



Detailed Lab Testing Report
DR140802D

**Buffer Performance Testing
Results and Analysis
Cisco Nexus 9396PX Switch
Arista 7150S Switch**

26 August 2014

Miercom

<http://www.miercom.com>

Contents

| | | |
|-----|--|----|
| 1 | Executive Summary..... | 3 |
| 2 | Buffer Management Study | 4 |
| 3 | Products Tested..... | 5 |
| 3.1 | Cisco Systems 9396PX | 5 |
| 3.2 | Arista Networks 7150S..... | 6 |
| 4 | Test Bed | 7 |
| 5 | How We Did It..... | 9 |
| 5.1 | Finding Max User Accessible Buffer Capacity | 9 |
| 5.2 | Finding Max Bursts without Packet Loss | 9 |
| 5.3 | DUT Switch Configuration | 10 |
| 6 | Max Buffer Space for User Traffic..... | 11 |
| 7 | Max Burst Size -- North-South Bound Traffic | 13 |
| 7.1 | Max Burst Size -- One Port, North-South Bound Traffic..... | 13 |
| 7.2 | Max Burst Size -- Multiple Ports, North-South Bound..... | 14 |
| 8 | Max Burst Size -- East-west Bound Traffic | 17 |
| 8.1 | Max Burst Size -- Single Destination 1GE Port, East-West Bound Traffic | 17 |
| 8.2 | Max Burst Size -- Multiple Destination 1GE Ports, East-West Bound Traffic..... | 18 |
| 9 | Max Burst Size -- Mixed North-South and East-West Congestion | 20 |
| 10 | Dropped Packet Observations | 22 |
| 11 | Conclusion..... | 22 |
| 12 | About Miercom..... | 23 |
| 13 | Use of This Report..... | 23 |
| 14 | Fair Test Notification..... | 23 |

1 Executive Summary

"Even the most simple architectural choices can impact mission-critical application performance," say Mohammad Alizadeh and Tom Edsall, of San Jose, California-based Insieme Networks (now Cisco), in their eye-opening paper, *On the Data Path Performance of Leaf-Spine Datacenter Fabrics*, published in the "2013 IEEE 21st Annual Symposium on High-Performance Interconnects."

Alizadeh and Edsall highlighted the trend in data center infrastructure design, in which backbone (or "spine") switch/routers, and access (or "leaf") switches are all fully interconnected in what resembles a large, non-blocking switch. However, with increasingly higher speed links in the backbone, data flows can become asymmetrical and intermittent data bursts are increasingly contending -- within a leaf switch -- for bandwidth to the same server connection.

Cisco sought to address the bursty-data issue in the architectural design of the Nexus 9300 Series Switches. The vendor implemented extended buffers and improved buffer management for better accommodating intermittent data bursts in the data center network access layer.

Cisco engaged Miercom to independently test whether the innovative buffer architecture and management built into the Nexus 9300 does indeed make a difference. A Cisco Nexus 9396PX model was tested. For comparison purposes the same tests were run on an Arista 7150S switch, representing a typical data center switch, featuring a traditional buffer design.

Key Findings and Conclusions:

- With a typical or default configuration, using only the default queue, the Arista 7150S-52 offers a maximum buffer space of 4.66 MB (Megabytes) that is accessible to user traffic. The Cisco Nexus 9396PX switch exhibited a combined 31.74 MB of buffer space available for user traffic -- about seven times the buffer size of the Arista 7150S-52.
- Both the Cisco Nexus 9396PX switch and the Arista 7150S-52 switches exhibited burst-absorption capacities that increased with packet size and the number of egress ports, until reaching a peak capacity.
 - The Arista 7150S-52 can absorb a burst up to 2.59 MB on a single egress port and a system maximum of 4.66 MB in aggregated bursts across four or more egress ports.
 - North-south bound bursts and east-west bound bursts were applied and studied on the Cisco Nexus 9396PX switch. For north-south bound traffic, the switch showed a 3.37 MB per-port burst capacity and a system max of 17.61 MB with eight or more egress ports. For east-west bound traffic, between local ports, the Cisco Nexus 9396PX switch can absorb bursts up to 2.45 MB on a single egress port or a system aggregate max of 7.88 MB with four or more egress ports. Combined, the Nexus 9396PX switch provides 25.5 MB of burst capacity for traffic in these two directions.

Test results clearly show that the buffer architecture and management of the Cisco Nexus 9396 Switch series do indeed make a difference. Miercom independently substantiates the superior performance of the Nexus 9300 with regards to buffer management and the handling of bursty traffic flows, accommodating longer-duration bursts and, in so doing, minimizing packet loss.

Robert Smithers
CEO
Miercom

2 Buffer Management Study

The proliferation of asymmetric link speeds in today's data center is resulting in more traffic contention and, increasingly, intermittent bursts of data, waiting in queue for oversubscribed links. This has raised concerns and prompted research, including the recent Alizadeh and Edsall paper cited earlier. The below excerpt from their report summarizes the issue and their assessment.

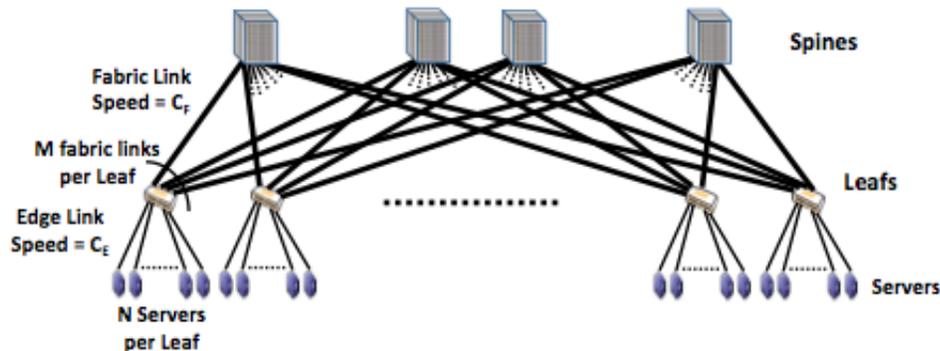


Fig. 1. Leaf-Spine datacenter network architecture. Multiple leaf and spine switches are connected in a full bipartite graph. The edge link speed (C_E) may be different from the fabric link speed (C_F) and there may be oversubscription at the leaf fabric links ($MC_F < NC_E$).

The authors state: "The size of switch buffers can have a significant impact on the data path performance and cost of datacenter fabrics. If buffers are too large, they increase cost and power consumption and may increase network latency thus adversely impacting latency-sensitive flows. On the other hand, if buffers are too small, bursty traffic patterns that are common in datacenters can cause excessive packet drops leading to TCP timeouts and high application latencies."

"This adverse effect of small buffers on bursty traffic patterns is known as the TCP Incast."

"Incast events are caused when many senders simultaneously transmit data to a single receiver. Hence, intuitively, the largest bursts should occur at the front-panel ports attached to the receivers where there is a single point of convergence coupled with the minimum link speed. Traffic bursts should be less severe in the spine since they get split across multiple paths and link speeds are typically higher. Our simulations verify that this is indeed true. Incast is more severe at the edge than in the spine; in particular, our results show that having larger buffers at the leaf switches is more effective at mitigating Incast than in the spine switches."

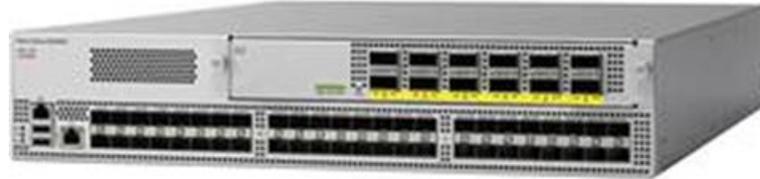
Cisco designed the Nexus 9300 Series "leaf" switches in response to this emerging environment – a higher speed "spine" network and lower-speed links to servers and endpoints. As discussed in this report, the Cisco 9396 switch distinguishes between high-speed (40GE) uplinks and lower-speed (1GE or 10GE) server links, and employs an innovative buffer architecture to accommodate the differences. In addition, the switch's buffer-management policies can be tailored for different traffic environments.

3 Products Tested

3.1 Cisco Systems 9396PX

The Cisco Nexus 9396PX Switch, shown in *Figure 1*, is a compact 2RU (two-rack-unit-high, or 3.5 inches) switch that supports 960 Gbps of bandwidth across 48 fixed 10GE SFP+ ports and 12 fixed 40GE QSFP+ "uplink" ports. The 40GE ports are on a removable uplink module that can be readily serviced or replaced by the user.

Figure 1: Cisco Nexus 9396PX Switch

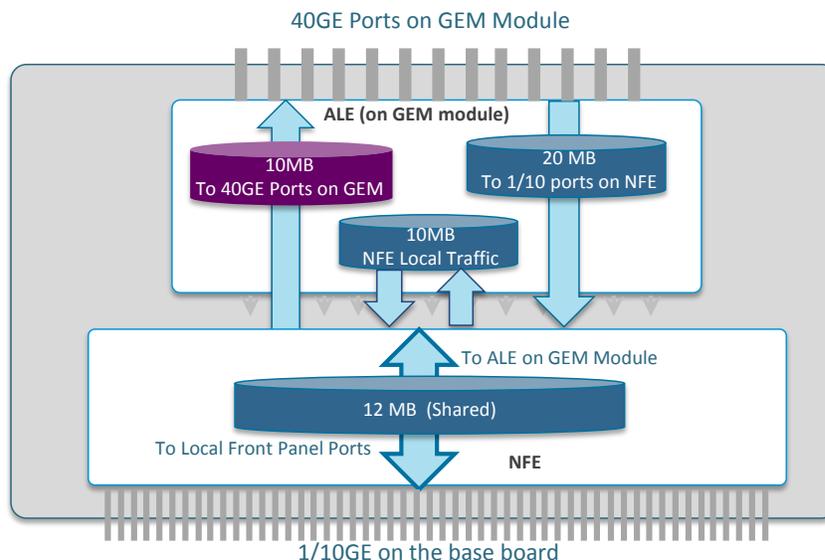


The Cisco 9396 is a "Top of Rack" (ToR) switch, designed to interface on one side to high-capacity "Spine" switches via the 12 x 40GE high-speed uplinks, and on the other side to dozens of "Leaf" nodes, typically servers. The design embraces the concept of data center switches with a non-blocking switch architecture.

The Cisco 9396 Switch incorporates two discrete packet-forwarding engines: the Application Leaf Engine (ALE) and the Network Forwarding Engine (NFE). Both are built with on-chip buffer.

Figure 2 depicts the innovative buffer layout on the Cisco Nexus 9396 switches, with multiple affiliated buffers designed to accommodate intermittent data bursts – typical of high-speed data flows in such a mixed-link-speed environment.

Figure 2: Cisco Nexus 9396PX Switch Buffer Architecture



The NFE has a 12-MB buffer for data egressing to connected servers through the 48 1/10GE ports. The NFE uses this buffer to send traffic and to signal the ALE to stop or resume sending packets.

The Application Leaf Engine (ALE) has three separate buffer regions:

- A 20-MB buffer that is shared by 40GE-to-1/10GE traffic for the extended output queuing in a typical network design. This represents the north-south bound direction, or from network to hosts.
- A 10-MB buffer for 1/10GE-to-40GE traffic. Typically, this is the south-north bound traffic in a data center network, or traffic from hosts to the network.
- A 10-MB buffer for “local” traffic between two 1/10GE ports that is hair-pinned onto the ALE from the NFE. This is normally east-west bound traffic between local host ports.

This report focuses on the buffer functions, their impact on these directional flows, and additional capacity provided by the ALE buffers.

3.2 Arista Networks 7150S

For comparing buffer performance, a Model 7150S-52 switch from Arista networks, shown in [Figure 3](#), was included in the test bed. The 1RU Model 7150S-52 switch tested featured 52 x 1/10GE (1 or 10 Gigabits/sec Ethernet) SFP+ ports, each equally suited for either uplink or downlink/server-access applications. All ports were tested as full 10GE ports. Sets of four ports could be grouped to form 40-Gbps links.

The Arista 7150S has a single 9.5 MB buffer that is shared by all ports on the switch.

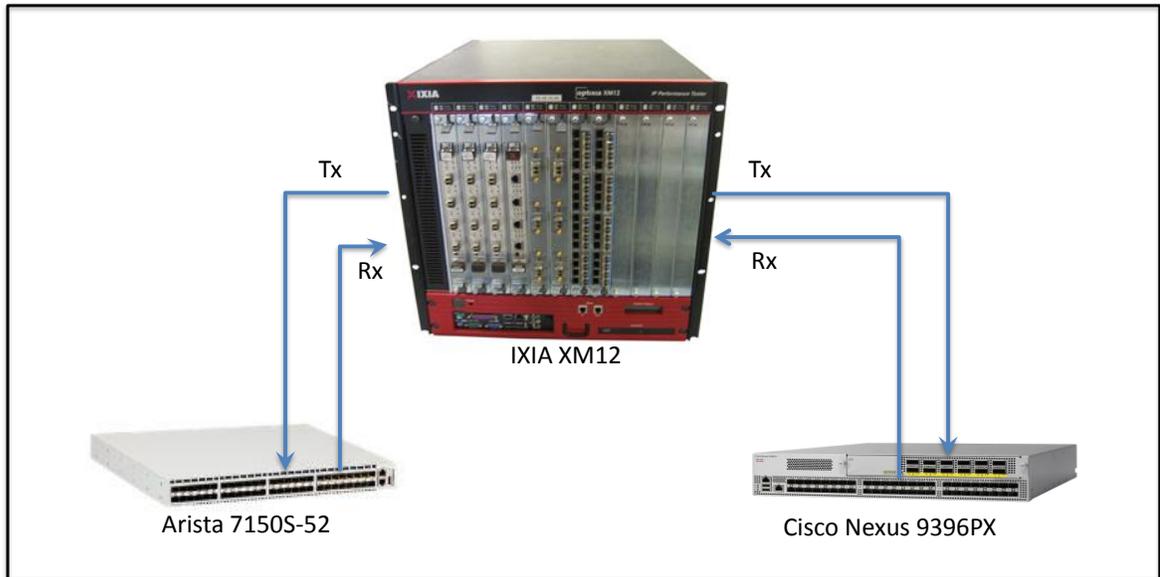
Figure 3: Arista 7150S-52 Switch



4 Test Bed

The tests were all driven by an Ixia XM12 traffic generator and traffic-analysis system. As shown in *Figure 4*, the Ixia system would send unidirectional, unicast flows through the Cisco Nexus 9396PX and Arista 7150S-52 switches.

Figure 4: Test Bed Setup



Data flows were applied using cardinal directions to describe the port relationships on the device under test (DUT) switches. The data flow directions emulate the actual traffic flow directions in real-world data center networks, including:

- North-to-South Bound:

From the data center access switches' perspective, north-South bound traffic is the traffic coming from the network and going to locally connected hosts or servers.

In the test, a group of ports was selected as network uplink ports and another group of ports as host ports. On the Cisco 9396PX switch, the uplinks were the 40GE ports on the switch GEM module and the host ports were the front panel 1/10GE ports. On the Arista 7150S-52 switch, the uplinks were a subset of the 1/10GE ports, and the host ports were another subset of the 1/10GE ports.

The testing studied the buffer management behavior and performance for the north-south bound traffic on the DUT switches

- South-to-North Bound:

The reverse of the above. Data was delivered by the Ixia test tool into one of the 1/10GE host ports (South) and forwarded to one of the network uplink ports (North).

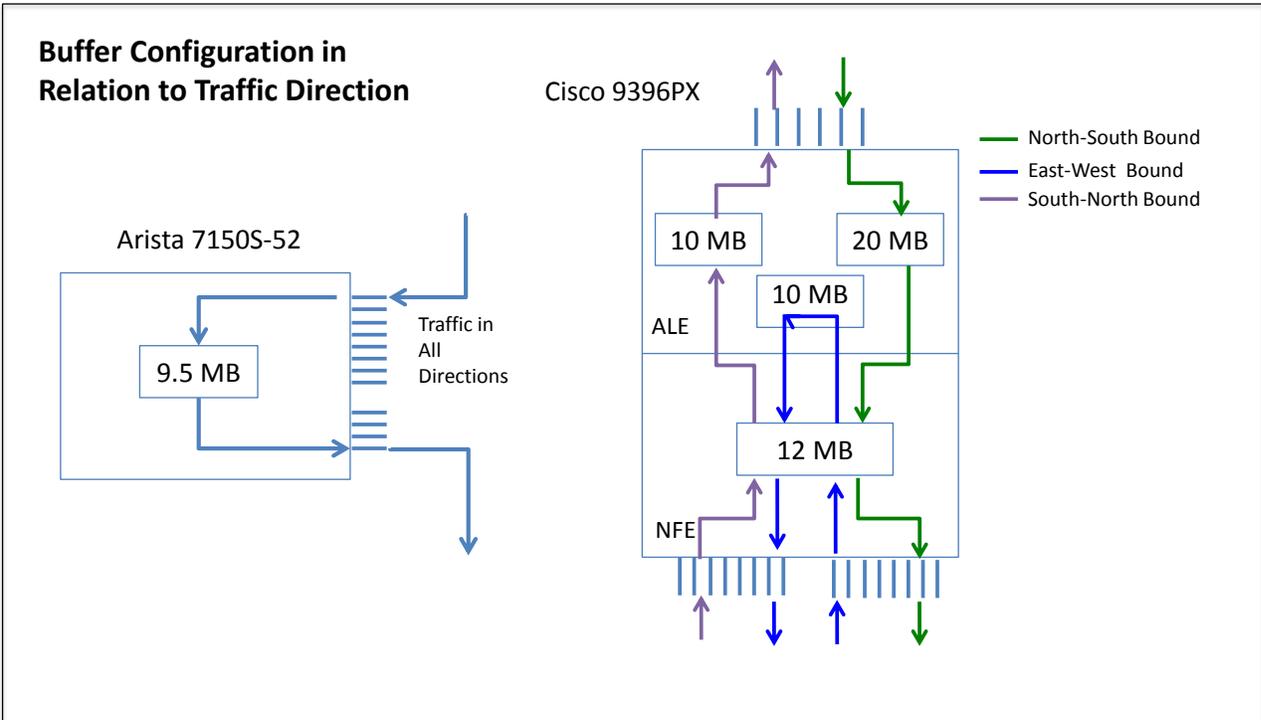
- East-West Bound:

East-west bound traffic through a data center access switch represents traffic between locally attached hosts. Data was delivered into one of the 1/10GE host ports and

forwarded to another 1/10GE host port. Since host devices often connect via different port speeds, an access switch might well be handling 1GE- and 10GE-attached hosts at the same time. To simulate that environment our testing incorporated traffic flows with 10GE source ports and 1GE destination ports. This represents perhaps the most demanding case for packet buffering.

Figure 5 illustrates the buffer structure in relation to these cardinal traffic directions within the Cisco Nexus 9396PX switch and the Arista 7150S-52 switch.

Figure 5: Cisco and Arista Output Buffer Disposition



We observed that the Arista 7150S-52 switch does not exhibit significant differences for these different traffic flows as all equally access the commonly shared 9.5 MB packet buffer memory.

However, on the Cisco Nexus 9396PX switch, the ALE provides separate buffer extensions for these three different traffic-flow directions, yielding different results for each. What's more, because the three extended buffer spaces on the ALE are independent of one another, they can handle traffic congestion for their corresponding directions simultaneously, collectively resulting in considerably more total available buffer space.

5 How We Did It

The best metrics for measuring the buffer capacity and performance of a switching system are:

- The maximum user accessible buffer space, and
- The maximum burst size accommodated without packet loss

The maximum user accessible buffer space is the maximum amount of packet buffer memory space that data flows can use when the switch ports are encountering continuous congestion.

The maximum burst size without packet loss is the size of traffic burst beyond the switch egress port speeds that can be buffered and eventually transmitted out of the egress ports without any packet loss.

5.1 Finding Max User Accessible Buffer Capacity

To identify the maximum user accessible buffer space, constant traffic flows are sent to the selected destination ports such that the aggregate traffic amount to each destination port continually exceeds its port speed. As a result, part of the traffic in the flow is queued in the packet buffer memory before being transmitted out of the egress ports.

Due to the continuous overflow on the egress port, more and more buffer space is consumed to queue the excess, until the maximum accessible buffer space is reached. At that point additional packets that need to be buffered will be dropped due to exhaustion of the packet buffer space. Our process is to note when packet drops occur, then monitor the buffer space utilization on the switch and record the maximum size of used buffer space. The result is the maximum user traffic accessible buffer space size.

5.2 Finding Max Bursts without Packet Loss

The first step in achieving this measurement is to set up the baseline flow. This is a unidirectional flow that fills the egress port by sending continuous traffic at the line rate of the egress port. The next step is to insure the base flows are going through the switch on a steady-state, continuous basis, with no packet drops.

Then, with the baseline flow running, a burst is sent onto each of the egress ports. The packets in the bursts are sent at the full line rate of the egress ports. The burst size varies with frame size and the number of packets being sent. The size of the burst is then adjusted, through repeated iterations, to find the maximum burst size for a particular frame size, with no packet drops. Each burst is sent only once. Any packet drops are noted by the Ixia system, which compares the burst sent with the packets received.

The same test is repeated for packet sizes ranging from 64 bytes to 9216 bytes and results are recorded. (A max packet size of 9214 bytes is used for the Arista 7150S-52 switch, as the maximum jumbo frame size supported by this switch is 9214-bytes.)

5.3 DUT Switch Configuration

The tests in this report were conducted with typical or default settings for buffer management on the DUT switches. All traffic would go into the default queue for queuing purpose upon egress port congestion.

Since each vendor may provide different methods for buffer tuning, it is possible that further customizing the buffer configuration of any switch could yield improved results from what we observed in this report. We decided, however, to stay with typical or default settings – as most users would employ -- and deemed possibly different customization results as beyond the scope of this report. If a vendor modifies its typical or default settings in a newer software release, resulting in better buffer capacity and/or performance for user traffic, these tests can be repeated with the new software, and a new report issued with the improved results.

6 Max Buffer Space for User Traffic

Description

This set of tests identifies the maximum buffer space accessible by user traffic. To determine this, the DUT switch is put into a constantly congested condition.

On the Arista 7150S-52 switch, four 10GE ports were used as destination host ports, and eight 10GE ports were used as source ports. Two source ports sent an aggregate 15 Gbps of traffic to a single destination port. The switch had up to four egress ports congested in this manner.

On the Cisco Nexus 9396PX switch, different traffic flows were used to test the different buffer spaces on the ALE:

- The 20MB buffer for north-south bound traffic (from 40GE ports to 1/10GE ports)

Four 40GE ports were used as source ports and eight 10GE ports were used as destination host ports. Each destination port received a constant 15 Gbps of traffic, causing continuous congestion on the port.

- The 10MB buffer for south-north bound traffic (from 1/10GE ports to 40GE ports)

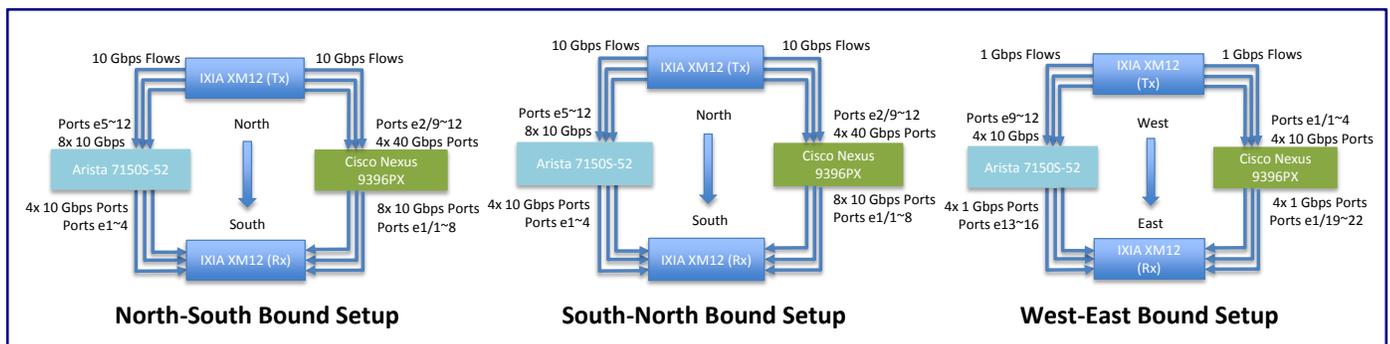
Twenty 10GE ports were used as source ports and four 40GE ports were used as destination ports. This direction is typically for traffic coming from the local hosts and going out to the network through the 40GE uplinks. In the test, each 40GE port received 45 Gbps of traffic.

- The 10MB buffer for east-west bound local traffic (from 1/10GE ports to 1/10GE ports)

This is the direction for traffic between local hosts. In the test four 10GE ports were used as source ports and four 1GE ports were used as destination ports. This test configuration was done to introduce the effect of port-speed mismatch into the study.

Figure 6 shows the test bed setup for different flow directions on the Nexus 9396PX switch.

Figure 6: Test Bed Setup for Different Flow Directions on Nexus 9396PX Switch



The results were captured on the DUT switches using their respective CLI commands for buffer monitoring or queue length monitoring. Both Cisco Nexus 9396PX and Arista 7150S-52 manage their packet buffer memory by dividing it into fixed-length cells and store data packets into these cells. The output data from both switches' CLI monitoring commands display the buffer use in

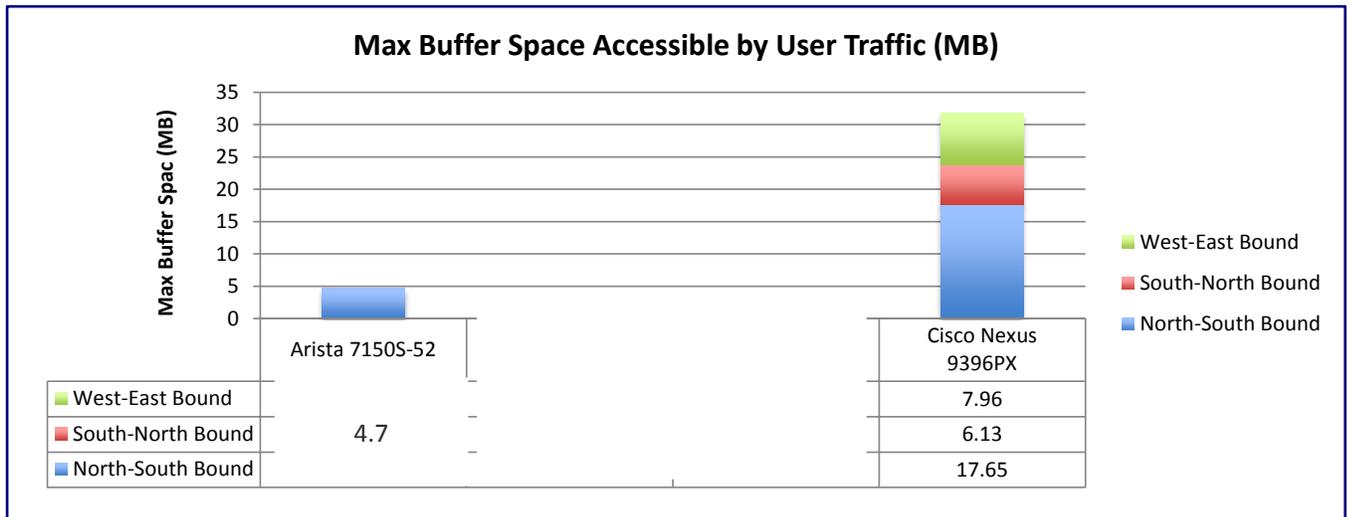
number of cells. It can be readily converted into the number of Bytes by multiplying by the cell length.

We ran these tests for typical packet sizes between 64 bytes and 9126/4 bytes. (As noted, the Arista switch supports a maximum Jumbo packet size of 9124 bytes.)

Observations and Analysis

As shown in [Figure 7](#), the Arista 7150S-52 switch has a max user accessible buffer size of 4.7 MB. The Cisco Nexus 9396PX offers a max user accessible buffer size of 17.65 MB for North-South bound traffic, 6.13 MB for South-North bound traffic and 7.96 MB for East-West bound traffic. Additionally, since the three buffer regions are independent to one another, they can handle congestions in their respective directions simultaneously. The test results show that the Cisco Nexus 9396PX switch can provide up to 31.7 MB buffer space in total for mixed traffic flows in all three directions.

Figure 7: Max Buffer Space Accessible by User Traffic



7 Max Burst Size -- North-South Bound Traffic

7.1 Max Burst Size -- One Port, North-South Bound Traffic

Description

This test emulated traffic flows coming from the network through the switch uplinks, and destined for host ports. On the Cisco Nexus 9396PX the uplinks are 40GE ports on the GEM module. On the Arista 7150S-52 the uplinks were selected 10GE ports on the switch. The host ports were 10GE ports on both DUT switches.

For each frame size, a single steady-state baseline flow of 10 Gbps was delivered to an uplink port and out through a 10GE egress host port. The flow is continuous, with no packets queued or dropped. The egress port was subsequently filled to capacity.

Then, a single 10-Gbps burst of the same-frame-size packets was sent through the same port while the baseline data flow was running. The burst was generated by sending a specific number of packets at the line rate of the DUT egress port. The burst size was calculated by frame size (bytes) x number of packets and then converted to MB (Megabytes). Successive trials were run to determine the Max Burst Size that showed no packets dropped in either the baseline flow or the burst flow after the baseline flow is turned off and all queues have emptied.

Configuration

The Ixia traffic generator was configured to:

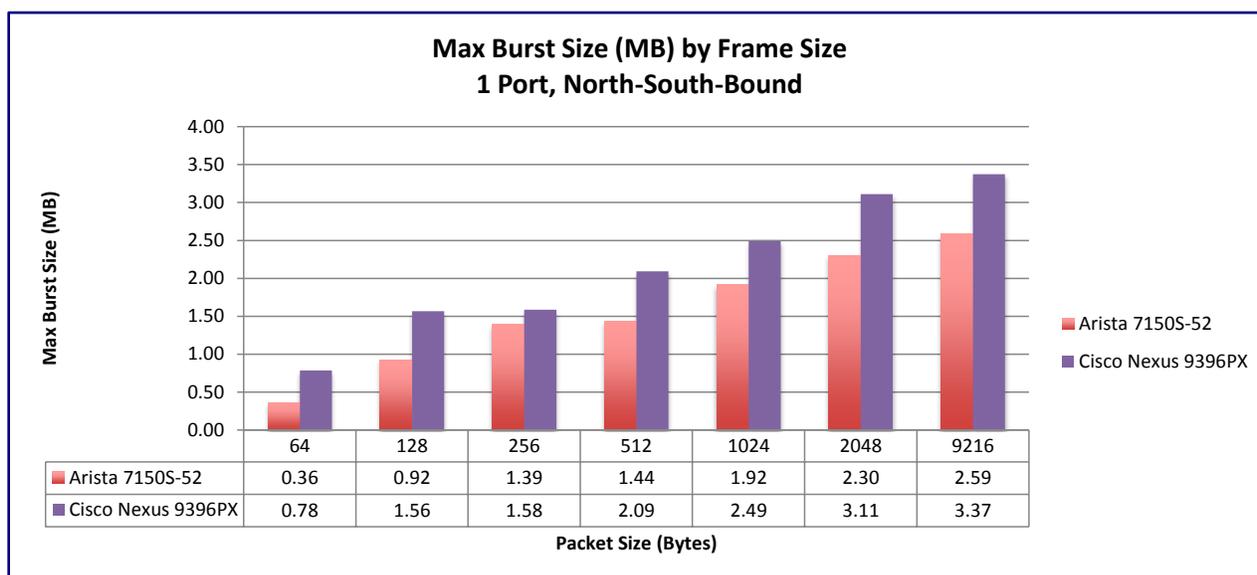
- 1) send a baseline flow continuously
- 2) send one burst of a specific number of the same-frame-size packets at the line rate of the egress port

The Cisco Nexus 9396PX was configured using one 40GE ingress uplink port, set to forward the data to one 10GE egress host port. The Arista 7150S was configured with two 10GE ingress uplink ports and one 10GE egress host port. The destination 10GE port received a constant 10-Gbps flow and a 10-Gbps burst.

Observations and Analysis

As [Figure 8](#) shows, the Maximum Burst Size grows roughly linearly with frame size. The larger frame sizes fit more efficiently into the buffer cell sizes, before filling up the respective output buffers and dropping packets.

Figure 8: Single-Port North-South-Bound Max Burst Size by Packet Size



In this single port-to-single port test case, testing found the Maximum Burst Sizes supported by the Cisco Nexus 9396PX to be considerably more than Arista 7150S-52.

7.2 Max Burst Size – Multiple Ports, North-South Bound

Description

In this test up to eight baseline traffic flows and bursts were applied. The purpose of the test was to see if the buffer memory on the device under test (DUT) switches can handle bursts on multiple ports, and to identify the system maximum aggregate burst size. Flow patterns to each destination port were the same as in the previous Test 7.1, but the baseline flows and the same size of bursts were sent to multiple destination ports at the same time.

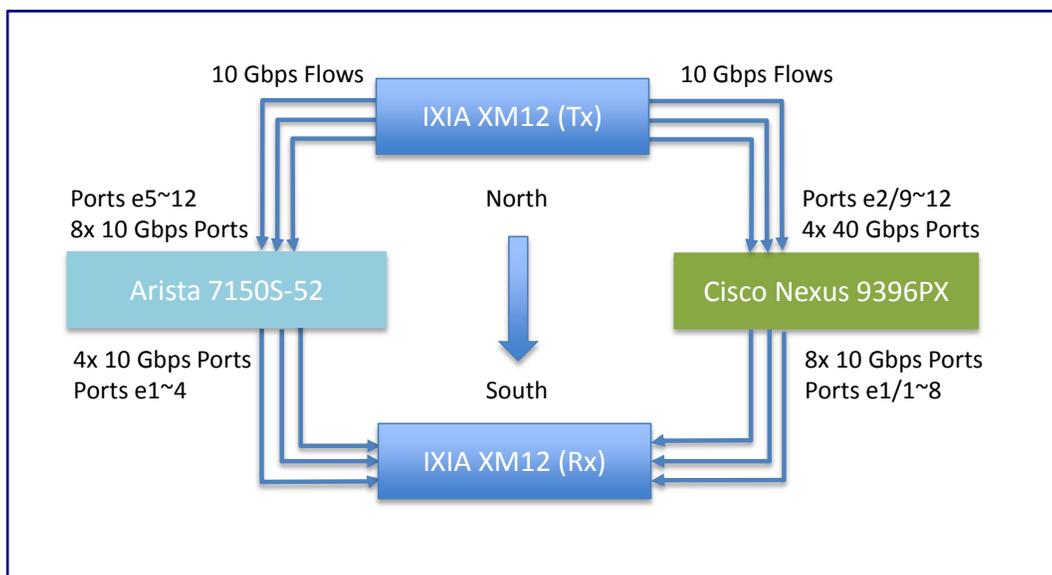
Configuration

On the Cisco Nexus 9396PX switch, traffic was sent from one to eight 40GE ports, to a corresponding number of 10GE ports. Each egress port received a constant 10-Gbps flow and a 10-Gbps burst. The testers observed and adjusted the burst size until finding the maximum burst size, with no packet loss in either the constant 10-Gbps flows or in the 10-Gbps bursts.

On the Arista 7150S switch, the same traffic was sent from 10GE ports to 10GE ports. Again, each destination 10GE port received a constant 10-Gbps flow and a 10-Gbps burst. Similarly, the burst size was observed and adjusted until finding the maximum burst size, with no packet loss in either the constant 10-Gbps flows or in the 10-Gbps burst. Up to eight 10GE source ports and four 10GE destination ports were used in this test.

The configuration for this test series is shown in the diagram in [Figure 9](#).

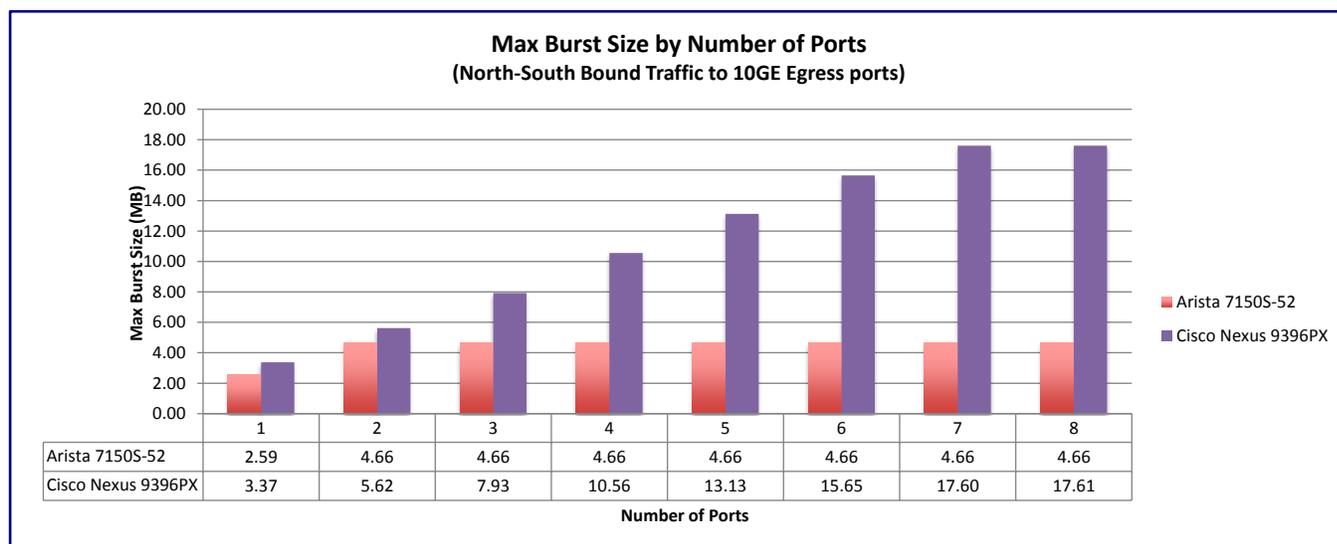
Figure 9: Multiple-Port North-South-Bound Burst Size Test Configuration



Observations and Analysis

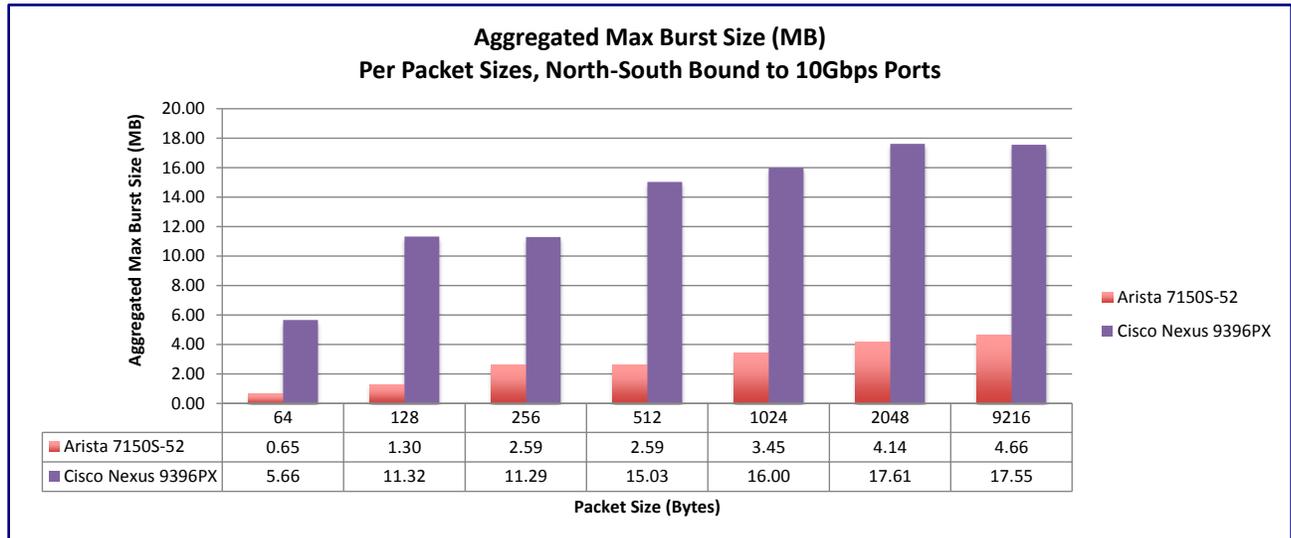
In this test configuration the Arista 7150S Max Burst Size peaked at two ports with a 4.6 Maximum Burst Size. The Max Burst Size of the Cisco 9396PX slightly exceeded the Arista 7150S at two ports, and then grew linearly as additional ports required buffer space for bursts, up to 17.6 MB. *Figure 10* displays the max burst sizes observed on both switch platforms with different numbers of destination ports.

Figure 10: North-South-Bound Max Burst Size by Number of Destination Ports



Different packet sizes in the range from 64 to 9216/4 bytes are tested for the Max Burst Size. *Figure 11* shows the results for different packet sizes. The Arista 7150S accommodates the largest Max Burst Size with the largest packet size, but only up to the 4.66-MB limit. The Cisco 9396PX's Max Burst Size scales up to 17.61 MB.

Figure 11: North-South-Bound Aggregated Max Burst Size by Packet Size



8 Max Burst Size – East-west Bound Traffic

8.1 Max Burst Size – Single Destination 1GE Port, East-West Bound Traffic

Description

This test emulated east-west bound traffic flows and bursts, common in data center networks. Here we mixed port speeds by sending traffic from a 10GE port to a 1GE port, simulating a worst case scenario where buffer is needed to address port-speed mismatch.

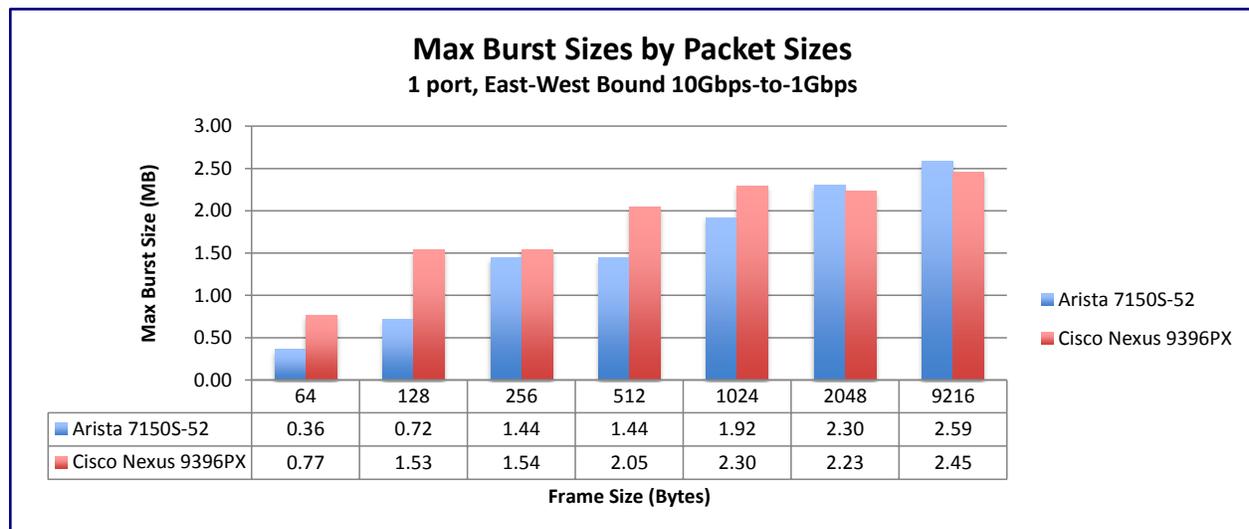
For each frame size, a single 1-Gbps steady-state baseline flow was delivered to an uplink port and out through a 1GE egress host port. The flow was continuous, with no packets queued or dropped. The egress port was filled to capacity.

Then, a single 1-Gbps burst of the same-frame-size packets was sent through the same ports while the baseline data flow was running. The burst was generated by sending a specific number of packets at the line rate of the DUT egress port. The burst size was calculated by frame size (bytes) x number of packets and then converted to MB (Megabytes). Successive trials were run to determine the Max Burst Size with no packets dropped, after the baseline flow is turned off and all queues have emptied.

Observations and Analysis

As [Figure 12](#) shows, the Maximum Burst Size grows roughly linearly with frame size. It is believed this is because the larger frame sizes fit more efficiently into the buffer cell sizes, before filling up the respective output buffers and dropping packets.

Figure 12: Single-Port East-West-Bound Max Burst Sizes by Packet Sizes



In this single-destination-port test case, the Maximum Burst Sizes supported by the Cisco Nexus 9396PX switch and the Arista 7150S-52 switch are very close to each other.

8.2 Max Burst Size – Multiple Destination 1GE Ports, East-West Bound Traffic

Description

In this test, up to four baseline traffic flows and bursts were applied. The purpose of the testing was to see if the buffer memory on the DUT switches would accommodate Max Burst Sizes on multiple ports, and to identify the system maximum aggregate burst size for traffic in the east-west-bound direction. Flow patterns to each destination port was the same as Test 8.1, but the baseline flows and the same size of bursts were sent to all selected destination ports at the same time.

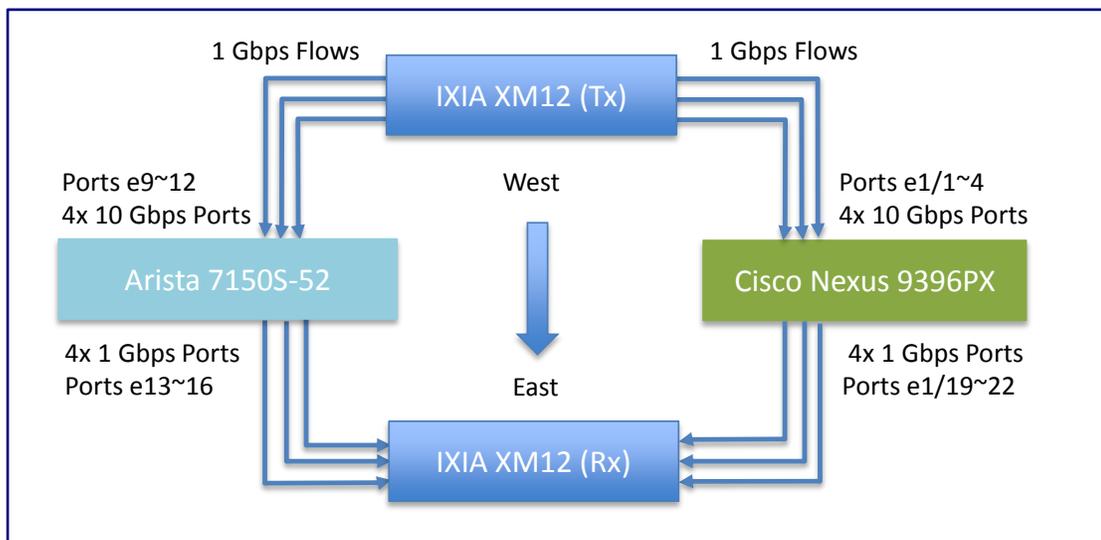
Configuration

On the Cisco Nexus 9396PX, traffic was sent to up to four 1GE ports. Each destination 1GE port received a constant 1-Gbps flow and a 1-Gbps burst. Four 10GE ports were used as source ports. Results were taken by observing and adjusting the burst size, for each frame size, until finding the Maximum Burst Size, with no packet loss in either the constant 1-Gbps flows or in the 1-Gbps burst.

Similarly, on the Arista 7150S, traffic was sent from four 10GE source ports to four 1GE destination ports. Each destination 1GE port received a constant 1-Gbps flow and then a 1-Gbps burst. Results were taken by observing and adjusting the burst size, until finding the maximum burst size, with no packet loss in either the constant 1-Gbps flows or in the 1-Gbps burst.

The configuration for this testing is shown in the *Figure 13*.

Figure 13: Multi-Port East-West-Bound Max Burst Size Test Configuration

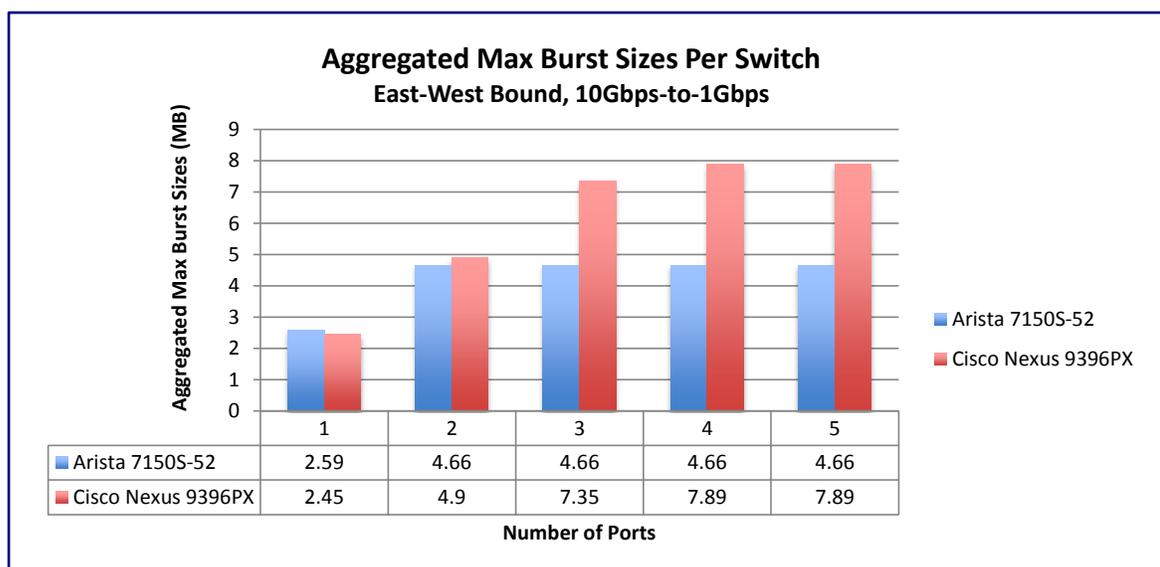


Observations and Analysis

The results shown in *Figure 14* indicate that the Max Aggregated Burst Size for the Arista switch peaks at 4.66 MB, and does not increase after the second port pair.

The Cisco Nexus 9396PX switch reaches the Max Aggregated Burst Size after adding the fourth port pair. The results indicate that the 9396PX can accommodate a Max Aggregated Burst Size of 7.89 MB.

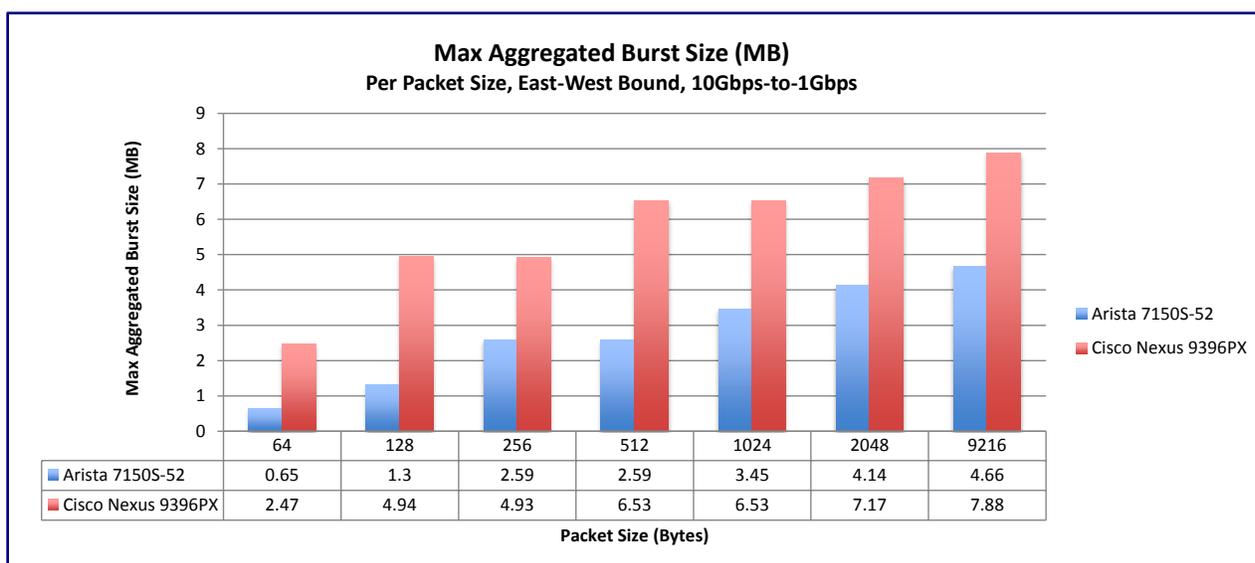
Figure 14: East-West-Bound Aggregated Max Burst Size by Number of Destination Ports



Note: The data shown above is based on 9,216-byte (Jumbo) burst-data packet size.

The graph in *Figure 15* shows the results of the 10GE-to-1GE flow test by frame size. Again, the Arista switch's Max Burst Size peaks at 4.66 MB, with jumbo frames. The Cisco Nexus 9396 accommodates a Max Burst Size of 7.88 MB.

Figure 15: East-West-Bound Max Aggregated Burst Size by Packet Size



9 Max Burst Size – Mixed North-South and East-West Congestion

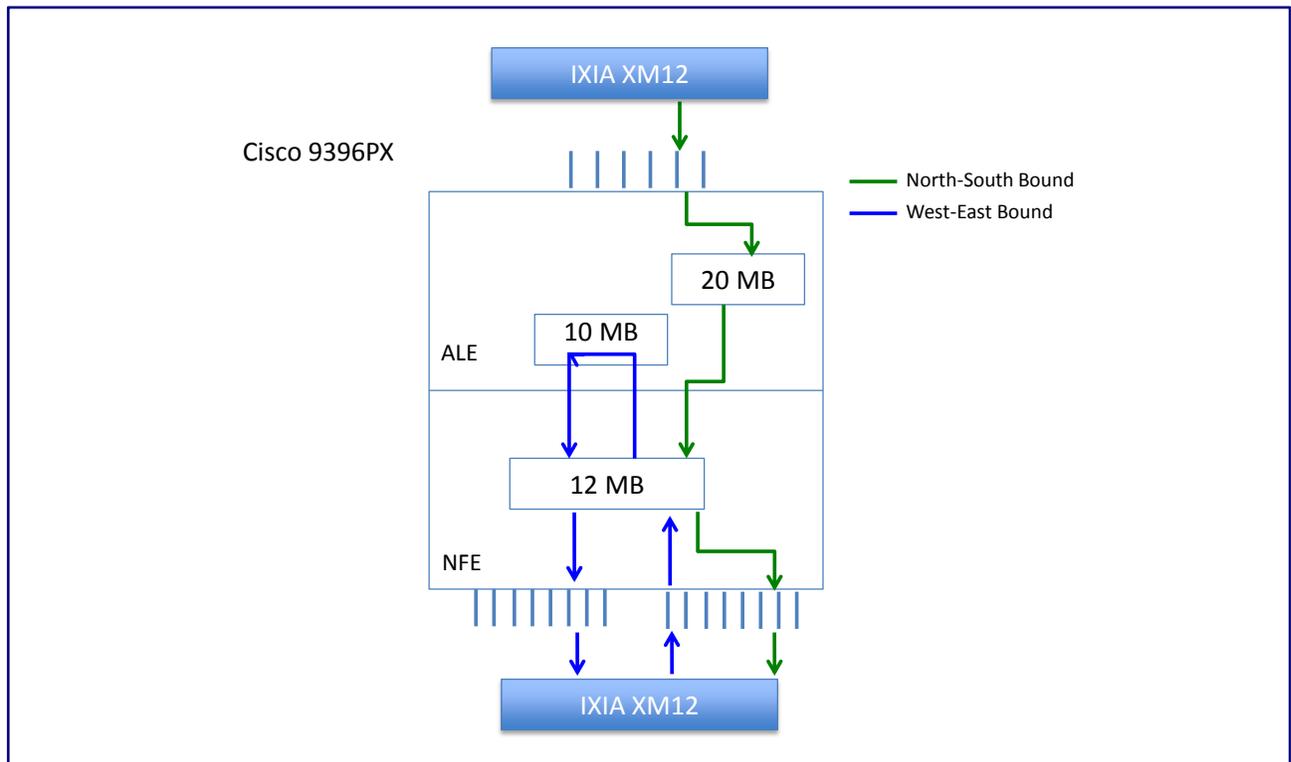
Description

In this test we examined buffer capacity and utilization where the switch experiences mixed North-South-bound and East-West-bound congestion simultaneously.

Configuration

The traffic flows on Cisco Nexus 9636PX switch for this burst tests are shown in *Figure 16*.

Figure 16: Traffic Flow Design on Nexus 9396PX



The north-south bound flows are from 40GE ports to 10GE ports.

The east-west bound flows are from 10GE ports to 1GE ports.

Because the Arista 7150S-52 switch uses the same 9.5-MB shared buffer space for traffic in all directions, and all ports are treated identically, the traffic flow directions do not bare any significance for the Arista 7150S-52. Therefore, no additional tests were needed on the Arista 7150S-52 switch for this test case.

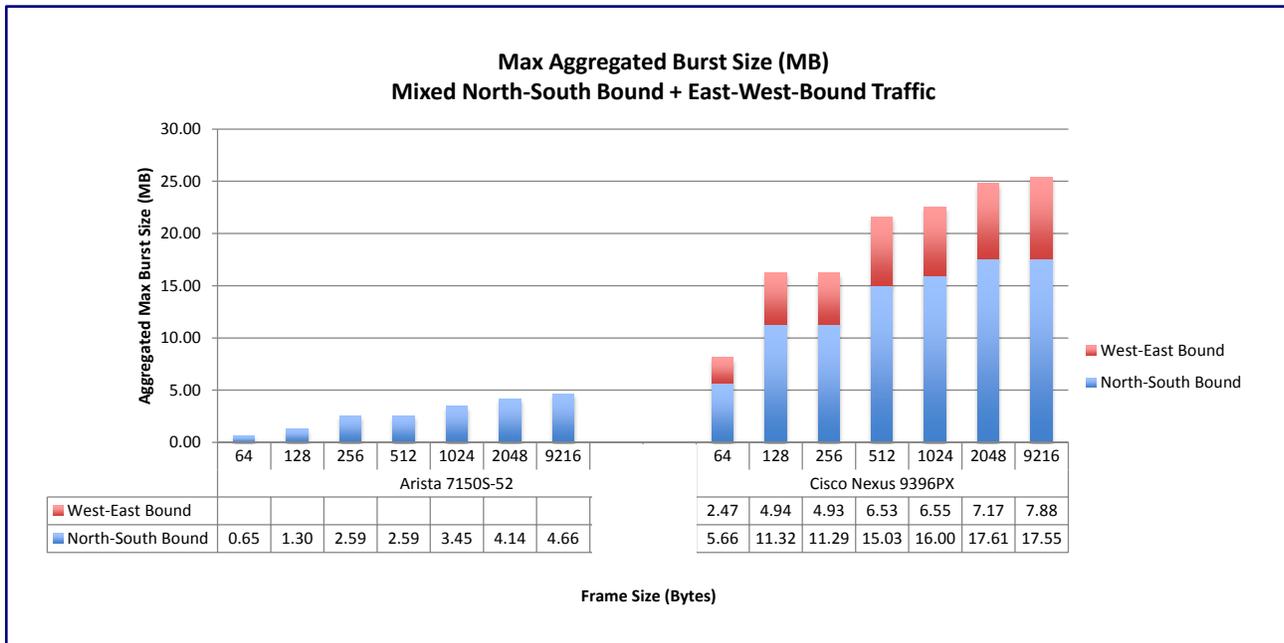
Observations and Analysis

On the Cisco Nexus 9396PX, there are three separate ALE buffer spaces dedicated for north-south-bound traffic (40GE to 1/10GE), south-north-bound traffic (1/10GE to 40GE) and east-west-bound traffic (1/10GE to 1/10GE). In our burst-size tests we focused on the north-south-bound and the east-west-bound directions as traffic bursts have more significant impact in these two directions on data center access switches.

In this test, the combination of north-south-bound and east-west-bound bursts were studied. We observed with the Cisco Nexus 9396PX switch that traffic could be sent into both buffer spaces at the same time, and a total Max Burst-handling up to 25.43 MB was accommodated.

The chart in [Figure 17](#) shows the aggregated maximum burst size for the combined north-south-bound and east-west-bound traffic. Nexus 9396PX switch supports up to 25.43-MB bursts whereas Arista 7150S-52 switch caps at 4.66 MB.

Figure 17: Max Aggregate Burst Size for Mixed North-South and West-East-Bound Traffic



10 Dropped Packet Observations

Finding the maximum burst size was done via an iterative process based on burst size, to find the largest burst with no packets dropped in either the burst flow or the constant base flow.

During this process we observed that when congestion occurs, the Arista 7150S-52 switch would drop packets equally from the constant baseline flow and from the burst flow. The majority of packet drops by the Cisco Nexus 9396 switch, however, were from the baseline flow. In most cases, the burst flow passed through the switch without drops, as shown in *Figure 18*.

On investigation we learned that this is due to buffer-management intelligence on the Cisco switch's ALE that can differentiate between long-lived flows and short burst flows. By default the delivery of short-lived bursts is favored over regular long-lived flows. So, when the switch starts to experience congestion, it tries to keep the short flows intact by dropping packets from the long-lived flows first.

Favoring short flows is desirable in many data center applications, where long flows are more tolerant to packet drops. This particularly benefits latency-sensitive applications with a large number of small flows, such as Memcached, where packets dropped from small flows can significantly impede application performance.

Figure 9: Packet Drops Observation on Nexus 9396PX Switch

| | Tx Port | Rx Port | Traffic Item | Tx Frames | Rx Frames | Frames Delta | Loss % | Tx Frame Rate | Rx Frame Rate | Tx L1 Rate (bps) | Rx L1 Rate (bps) | Rx Bytes |
|---|----------------|---------------|--------------|-------------|-------------|--------------|--------|---------------|---------------|------------------|------------------|--------------|
| 1 | 40GE-9396-2/9 | 10GE-9396-1/1 | const-9396 | 388,470,164 | 388,469,183 | 981 | 0.000 | 2,349,389.151 | 2,349,389.651 | 9,999,000,225... | 9,923,821,884... | 197,342,3... |
| 2 | 40GE-9396-2/9 | 10GE-9396-1/1 | burst-9396 | 5,000 | 5,000 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 2,540,000 |
| 3 | 40GE-9396-2/10 | 10GE-9396-1/2 | const-9396 | 388,470,185 | 388,469,842 | 343 | 0.000 | 2,349,388.571 | 2,349,389.571 | 9,998,997,757... | 9,923,821,548... | 197,342,6... |
| 4 | 40GE-9396-2/10 | 10GE-9396-1/2 | burst-9396 | 5,000 | 5,000 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 2,540,000 |
| 5 | 40GE-9396-2/11 | 10GE-9396-1/3 | const-9396 | 388,471,352 | 388,470,940 | 412 | 0.000 | 2,349,388.707 | 2,349,388.707 | 9,998,998,335... | 9,923,817,896... | 197,343,2... |
| 6 | 40GE-9396-2/11 | 10GE-9396-1/3 | burst-9396 | 5,000 | 5,000 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 2,540,000 |

11 Conclusion

Data center network-access switches need sufficient packet buffer capacity and performance to accommodate transient congestions and bursts caused by incast traffic and port-speed mismatches.

Industry research shows that a balanced distribution of packet buffer resources between network leaf and spine layers provides the optimal solution, and that placing larger buffers at the leaf switches is more effective at mitigating Incast than in the spine switches.

Cisco Nexus 9396 Switches are designed with extended buffer spaces and advanced buffer management mechanisms. The 9396 demonstrated superior buffer capacity and burst-absorption performance over a typical data center access switch with traditional buffer architecture and capacity, represented by the Arista 7150S-52 switch in this report.

12 About Miercom

Miercom has published hundreds of network-product-comparison analyses in leading trade periodicals and other publications. Miercom's reputation as the leading, independent product test center is undisputed.

Miercom's private test services include competitive product analyses, as well as individual product evaluations. Miercom features comprehensive certification and test programs including: Certified Secure, Certified Green, Certified Interoperable and Certified Reliable. Products may also be evaluated under the Performance Verified program, the industry's most thorough and trusted assessment for product usability and performance.

13 Use of This Report

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

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

No part of any document may be reproduced, in whole or in part, without the specific written permission of Miercom or Cisco. All trademarks used in the document are owned by their respective owners.

All vendors with products featured in this report were afforded the opportunity before, during, and after testing was complete to comment on the results and demonstrate the performance of their product(s). Any vendor with a product tested by Miercom in one of our published studies that disagrees with our findings is extended an opportunity for a retest and to demonstrate the performance of the product(s) at no charge to the vendor.

14 Fair Test Notification

All vendors with products featured in this report were afforded the opportunity before, during, and after testing was complete to comment on the results and demonstrate the performance of their product(s). Any vendor with a product tested by Miercom in one of our published studies that disagrees with our findings is extended an opportunity for a retest and to demonstrate the performance of the product(s) at no charge to the vendor.

Arista reviewed preliminary data from this testing before the results were published. Arista was afforded the opportunity to repeat this testing in their own facilities with their own test equipment and engineers, but declined the opportunity.