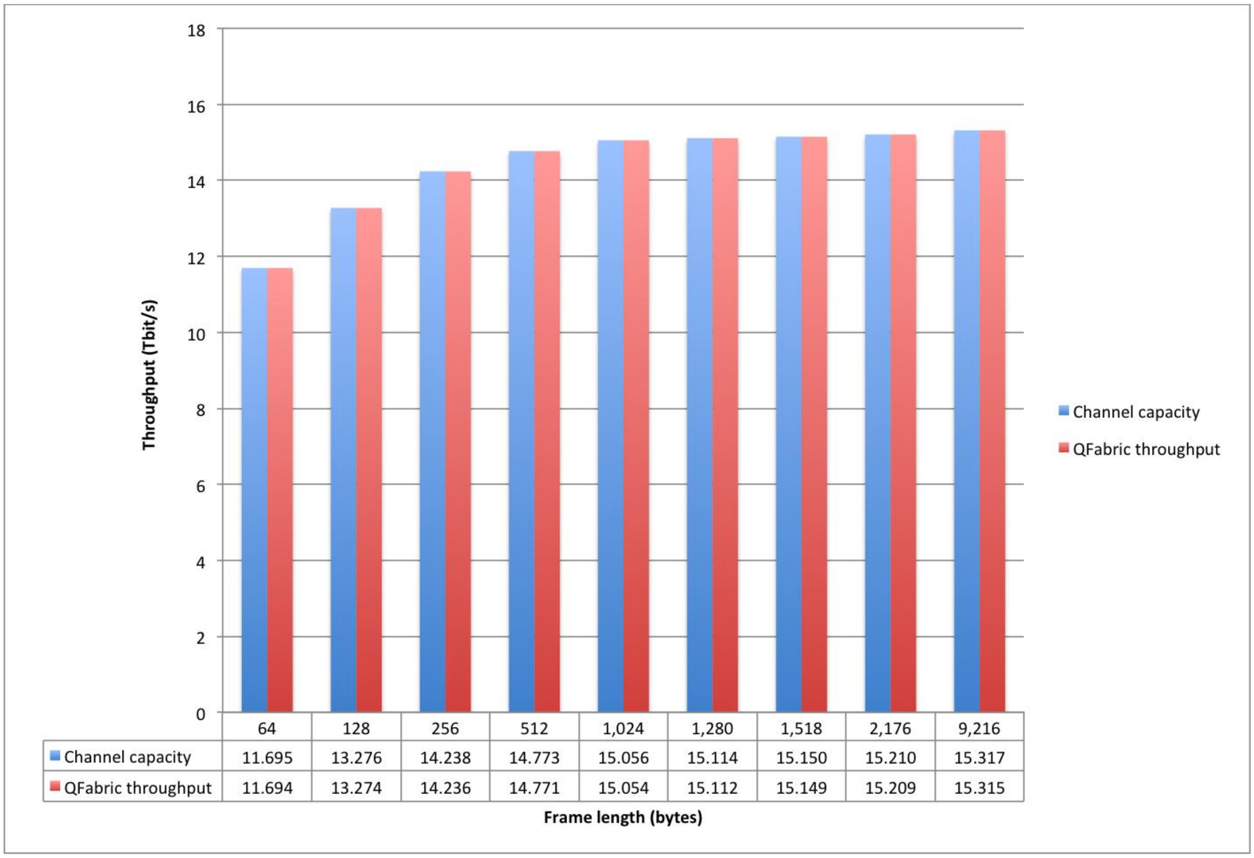
Layer 2 Multicast Performance

With the highest throughput and lowest latency of all tests in this project, IP multicast traffic really showed off the potential of the QFabric System. That's important for high-bandwidth applications that use multicast, such as streaming media, telepresence, and videoconferencing. **The multicast performance results also validate Juniper's claim that the QFabric Interconnect component is nonblocking, no matter how many edge ports are involved.**

To assess multicast performance, Network Test used the standard throughput and latency benchmarks defined in RFC 3918. As with the unicast tests, these benchmarks determine the highest forwarding rate with zero packet drops, and then measure latency at that rate. The QFabric Nodes ran in their default store-and-forward mode for these tests.

Engineers forced the maximum amount of multicast replication by configuring the Spirent TestCenter traffic generator to emulate one transmitter and 1,535 receivers, all subscribed to the same 100 multicast groups. The QFabric System used IGMPv2 snooping to keep track of multicast state.

Multicast throughput was close to line rate in all tests, regardless of frame length, with the system moving traffic at speeds of up to 15.3 terabits per second. The minor difference between observed rates and the theoretical maximum is explained by clocking differences between the QFabric Nodes and the Spirent TestCenter traffic generator. To compensate for these clocking differences, engineers reduced the intended load to 99.99 percent of nominal line rate. Figure 7 shows throughput for IP multicast traffic.



	64	128	256	512	1,024	1,280	1,518	2,176	9,216
🟦 Channel capacity	11.695	13.276	14.238	14.773	15.056	15.114	15.150	15.210	15.317
🟥 QFabric throughput	11.694	13.274	14.236	14.771	15.054	15.112	15.149	15.209	15.315

Figure 6: Multicast Throughput

Multicast latency was low and consistent across all frame sizes tested. Low, deterministic latency is essential for video and voice applications, both of which often use IP multicast. **In these tests, multicast average latency never exceeded 4 microseconds, even when forwarding traffic on all ports at the maximum zero-drop rate.** Figure 6 summarizes multicast latency measurements.

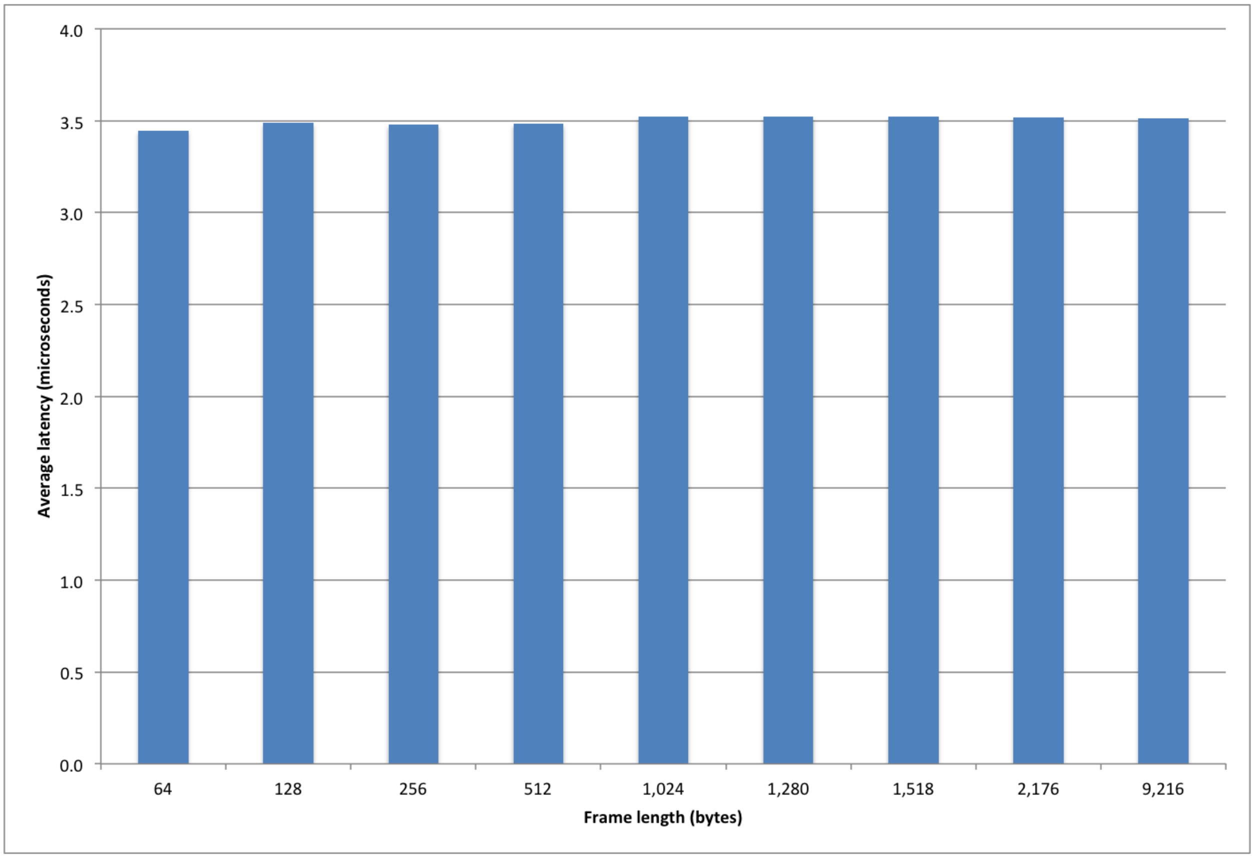


Figure 7: Layer 2 Multicast Latency

Integrating the QFabric System Into Existing Data Centers

Although the QFabric System introduces a new approach to data center network design, it is not a rip-and-replace technology. To verify Juniper's claim of interoperability with existing devices, Network Test used three common protocols to link a QFabric System with other devices. These protocols were IEEE 802.3ad link aggregation; OSPF equal cost multipath (OSPF ECMP); and BGP. For all three protocols, Network Test verified interoperability between the QFabric System and two other switch/routers: A Cisco Nexus 7010 and a Cisco Catalyst 6506-E.

All interoperability tests used the same topology, with two physical links on the Juniper and Cisco devices load-sharing traffic. In the link aggregation case, the two physical links formed one logical link, negotiated using the link aggregation control protocol (LACP). In the OSPF ECMP case, engineers established separate OSPF adjacencies on each pair of physical interfaces, with the same routing metric used for each.

For both link aggregation and OSPF ECMP, the Juniper QFabric System correctly interoperated with both Cisco switches. After every test, test engineers checked the counters on each interface linking the

Juniper and Cisco devices and observed roughly equivalent numbers of frames forwarded on each physical interface.

Besides supporting link aggregation and OSPF ECMP, **Network Test also verified that the Juniper QFabric System can interoperate with other routers using BGP, an important consideration when a QFabric System exchanges routing information about the global Internet.** Once again, these tests were conducted using the Cisco Nexus 7010 and Cisco Catalyst 6506-E.

The BGP tests used both the external and internal variations of BGP (eBGP and iBGP). Here, engineers configured the Spirent TestCenter traffic generator and one interface on a Cisco device to use eBGP, with the Spirent and Cisco devices using different autonomous system numbers (ASNs). The rest of the test bed – including the Juniper QFabric System and the remaining Cisco and Spirent TestCenter interfaces – used iBGP, as would be the case when BGP distributes routing information within a single autonomous system. Test engineers also used link aggregation to bond two interfaces between the Juniper QFabric System and Cisco devices into single logical interfaces. Figure 9 shows the configuration used to verify BGP interoperability.

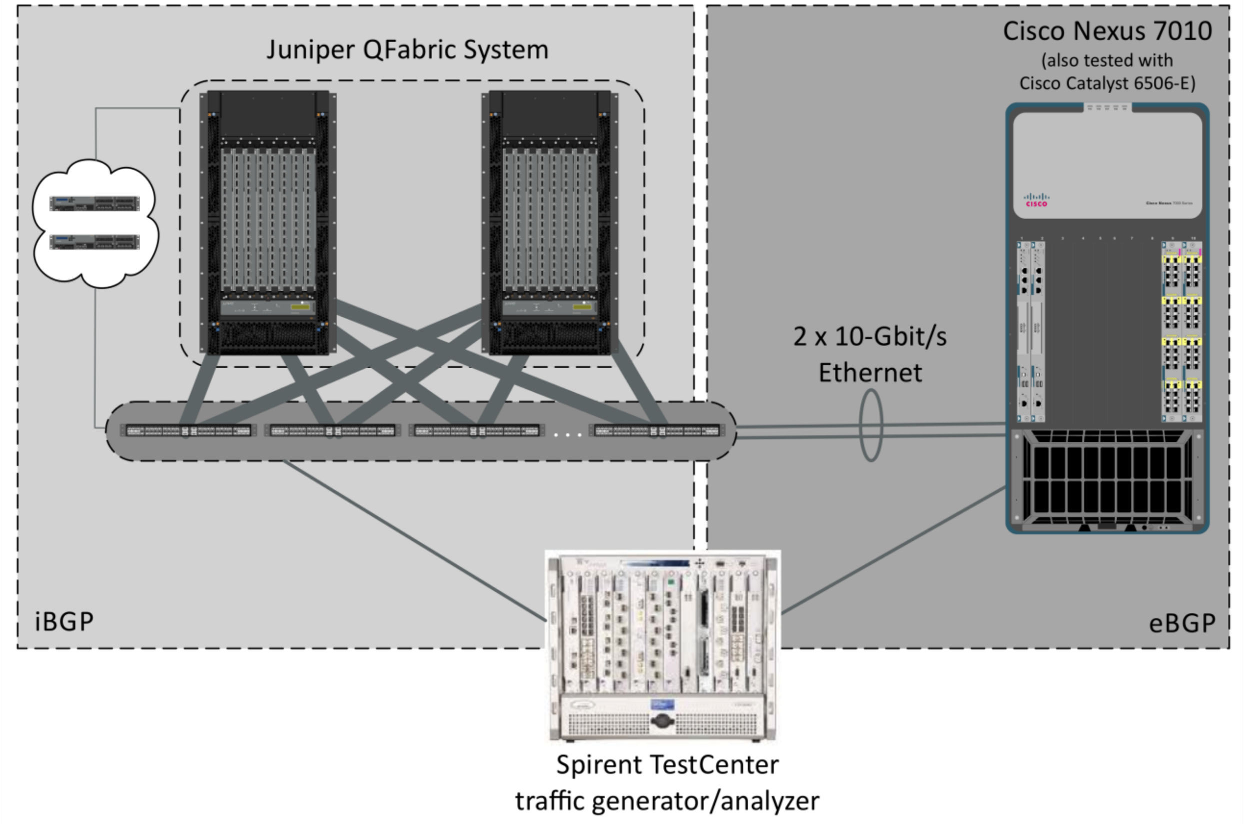


Figure 8: BGP Interoperability Test Bed

As in the previous tests, **the Juniper QFabric System successfully established BGP sessions and forwarded all routing and data traffic.** Test engineers verified BGP interoperability both by observing frame counters on the Spirent TestCenter test instrument and interface counters on the Juniper QFabric and Cisco devices. **The Juniper QFabric System demonstrated BGP interoperability in a test bed running both eBGP and iBGP, an important consideration when data center networks must exchange routing information about the global Internet.**

Conclusion

These tests validated the performance of the Juniper QFabric System on an unprecedented scale. In this largest public switch test ever conducted, engineers built a QFabric test bed comprised of 1,536 10-Gbit/s Ethernet edge ports. **In tests involving a fully meshed pattern of all 1,536 edge ports – the most stressful test case possible – QFabric moved Ethernet frames at rates approaching the channel capacity.** Further, **Layer 2 average latency is 5 microseconds or less for frame sizes up to 512 bytes at intended loads of up to 20 percent of line rate.**

Multicast throughput was higher still, since there's no oversubscription involved. **The multicast tests validated Juniper's claim that its QFabric Interconnect component is nonblocking, in this case moving traffic at more than 15 Tbit/s. Multicast latency also is low and consistent, with average latency never exceeding 4 microseconds.**

A final set of tests examined interoperability between the Juniper QFabric System and other switch/routers using common data center protocols such as link aggregation, OSPF equal cost multipath, and BGP. **Interoperability between the Juniper QFabric System and Cisco Systems equipment was successful in all cases, demonstrating that QFabric technology can be added incrementally** without the need to replace existing data center devices.

Finally, for all the complexity of this test bed – again, the largest used in any public switch test – **the QFabric System was managed as one single device.** The implication for network managers is twofold: First, this means **the QFabric System helps reduce operational complexity by presenting far fewer devices to be managed**. Second, **even in very large cloud and data center applications, the QFabric System allows the entire data center network infrastructure to be managed as a single entity.**

Appendix A: About Network Test

Network Test is an independent third-party test lab and engineering services consultancy. Our core competencies are performance, security, and conformance assessment of networking equipment and live networks. Our clients include equipment manufacturers, large enterprises, service providers, industry consortia, and trade publications.

Appendix B: Hardware and Software Releases Tested

This appendix describes the software versions used on the test bed. All tests were conducted in November-December 2011 at Juniper's headquarters facility in Sunnyvale, CA, USA.

Component	Version
Juniper QFabric System (including QFX3100 QFabric Director, QFX3008 QFabric Interconnect, and QFX3500 QFabric Node top-of-rack switches)	Junos 11.3X30.9 (performance tests); Junos 11.3I20111015_1217 (interoperability tests)
Cisco Nexus 7010	NX-OS 5.2(1)
Cisco Catalyst 6506-E	IOS 12.2(33)SXI4a
Spirent TestCenter	3.90.0293.0000

Appendix C: Disclaimer

Network Test Inc. has made every attempt to ensure that all test procedures were conducted with the utmost precision and accuracy, but acknowledges that errors do occur. Network Test Inc. shall not be held liable for damages which may result for the use of information contained in this document. All trademarks mentioned in this document are property of their respective owners.

Network Test Inc.
31324 Via Colinas, Suite 113
Westlake Village, CA 91362-6761
USA
+1-818-889-0011
http://networktest.com
info@networktest.com