Performance Analysis:

Cisco Application Centric Infrastructure (ACI) Fabric

DR150417C

June 2015

Miercom
www.miercom.com

Contents

1 - Executive Summary

Miercom conducted among the first independent performance tests of Cisco's new, next-generation data center networking architecture: the Application Centric Infrastructure, or ACI.

The ACI fabric is described in more detail in section two. Briefly, Cisco's ACI is an architecture for binding many switches into a scalable, highly resilient and logical whole for centralized and automated provisioning and management. The APIC (Application Policy Infrastructure Controller), a centralized group-based policy controller, tracks, and dynamically adjusts the fabric to accommodate changes in, endpoint locations, traffic patterns, applications, services and policies.

The ACI fabric tested consisted of more than two dozen high-capacity ACI leaf switches and four ACI spine switches, interconnected in a two-tier CLOS fabric architecture. This was not a conventional single-switch test, as throughput, latency, and convergence in this case were measured end-to-end across the fabric network. Resiliency and failed-link convergence testing also looked at the failure of switch and control nodes.

This report summarizes the results of the testing in these key areas:

- Throughput between switch pairs across the fabric, and the full-mesh throughput over nearly a thousand 10GE links, between 20 switches through backbone nodes
- The latency experienced by data traversing multiple fabric switches
- Link-failure convergence and fabric recovery from a failed backbone or control node
- Traffic counting metrics and other fabric-management aspects.

Key findings

- ACI achieves 99+ percent of theoretical maximum throughput in end-to-end network throughput testing
  In tests of L2 and L3 line-rate throughput between switch pairs, all frame sizes consistently achieved 99+ percent of the theoretical maximum.
- Multicast and full-mesh performance like a single switch
  The ACI fabric delivered performance over 960 line-rate 10GE links across 20 switches similar to the performance of a single high-capacity switch.
- Low and consistent latency
  For traffic traversing the entire fabric network, with average end-to-end latency under 8 µsecs for all but Jumbo-size frames.
- Fast convergence around failed link and switch
  The ACI fabric exhibited impressive resiliency to, and quick re-convergence around, failed links and switches. Failed controllers have no impact on data flows; no data is lost.

Miercom has independently tested key performance aspects of the Cisco Application Centric Infrastructure (ACI) fabric. Throughput and latency across the fabric rival single-switch performance, and built-in redundancy and safeguards mitigate link and switch failures. In light of these findings we present the *Miercom Performance Verified* certification to the Cisco Application Centric Infrastructure.

Robert Smithers
CEO
Miercom

2 - About Application Centric Infrastructure (ACI)

This was not a test of single switch platform, but a collection of switch platforms that collectively deliver what Cisco calls its Application Centric Infrastructure, or ACI.

With ACI, Cisco has advanced the network infrastructure beyond big, powerful switches, which do their switching and routing more or less autonomously. The ACI fabric consists of many powerful switches, interconnected redundantly by high speed links, typically 40 Gigabit Ethernet, and managed, coordinated and controlled by an Application Policy Infrastructure Controller, or APIC, which will typically also be redundant and database synchronized.

An ACI fabric, as we observed in our testing, performs similarly in most respects to a single powerful switch, even though it consists of many interconnected and coordinated switches. And with the addition of a new overlay construct – the protocol mechanism used to track endpoints, data flows, application traffic and policies – this collective ACI fabric may indeed be greater than the sum of its parts.

Why ACI? In a nutshell, big iron has certain inherent limitations that ACI seeks to transcend.

For example:

Scalability. A large multi-slot switch chassis can still only grow to a point. ACI enables switches to be added, upgraded and replaced as needs grow. Cisco says a single ACI fabric can grow to encompass 1 million endpoints, upwards of 200,000 10GE ports and up to 64,000 tenants.

Decoupling of endpoints from the underlying network topology. A single switch remains limited by the static nature of its MAC and routing tables. The network infrastructure needs to go beyond the confines of IP subnetting if it is to support new technology trends – cloud computing, mobility and virtualization. ACI does this with different characterizations of the users of network bandwidth, such as "endpoint groups" and "application profiles."

Application-centric reconfiguration. ACI fabric management is policy based. As policies change, such as for access, security or QoS, these can be applied instantly throughout the fabric. Data paths are learned and fabric changes are automatically discovered.

The ACI architecture thinks in terms of "leaf" nodes – for endpoint access – and "spine" nodes – for backbone connectivity and for enabling redundant paths between all nodes. The ACI fabric we tested was composed of all Nexus 9000 switches. This is one of the first Cisco switch families made expressly with ACI in mind. At present all nodes in an ACI fabric need to support the unique operational aspects of ACI, although other "non-ACI" switches can still be linked to the fabric. They just won't be ACI fabric participants.

In a real sense ACI allows application requirements to define the network. Because of this, the ACI architecture simplifies, optimizes, and accelerates the entire application deployment life cycle. A look at how ACI works helps in explaining this.

The ACI fabric defines application communication paths – tunnels – between user-defined application services and endpoint groups, or EPGs. These are then overlaid on the network

definitions of the fabric. This allows the network definitions to be standardized and scalable across the fabric, and to be managed by the APIC controller.

When application requirements change, or new applications or services are added, the changes and additions are added to the ACI object model. The APIC controller, then, is responsible for deploying the changes in the overlay architecture – using the underlying network for communication with the appropriate hosts – without having to change the underlying network structures.

The APICs manage and monitor the fabric from an application-centric view. They are responsible for updating the overlay provisioning when changes need to be made to the application structures. Once the set-up is complete, these controllers are not involved in routing data, so if they go offline for any reason, traffic flow is not affected. An off-line controller affects just the configuring of the application object model on the network and application-centric monitoring.

A paper that provides more detail on the ACI architecture can be found at:
http://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/1-x/aci-fundamentals/b_ACI-Fundamentals.pdf

3 - Theoretical Maximum ACI Throughput

Much of the added network functionality and capability in the Application Centric Infrastructure comes from an overlay protocol layer, in which VXLAN (Virtual Extensible LAN) is used as the overlay technology to provide an integrated fabric solution. The MAC-in-UDP encapsulation of VXLAN introduces a 50-byte or 54-byte overhead, depending on whether the underlay network link uses 802.1q VLAN tag or not, to every packet, and so it becomes a factor to consider in calculating throughput for the end-to-end ACI fabric performance tests we conducted.

In general, maximum throughput, or wire speed, is a function of packing as many bytes of data on a transmission link as possible, and having all that data pass through a device under test (DUT) and delivered back out with no loss. In this case, though, the DUT is a fabric consisting of many links and switches.

A number of Spirent test systems were assembled to perform the substantial throughput tests required for this review. Indeed, a total of 960 x 10GE ports were used to generate bi-directional, wire-speed data flows and the results were carefully compared and analyzed.

In testing of a single switch, the Spirent test system sends data in on one port and compares it with what it receives back out on another port. For a 10-Gigabit/s Ethernet (10GE) port, this equates to 10,000,000,000 bits, or 1,250,000,000 GB/s (Gigabytes/second).

But this is not all user data. A minimum-size packet, including MAC, IPv4 header, payload and CRC, is normally regarded as 64 bytes. On all forms of Ethernet there is also a minimum gap between packets (an IFG, Inter Frame Gap) of 12 bytes and a preamble, preceding every packet, of 8 bytes. Often there is also a VLAN identifier, per IEEE 802.1q, which adds 4 more bytes. As a rule, though, throughput tests do not usually include the 4-byte VLAN tags.

So a minimum, 64-byte packet actually consumes 84 bytes of bandwidth on an Ethernet wire. The theoretical maximum packets per second (PPS) throughput, then, for a conventional 10GE link (in each direction) is:

For 64-byte packets, wire speed = 14,880,960 pps or 14.88 Mpps

For 1518-byte packets, wire speed = 812,274 pps or 0.812 Mpps

However, for all packets traversing a Cisco ACI fabric, the actual throughput will be different from the above due to the VXLAN encapsulation.

When Ethernet frames arrive on the ingress ACI leaf switch, the leaf switch will encapsulate the Ethernet frames into VXLAN packets and forward the encapsulated packets onto the ACI fabric. If there is a 802.1q VLAN tag in the original Ethernet header, it will be stripped off. The VXLAN encapsulation itself adds 54 bytes to the original packet.

A 64-byte packet normally consumes 84 bytes of bandwidth:

64 +12 (IFG) + 8 (Preamble) = 84

Within a Cisco ACI, however, the same minimum-size packet requires:

138 bytes if it does not have an 802.1q VLAN tag:

64 +12 (IFG) + 8 (Preamble) +54 (added by VXLAN encapsulation)

or 134 bytes if it has an 802.1q VLAN tag:

64 +12 (IFG) + 8 (Preamble) -4 (VLAN stripped) +54 (added by VXLAN encapsulation)

The additional bytes for VXLAN encapsulation is a cost that needs to be taken into account when computing an accurate Theoretical Maximum for the line rate of the ACI fabric. This adjustment is done as follows:

ACI Percent of Theoretical Maximum Throughput =

(PacketSize +12 +8) / (PacketSize +12 +8 +54) for non-802.1q packets

(PacketSize +12 +8) / (PacketSize -4 +12 +8 +54) for 802.1q packets

When running Spirent tests at line rate, this is how to convert to the Cisco ACI Theoretical Maximum Line Rate.

Spirent Theoretical Max Line Rate (%) x (Spirent Packet Size / Cisco ACI Packet Size)

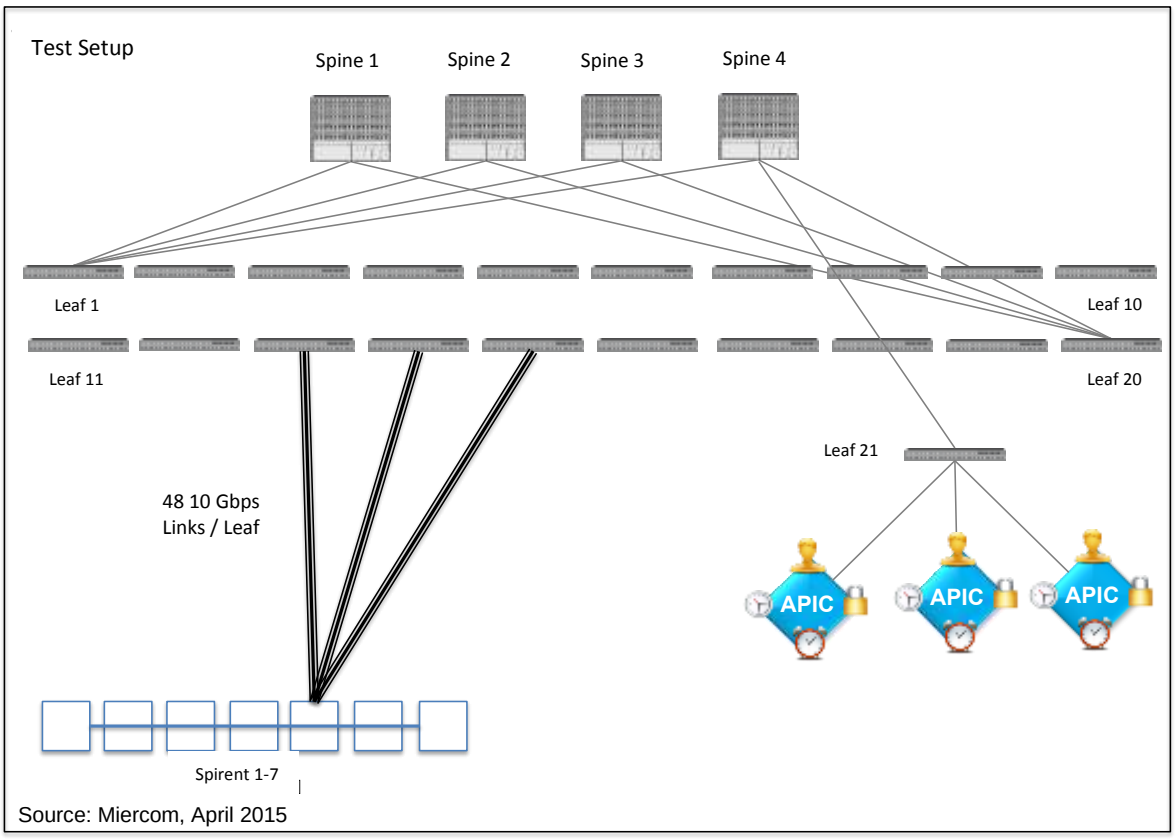
For all 64-byte packets with an 802.1q VLAN tag, then, the Cisco ACI Maximum Theoretical Throughput (ACI wire speed) is 63.77 percent of the conventional Spirent Maximum Theoretical Throughput.

For all 1,518-byte packets with an 802.1q VLAN tag, the Cisco ACI Maximum Theoretical Throughput is 96.86 percent of the conventional Spirent Maximum Theoretical Throughput.

For larger 9,000-byte Jumbo packets, the difference narrows even more. The Cisco ACI Maximum Theoretical Throughput is 99.34 percent of the conventional Spirent Maximum Theoretical Throughput.

4 - Test Bed – How We Did It

A substantial test bed was assembled for this testing. As the Test Setup diagram below shows, a total of 25 Nexus 9000 switches were deployed – four Spine nodes and 21 Leaf nodes. The Spine nodes were all Nexus 9504 four-slot chassis. Each contained two 36-port, 40 Gigabit/s Ethernet (40GE) ACI Spine I/O Modules with QSFP+ interfaces, collectively offering 288 x 40GE ports.



Spine and Leaves. Each of the 20 Leaf switches connected to each of the four Spine switches via three 40GE links. For simplification just the Spine connections from Leaf switches 1 and 20 are shown above. In total, there were 240 40GE links between the leaf-layer and the spine-layer. In addition, three redundant, database-synchronized ACI controllers (APICs) were connected to a 21st Leaf switch, also linked into the Spine. Using the 21th leaf node for APIC controller connectivity is to guarantee that the first 20 leaf nodes can be fully loaded with test data traffic.

The 21 Leaf nodes were all Cisco Nexus 9396PX switches. As the picture below shows, the 9396PX's feature 48 x 10GE ports (SFP+) and twelve 40GE uplink ports. The 40GE ports connected the Leaf switch to each Spine switch, three links to each Spine node.

The ACI architecture calls for the interconnection of switches in this manner to maximize bandwidth and alternate-route redundancy. Note that no Leaf switches are directly connected. Rather, all communications – other than local intra-switch – are done in two hops, traversing a Spine switch. This Leaf switch-to-Leaf switch flow is referred to as East-West traffic.

The ACI fabric, then, consisted of a high-speed backbone network, connecting the Spine switches with the Leaf switches via some 240 x 40GE links, and 20 switches yielding a total of 960 x 10GE ports. The ACI fabric we assembled and tested, then, behaved like a single, massive 960 x 10GE-port switch.

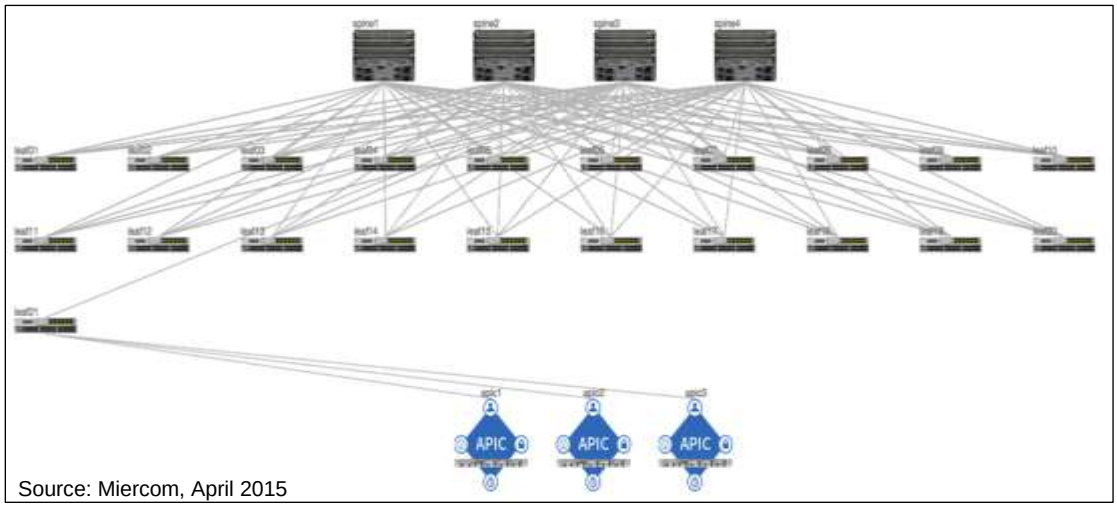
In our test-bed configuration, three Application Policy Infrastructure Controllers, or APICs, were deployed, connecting into the fabric via another Leaf switch (Leaf 21), since all the ports on all the other 20 Leaf switches were used to carry test traffic. The APICs were running version 1.0 (3i) software.

All of the 48 x 10GE ports on each of the 20 Leaf switch were connected to the Spirent test system. A total of seven Spirent test-system chassis were employed to generate and deliver wire speed, bidirectional traffic to all 960 x 10GE ports on the 20 Leaf switches. Six were SPT-9000A Test Center chassis; the seventh was an SPT-N11U Mainframe Chassis. All were loaded with cards to generate (and receive and analyze) 10GE test traffic streams.

The Spirent test systems all ran firmware 4.47.5296. The systems were chained together to synchronize their traffic generation and collect and assemble results.

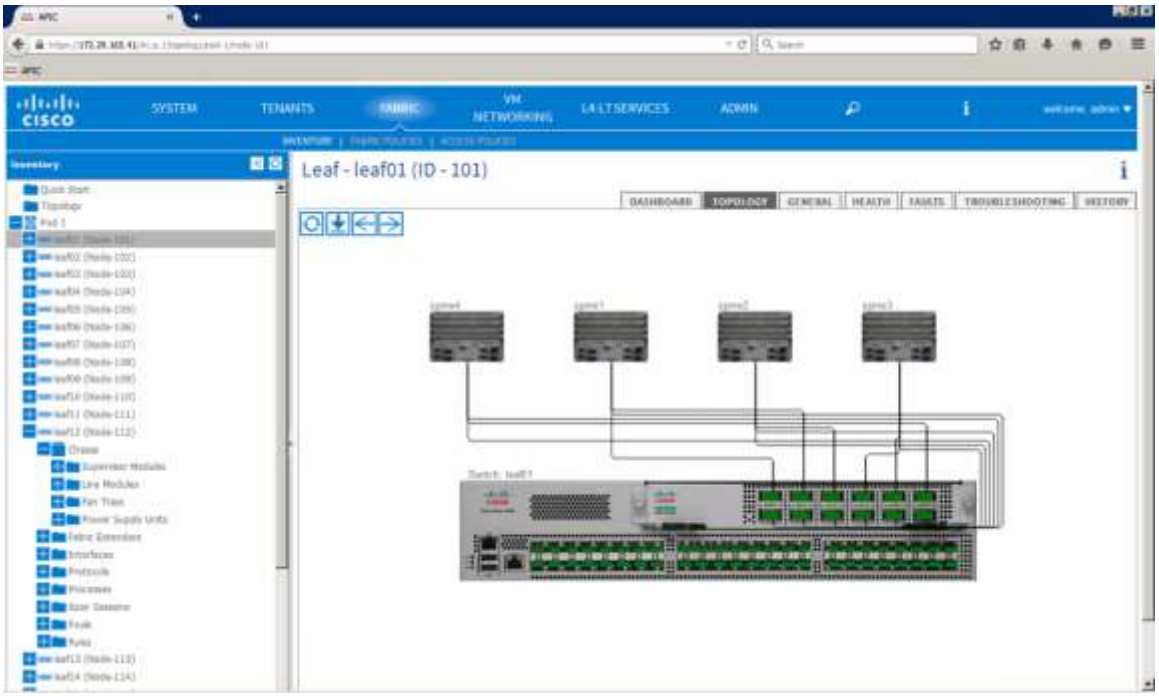
View from the APIC

Below is the display of the ACI-fabric test bed provided by the graphical user interface of the APIC controller. On an active system, the links are not all shown until the user hovers over particular leaf nodes; then those nodes' connections with link details are shown.



Source: Miercom, April 2015

View from the Leaf. As a centralized point for monitoring the entire ACI fabric, the APIC controller can readily retrieve a view of the fabric from any particular Leaf switch. Leaf switch #1's view of the fabric, as viewed on the APIC controller, is shown below.



Source: Miercom, April 2015

5 - Layer-2 Throughput and Latency Tests

Layer-2 testing involves the transmission and switching of frames at the MAC layer. Tests were conducted in accordance Internet RFC specifications for the standardized conduct of these tests. RFC-based testing included:

- RFC 2544, for port-pair testing, where in our case all traffic traversed the ACI fabric. Throughput and latency were measured.
- RFC 2889, for full-mesh testing, which in the case of the Cisco ACI means that wire-speed frames delivered by the Spirent tester to one Leaf switch's 10GE port were distributed by the switch and fabric to each of the other 959 Leaf switch 10GE ports in round-robin fashion. As with port-pair testing, throughout and latency are measured.
- RFC 3918, for multicast testing, where one-way traffic received on one port is replicated by the switch (or in this case, fabric) and distributed to all other 959 receive ports. Throughput and latency were again measured.

Some of the test results, where noted, have been normalized to percent of maximum (or percent of line rate) within the ACI fabric. As discussed in the previous section, the line-rate maximum for traffic within the fabric is less than outside the fabric because of the additional overhead added to each frame by the ACI overlay protocol.

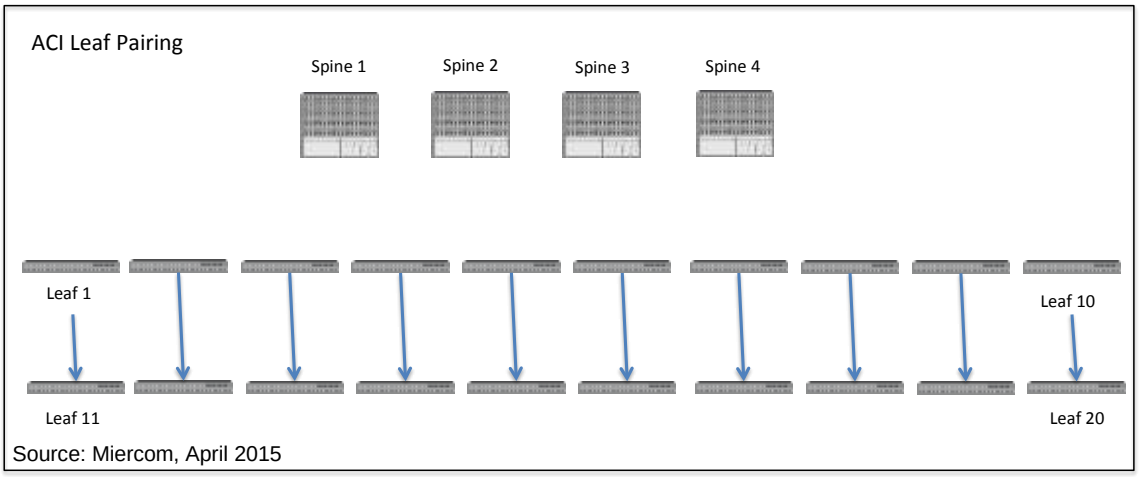
L2 - Leaf Pairing, Unicast Throughput

What's Measured

Bidirectional throughput, as a percent of the theoretical ACI-fabric maximum, for traffic between the 48 x 10GE ports of one Leaf switch, with the corresponding 48 ports on another Leaf switch.

How We Did It

As shown in the diagram on page 12, Leaf switches were paired, such that traffic sent in on ports 1 through 48 of Leaf Switch 1 were forwarded to ports 1 through 48 of Leaf Switch 11, and vice versa in the opposite direction. Traffic was then sent, bidirectionally, between 10 pairs of Leaf switches. In reality, every frame went through a Spine switch to its destination port.



For the ACI-fabric set-up for this test, a single endpoint group (EPG) was used, a single VLAN, and a single subnet.

The traffic delivered to each port consisted of 4,000 flows – 4,000 source and destination UDP combinations. The source UDP port numbers started at 16000 and incremented by one per flow. Similarly, the destination UDP port numbers started at 51001 and incremented by one per flow.

Results

The table below summarizes the throughputs achieved, by frame size, over the Cisco ACI fabric, as a percent of the line rate delivered by the Spirent testers.

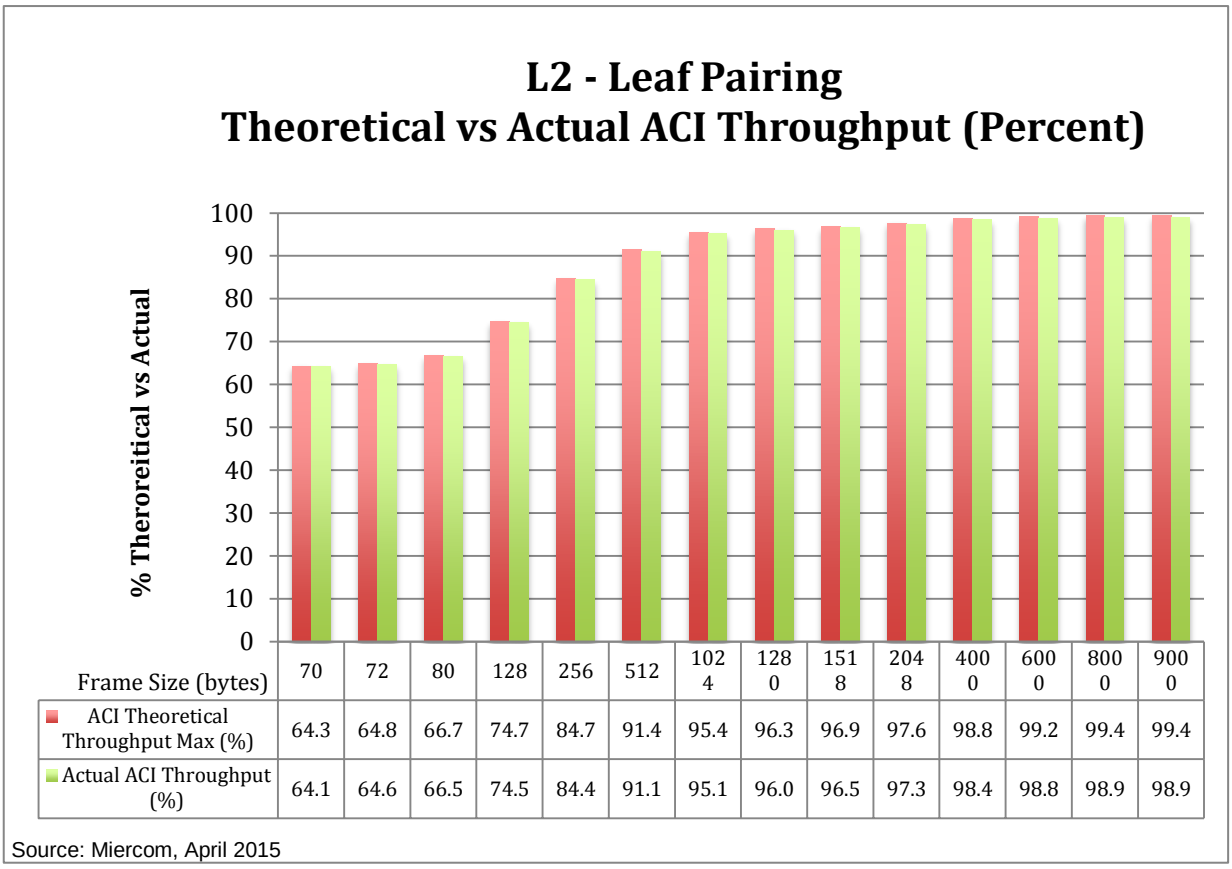
L2 - Leaf Pairing – Unicast Throughput (per RFC 2544)

Frame Size (Bytes)	Tx Line Rate (Percent)	ACI Theoretical Throughput Max (Percent)	Actual ACI Throughput (Percent)	Percent of Theoretical ACI Max Throughput Achieved
70	100	64.3	64.1	99.7
72	100	64.8	64.6	99.7
80	100	66.7	66.5	99.7
128	100	74.7	74.5	99.7
256	100	84.7	84.4	99.7
512	100	91.4	91.1	99.7
1024	100	95.4	95.1	99.7
1280	100	96.3	96.0	99.7
1518	100	96.9	96.5	99.7
2048	100	97.6	97.3	99.7
4000	100	98.8	98.4	99.7
6000	100	99.2	98.8	99.6
8000	100	99.4	98.9	99.5
9000	100	99.4	98.9	99.5

The results indicate that virtually all of the available bandwidth – from 99.5 to 99.7 percent of the wire speed maximum within the ACI fabric – was effectively filled.

Take 512-byte frames, for example. The maximum that could be achieved within the ACI fabric is 91.4 percent sent at line rate – from outside the ACI fabric. 99.7 percent of the maximum possible throughput was achieved for 512-byte frames traversing the ACI fabric.

This data is represented graphically below.



Frame Size (bytes)	70	72	80	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
ACI Theoretical Throughput Max (%)	64.3	64.8	66.7	74.7	84.7	91.4	95.4	96.3	96.9	97.6	98.8	99.2	99.4	99.4
Actual ACI Throughput (%)	64.1	64.6	66.5	74.5	84.4	91.1	95.1	96.0	96.5	97.3	98.4	98.8	98.9	98.9

Source: Miercom, April 2015

L2 - Full Mesh Across All Ports, Unicast Throughput

What's Measured

Bidirectional throughput, by frame size, from wire speed traffic delivered to all 960 x 10GE ports on all 20 Leaf switches, with frames forwarded in round-robin fashion to all other 959 ports.

How We Did It

For each frame-size iteration, test traffic was issued at 10 Gigabit/s across all 960 Leaf switch ports. Each port on every Leaf forwards traffic in round-robin distribution to all the 959 other ports – including 47 local ports on the same Leaf switch, and 912 ports on all the remote Leaf switches. The Spirent testers were chained together so that the test traffic streams were synchronized.

Results

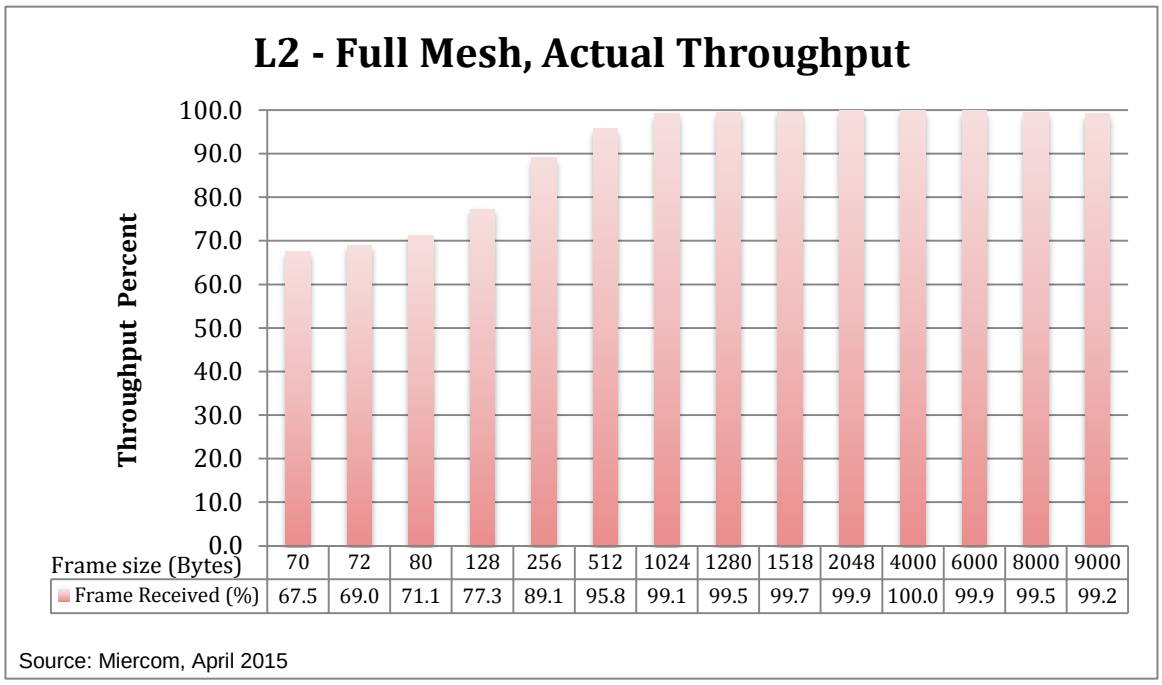
The table below summarizes the throughputs achieved, by frame size, over the Cisco ACI fabric, and as a percent of the line rate delivered by the Spirent testers.

The throughput percent (of the Spirent-delivered wire rate) is higher than in the port-pair testing due to the 4.9 percent leaf-switch local traffic.

L2 - Full Mesh Across All Ports – Unicast Throughput per RFC 2889

Frame Size (Bytes)	Tx Line Rate (Percent)	Tx Frame Count	Rx Frame Count	Actual Throughput (Percent)
70	100	400000000320	270016787405	67.5
72	100	391304348160	269843094756	69.0
80	100	360000000025	256120097469	71.1
128	100	243243243840	187973776017	77.3
256	100	130434783360	116176947696	89.1
512	100	67669173120	64813421333	95.8
1024	100	34482759360	34177186739	99.1
1280	100	27561936909	27561936909	99.5
1518	100	23339224796	23339224796	99.7
2048	100	17382840567	17382840567	99.9
4000	100	8950908943	8950908943	100.0
6000	100	5980066560	5972499070	99.9
8000	100	4488778560	4465057826	99.5
9000	100	3991130880	3957335696	99.2

The key results are shown graphically below.



L2 - Multicast Throughput Across All Ports

What's Measured

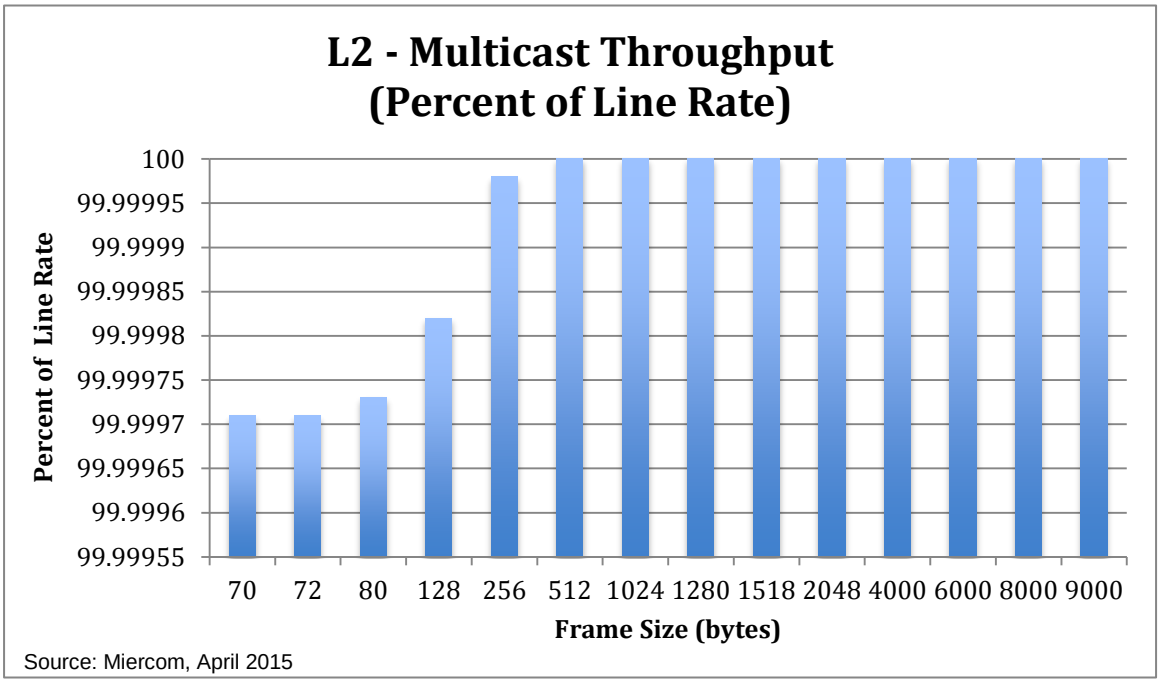
Unidirectional throughput, by frame size, from a single wire-speed source traffic stream, replicated and delivered to all the other 959 x 10GE ports on all 20 Leaf switches.

How We Did It

For each frame-size iteration, unidirectional test traffic was issued at 10 Gigabit/s to Port 1 on Leaf switch 1 (source), where it was replicated and delivered to all the other 959 "receive" Leaf switch ports – the 47 other ports on Leaf switch 1, and all 48 ports on Leaf switches 2 through 20 across the ACI fabric. The input traffic used 100 multicast groups.

Results

The graph below summarizes the throughputs achieved, by frame size, as a percent of wire speed. Note that more than 99.97 percent of wire speed was achieved even at the smallest, 70-byte packet size. The throughputs were slightly less than 100 percent due to the added ACI overhead for tunneling (the VXLAN header).



Source: Miercom, April 2015

L2 - Leaf Pairing, Unicast Latency

What's Measured

The minimum, average and maximum latency resulting from RFC 2544-based Leaf-pairing throughput testing.

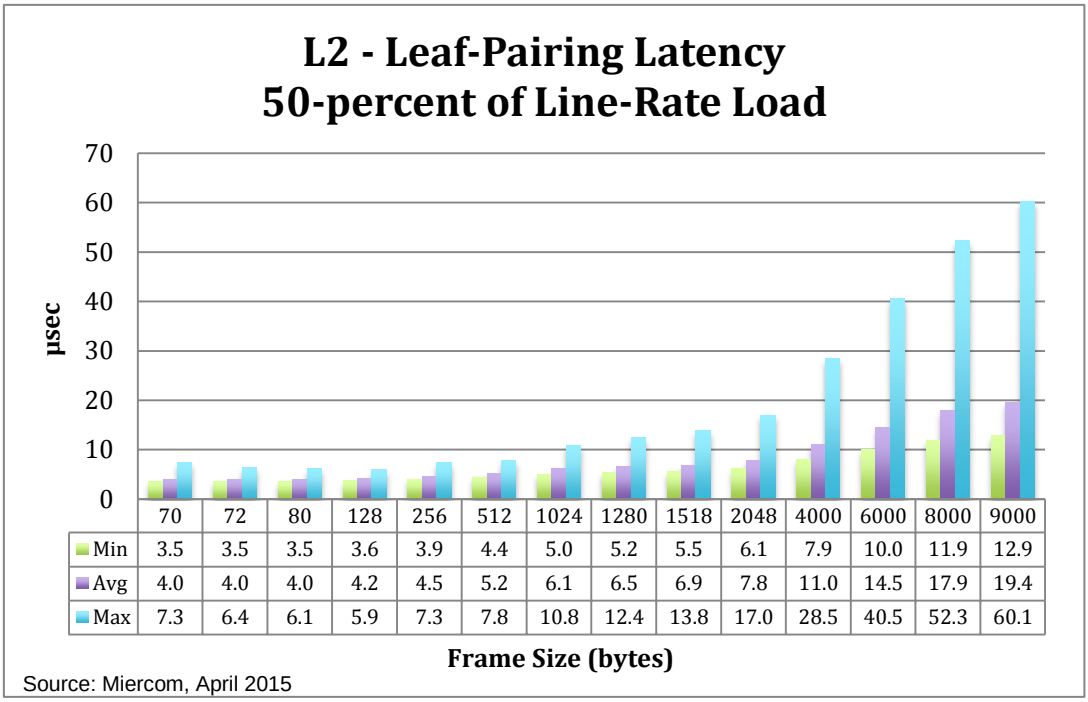
How We Did It

The set-up for these latency measurements is the same as in the earlier paired-Leaf-switch throughput testing: Leaf switches were paired, such that traffic sent in on ports 1 through 48 of Leaf Switch 1 were forwarded to ports 1 through 48 of Leaf Switch 11, and vice versa in the opposite direction. Traffic was thus sent, bidirectionally, between 10 pairs of Leaf switches, and in all cases through Spine switches and the ACI fabric.

Latency is measured automatically by the Spirent testers for all throughput tests. The latency measurements are reported here for 50-percent and 70-percent of line-rate throughputs. There was some small frame loss at 70 percent of bidirectional throughput, which is because that the traffic load exceeded the theoretical max throughput of the fabric links between the spine and the leaf switches after the VXLAN encapsulation.. Generally, latency measurements are not valid where loss occurs because the average and maximum latencies are skewed very high due to the dropped frames.

Results

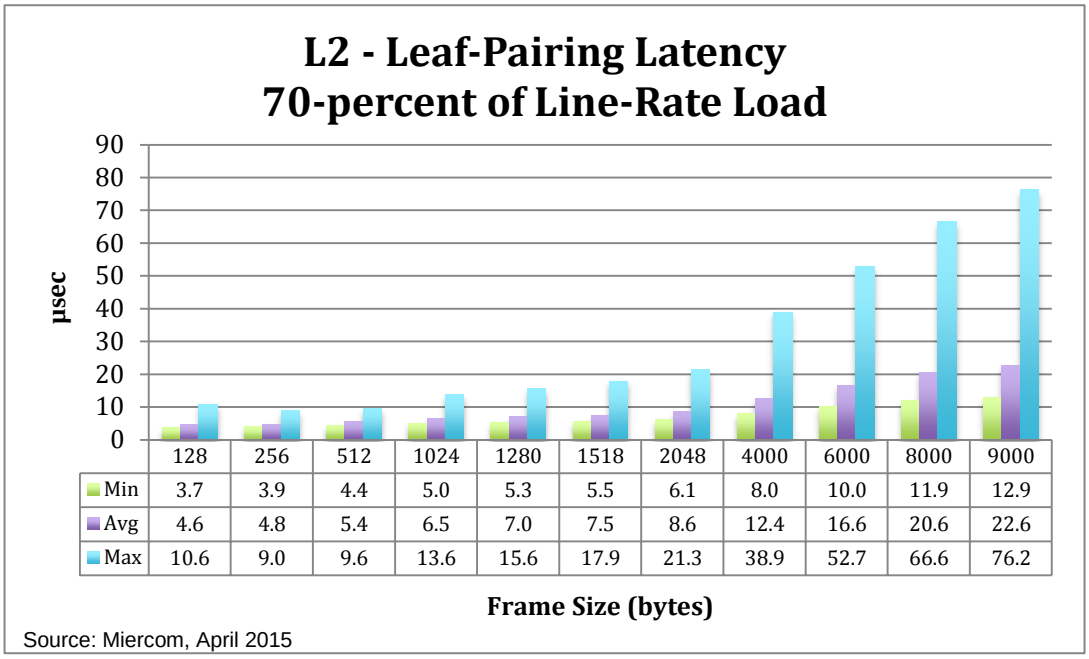
The first graph below shows latencies experienced, by frame size, over the Cisco ACI fabric, in microseconds (µsec), during the paired-Leaf-switch testing for 50-percent line-rate load.



	70	72	80	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
Min	3.5	3.5	3.5	3.6	3.9	4.4	5.0	5.2	5.5	6.1	7.9	10.0	11.9	12.9
Avg	4.0	4.0	4.0	4.2	4.5	5.2	6.1	6.5	6.9	7.8	11.0	14.5	17.9	19.4
Max	7.3	6.4	6.1	5.9	7.3	7.8	10.8	12.4	13.8	17.0	28.5	40.5	52.3	60.1

Source: Miercom, April 2015

There was no loss in the 50-percent-of-line-rate, paired-Leaf testing, so the minimum, average and maximum latency values are all valid. We note that the average latency for frame sizes up through 2,048 bytes is less than 8 microseconds, which is remarkable since all packets traversed at least three switches in the ACI fabric. Average latency starts to rise above 10 microseconds only for Jumbo frames of 4,000 bytes or more, and even then never exceeds 20 microseconds.



	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
Min	3.7	3.9	4.4	5.0	5.3	5.5	6.1	8.0	10.0	11.9	12.9
Avg	4.6	4.8	5.4	6.5	7.0	7.5	8.6	12.4	16.6	20.6	22.6
Max	10.6	9.0	9.6	13.6	15.6	17.9	21.3	38.9	52.7	66.6	76.2

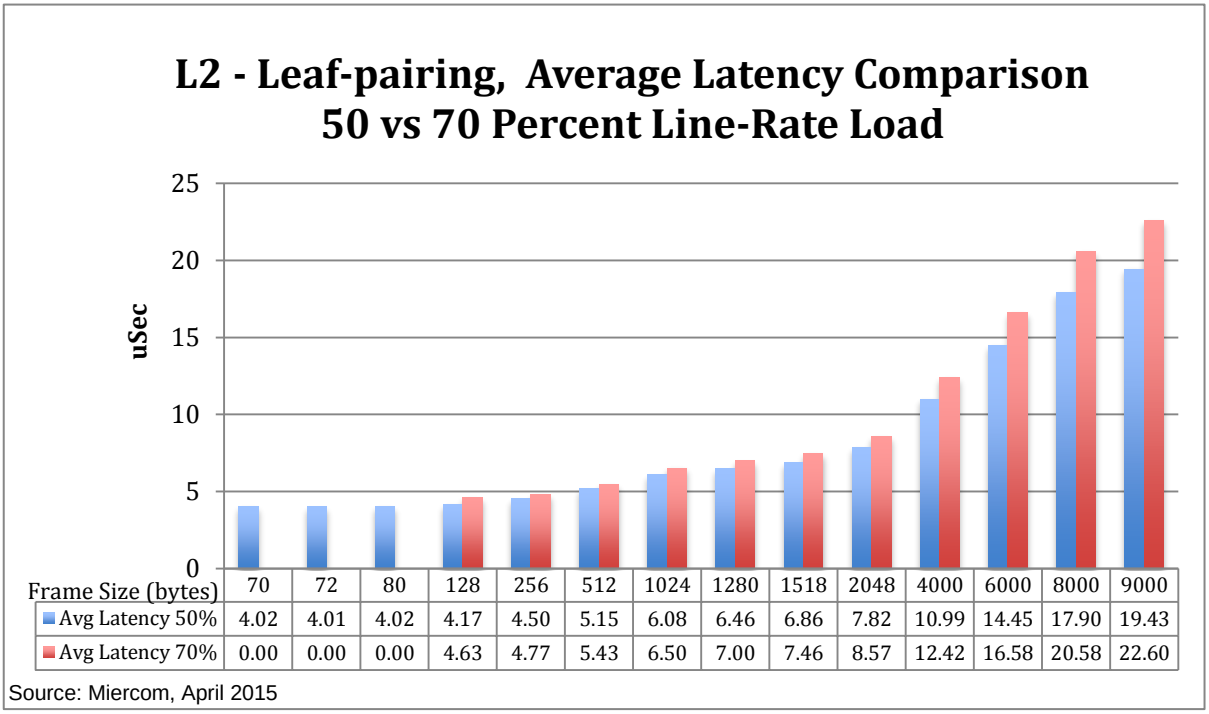
Source: Miercom, April 2015

The above graph shows latencies in the same test environment (paired Leaf switches), but with 70 percent of bidirectional traffic load. Note that the smallest frame sizes – 70, 72 and 80 bytes – are not included in the latency graph, since some amount of loss occurred at those frame sizes at 70-percent load which exceeded the theoretical max throughput for VXLAN traffic for those packet sizes.

The below table shows the amount of frame losses experienced by the smallest frame sizes during the 70-percent load test.

Frame Size (bytes)	Minimum Frame Loss (%)
70	8.46
72	7.74
80	5.07

The following graph shows the average latencies, by frame size, side by side for the 50- and 70-percent loads. Again, the smallest frame sizes have been omitted from the 70-percent bidirectional traffic load results since some loss occurred with those frame sizes.



Frame Size (bytes)	70	72	80	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
Avg Latency 50%	4.02	4.01	4.02	4.17	4.50	5.15	6.08	6.46	6.86	7.82	10.99	14.45	17.90	19.43
Avg Latency 70%	0.00	0.00	0.00	4.63	4.77	5.43	6.50	7.00	7.46	8.57	12.42	16.58	20.58	22.60

Source: Miercom, April 2015

What the comparison graph shows is that there is little load-based variation in average latency for traffic traversing the ACI fabric. For all frame sizes less than 2,048 bytes, the average latencies are under 9 microseconds – again, remarkable, since all traffic crossed three switches and two links through the Cisco ACI fabric. The variation in average latency based on load was miniscule: less than one microsecond for frames up to the standard 1,518-byte frame size, to just a few microseconds for the largest Jumbo frames.

L2 - Multicast Latency

What's Measured

The minimum, average and maximum latency resulting from RFC 3918-based Multicast throughput testing.

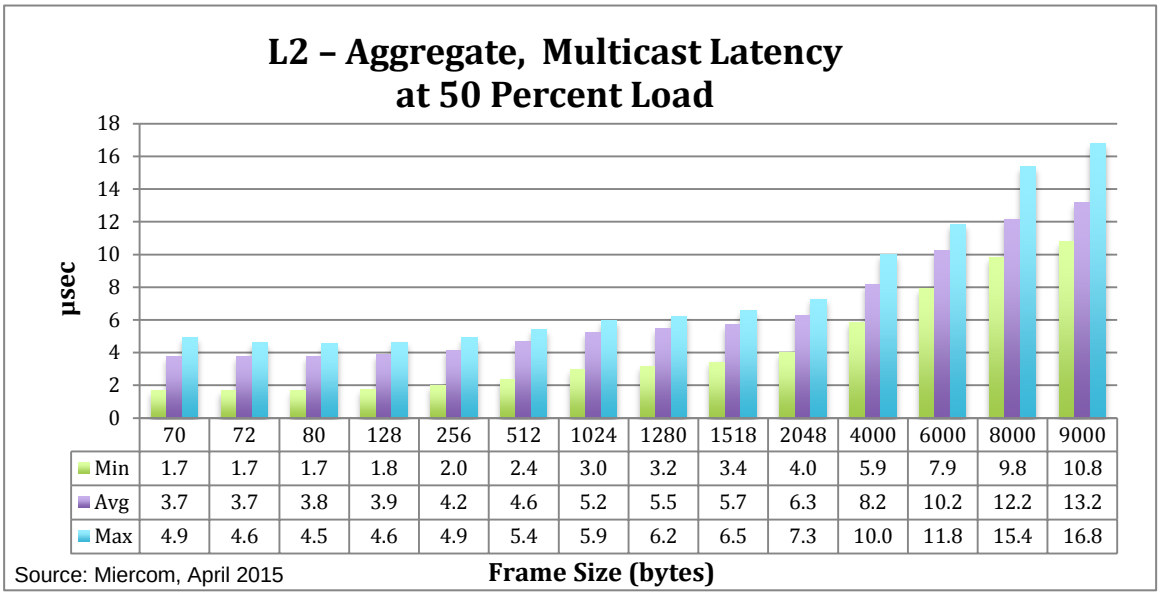
How We Did It

The set-up for these latency measurements is the same as for the earlier Multicast throughput testing: Unidirectional test traffic was issued to Port 1 on Leaf switch 1 (source), where it was replicated and delivered to all the other 959 "receive" Leaf switch ports – the 47 other ports on Leaf switch 1, as well as all 48 ports on Leaf switches 2 through 20, across the ACI fabric. The input traffic used 100 multicast groups.

The multicast latency measurements are reported here for 50-percent and 70-percent of line-rate source input (port 1 on Leaf 1 switch).

Results

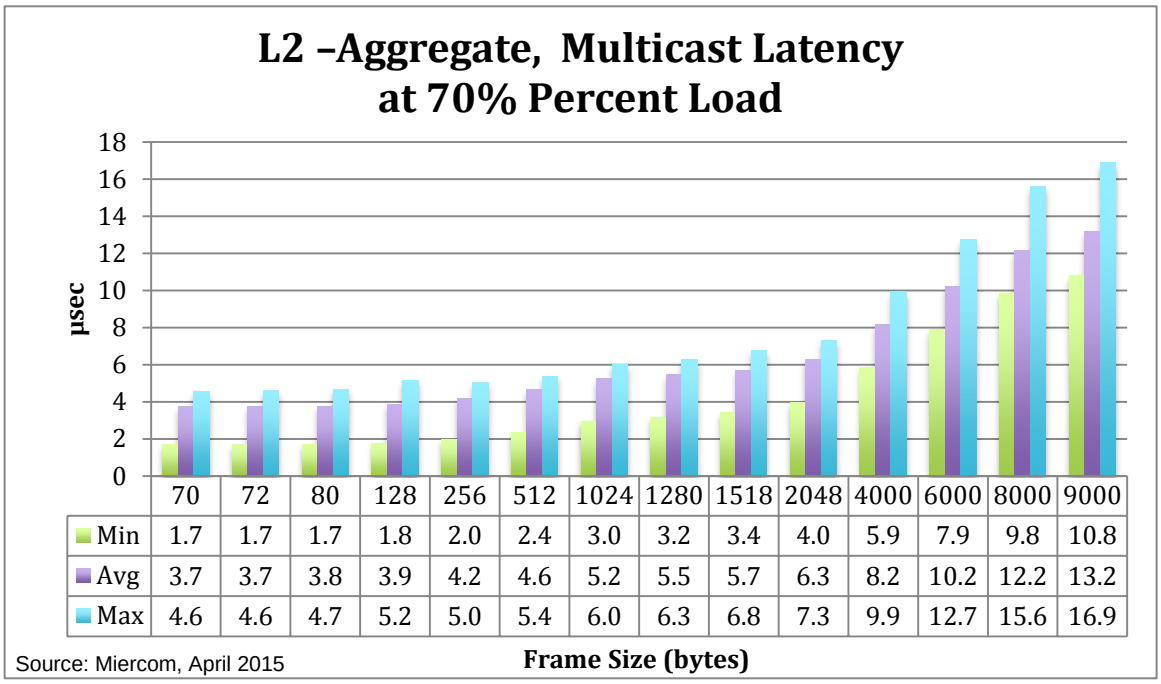
The graph below summarizes the minimum, average and maximum latencies experienced by frame size, for all multicast traffic in aggregate – traffic delivered to the other ports of Leaf switch 1, as well as all the traffic delivered through the ACI fabric to the other 912 Leaf switch ports – for a 50-percent of wire speed, source traffic stream. All traffic was replicated and forwarded to all ports without loss.



	70	72	80	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
Min	1.7	1.7	1.7	1.8	2.0	2.4	3.0	3.2	3.4	4.0	5.9	7.9	9.8	10.8
Avg	3.7	3.7	3.8	3.9	4.2	4.6	5.2	5.5	5.7	6.3	8.2	10.2	12.2	13.2
Max	4.9	4.6	4.5	4.6	4.9	5.4	5.9	6.2	6.5	7.3	10.0	11.8	15.4	16.8

The multicast latency results for 50-percent load are similar to the latencies seen for Leaf-pair and full-mesh traffic: latencies are consistent, with fairly little variability, and low. As seen before, the average latency for all frames smaller than 4,000 bytes is under 8 microseconds. Average latency for even the largest Jumbo frames is still just 13 microseconds.

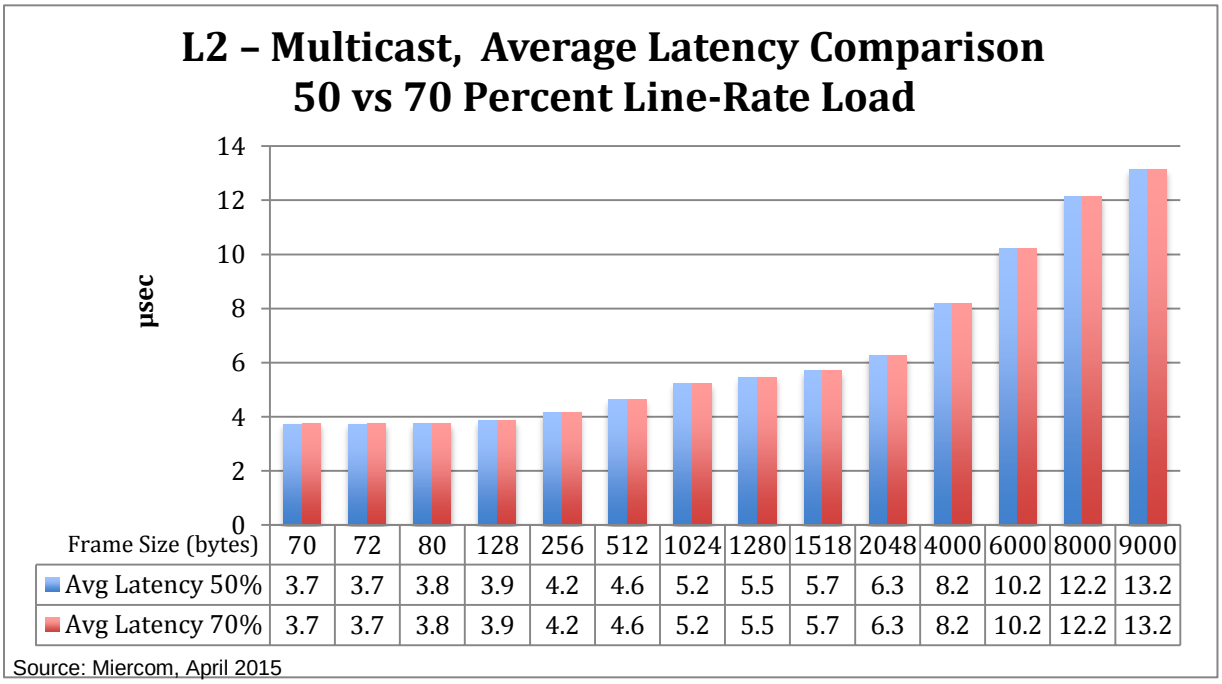
The below graph shows minimum, average and maximum latencies by frame size for aggregate multicast traffic – within the same Leaf switch 1, as well as across the ACI fabric to all other receive switch ports – for a 70-percent of wire-speed load.



	70	72	80	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
Min	1.7	1.7	1.7	1.8	2.0	2.4	3.0	3.2	3.4	4.0	5.9	7.9	9.8	10.8
Avg	3.7	3.7	3.8	3.9	4.2	4.6	5.2	5.5	5.7	6.3	8.2	10.2	12.2	13.2
Max	4.6	4.6	4.7	5.2	5.0	5.4	6.0	6.3	6.8	7.3	9.9	12.7	15.6	16.9

Source: Miercom, April 2015

As with the 50-percent multicast load, there was no frame loss with the 70-percent load. The average multicast latency seems unaffected by load. The graph below compares the average latency for multicast traffic at 50- and 70-percent line-rate loads.



Frame Size (bytes)	70	72	80	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
Avg Latency 50%	3.7	3.7	3.8	3.9	4.2	4.6	5.2	5.5	5.7	6.3	8.2	10.2	12.2	13.2
Avg Latency 70%	3.7	3.7	3.8	3.9	4.2	4.6	5.2	5.5	5.7	6.3	8.2	10.2	12.2	13.2

Source: Miercom, April 2015

6 - Layer-3 Throughput and Latency Tests

The same extensive battery of tests applied with Layer-2 frames over the Application Centric Infrastructure fabric was then repeated with Layer-3 traffic – that is, IP packets. The results of the Layer-3 throughput and latency tests are reported in this section

L3 - Leaf Pairing, Unicast Throughput

What's Measured

Bidirectional throughput, as a percent of the theoretical ACI-fabric maximum, for traffic between the 48 x 10GE ports of one Leaf switch, with the corresponding 48 ports on another Leaf switch.

How We Did It

As with Layer-2 testing, Leaf switches were paired, such that L3 traffic sent in on ports 1 through 48 of Leaf Switch 1 were forwarded to ports 1 through 48 of Leaf Switch 11, and vice versa in the opposite direction. Traffic was thus sent, bidirectionally, between 10 pairs of Leaf switches. Every packet went through a Spine switch to its destination port.

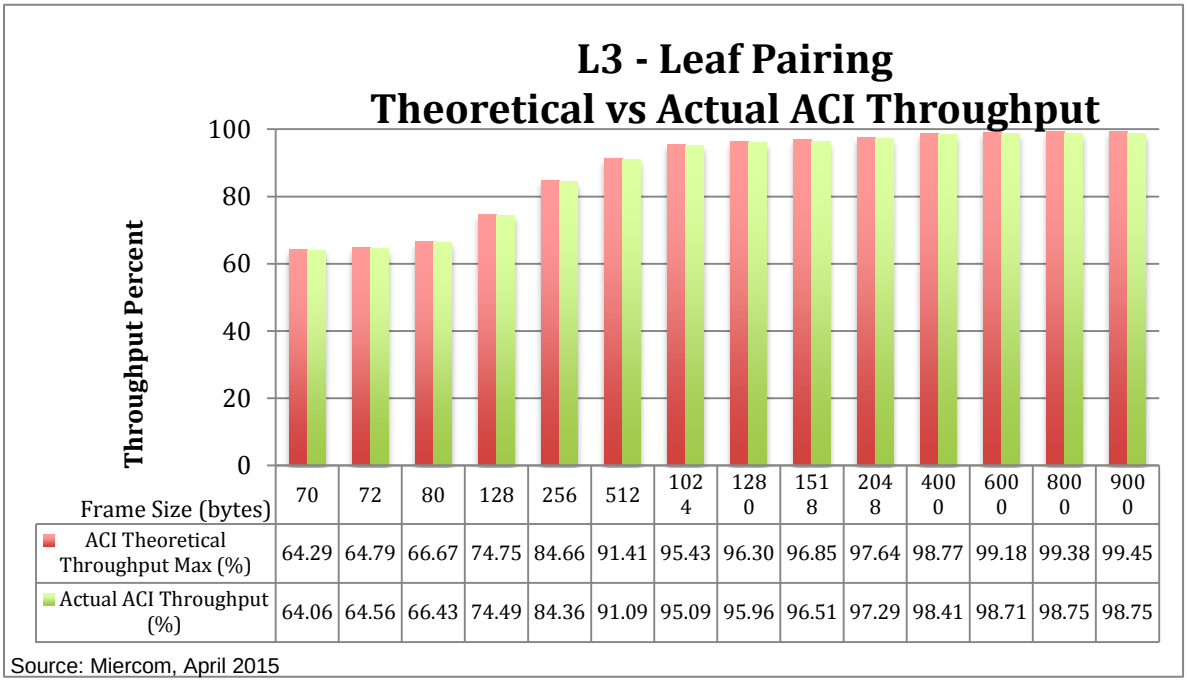
Results

The table below summarizes the throughputs achieved, by packet size, over the Cisco ACI fabric, as a percent of the line rate delivered by the Spirent testers.

L3 - Leaf Pairing – Unicast Throughput (per RFC 2544)

Frame Size (Bytes)	Tx Line Rate (Percent)	ACI Theoretical Throughput Max (Percent)	Actual ACI Throughput (Percent)	Percent of Theoretical ACI Max Throughput Achieved
70	100	64.3	64.1	99.7
72	100	64.8	64.6	99.7
80	100	66.7	66.5	99.7
128	100	74.7	74.5	99.7
256	100	84.7	84.4	99.7
512	100	91.4	91.1	99.7
1024	100	95.4	95.1	99.7
1280	100	96.3	96.0	99.7
1518	100	96.9	96.5	99.7
2048	100	97.6	97.3	99.7
4000	100	98.8	98.4	99.7
6000	100	99.2	98.8	99.6
8000	100	99.4	98.9	99.5
9000	100	99.4	98.9	99.5

The results indicate that from 99.5 to 99.7 percent of the wire speed maximum throughput within the ACI fabric was achieved.



Frame Size (bytes)	70	72	80	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
ACI Theoretical Throughput Max (%)	64.29	64.79	66.67	74.75	84.66	91.41	95.43	96.30	96.85	97.64	98.77	99.18	99.38	99.45
Actual ACI Throughput (%)	64.06	64.56	66.43	74.49	84.36	91.09	95.09	95.96	96.51	97.29	98.41	98.71	98.75	98.75

Source: Miercom, April 2015

L3 - Leaf Pairing, Unicast Latency

What's Measured

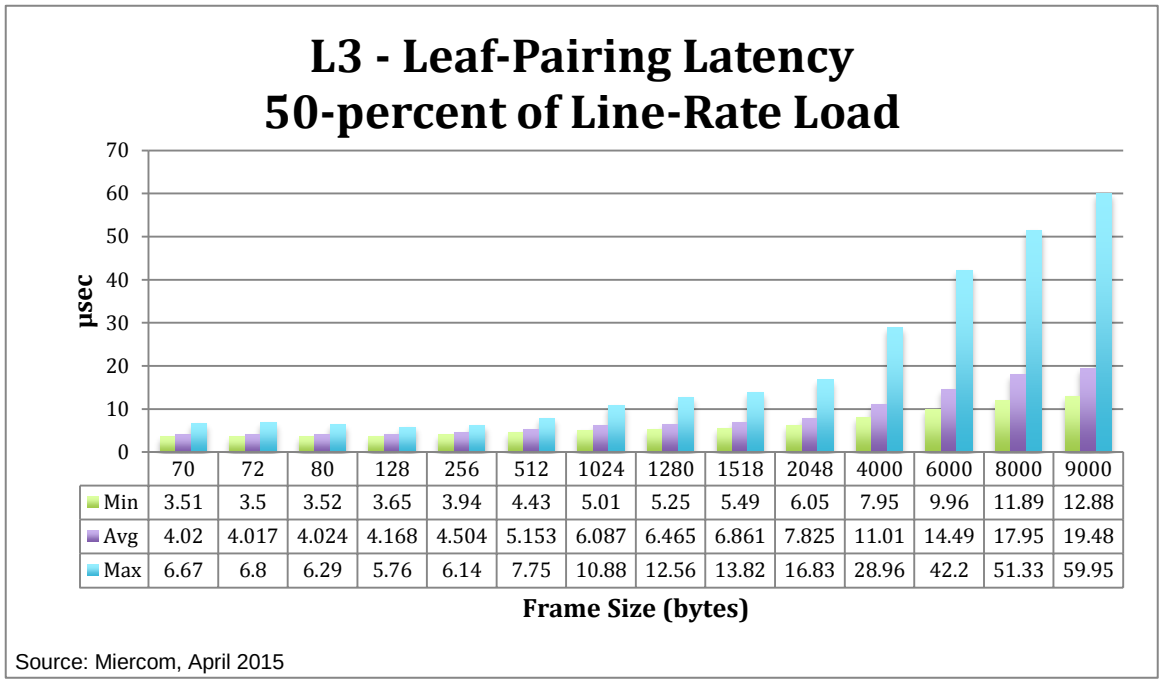
The latency, by frame size, of bidirectional L3 IP traffic delivered between all ports of paired Leaf switches. All traffic traversed two links and a Spine switch across the ACI fabric. Latency results are shown below for traffic loads of 50 and 70 percent of wire speed.

How We Did It

Leaf switches were paired, such that L3 traffic sent in on ports 1 through 48 of Leaf Switch 1 were forwarded to ports 1 through 48 of Leaf Switch 11, and vice versa in the opposite direction. L3 IP traffic was thus sent, bidirectionally, between 10 pairs of Leaf switches.

Results

The following graph summarizes the latencies experienced, by packet size, for a 50-percent-of-bidirectional wire-speed load of IP packets traversing the Cisco ACI fabric. At the 50-percent load level there was no packet loss.

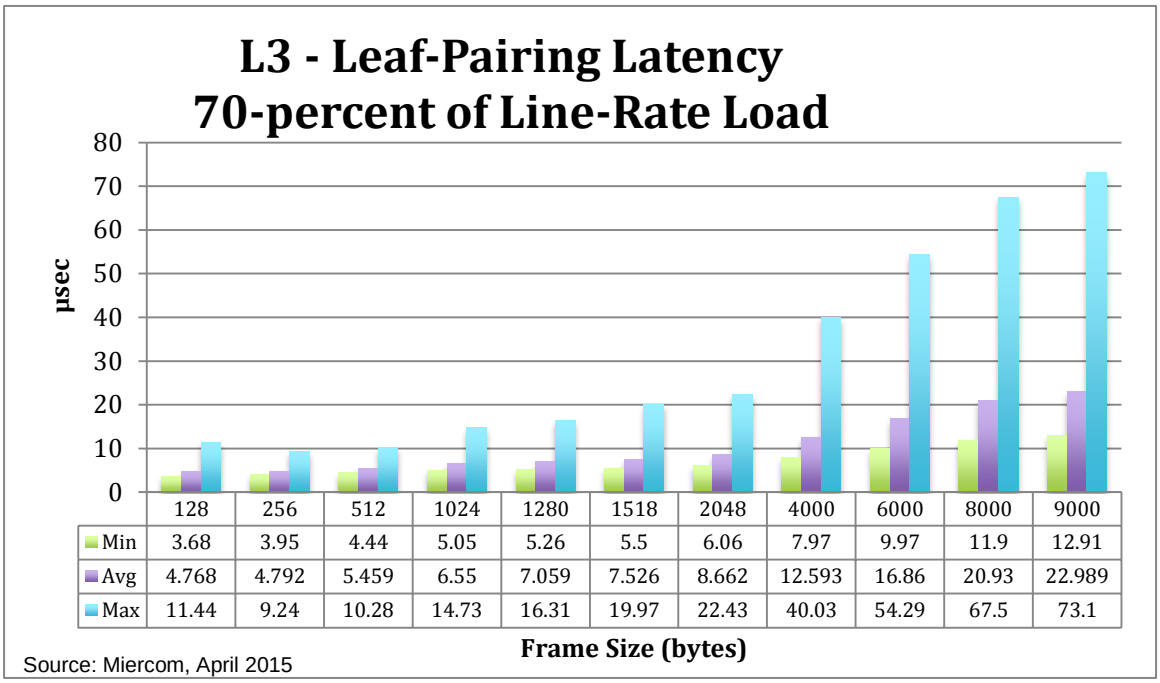


	70	72	80	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
Min	3.51	3.5	3.52	3.65	3.94	4.43	5.01	5.25	5.49	6.05	7.95	9.96	11.89	12.88
Avg	4.02	4.017	4.024	4.168	4.504	5.153	6.087	6.465	6.861	7.825	11.01	14.49	17.95	19.48
Max	6.67	6.8	6.29	5.76	6.14	7.75	10.88	12.56	13.82	16.83	28.96	42.2	51.33	59.95

Source: Miercom, April 2015

The latencies of L3 IP packets traversing the ACI fabric in leaf-paired testing is nearly identical to the latency of L2 frames in the same test environment. As before, average latency is less than 8 microseconds for packet sizes up to 2,048 bytes – remarkable since all packets traversed two backbone links, a Spine switch and two Leaf switches.

The graph below shows packet latencies in the same environment for 70-percent load.



	128	256	512	1024	1280	1518	2048	4000	6000	8000	9000
Min	3.68	3.95	4.44	5.05	5.26	5.5	6.06	7.97	9.97	11.9	12.91
Avg	4.768	4.792	5.459	6.55	7.059	7.526	8.662	12.593	16.86	20.93	22.989
Max	11.44	9.24	10.28	14.73	16.31	19.97	22.43	40.03	54.29	67.5	73.1

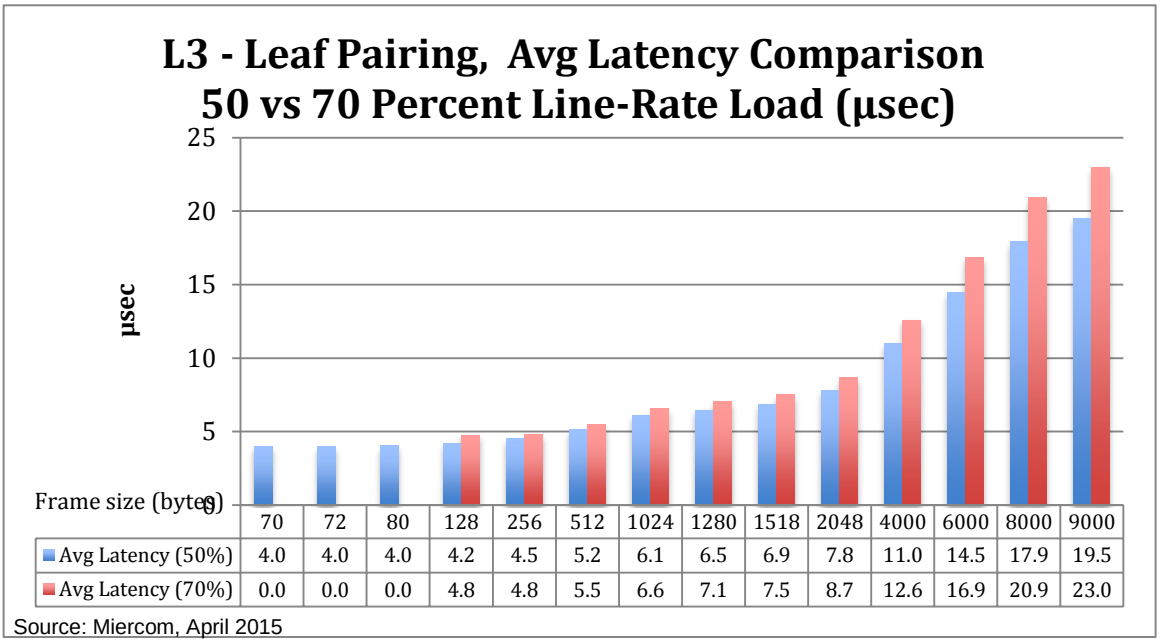
Source: Miercom, April 2015

Note that latencies for the smallest packet sizes – 70, 72 and 80 bytes – have been omitted from the 70-percent-load graph. That is because some packet loss occurred with these small packet sizes. Latency measurements when loss occurs are invalid because lost packets skew the automated calculations of latency very high.

The table below shows the amount of loss that occurred with these very small packet sizes, and the effect the loss had on raising the average and maximum latency values.

	Percent		µsec		
Frame Size	Intended Load	Frame Loss	Minimum Latency	Average Latency	Maximum Latency
70	70	8.48	3.72	110.31	299.60
72	70	7.77	3.67	111.74	304.97
80	70	5.09	3.72	117.67	315.89

Put side-by-side, the graph below shows average latency for the Leaf-paired environment for 50- and 70-percent line-rate packet loads. We note, again, that for 70-percent load the average latencies for the smallest packet sizes have been omitted because loss occurred for those tests.



The results show slightly higher average latencies for packets in the 70-pecent load environment, but the differences are nominal. The difference in average latency was less than one microsecond for all packet sizes up to and including 2,048 bytes, and just a few microseconds for the largest Jumbo packets.

7 - Convergence Tests

This battery of tests was devised to ascertain the resiliency of the Cisco Application Centric Infrastructure fabric. Specifically, tests were conducted to quantify the extent of data loss resulting from:

- Failure of a fabric link
- Failure of a Spine switch
- Failure of one or more APICs

Fabric Link Failure and Recovery

What's Measured

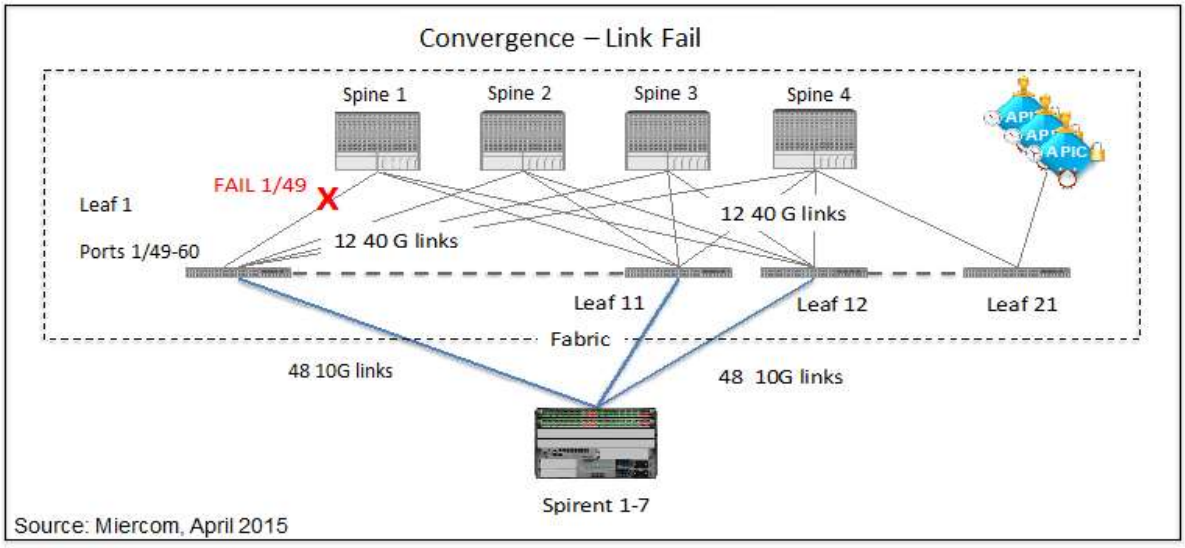
The amount of time that a transmission path is unavailable within the ACI fabric as a result of a failed link, before reconvergence (an alternate path and re-routing is implemented).

How We Did It

Several link-outage scenarios were created:

- Shut-down of a backbone link (40GE uplink) from the Leaf switch
- Shut-down of a backbone link from the Spine switch
- Physical unplugging of a backbone link.

To determine the effect of the link outage, the Spirent test system sent bidirectional traffic at 50 percent of line rate between the Leaf 1 switch and the Leaf 12 switch (see diagram below). Then test traffic was applied between all 48 x 10GE ports on both switches, through all four Spine switches.



Then port 1/49 on Leaf Switch 1 (a 40GE uplink to Spine switch 1) was failed, and the packet-loss duration for each direction was measured – for three different scenarios:

1. Shutting down the link on Leaf switch 1.
2. Shutting down the link on Spine switch 1.
3. Physically disconnecting the link.

The path outage from Leaf switch 1 to Leaf switch 12 is termed a direct spine link failover (see below table), while the reverse path, from Leaf switch 12 back to Leaf switch 1, is called an indirect spine link failover. Three test trials were run for each link-failover scenario.

Results

The table below summarizes the duration of the path outage before reconvergence for each of the three scenarios, for the three trials. All times shown are in milliseconds.

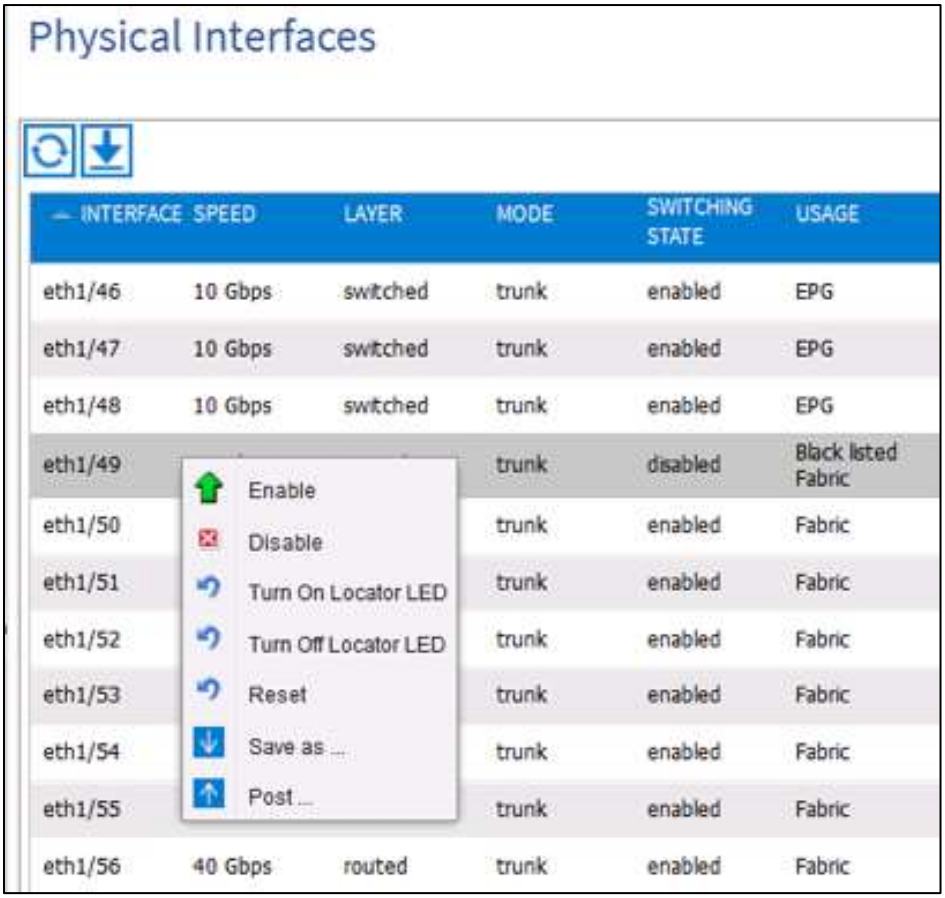
	1. Shutdown link on Leaf switch	2. Shutdown link on Spine switch	3. Physically disconnect link
	Packet loss duration (msec)	Packet Loss duration (msec)	Packet Loss duration (msec)
Trial 1			
Direct Spine Link Failover	4.296	0.233	0.002
Indirect Spine Link Failover	419.465	136.873	346.000
Trial 2			
Direct Spine Link Failover	6.439	0.234	0.002
Indirect Spine Link Failover	415.037	101.614	368.429
Trial 3			
Direct Spine Link Failover	5.142	0.236	0.002
Indirect Spine Link Failover	411.923	152.258	362.264

The results show that the duration of the outage is very dependent on the direction of data flow in relation to the failed link.

The shortest outages occur in the Leaf 1-to-Leaf 12 direction in this test environment. Physically unplugging the link (an unexpected loss of connectivity) is resolved in just two microseconds. Reconvergence takes a little longer after shutting down the link on the Spine node – an average of just 0.234 milliseconds (234 microseconds), and about 5 milliseconds when shut down from Leaf switch 1.

For the reverse direction, from Leaf switch 12 back to Leaf switch 1 through Spine switch 1, reconvergence takes a bit longer, but in no case longer than a half-second. Shutting down the link from the Spine switch incurs a path interruption of about 130 milliseconds. Unplugging the link incurs about a one-third second outage, and shutting it down at Leaf switch 1 results in the longest path outage, about 415 milliseconds.

Shutting down the interface for this testing is a straightforward process done via the APIC (Application Policy Infrastructure Controller). The screen for shutting down link 1/49 (eth1/49) on Leaf switch 1 is shown below:



Screen Shot of shut down

Spine Node Failure and Recovery

What's Measured

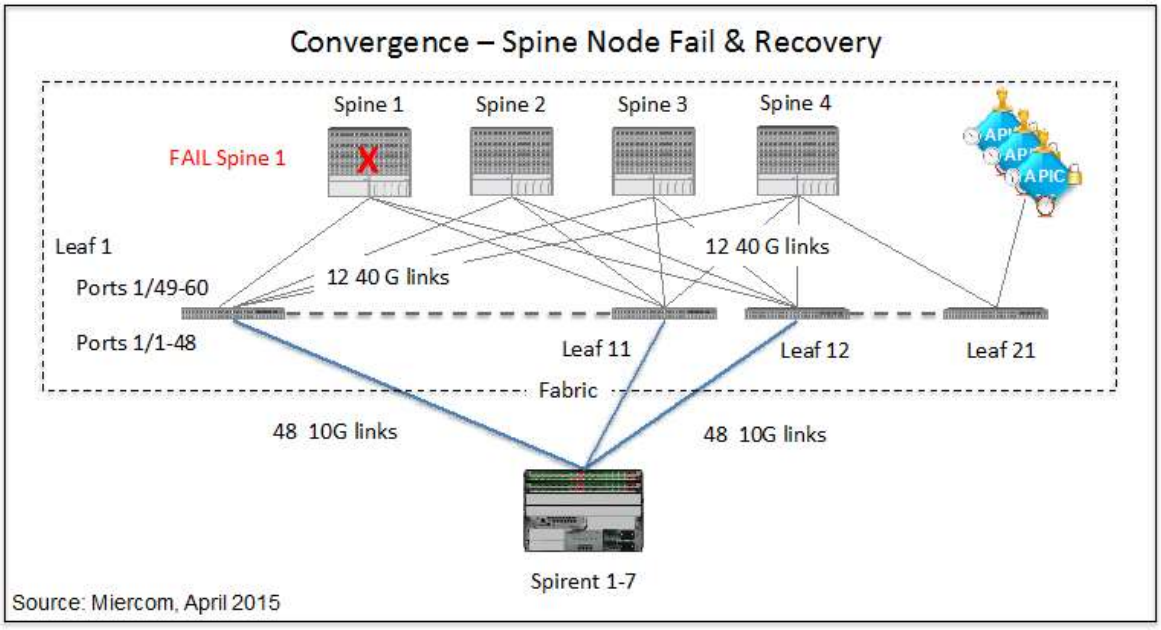
The amount of time that transmission paths are unavailable within the ACI fabric as a result of a failed Spine switch, before reconvergence (alternate paths and re-routing are implemented).

Afterwards, the same Spine switch is brought back up with a reload and its re-integration into the ACI fabric observed.

How We Did It

The Spirent test system was set to deliver a bidirectional traffic flow at 10 percent of line rate between Leaf switch 1 and Leaf switch 12. The flow of traffic in both directions through Spine switch 1 was confirmed (see below diagram).

Spine switch 1 is then powered down. The path-unavailability duration is then calculated based on packet loss before reconvergence.



Afterwards, the node is brought up with a reload, and any data-flow interruptions to the ACI fabric are noted.

Results

After Spine switch 1 was powered down, the path outage time before reconvergence was very brief, just 0.3 milliseconds.

The Spine switch was brought back up with a reload. We observed that the Spine switch did not accept traffic again until all connected links were ready to advertise the recovered routes. Then traffic through Spine switch 1 resumed. There was no loss of data from active traffic flows during the switch's recovery and re-integration into the fabric.

APIC (Controller) Failure – Two of Three Nodes

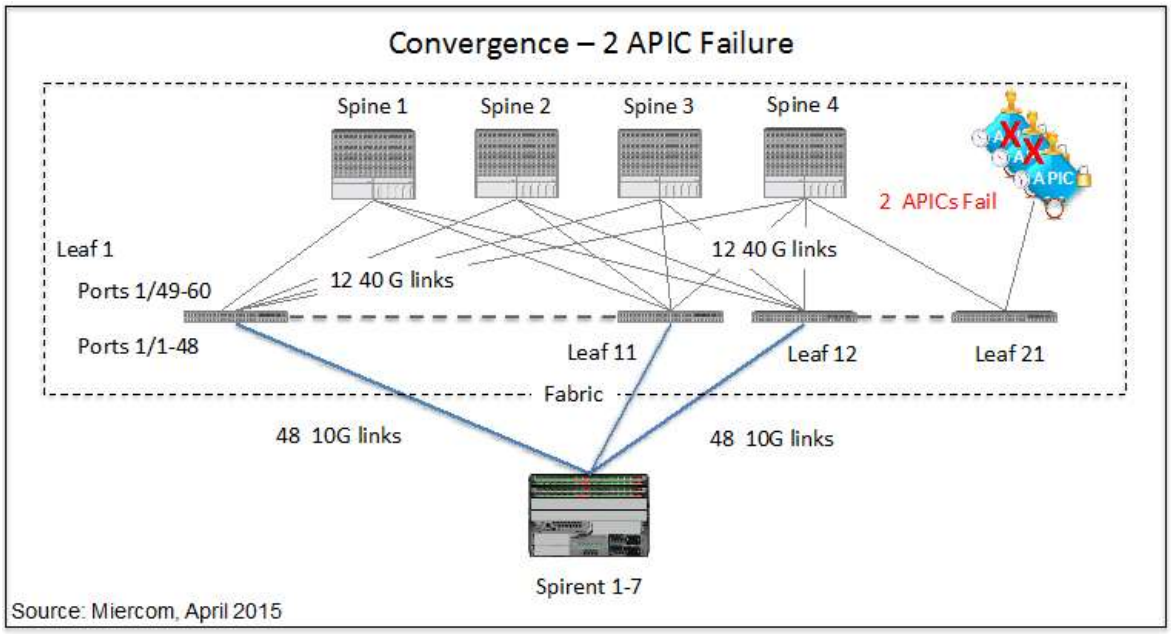
What's Measured

Any interruption to active traffic flows as a result of two Application Policy Infrastructure Controllers (APICs) being unexpectedly removed from the ACI fabric.

Afterwards, the APICs are reloaded and their re-integration into the ACI fabric observed.

How We Did It

Three APICs, their databases synchronized, are collectively running the ACI fabric. Traffic at 10 percent of line rate is run bidirectionally between paired Leaf switches. Then, two of the three APICs are disconnected from the network (see diagram).



Results

The disconnection of the two APICs had no impact on active traffic flows. However, via the third APIC we learned that the ACI fabric had become read-only. It turns out that a solitary APIC does not make configuration- change decisions by itself. A majority of APICs must decide on configuration changes. Since the third APIC sees that the other two APICs are down, the fabric is treated as read-only.

The third, remaining APIC does continue to monitor traffic in the fabric.

Once the two APICs are reconnected, they check with each other and re-synchronize their databases. There is no effect on active traffic flows when the two APICs come back on line.

All APIC Failure

What's Measured

Any interruption to active traffic flows as a result of all Application Policy Infrastructure Controllers (APICs) being unexpectedly removed from the ACI fabric.

How We Did It

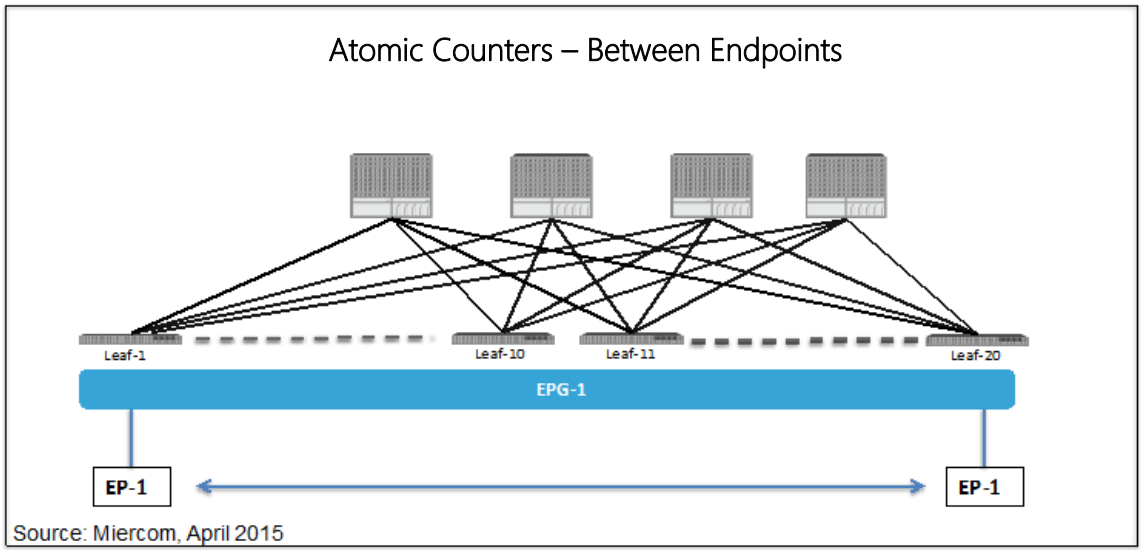
Three APICs, their databases synchronized, are collectively running the ACI fabric. Traffic at 10 percent of line rate is run bidirectionally between paired Leaf switches. Then, all three APICs are disconnected from the network (see diagram).

Results

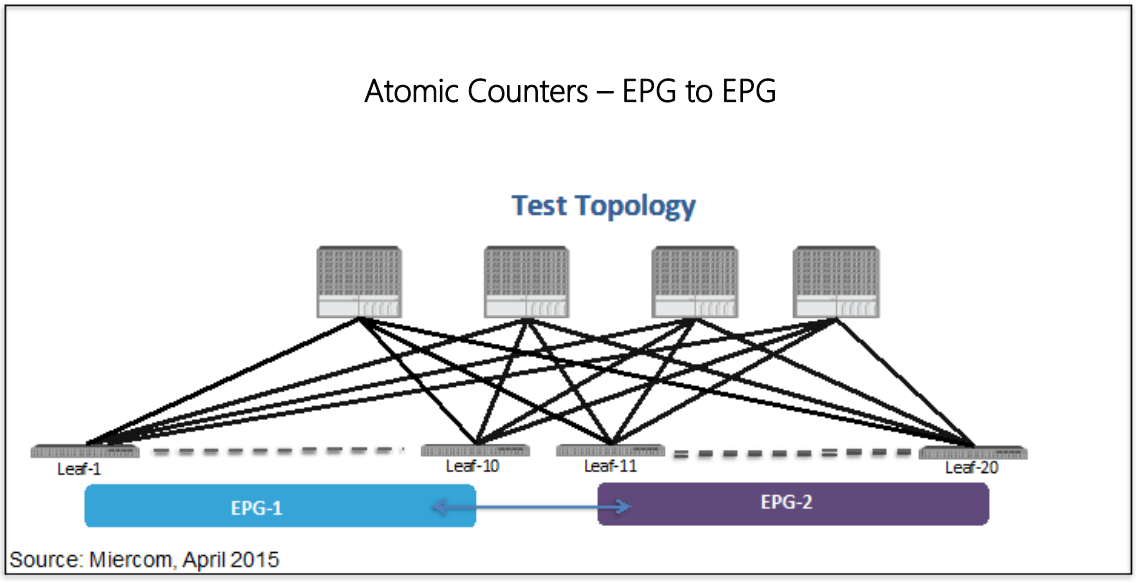
The ACI fabric continues to forward traffic, even with no APIC controllers connected to the fabric. There is no packet loss.

8 - Atomic Counters

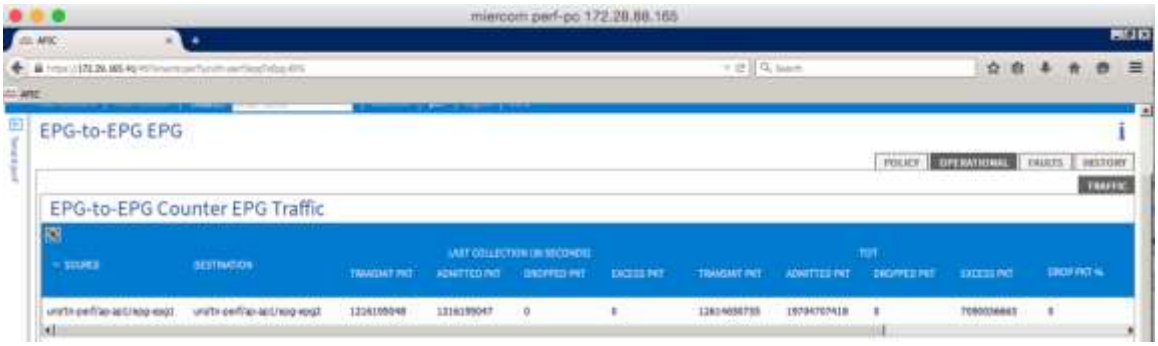
A key operational capability of the Cisco ACI is the ability to track traffic flow from many different perspectives. These "Atomic Counters" provide a way to gather statistics about traffic between a given pair of leaf switches, along a chosen network path, or even between chosen endpoint groups or endpoints for application specific counters. *Between any two endpoints, as well as between endpoint groups, Atomic Counters tally transmitted packets, admitted packets, dropped packets and excess packets.*



Atomic Counters also help APICs discover and maintain associations of endpoints in what are called Endpoint Groups, or EPGs.



The atomic counters provide troubleshooting tools for ACI's application layers. It enables quick debugging and isolation of application connectivity problems. Below is the APIC's view of traffic between a pair of endpoint groups.



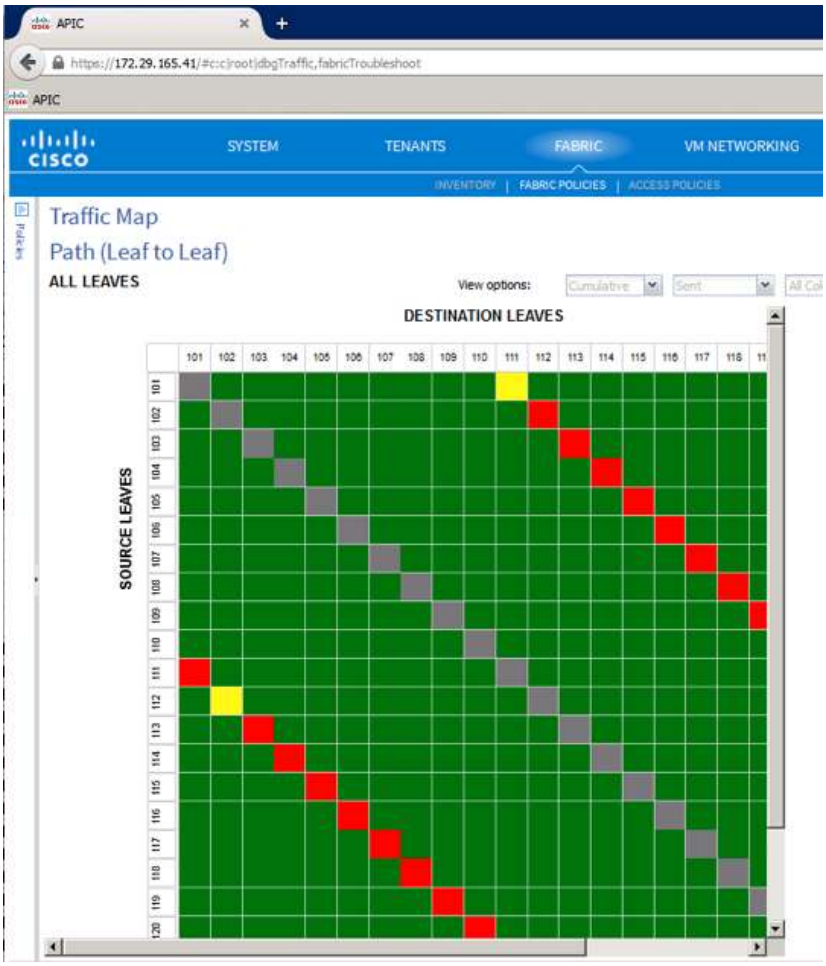
Our testing exchanged known traffic volumes between endpoints and between endpoint groups. The APIC interface lets users readily view traffic counts. Our tests found the Atomic Counters tracking of activity within the ACI fabric to be accurate and extensive.

9 – APIC Tools

Throughout the testing we had ample opportunity to exercise the Application Policy Infrastructure Controller (APIC). Some of the APIC software interfaces and tools are especially noteworthy in this report.

Traffic Map

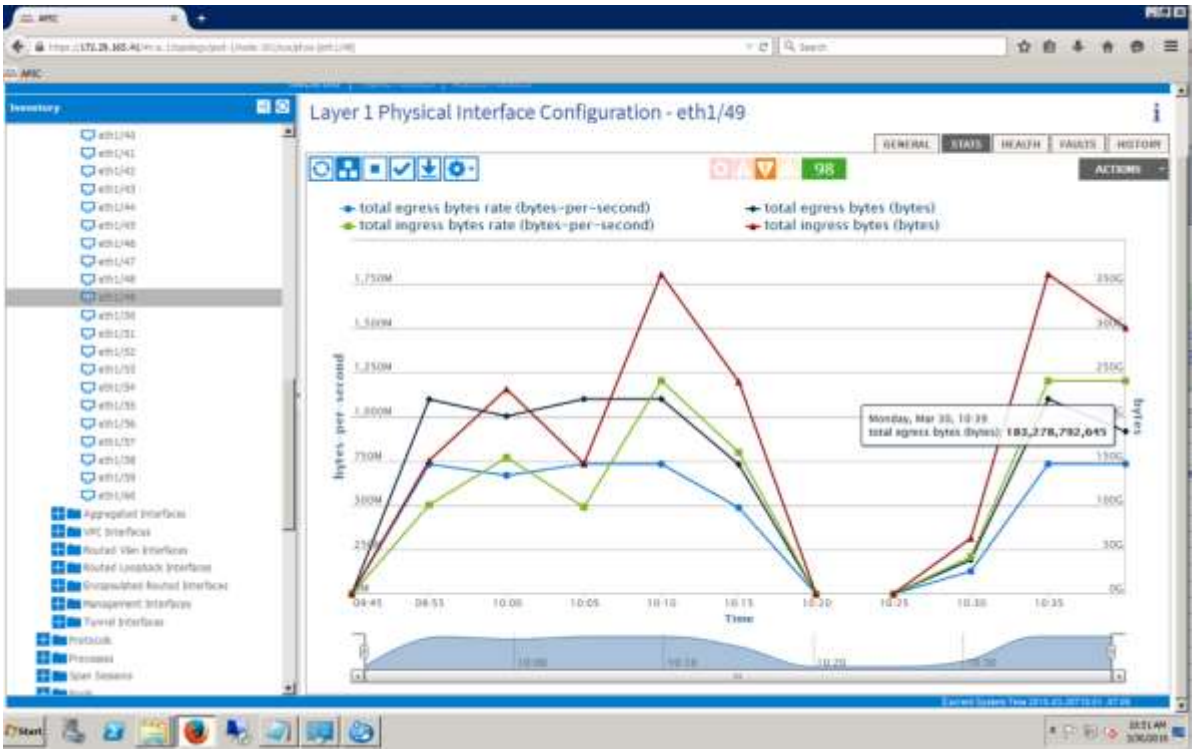
Among its many chores, the APIC manages the configuration changes for the VXLAN overlay and monitors traffic at the network level. The Traffic Map is one of the novel APIC monitoring tools.



The screen shot of the Traffic Map display shows traffic-flow activity during our Leaf-Pair testing. The easy-to-understand graphic shows active traffic flows between Leaf-switch port pairs. The connectivity squares are color coded, where red indicates greater than 70 percent path-capacity utilization, yellow is greater than 20 percent and green is under 20 percent. Port pair 111-101, for example, colored red, has a traffic flow greater than 70 percent of line rate.

Physical Interface Configuration - Port Disable / Re-enable

This APIC tool below shows inbound and outbound traffic history through port "eth1/49," used in our failover and reconvergence testing. The chart display shows link activity and traffic volume as it goes through the "Link Failure and Restore" test.

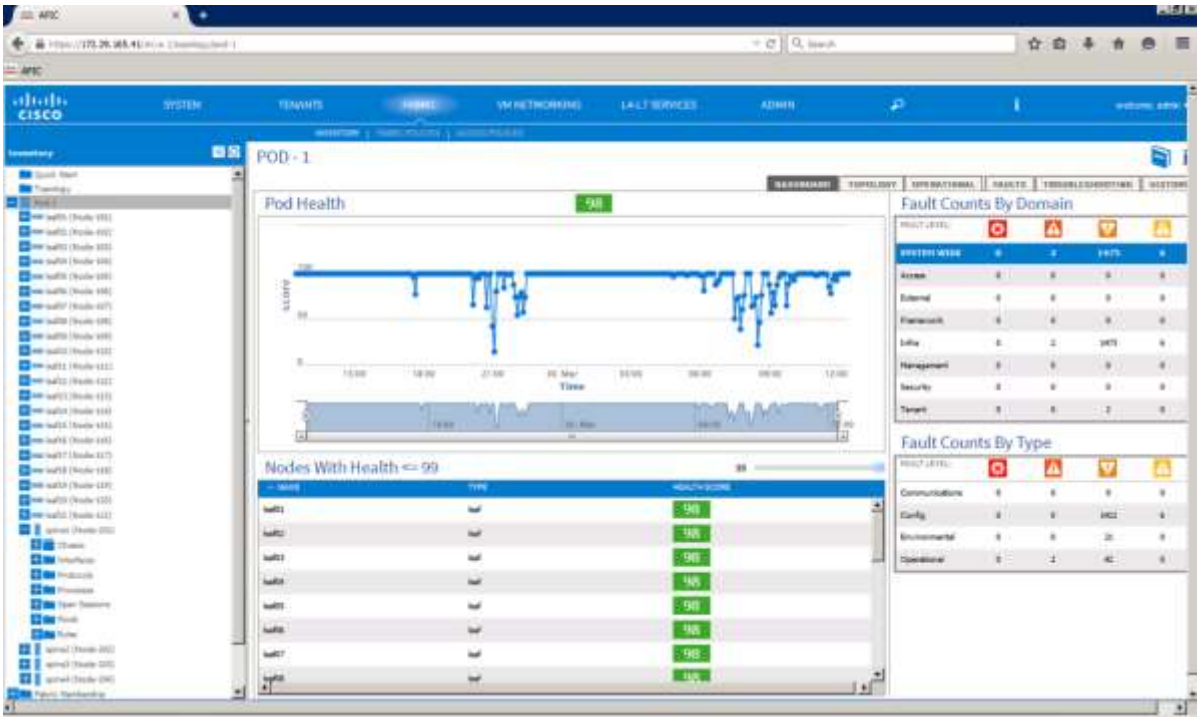


Screen shot of traffic history.

ACI Health Score

The ACI Health Score tool below is another application we used extensively and is deserving of special note. The high-level view provides a good graphical representation of the fabric's overall health, or any specific component.

A timeline readily shows when noteworthy events occurred. The user can drill down, by node or fault, for additional technical details and follow-up.



Time line screen shot that can be expanded as needed.

10 - Independent Evaluation

This report was sponsored by Cisco Systems, Inc. The data was obtained completely and independently as part of Miercom's network-product-performance analyses.

11 - About Miercom

Miercom has published hundreds of network-product-comparison analyses – many made public, appearing in leading trade periodicals and other publications, and many confidential, for internal use only. Miercom's reputation as the leading, independent product test center is undisputed.

Private test services available from Miercom include competitive product analyses, as well as individual product evaluations. Miercom test methodologies are generally developed collaboratively with the client, and feature comprehensive certification and test programs including: Certified Interoperable, Certified Reliable, Certified Secure and Certified Green. Products may also be evaluated under the Performance Verified program, the industry's most thorough and trusted assessment for product usability and performance.

12 - Use of This Report

Every effort was made to ensure the accuracy of the data in this report. However, errors and/or oversights can nevertheless occur. The information documented in this report may depend on various test tools, the accuracy of which is beyond our control. Furthermore, the document may rely on certain representations by the vendors that were reasonably verified by Miercom, but are beyond our control to verify with 100-percent certainty.

This document is provided "as is" by Miercom, which 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. Miercom is not liable for damages arising out of or related to the information contained in this report.