Lippis Report

Cisco Nexus® 9508 Switch Performance Test

A Report on the:

Cisco Nexus® 9508 Switch

October, 2013

Acknowledgements

We thank the following people for their help and support in making possible this firstevaluation of the fastest 40GbE data center Ethernet switch:

We thank the following people for their help and support in making possible this first evaluation of the fastest 40GbE data center Ethernet switch:

Bradley Wong, Distinguished Engineer at Cisco, Lilian Quan, Technical Marketing Engineer at Cisco and Arun Selvarajan Technical Marketing Engineer for their help and engineering support.

Michael Githens, Lab Program Manager at Ixia, for his technical competence, extra effort and dedication to fairness as he executed test week and worked with participating vendors to answer their many questions.

Jim Smith, VP of Marketing at Ixia, for his support and contributions to creating a successful industry event.

Bill Nicholson for his graphic artist skills that make this report look amazing.

Jeannette Tibbetts for her editing that makes this report read as smooth as a test report can.

Steven Cagnetta for his legal review of all agreements.

Cisco Systems Nexus® 9508

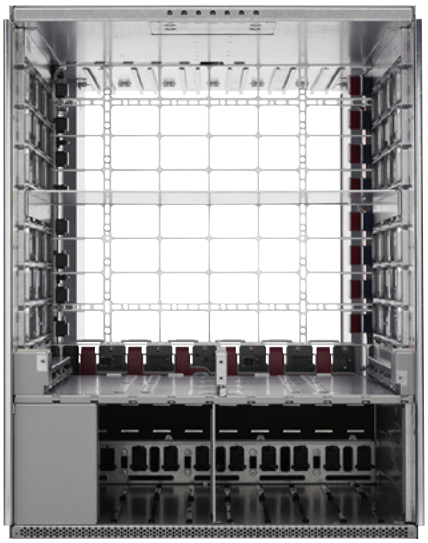
Cisco recently launched its Nexus® 9500 series of data center switches which Cisco promises offers the highest port density of 10/40GbE and future 100GbE—the most power efficient, the most programmable and fastest packet forwarding modular data center switch in the industry. To verify these claims, Cisco engaged the Lippis/Ixia team to test the new Nexus® 9508 modular data center switch.

The Nexus® 9508 boasts impressive density and speed metrics. By leveraging a combination of merchant and custom silicon, the chassis supports an industry-leading 288 non-blocking 40GbE ports in its eight-slot chassis format. Each of the eight slots within the chassis is populated with a 36 40GbE line card, equating to 288 40GbE. In addition, the Nexus® 9508 supports 1,152 non-blocking 10GbE ports with 40GbE-to-10GbE breakout cables.

In addition to its density, the Nexus® 9508 has an efficient architecture, including an eight-slot chassis supporting eight line cards, two supervisor modules located underneath the line card modules for 1:1 redundancy, eight power supply slots (only two are needed to fully power the chassis) and lastly, an innovative design that removes the mid plane. By removing the mid plane, the Nexus® 9508 allows for direct port-to-fabric module connectivity and unobstructed airflow, which contributes to the chassis energy efficiency.

The Nexus® 9508 sports a compelling software architecture that has seen Nexus® OS (NX-OS) become a 64-bit Linux-based system supporting switch programmability, automation, orchestration and visibility. The NX-OS now allows for a level of programmability unseen in the industry; not only can the switch be programmed via CLI commands but now DevOps and NetOps teams can utilize Bash, REST, NETCONF, XML, JSON, onePK, and the OpenFlow protocol to program the switch externally. In terms of automation and orchestration, NX-OS includes XMPP and OpenStack integration as well as agents for popular DevOps tools, such as Puppet and Opscode Chef. Lastly, the NX-OS maximizes its options for switch visibility via features such as vTracker, dynamic buffer monitoring, consistency checkers, embedded event manger and Wireshark. These enhancements in NX-OS signal that Cisco understands the increasing need for administrators to automate network functions.

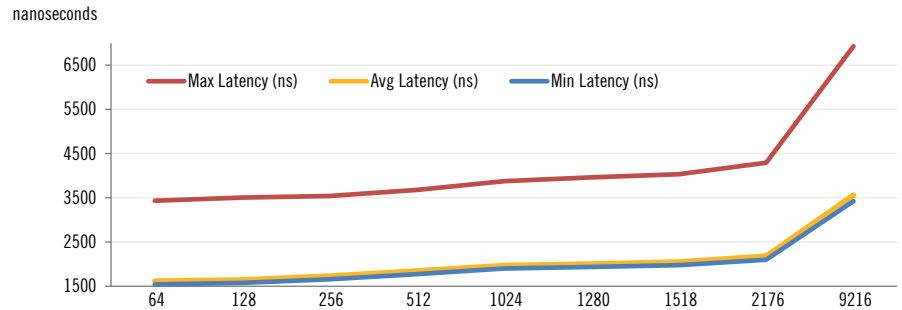
While the new features of the Nexus® 9508 chassis offer countless configuration options to test and implement, we have run exhaustive tests of the chassis core competencies in switching latency, throughput, congestion and IP multicast to ensure it can perform as stated. What we found is that the Nexus® 9508 is the most powerful data center switch engineers have built to date. As of the time of writing, the Nexus® 9508 is the fastest modular data center switch at a scale that no others have tested.



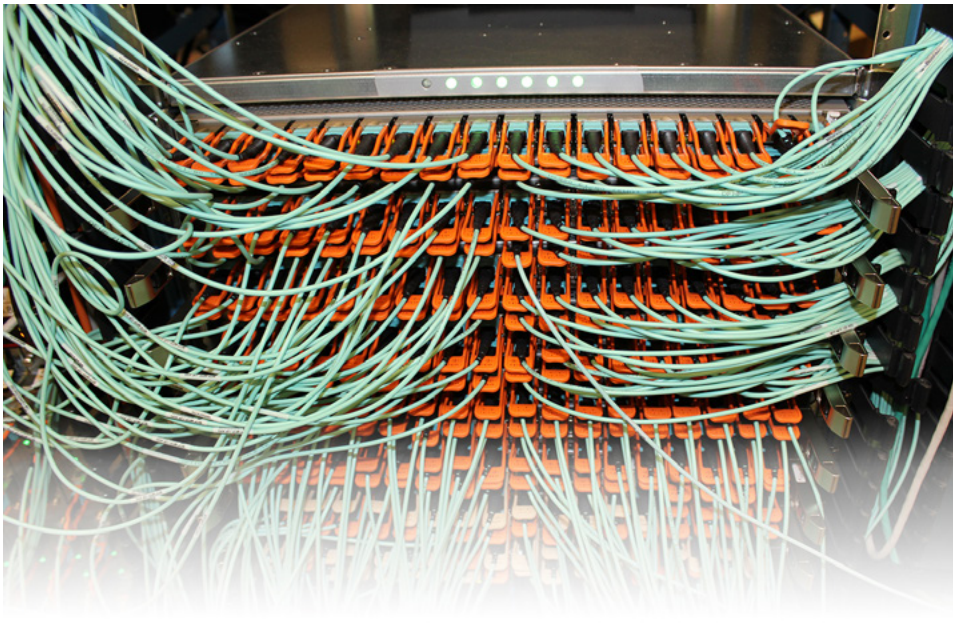
Cisco Nexus® 9508 Test Configuration

	Hardware	Software Version	Port Density
Device under test	Cisco Nexus® 9508 http://www.cisco.com/en/US/products/	NXOS 6.1(2)I1(1)	288-40GbE
Test Equipment	(3) Ixia XG12 High Performance Chassis	IxOS 6.40 EA IxNetwork 7.10 EASP1	
	(32) Xcellon-Multis XM40GE12 QSFP+FAN 40GE Load Module		
	http://www.ixiacom.com/		
Cabling	Ixia's Multis CXP-to-3-40GE QSFP Active Optical Cable (AOC)		
	http://www.ixiacom.com/pdfs/datasheets/xcellon-multis-100-40-10-loadmodule.pdf		

Cisco Nexus® 9508 RFC2544 Layer 3 Latency Test



	64	128	256	512	1024	1280	1518	2176	9216
Max Latency	3437	3505	3542	3677	3877	3962	4032	4292	6927
Avg Latency	1617.21	1645.96	1728.14	1848.62	1967.63	2003.40	2047.39	2180.06	3562.59
Min Latency	1540	1577	1657	1775	1900	1937	1977	2102	3427
Avg Delay Variation	3	3	3.014	3	3	3.177	1	3	3

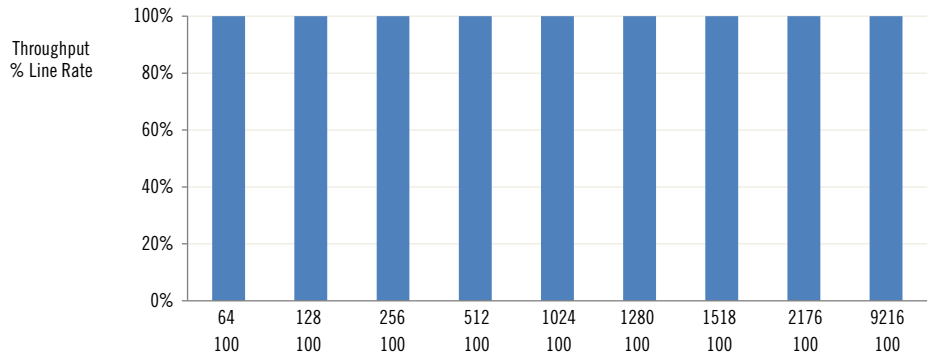


No data center switch has been tested at full 288 40GbE ports scale, thus Ixia engineers designed a new approach to test at this scale. To test the Nexus® 9508, we connected 288 40GbE ports to three Ixia XG12 High Performance Chassis running its IXOS 6.50 and IxNetwork 7.10. To deliver 288 40GbE ports to the Nexus® 9508, an Ixia CXP port capable of supplying 120Gbps of network traffic was split into three 40GbE QSFP+ ports. This was accomplished via the Ixia Xcellon-Multis XM40GE12 QSFP+FAN 40GE Load Modules, which support four CXP ports. Eight Ixia Xcellon-Multis load modules populated each XG12 chassis, delivering 96 40GbE streams of line rate traffic. Three XG12s populated with 96 40GbE each delivered 288 40GbE of line rate traffic flow into the Nexus® 9508 at varying packet sizes.

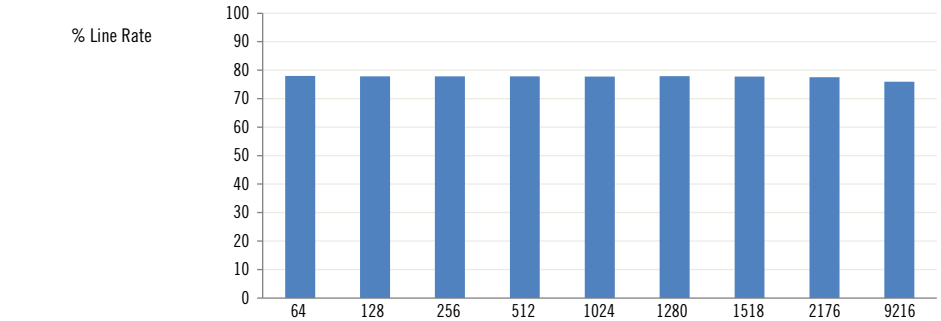
Video feature: Click here to watch Nick Lippis and Bradly Wong of Cisco detail the Nexus 9508 performance test methodology and results

The Nexus® 9508's 288 40GbE ports were connected via Ixia Multis CXP-to-3-40GE QSFP Active Optical Cable (AOC). The optical wavelength was 850NM, and 3-meter optical cables connected the test gear and the Nexus® 9508. The Ixia modules connected to the Nexus® 9508 via its 9636PQ 40GbE modules that support 36 40GbE QSFP+

Cisco Nexus® 9508
RFC 2544 L3 Throughput Test



Cisco Nexus® 9508
RFC 2889 48-port Congestion Test



	64	128	256	512	1024	1280	1518	2176	9216
Agg Forwarding Rate (% Line Rate)	77.969	77.845	77.848	77.846	77.776	77.871	77.751	77.525	75.960
L3 Head of Line Blocking	no	no	no	no	no	no	no	no	no
L3 Back Pressure	yes	yes	yes	yes	yes	yes	yes	yes	yes
Agg Flow Control Frames	0	0	0	0	0	0	0	0	0

ports. Lastly, the Nexus® 9508 ran a pre-release version of NXOS 6.1(2)I1(1) operating system.

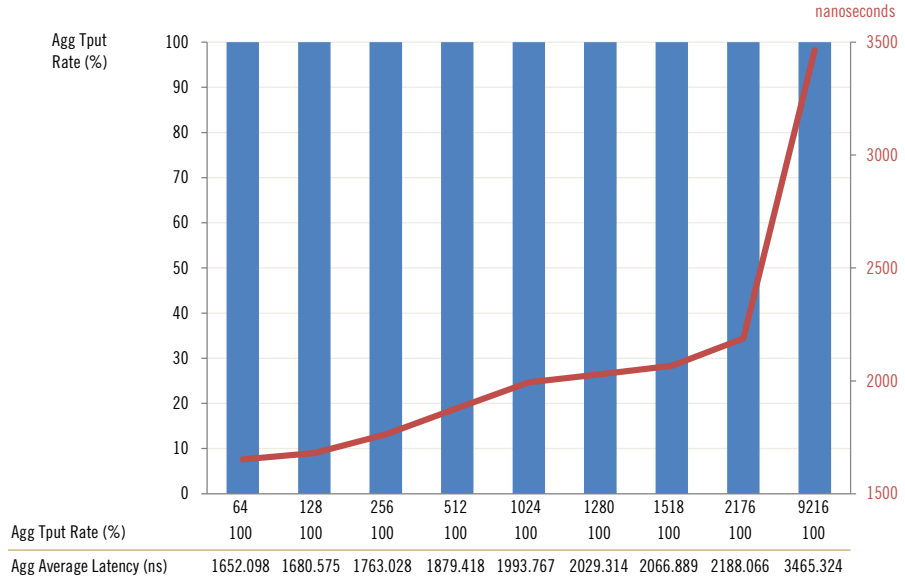
During layer 3 unicast traffic latency test, the Cisco Nexus® 9508 measured average store-and-forward latency ranging from a low of 1.6 microseconds (1617.2 ns) at 64 byte size packets to a high of 3.5 microseconds (3562.6 ns) at 9216 byte size packets. The average delay variation ranged between from 1 to 3 ns, providing consistent latency across all packet sizes at line rate. These are by far the lowest latency measurements we have observed in core switches to date. The previous record for modular switch latency was 2.2 to 11.9 microseconds, at the same packet range, however at much less density. The previous core switch record was tested at 352 10GbE ports while the Nexus® 9508 was populated with 1,152 equivalent 10GbE ports. That is, the Cisco Nexus® 9508 forwarded 9,216 byte size packets in a third of the time of the previous latency record holder but while processing three times the traffic load!

The Cisco Nexus® 9508 demonstrated 100% throughput as a percentage of line rate across all 288 40GbE ports. In other words, not a single packet was dropped while the Cisco Nexus® 9508 was presented with enough traffic to populate its highly dense 288 40GbE ports at line rate. Said another way, approximately 11.5Tbs of wire speed traffic was forwarded by the Cisco-Insieme Nexus® 9508 via a wide range of packet sizes and not a single packet was dropped!

To understand how the Cisco Nexus® 9508 behaves during periods of congestion, we stressed the Nexus 9508 switch processing and buffer architecture's congestion management subsystems. To achieve this goal, one group of four 40GbE ports was configured for congestion testing. The Nexus® 9508 demonstrated nearly 78% of aggregated forwarding rate as percentage of line rate during conges¬tion conditions for L3 traffic flows. A single 40GbE port was flooded at 150% of line rate. There was no Head of Line Blocking (HOLB) observed, which means that as a 40GbE port on the Nexus® 9508 became congested, it did not impact the performance of other switch ports.

Note that Cisco Nexus® 9508 congestion results show back pressure detection, when in fact, this is a phantom reading. Ixia and other test equipment calculate back pressure per RFC 2889 paragraph 5.5.5.2., which states that if the total number of received frames on the congestion port surpasses the number of transmitted frames at MOL (Maximum Offered Load) rate then back pressure is present. Thanks to the Nexus® 9508 packet buffer memory, it can overload ports with more pack¬ets than the MOL; therefore, the Ixia or any test equipment "calculates/sees" back pressure, but in reality, this is an anomaly of the RFC testing method and not the Nexus® 9508.

Cisco Nexus® 9508
RFC 3918 Multicast Test



	64	128	256	512	1024	1280	1518	2176	9216
Agg Tput Rate (%)	100	100	100	100	100	100	100	100	100
Agg Average Latency (ns)	1652.098	1680.575	1763.028	1879.418	1993.767	2029.314	2066.889	2188.066	3465.324

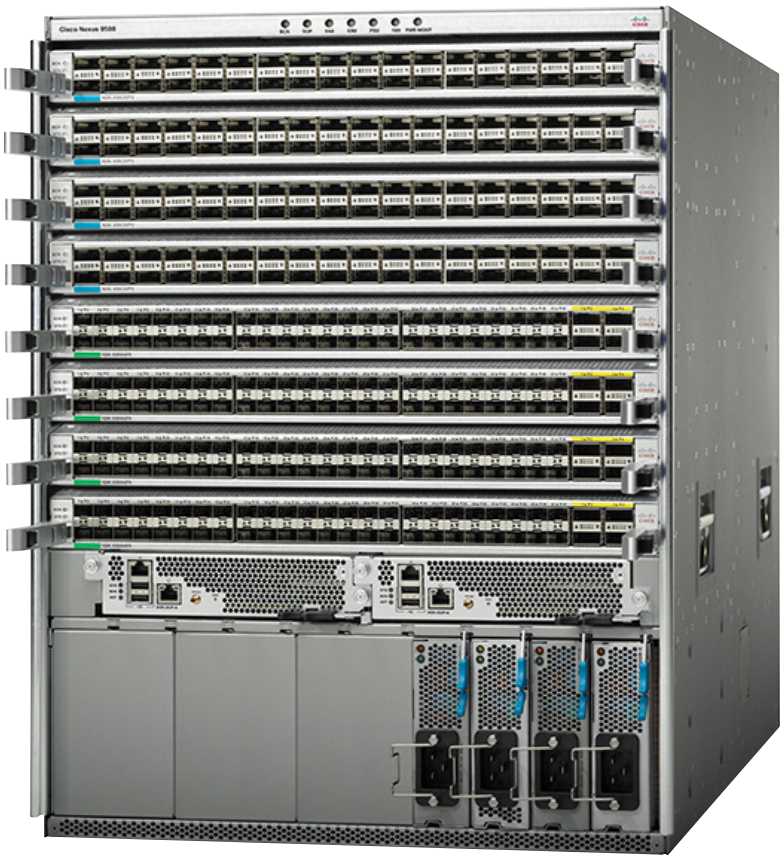
The Cisco Nexus® 9508 switch does not use back pressure (PAUSE) to manage conges¬tion, while other core switches in the industry transmit pause frames under congestion. The Cisco Nexus® 9508 leverages egress buffers and queue management on its line card modules to manage congestion, and Cisco claims will also offer up to three No-Drop classes of service per port to service drop-sensitive workload such as RDMA over Converged Enhanced Ethernet (RoCEE) or Fibre Channel over Ethernet (FCoE). It should be mentioned that the Cisco Nexus® 9508 does honor pause frames and stops transmitting, as many hosts and other switches still cannot support wire speed 10 and 40GbE traffic.

No control/pause frames were detected and observed by Ixia test gear while testing the Nexus® 9508, which is the norm in core switches.

The Cisco Nexus® 9508 demonstrated 100% aggregated throughput for IP multicast traffic with store-and-forward latencies ranging from a low of 1.6 microseconds (1652 ns) at 64 byte size packets to a high of 3.5 microseconds (3465.3 ns) at 9216 byte size packets—a very impressive result for such high density. The Nexus® 9508 forwards IP multicast traffic faster than any other core switch we have observed in these Lippis/Ixia tests. The previous record for IP multicast latency in modular switches was 2.4 to 11.6 microseconds, at the same packet range, however at much less density. The previous core switch record was tested at 352 10GbE ports while the Nexus® 9508 was populated and tested with 1152 equivalent 10GbE ports.

Discussion

The Cisco Nexus® 9508 offers the lowest latency and highest throughput for L3 unicast and IP multicast traffic, and demonstrated efficient congestion management behavior. The Nexus 9508 was designed for 40GbE aggregation in data centers; as server connectivity transitions from 1GbE to 10GbE, the core of the network is also transitioning from 10GbE to 40GbE at scale. Additionally, as leaf/spine scale-out data center network designs become increasingly popular, thanks to their potential for non-blocking, bi-partite graph connectivity that provides one-to-two hop forwarding between servers, the Cisco Nexus® 9508 is well positioned as the spine switch of choice. In hyperscale enviroments where the number of physical servers are measured in the tens to hundreds of thousands, the Cisco Nexus® 9508 is well suited for this use case too, thanks to its high density of 1152 10GbE and 288 40GbE ports and consistent low latency of both unicast and IP multicast traffic flows.



The Cisco Nexus® 9508 demonstrated its engineering prowness at this Lippis/Ixia test. It delivered the lowest latency observed for modular switching over the past three years at the Lippis/Ixia industry test and at the highest 40GbE port density. Never has a modular switch been tested for performance while processing 288 ports of 40GbE or 11.5Tbps of wire speed traffic. In addition, the Cisco Nexus® 9508 was not shipping when tested as most firms require time to scale up their products. The Cisco Nexus® 9508 forwarded 9216 byte size packets in a third of the time of the previous latency Lippis/Ixia test record holder but while processing three times the traffic load! That's the pinnacle of engineering in computer network switch design.

For IP multicast traffic forwarding, the Cisco Nexus® 9508 demonstrated 100% aggregated throughput with store-and-forward latencies ranging from a low of 1.6 microseconds (1652 ns) at 64 byte size packets to a high of 3.5 microseconds (3465.3 ns) at 9216 byte size packets—a very impressive result for such high density. This was another record for the Cisco Nexus® 9508 as it forwards IP multicast traffic nearly three times faster than any other core switch we have observed in these Lippis/Ixia tests.

The Cisco Nexus® 9508's congestion management is excellent at nearly 78% of aggregated forwarding rate as percentage of line rate during conges¬tion conditions for L3 traffic flows, but when considering the density of ports supported and sheer magnitude of the traffic flow, the Cisco engineers achieved congestion management at a scale never before attempted.

For L2/L3 40GbE aggregation, hyperscale and spine switch use cases, the Cisco Nexus® 9508 proved during these Lippis/Ixia tests that it's well engineered for these environments.

The Lippis Report Test Methodology

To test products, each supplier brought its engineers to configure its equipment for test. An Ixia test engineer was available to assist each supplier through test methodologies and review test data. After testing was concluded, each supplier's engineer signed off on the resulting test data. We call the following set of testing conducted "The Lippis Test." The test methodologies included:

Throughput Performance: Throughput, packet loss and delay for L2 unicast, L3 unicast and L3 multicast traffic was measured for packet sizes of 64, 128, 256, 512, 1024, 1280, 1518, 2176 and 9216 bytes. In addition, a special cloud computing simulation throughput test consisting of a mix of north-south plus east-west traffic was conducted. Ixia's IxNetwork RFC 2544 Throughput/Latency quick test was used to perform all but the multicast tests. Ixia's IxAutomate RFC 3918 Throughput No Drop Rate test was used for the multicast test.

Latency: Latency was measured for all the above packet sizes plus the special mix of north-south and east-west traffic blend. Two latency tests were conducted: 1) latency was measured as packets flow between two ports on different modules for modular switches, and 2) between far away ports (port pairing) for ToR switches to demonstrate latency consistency across the forwarding engine chip. Latency test port configuration was via port pairing across the entire device versus side-by-side. This meant that a switch with N ports, port 1 was paired with port (N/2)+1, port 2 with port (N/2)+2, etc. Ixia's IxNetwork RFC 2544 Throughput / Latency quick test was used for validation.

Jitter: Jitter statistics was measured during the above throughput and latency test using Ixia's IxNetwork RFC 2544 Throughput/Latency quick test.

Congestion Control Test: Ixia's IxNetwork RFC 2889 Congestion test was used to test both L2 and L3 packets. The objective of the Congestion Control Test is to determine how a Device Under Test (DUT) handles congestion. Does the device implement congestion control and does congestion on one port affect an uncongested port? This procedure determines if HOL blocking and/or if back pressure are present. If there is frame loss at the uncongested port, HOL blocking is present. Therefore, the DUT cannot forward the amount of traffic to the congested port, and as a result, it is also losing frames destined to the uncongested port. If there is no frame loss on the congested port and the port receives more packets than the maximum offered load of 100%, then back pressure is present.

Video feature: **Click to view a discussion on the Lippis Report Test Methodology**

RFC 2544 Throughput/Latency Test

Test Objective: This test determines the processing overhead of the DUT required to forward frames and the maximum rate of receiving and forwarding frames without frame loss.

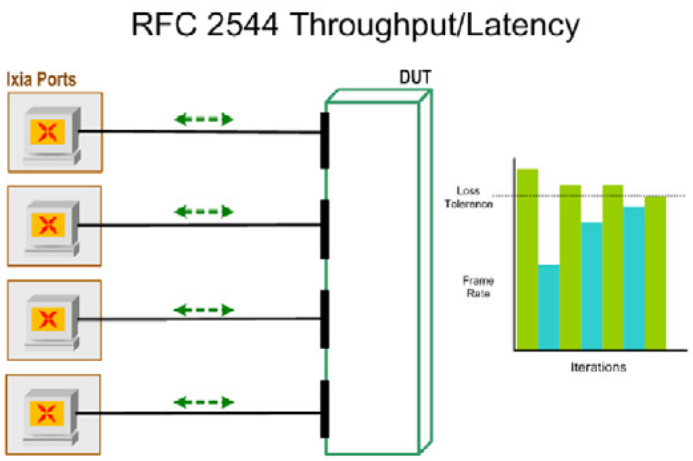
Test Methodology: The test starts by sending frames at a specified rate, usually the maximum theoretical rate of the port while frame loss is monitored. Frames are sent from and received at all ports on the DUT, and the transmission and reception rates are recorded. A binary, step or combo search algorithm is used to identify the maximum rate at which no frame loss is experienced.

To determine latency, frames are transmitted for a fixed duration. Frames are tagged once in each second and during half of the transmission duration, then tagged frames are transmitted. The receiving and transmitting timestamp on the tagged frames are compared. The difference between

the two timestamps is the latency. The test uses a one-to-one traffic mapping. For store and forward DUT switches, latency is defined in RFC 1242 as 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. Thus latency is not dependent on link speed only, but processing time too.

Results: This test captures the following data: total number of frames transmitted from all ports, total number of frames received on all ports, percentage of lost frames for each frame size plus latency, jitter, sequence errors and data integrity error.

The following graphic depicts the RFC 2554 throughput performance and latency test conducted at the iSimCity lab for each product.



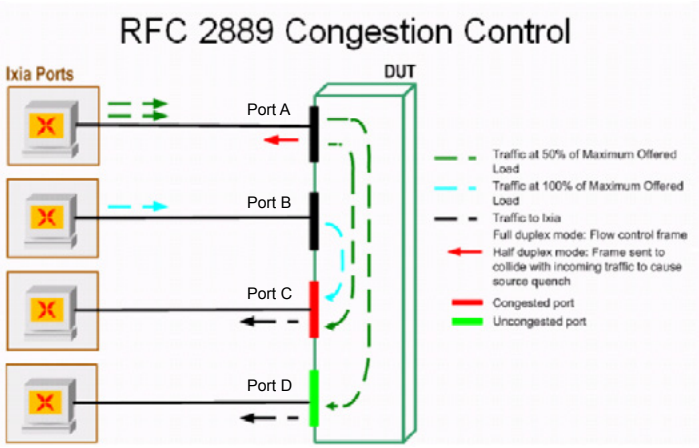
RFC 2889 Congestion Control Test

Test Objective: The objective of the Congestion Control Test is to determine how a DUT handles congestion. Does the device implement congestion control and does congestion on one port affect an uncongested port? This procedure determines if HOL blocking and/or if back pressure are present. If there is frame loss at the uncongested port, HOL blocking is present. If the DUT cannot forward the amount of traffic to the congested port, and as a result, it is also losing frames destined to the uncongested port. If there is no frame loss on the congested port and the port receives more packets than the maximum offered load of 100%, then back pressure is present.

Test Methodology: If the ports are set to half duplex, collisions should be detected on the transmitting interfaces. If the ports are set to full duplex and flow control is enabled, flow control frames should be detected. This test consists of a multiple of four ports with the same MOL. The custom port group mapping is formed of two ports, A and B, transmitting to a third port C (the congested interface), while port A also transmits to port D (uncongested interface).

Test Results: This test captures the following data: intended load, offered load, number of transmitted frames, number of received frames, frame loss, number of collisions and number of flow control frames obtained for each frame size of each trial are captured and calculated.

The following graphic depicts the RFC 2889 Congestion Control test as conducted at the iSimCity lab for each product.



RFC 3918 IP Multicast Throughput No Drop Rate Test

Test Objective: This test determines the maximum throughput the DUT can support while receiving and transmitting multicast traffic. The input includes protocol parameters Internet Group Management Protocol (IGMP), Protocol Independent Multicast (PIM), receiver parameters (group addressing), source parameters (emulated PIM routers), frame sizes, initial line rate and search type.

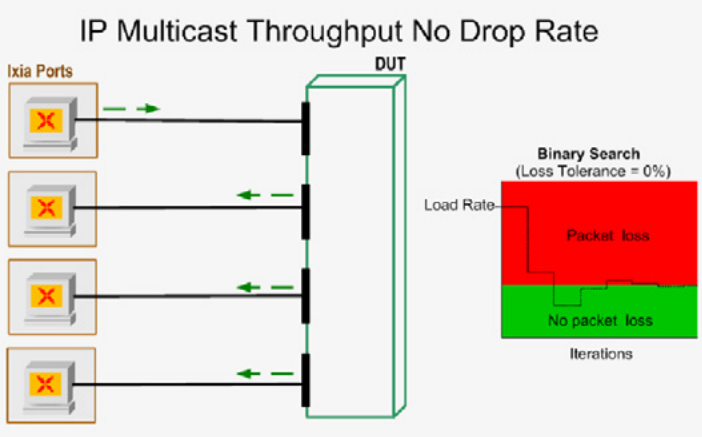
Test Methodology: This test calculates the maximum DUT throughput for IP Multicast traffic using either a binary or a linear search, and to collect Latency and Data

Integrity statistics. The test is patterned after the ATSS Throughput test; however this test uses multicast traffic. A one-to-many traffic mapping is used, with a minimum of two ports required.

If choosing OSPF (Open Shortest Path First) or ISIS (Intermediate System-Intermediate System) as IGP (Internet Gateway Protocol) protocol routing, the transmit port first establishes an IGP routing protocol session and PIM session with the DUT. IGMP joins are then established for each group, on each receive port. Once protocol sessions are established, traffic begins to transmit into the DUT and a binary or linear search for maximum throughput begins.

If choosing "none" as IGP protocol routing, the transmit port does not emulate routers and does not export routes to virtual sources. The source addresses are the IP addresses configured on the Tx ports in data frame. Once the routes are configured, traffic begins to transmit into the DUT and a binary or linear search for maximum throughput begins.

Test Results: This test captures the following data: maximum throughput per port, frame loss per multicast group, minimum/maximum/average latency per multicast group and data errors per port.The following graphic depicts the RFC 3918 IP Multicast Throughput No Drop Rate test as conducted at the iSimCity lab for each product.



Terms of Use

This document is provided to help you understand whether a given product, technology or service merits additional investigation for your particular needs. Any decision to purchase a product must be based on your own assessment of suitability based on your needs. The document should never be used as a substitute for advice from a qualified IT or business professional. This evaluation was focused on illustrating specific features and/or performance of the product(s) and was conducted under controlled, laboratory conditions. Certain tests may have been tailored to reflect performance under ideal conditions; performance may vary under real-world conditions. Users should run tests based on their own real-world scenarios to validate performance for their own networks.

Reasonable efforts were made to ensure the accuracy of the data contained herein but errors and/or oversights can occur. The test/ audit documented herein may also rely on various test tools, the accuracy of which is beyond our control. Furthermore, the document relies on certain representations by the vendors that are beyond our control to verify. Among these is that the software/ hardware tested is production or production track and is, or will be, available in equivalent or better form to commercial customers. Accordingly, this document is provided "as is," and Lippis Enterprises, Inc. (Lippis), gives no warranty, representation or undertaking, whether express or implied, and accepts no legal responsibility, whether direct or indirect, for the accuracy, completeness, usefulness or suitability of any information contained herein.

By reviewing this document, you agree that your use of any information contained herein is at your own risk, and you accept all risks and responsibility for losses, damages, costs and other consequences resulting directly or indirectly from any information or material available on it. Lippis is not responsible for, and you agree to hold Lippis and its related affiliates harmless from any loss, harm, injury or damage resulting from or arising out of your use of or reliance on any of the information provided herein.

Lippis makes no claim as to whether any product or company described herein is suitable for investment. You should obtain your own independent professional advice, whether legal, accounting or otherwise, before proceeding with any investment or project related to any information, products or companies described herein. When foreign translations exist, the English document is considered authoritative. To assure accuracy, only use documents downloaded directly from www.lippisreport.com .

No part of any document may be reproduced, in whole or in part, without the specific written permission of Lippis. All trademarks used in the document are owned by their respective owners. You agree not to use any trademark in or as the whole or part of your own trademarks in connection with any activities, products or services which are not ours, or in a manner which may be confusing, misleading or deceptive or in a manner that disparages us or our information, projects or developments.

About Nick Lippis

Nicholas J. Lippis III is a world-renowned authority on advanced IP networks, communications and their benefits to business objectives. He is the publisher of the Lippis Report, a resource for network and IT business decision makers to which over 35,000 executive IT business leaders subscribe. Its Lippis Report podcasts have been downloaded over 200,000 times; ITunes reports that listeners also download the *Wall Street Journal's* Money Matters, *Business Week's* Climbing the Ladder, *The Economist* and The *Harvard Business Review's* IdeaCast. He is also the co-founder and conference chair of the Open Networking User Group, which sponsors a bi-annual meeting of over 200 IT business leaders of large enterprises. Mr. Lippis is currently working with clients to design their private and public virtualized data center cloud computing network architectures with open networking technologies to reap maximum business value and outcome.

He has advised numerous Global 2000 firms on network architecture, design, implementation, vendor selection and budgeting, with clients including Barclays Bank, Eastman Kodak Company, Federal Deposit Insurance Corporation (FDIC), Hughes Aerospace, Liberty Mutual, Schering-Plough, Camp Dresser McKee, the state of Alaska, Microsoft, Kaiser Permanente, Sprint, Worldcom, Cisco Systems, Hewlett Packet, IBM, Avaya and many others. He works exclusively with CIOs and their direct reports. Mr. Lippis possesses a unique perspective of market forces and trends occurring within the computer networking industry derived from his experience with both supply- and demand-side clients.

Mr. Lippis received the prestigious Boston University College of Engineering Alumni award for advancing the profession. He has been named one of the top 40 most powerful and influential people in the networking industry by *Network World*. *TechTarget*, an industry on-line publication, has named him a network design guru while *Network Computing Magazine* has called him a star IT guru.

Mr. Lippis founded Strategic Networks Consulting, Inc., a well-respected and influential computer networking industry-consulting concern, which was purchased by Softbank/Ziff-Davis in 1996. He is a frequent keynote speaker at industry events and is widely quoted in the business and industry press. He serves on the Dean of Boston University's College of Engineering Board of Advisors as well as many start-up venture firms' advisory boards. He delivered the commencement speech to Boston University College of Engineering graduates in 2007. Mr. Lippis received his Bachelor of Science in Electrical Engineering and his Master of Science in Systems Engineering from Boston University. His Masters' thesis work included selected technical courses and advisors from Massachusetts Institute of Technology on optical communications and computing.