
European Advanced Networking Test Center

# *Competitive Test Report*

## *Force10 TeraScale E-Series*
## *versus*
## *Cisco Catalyst 6500*

**EANTC AG**

**September 2005**

# *Table of Contents*

# Introduction

In June 2005, Cisco Systems commissioned the European Advanced Networking Test Center (EANTC) to independently validate the performance, scalability, and availability of Cisco Catalyst 6500 versus Force10 TeraScale E-Series switches. Force10 promotes network designs where large numbers of ports, normally spread across multiple switches in a redundant network configuration are instead "collapsed" into a single high port-density solution comprising a single switch with up to 1,260 Gigabit Ethernet ports. Clearly this is a major departure from current industry accepted network design best practices and one that raises significant questions with regard to service scalability and availability.

To answer these and many other questions raised by the Force10 TeraScale architecture, we conducted a number of tests to investigate just how well the Force10 switches compare to the Cisco Catalyst 6500.


**EANTC extensively tested the following categories:**

- **Force10's Performance claims made in the Tolly Group Test Reports**
- **Force10's High Availability Claims**
- **Routing Protocol Scalability, Stability, & Resilience**
- **IP Multicast Scalability, Stability, & Resilience**

Our tests are divided into the following broad categories:

- **Investigate Force10 claims made in the Tolly Group reports:** We compare the performance, availability, and scalability claims made by Force10 as validated by the Tolly Group with our own findings. Force10's tests can be found at http://www.tolly.com/Search.aspx?VendorID=27 and are contained in Test Reports 204147 and 204148.

- **Test Force10's High Availability Claims:** We specifically tested the high availability claims made by Force10 as this is such an important part of Force10's proposed network designs. We used the claims made in the Tolly Group test reports and Force10's whitepapers entitled "High Availability in the Force10 Networks E-Series" (Version 1.3) and "Guaranteed Access to System Management even During Processor Overload" as our guide.

- **Routing Protocol Scalability, Stability, & Resilience:** In this section of tests, we focused on the scalability, stability, and resilience of the control plane, the brains of the switch. If the control plane is not capable of scaling to support the high port density of the Force10 switches, then the higher port density offered by the E-Series is of questionable value. For instance, can the routing protocols support the same number of neighbors as there are ports on the switch? How long does it take to learn routes from such a large number of neighbor routers? What about convergence times? A very powerful control plane matching the very large number of interfaces is required to ensure the switch remains stable.

- **IP Multicast Scalability, Stability, & Resilience:** IP multicast traffic is becoming an increasingly important capability of many networks, applications range from remote learning, distribution of market data feeds in trading rooms to the data streams used in education for remote learning applications etc. How does the high availability claims for the Force10 switches apply to mission critical multicast traffic? What multicast scalability can I expect? How effective are Force10's multicast resilience features?

European Advanced Networking Test Center

# Executive Summary — Verification of Tolly Group Report 204147

In June 2005, Cisco Systems commissioned the European Advanced Networking Test Center (EANTC) to independently validate the performance, scalability, and availability of the Force10 TeraScale E-Series switches.

In this category of tests, we attempted to reproduce a number of the original Tolly Group tests found in Tolly Group Test Report 204147.

See http://www.tolly.com/Search.aspx?VendorID=27.

Like the Tolly Group, our tests were constrained by the availability of test equipment and the quantity of each vendor's switching equipment. In the main, our tests focused on Gigabit Ethernet, as Force10 acknowledges their sales of Gigabit Ethernet far outstrip those of 10-Gigabit Ethernet.

Cisco Systems supplied the SmartBits test equipment and the four Force10 TeraScale E-Series switches used in the tests.

Force10's Tolly Group test reports lack substantial detail, and on certain tests we found it difficult to understand how the tests were configured and the particular results achieved. After consulting with the Tolly Group, we obtained the Force10 switch configurations for three of the tests, to help us understand how the tests were actually run.

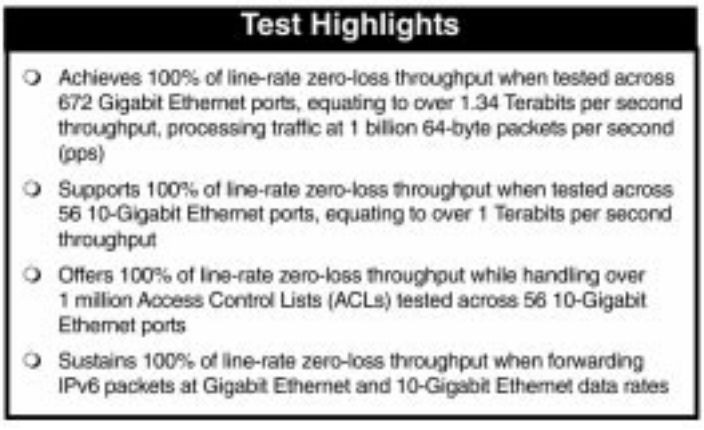



**Figure 1: Excerpt from Tolly Group Report 204147**

Tolly Group's reports state the tests were run using a beta software version. The Tolly Group states "According to Force10 Networks, this Beta release was replaced by Version 6.1.1.0." There is evidence that Force10 used a special engineering software to achieve these test results. (See the letter in the Appendix dated 9/28/2005 from Force10 to EANTC.) EANTC feels the test results question the performance that can be expected from an off-the-shelf Force10 system.

## Tolly Group Report 204147 – EANTC Findings:

### ➜ TeraScale E-Series is not non-blocking, wire-rate at all frame sizes:

Contrary to Force10's claims, our tests show the TeraScale E-Series does not provide 672 Gigabit Ethernet ports with non-blocking, zero-loss throughput at all frame sizes. We found frame loss increases with frame-size, and for unicast-only traffic frame loss peaked at a maximum loss of 37 % (4,472-byte frames).

### ➜ Latency vs. Throughput Compromise:

By default, we found the TeraScale E-Series has one of the highest latencies in the switch market, up to five times that of the Catalyst 6500. (For example, 64-byte latency: Cisco = 7 mSec, Force10 = 35 mSec.) Force10 provides commands that can alter the switch fabric scheduler and reduce latency: however, we found this also increases packet loss. (For example, for 256-bytes frames, frame-loss jumped increased from less than 1% to over 16% when the lower-latency setting was used.)

In the Tolly Group reports, it is unclear whether throughput and latency tests were run using the same switch fabric settings, or whether throughput tests were run using one setting, while latency tests used the other.

### ➜ Non-unicast traffic causes unicast packet loss to increase:

We found that unicast throughput is affected by small amounts of non-unicast traffic crossing the switch fabric at the same time.

In a reduced 624-port L2 Snake configuration, with an additional 3-Port VLAN on a separate line card used to inject non-unicast traffic at 10% load, we saw frame loss on the unicast flows in the snake increase. (For example, frame-loss for 256-byte traffic increased from under 1% to 20%, a twenty-fold increase.) This points to possible scheduling problems on the switch fabric.

### ➜ Force10's 1-Million ACL test was not verifiable:

Force10's Tolly Group tests claim to have seen no degradation in throughput even when active features such as an ACL with 1-Million entries were applied to 56 10-Gigabit interfaces. EANTC found that this test used an ACL with 20,993 entries in total. In addition, we found the following:

- Force10 used special software and/or hardware for this test. The test is not reproducible with current production software, because insufficient line card memory (CAM) is available.
- The Tolly Group's result does not follow industry accepted methods of counting ACLs. They claim the action of applying the same 20,993-entry ACL to 56 interfaces (56 x 20,993 = 1,175,608) represents a 1-Million entry ACL. EANTC says this only proves the TeraScale E-Series supports a 20K ACL; in fact, when we tested it, the TeraScale E-Series couldn't even scale to 20K.
- When we tried to apply the 20,993-entry ACL to a single 10GE interface, it instantly overloaded the CAM, which holds 13K entries. The Tolly Group claims this ACL was applied to 56 x 10-Gigabit Ethernet interfaces. We can't see how this was possible without either special software to manipulate CAM partitions or special hardware to increase the size of the CAM on the line cards.

### ➜ Force10's Native IPv6 Forwarding Test could not be verified:

Force10's Tolly Group test report claims that in an 8-Port test the TeraScale E-Series demonstrated non-blocking zero-loss throughput for "Native IPv6" traffic. Our tests, one year later could not reproduce the results for the reason the TeraScale E-Series does not support IPv6. Even though Force10 claimed the beta software they used was released as FToS Version 6.1.1.0 – our tests using Version 6.2.1.3, numerous releases after the Tolly tests, show the following:

- There is no support for IPv6, including no IPv6 routing protocols, and no IPv6 commands.
- Force10's marketing literature and line card manuals claim support for 64K IPv6 FIB and 24K IPv6 ACL CAM entries per line card. Our tests discovered a total of just eight IPv6 FIB entries and zero IPv6 ACL entries in the CAM. The CAM profile is fixed and cannot be modified by the user.

## Summary

Our tests of the Force10 TeraScale E-Series and correspondence with both the Tolly Group and Force10 confirms that many of the test results achieved at Tolly Group were the result of Force10 employing "specialised" non-production software, especially customized for these specific tests. See the letter in the Appendix dated 9/28/2005 from Force10 to EANTC. It is unclear from the Tolly Group test report and our correspondence with Force10, whether CAM partitions were modified between test cases or whether a single CAM profile was used for all tests.

The additional capabilities provided by this special software have not been released as a general production release and readers might question the validity of the Tolly Group reports as an accurate representation of Force10's performance and scalability.

# Background Information
# Force 10 Switch Architecture

The TeraScale E1200 was the focus of the Tolly Group test reports. The smaller E600 and E300 were not tested by Tolly, but share the same hardware and software.

**Switch Fabric Design.** Force10's switch fabric is based on a crossbar switch fabric design similar to Cisco's Catalyst 6500 switch. To allow packets to be transferred across the fabric from one card slot to another, a cross-connect must be established to connect the two card slots together momentarily.

As traffic consumes an increasing percentage of a crossbar fabric's capacity, the switch fabric scheduler must compute an increasingly complex series of cross-connects to service all the traffic. A saturation point is reached when the scheduler can't forward all traffic streams, at this point the fabric becomes blocking.

In the Force10 TeraScale architecture, the switch fabric scheduler uses a concept called an "Epoch." The Epoch is a time window during which a set of pre-planned cross-connects is executed in sequence. The exact set of cross-connects and their sequence varies during each Epoch period and is governed by which packets need to flow where.

**Crossbar Efficiency & Epoch Settings.** The Epoch not only defines a period of time in which cross-connects are executed in the switch fabric, it also defines how far ahead in time the switch fabric scheduler looks when planning the sequence in which the next series of cross-connects should be implemented for maximum fabric efficiency.

By default, the switch fabric scheduler works with an Epoch of 10.4 µSec. While cross-connects for the current Epoch are being executed by the switch fabric; in parallel, the scheduler is looking into the line card ingress queues feeding into the fabric and using the time to plan the most efficient way to set up cross-connects to service the largest number of queued packets possible in the next Epoch.

The larger the Epoch, the further the scheduler can look ahead and the more efficiently it can use the capacity of the fabric.

This scheme has one drawback. The packets queued at the ingress to the switch fabric (Force10 implements a technique called Virtual Output Queuing) must wait a longer time before they have an opportunity to be forwarded. The larger the Epoch; the longer the delay.

**Latency And Two Epoch Settings.** The Force10 TeraScale architecture supports two Epoch settings for the scheduler. The default is 10.4 µSec and offers the most efficient utilization of the switch fabric, at the expense of latencies ranging from 30 µSec upwards.

For certain applications where latency is an issue, the switch can be configured to use the second Epoch setting of 3.2 µSec. This reduces the delay (latency), but also reduces the scheduler's ability to make efficient use of the fabric capacity, resulting in significantly higher frame loss across all frame sizes.

Where possible, we have run tests at each Epoch setting so the reader is aware of the implications.

**Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test**

# GigE L2 Store & Forward Latency Test

## Test Objectives

These tests measure the Gigabit Ethernet store & forward latency of the Force10 TeraScale E-Series using LC-EF-GE-48T 48-port line cards and the Cisco Systems Catalyst 6500 employing WS-X-6748-GE-TX 48-port line cards.

On the Force10 E-Series, the tests are conducted with the switch configured for each available switch fabric "Epoch" setting as this affects the latency of the switch.

The tests compare the latency achieved by the Force10 TeraScale E1200 with the results of the Cisco Catalyst 6500.

**Test Highlights**

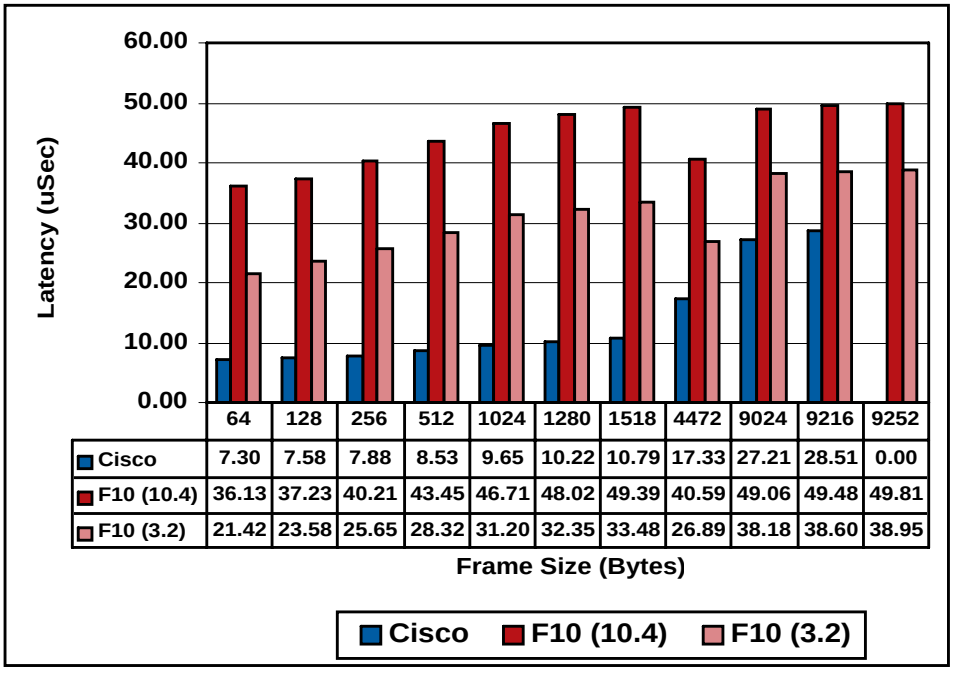
- ➜ **Force10's Gigabit Ethernet latency with default settings is three times that of the Catalyst 6500**
- ➜ **Force10's switch fabric Epoch setting can be used to reduce the latency by up to 40 %, but at the cost of higher packet loss**
- ➜ **Performance & latency tests should be carried out using a single Epoch-setting to gain accurate results.**

## EANTC Analysis

As can be seen from the results below, the switch fabric Epoch setting on the Force10 switch has a direct impact on the average latency. It is possible to adapt the Epoch setting to achieve lower latency (at the cost of lower throughput, though, as we will see in other test cases).

Even when using the shorter switch fabric Epoch, the Force10's latency is over double that of the Catalyst 6500.

Both switches deliver latencies for any packet size that are well within the standards requirements. Latency in the microsecond range plays a major role for storage area networks (iSCSI), high-energy physics networks and grid computing.




| | 64 | 128 | 256 | 512 | 1024 | 1280 | 1518 | 4472 | 9024 | 9216 | 9252 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cisco | 7.30 | 7.58 | 7.88 | 8.53 | 9.65 | 10.22 | 10.79 | 17.33 | 27.21 | 28.51 | 0.00 |
| F10 (10.4) | 36.13 | 37.23 | 40.21 | 43.45 | 46.71 | 48.02 | 49.39 | 40.59 | 49.06 | 49.48 | 49.81 |
| F10 (3.2) | 21.42 | 23.58 | 25.65 | 28.32 | 31.20 | 32.35 | 33.48 | 26.89 | 38.18 | 38.60 | 38.95 |

**Figure 1: Force10 E1200 vs. Cisco Catalyst Store & Forward Latency, Gigabit Ethernet**

## Test Configuration and Methodology

This test used all 48 port on a line card and sent traffic in a full-mesh traffic pattern at 10 % line rate. For the Force10 switch, the test was repeated at an Epoch setting of 10.4 µSec and 3.2 µSec.

The test measured the average store and forward latency introduced by the switch at a variety of frame sizes.

Force10 TeraScale E-Series Competitive Test

# Gigabit Ethernet Layer 2 Frame Loss

## Test Objectives

Confirm Force10's claim to provide zero-loss, wire-rate performance at all frame sizes. Run the test at both switch fabric Epoch settings, and record the impact of each setting on aggregate forwarding rate and packet loss.

Compare results with Force10's claims made in the Tolly Group Test Report 204147.

## EANTC Analysis

Results show that Force10's E1200 cannot provide zero-loss, wire-rate performance at any frame size.

With the default Epoch setting, we monitored a small packet loss for 64-byte through 256-byte frames, increasing to 5 % for larger frame sizes, latencies ranged from 35–50 µSec. When tested with the 3.2 µSec Epoch setting that reduces latency to between 20–37 µSec, we found packet loss jumped to 13–38 % as a result.

The selection of the Epoch value has strong influence on the results. The Epoch setting that reduces latency is subject to 13–38 % packet loss depending on the packet size.

There seems to be a compromise to be made between high performance or low latency.

### Test Highlights

- **In the same configuration as the Tolly Group tests, the TeraScale E1200 failed to deliver zero-loss, wire-rate performance at any packet size.**
- **The results measured with Epoch 3.2 µSec show significant performance degradation in comparison to Epoch 10.4 µSec.**
- **Force10's claim that "E1200 provides 56.25 Gigabits per second of non-blocking bandwidth to each line card slot" could not be verified. Packet loss was observed in all our test runs.**

## Test Configuration and Methodology

The DUT was configured with 336 2-Port VLANs; VLAN 2-337. VLANs were interconnected with external cross-over cables to form a 672-Port snake configuration. L2 forwarding-table aging time was disabled for all VLANs. All control plane protocols were disabled to ensure no control protocol frames were being sent by the DUT as this might interfere with the test.

This test conforms to the methodology for measuring frame-loss as specified in RFC 2544 and to the definition of frame-loss specified in RFC 1242.

Bidirectional traffic was transmitted between the two SmartBits ports attached to either end of the snake at 100% wire-rate for each of the frame-sizes. For each frame size, frame loss was recorded as a percentage of the number of frames transmitted. As the frames pass along the snake, they are effectively multiplied and are switched again in the next VLAN, until they exit the final VLAN in the snake. This exercises the switch as if all 672 ports were connected to separate traffic sources/sinks. This configuration is the same as that used in the Tolly Group test.

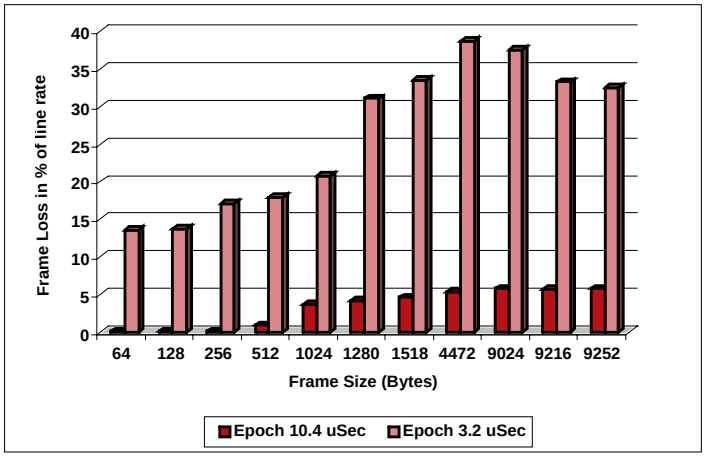



**Figure 1: Force10 Layer 2 Frame Loss**

Force10 TeraScale E-Series Competitive Test

# Gigabit Ethernet Frame Loss with Mixed Traffic Types

## Test Objectives

Confirm Force10's claim to provide zero-loss, wire-rate performance at all frame-sizes as stated in Tolly Group test Report 204147.

Verify the effects of small amounts of non-unicast background traffic in a separate VLAN on unicast performance within the 624-port L2 Snake.

## EANTC Analysis

Results show that the unicast performance offered by Force10's TeraScale E1200 degrades significantly as soon as the switch fabric is asked to simultaneously handle even small amounts of multicast, broadcast or flooded unicast frames.

Even small amounts (10 % rate on three ports) of non-unicast background traffic affect the TeraScale switch fabric's efficiency and degrades the frame loss results on the 624-port snake unicast traffic.

When we simulated a broadcast storm in the 3-port VLAN, sending 64-byte frames at 100 % load, the combination of Epoch 3.2 µSec and L2 broadcast storm results in 60 % frame loss.

Even with the best possible Epoch setting (10.4 µSec), this small amount of background non-unicast traffic causes 20 % additional frame loss for all frame sizes.

We found that even flooded unicast frames (no MAC address in L2 forwarding table) had the same effect.

Although the non-unicast traffic was contained in a three-port isolated VLAN configured on a totally separate line card from the unicast snake traffic, the non-unicast traffic had a dramatic effect on aggregate unicast performance.

**Test Highlights**

- **The addition of 10 % non-unicast traffic on three ports in a separate VLAN, increases aggregate unicast frame loss from 0.08 % to 20 % or more.**
- **Force10's claim that *"The E-Series switch fabric provides non-blocking connectivity along with advanced queuing, multicast, and jumbo frame support"* was not verifiable. Multicast traffic had significant impact on unicast frame-loss.**

## Test Configuration and Methodology

The Device under Test (DUT) was configured with a 624-port L2 snake, using the first 13 line cards in the chassis.

The snake comprised 312 two-port VLANs; VLAN 2-313 with each VLAN having two consecutive untagged ports as members. The individual VLANs were interconnected via external cables to form the snake; weaving traffic in and out of the switch and across the switch fabric. Force10 line cards have no local on-board switch fabrics, all traffic, even between adjacent ports traverses the main switch fabric.

On the 14th line card, not part of the snake, we configured three gigabit ports into a separate VLAN; VLAN 998. We used this VLAN to contain the non-unicast background traffic, so it was not flooded into the snake.

In each test run only one type of traffic (multicast, L2 broadcast etc.) was generated in parallel to the L2 unicast traffic on the 624-port snake.

L2 forwarding-table aging time was disabled for all VLANs. All control plane protocols were disabled on the ports used by the snake to ensure no control protocol frames are being sent by the DUT as this might interfere with the test.

This test conforms to the methodology for measuring frame-loss as specified in RFC 2544 and conforms to the definition of frame-loss specified in RFC 1242.

Bidirectional unicast traffic was transmitted between the two SmartBits ports at either end of the snake with the traffic rate set to 100 % wire-rate for each of the frame-sizes tested. For each frame size, frame loss was recorded as a percentage of the number of frames transmitted.

As the frames pass along the snake, they are effectively multiplied and are switched again in the next VLAN, until they exit the final VLAN in the snake;. This exercises the switch as if all 672 ports were connected to separate traffic sources/sinks.

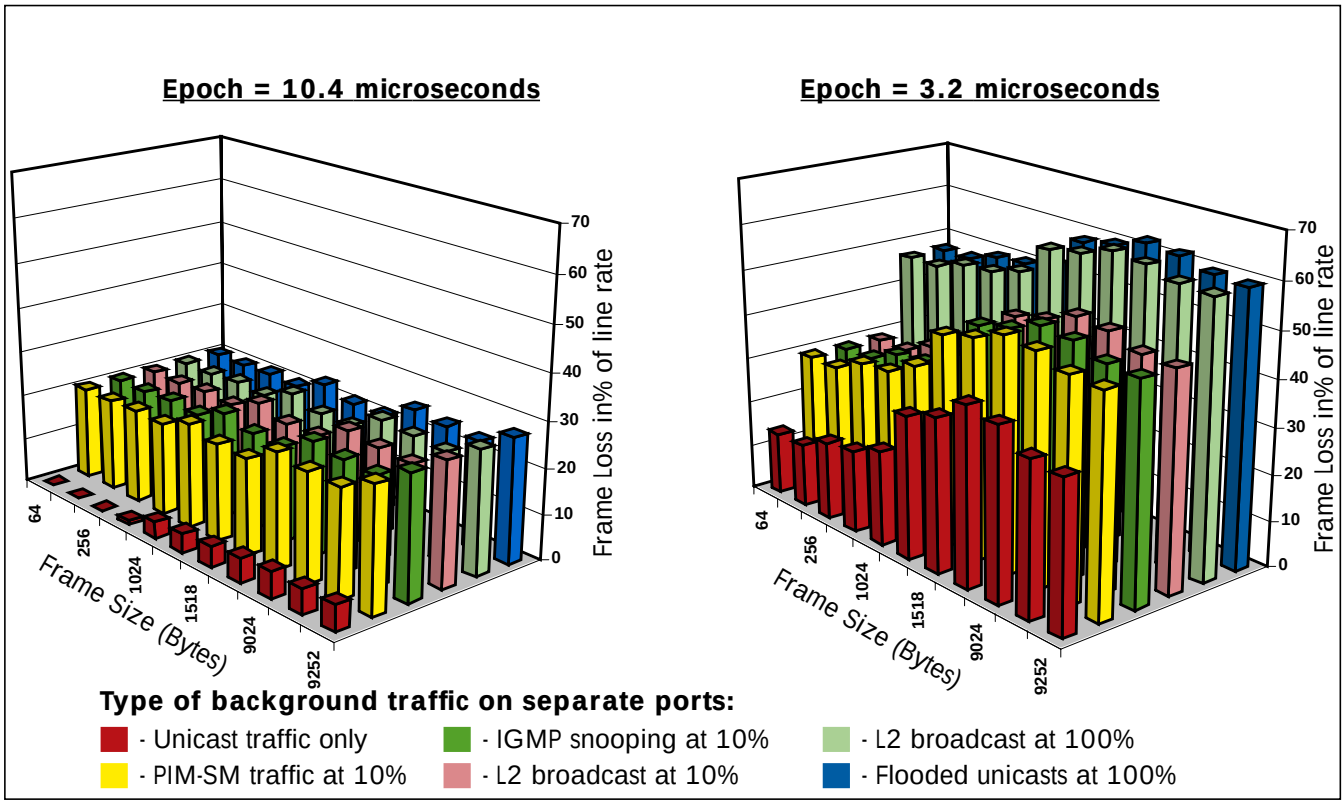



**Figure 1: Force10 – Effect of Mixed Background Traffic on Unicast Frame Loss**

**Force10 TeraScale E-Series Competitive Test**

# Gigabit Ethernet L2 Frame Loss with 1-million ACLs

## Test Objectives

This test assesses the validity of Force10's 1-million ACLs Test as described in the Tolly Group test Report 204147, which claims that the application of such a large ACL has no effect on the E1200's throughput.

According to Force10's line card installation manuals shown in *Figure 11: Excerpt from Force10 Manual: Line Card Feature Highlights*, the Gigabit and 10-Gigabit cards have identical ACL scalability. We therefore ran the test in two configurations; one applied to Gigabit Ethernet interfaces, the other applied to 10 Gigabit Ethernet interfaces (the same as the Tolly Group's test).

This test will also look further into the claims made by Force10 regarding other scalability metrics associated with their line cards.

## Force10/Tolly Group Claims:

According to Tolly Group Test Report 204147,

- *"Engineers generated over 1-million ACLs and configured the test suite for »snake« traffic".*
- *"Engineers measured the bidirectional zero-loss 10-Gigabit Ethernet throughput across 56 10-GbE ports (14 four-port 10GbE line cards) with over 1-million ACLs active".*
- *"Up to 1.4 million ACLs per system with real-time insertion capability, without creating a security issue".*

We believe it is unlikely that the Force10 E-Series was actually configured with 1-million separate Access Control Lists (ACLs). It seems more likely the Tolly Group test report actually meant to say a single ACL with 1-million Access Control List Entries (ACEs), also called ACL rules or statements.

An access control list comprises a number of statements (rules) in a list format. The statements are interpreted by the switch in sequence, one statement at a time, and compared to the current packet being processed. If there is a match, the rule is applied and no further ACL lookup continues for that packet.

**Test Highlights**

- ➜ **EANTC discovered the 1-Million ACL test results in the Tolly Group Test Report 204147 used a Force10 E1200 that was configured with a single access list of 20,993 entries, not 1-Million, as claimed.**
- ➜ **Tolly Group's ACL cannot be applied to a single 10GE interface. The E1200 runs out of CAM space. Tolly claims this ACL was applied to all four interfaces on each line card.**
- ➜ **Force10 claims 128K IPv4 ACL entries per port-pipe (Port-Pipe = 24 x GigE ports, or 2 x 10GE ports), in fact just 13K CAM entries are supported for IPv4 ingress ACLs and just 1K entries for IPv4 egress ACLs.**
- ➜ **A 1,000-entry ACL applied as an Ingress ACL will consume 1,000 CAM entries for each interface to which it is applied.**
- ➜ **When CAM is exhausted, the next interface is programmed with a partial ACL. All traffic from this interface is discarded by the partial ACL, which has an implicit DENY ALL statement as the last entry.**

By default, ACLs have an implicit DENY ALL statement at the end of the list. This is inserted automatically by the operating system of the switch. Unless the user specifically enters a PERMIT ALL statement to override this action, any packets not matching all the statements in the list will by default be discarded.

An ACL with 1,000 statements (rules) is said to be a 1,000-entry ACL. In the case of the Tolly Group tests, that would mean the switch was configured with an ACL comprising 1-million separate statements (rules).

Having centrally defined the rules in the ACL, the ACL is then applied to individual interfaces on the switch where the rules are to be enforced. The action of associating an ACL with an interface writes the ACL into the allotted CAM partition on the line card.

## Force10 TeraScale E-Series — CAM Scalability

As can be seen in the extracts from the Force10 line card installation guides, each line card (Gigabit or 10-Gigabit) supports an 18Mbyte Content Addressable Memory (CAM) per Port-Pipe. (A port-pipe is the fabric-channel into the switch fabric, and there are two port-pipes per card. On a 48-port Gigabit Ethernet Card, the first 24 ports are on Port-Pipe 0, the second 24 ports are on Port-Pipe 1. For a 4-port 10-Gigabit line card, the split is two ports per port pipe.)

```
Force10#sho cam profile all

-- Line card 0 --(LC-EF-GE-48T)
CamSize          : 18-Meg
                 : Current Settings
Profile Name     : STANDARD
L2 Fib           : 32K entries
L2 Acl           : 1K entries
Ipv4 Fib         : 256K entries
Ipv4 Acl         : 13K entries
Ipv4 Flow        : 24K entries
EgL2 Acl         : 1K entries
EgIpv4 Acl       : 1K entries
EPI              : 8184  entries
Ipv6 Fib         : 8  entries
Ipv6 Acl         : 0  entries
Ipv6 Flow        : 0  entries
EgIpv6 Acl       : 0  entries


-- Line card 2 --(LC-EF-10GE-4P)
CamSize          : 18-Meg
                 : Current Settings
Profile Name     : STANDARD
L2 Fib           : 32K entries
L2 Acl           : 1K entries
Ipv4 Fib         : 256K entries
Ipv4 Acl         : 13K entries
Ipv4 Flow        : 24K entries
EgL2 Acl         : 1K entries
EgIpv4 Acl       : 1K entries
EPI              : 8184  entries
Ipv6 Fib         : 8  entries
Ipv6 Acl         : 0  entries
```

**Figure 1: Force10 CAM Partition Sizes (Software Version 6.1.2.4)**

Unlike the Catalyst 6500, which has separate TCAM for FIB, ACLs and QoS, the Force10 TeraScale E-Series has a single monolithic piece of TCAM that is then logically partitioned to serve specific roles.

In a previous version of FTOS (Version 6.1.2.4), there was a CLI command that allowed the user to look at exactly how the 18Mbyte CAM on each line card (port-pipe) was partitioned, and how many entries were available for each specific partition.

We reverted the E1200 back to FTOS Version 6.1.2.4 and issued the SHOW CAM PROFILE command. The resulting display, captured below, shows the true scalability, not only of ACLs, but many other areas and differs from the published Force10 specifications.

This command output contradicts Force10's line card specifications. For instance, in their literature Force10 claims they support 256K L2 FIB entries per line card, whereas this shows they support 32K. A fact confirmed by the MAC address cache tests conducted as part of these EANTC-validated tests.

Force10 also claims to support 1-Million ACL entries, yet each line card can only support 13K IPv4 entries per port-pipe as confirmed in the EANTC ACL test.

Further, note how in the Tolly Group test, Force10 and Tolly claimed the E1200 demonstrated wire-rate zero loss IPv6 forwarding on a full-mesh of eight 10-gigabit ports. This test may have been limited to eight ports because they only support eight IPv6 FIB entries in the CAM.

Note also the profile names beginning with Eg - these relate to Egress ACL entries, and are substantially less in number than those available for Ingress ACLs.

In FTOS Version 6.2.1.3, the current software version at the time of testing, we issued the same command, but found that the command no longer shows any detailed information about CAM partitions:

```
F10-E1200#show cam profile all

-- Chassis Cam Profile --

  --- chassis cam profile not configured ---
```

**Figure 2: Force10 CAM Profile Command (Software Version 6.1.2.4)**

## Test Results

*Figure 3: Force10 E1200 Ingress Access Control List (ACL) Capacity* summarizes the test results.

**Run #1:** We applied the 5,000-entry ACL to GI0/0 and GI0/1 as an Ingress ACL without incident. At each stage, we checked that the ACL was correctly installed in CAM. *Figure 4: Number of CAM Entries Consumed* confirms the number of CAM entries consumed by the ACL for each port to which it is applied.

With the ACL applied to two interfaces, rather than just 5,000 CAM entries being consumed, the E1200 has consumed 10,000 out of the available 13,000 entries, as each interface needs its own personal copy of the whole ACL to work.

To prove the limitation shown by the `show cam profile all` command, we tried to apply the same ACL to interface gi0/2, the third interface on the line card.

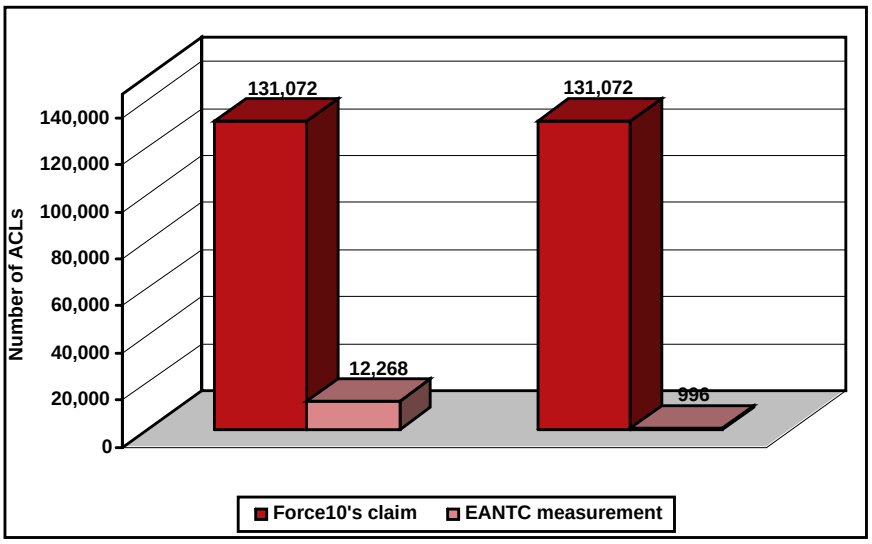



**Figure 3: Force10 E1200 Ingress Access Control List (ACL) Capacity**

```
F10-E1200#show ip accounting access-list TEST-5K-ACL | fi 0/0
Extended Ingress IP access list TEST-5K-ACL on GigabitEthernet 0/0
Total cam count 5000
 seq 5 deny tcp any any eq 26024
 seq 10 deny tcp any any eq 28863
```

**Figure 4: Number of CAM Entries Consumed**

The console log shown in *Figure 5: Error Messages When Applying 5,000 Entry ACL On Three Interfaces* confirmed that the third ACL did not fit into hardware completely.

The last 2,732 entries never made it into the CAM for gi0/2, therefore, the final `permit ip any any` statement is now missing. This means the ACL applied to gi0/2 now ended with an implicit `deny all` statement.

When we ran the full-mesh test traffic, all traffic from gi0/2 (SmartBits Port 1A3) was discarded by this partial ACL and the implicit `deny all` function.

In summary, applying a relatively small 5,000-entry ACL to all 48 ports on the line card failed after the 2nd interface was configured. The third interface was left with a partial ACL that denied everything, blocking even legitimate traffic from entering on that interface. The Catalyst 6500 refuses to apply a partial ACL if the size of the ACL exceeds the size of its available line card memory. Our testing did not find evidence to support the claim that Force10 supports a 1-million entry ACL.

```
00:41:03: %E48TF:0 %ACL_AGENT-2-ACL_AGENT_ENTRY_ERROR: Unable to apply seq 11445 onwards of access-list
TEST-5K-ACL applied on GigabitEthernet 0/2
00:41:03: %E48TF:0 %ACL_AGENT-2-ACL_AGENT_ENTRY_ERROR: Unable to apply seq 11535 onwards of access-list
TEST-5K-ACL applied on GigabitEthernet 0/2
```

**Figure 5: Error Messages When Applying 5,000 Entry ACL On Three Interfaces**

```
F10-E1200#show ip accounting access-list TEST-5K-ACL | fi 0/0
Extended Ingress IP access list TEST-5K-ACL on GigabitEthernet 0/0
Total cam count 5000

F10-E1200#show ip accounting access-list TEST-5K-ACL | fi 0/1
Extended Ingress IP access list TEST-5K-ACL on GigabitEthernet 0/1
Total cam count 5000

F10-E1200#show ip accounting access-list TEST-5K-ACL | fi 0/2
Extended Ingress IP access list TEST-5K-ACL on GigabitEthernet 0/2
Total cam count 2268
List is not in sync with running config
```

**Figure 6: Number of Installed ACL Entries in Run #1**

**Seeking assistance from the Tolly Group.** As a peer independent test lab, EANTC asked for assistance from the Tolly Group to understand the exact configuration used in the 1-Million ACL Test, plus two other tests that concerned us.

Tolly group provided EANTC with the switch configuration files for all three tests we queried.

When we received the configuration file for the 1-Million ACL test we found a single 20,993-entry ACL, not 1-Million entries as we'd expected.

The usual way ACL scalability is measured is by measuring the number of rules in an ACL the DUT is able to enforce. For instance, a 1,000-entry ACL applied to 100 interfaces shows the DUT can enforce a 1,000-entry ACL.

In the Tolly Group Test however, Force10 and the Tolly Group created the following way of reporting the ACL scalability of the TeraScale E1200.

1. The ACL was 20,993 entries in size.
2. The same ACL was applied to 56 x 10-Gigabit interfaces
3. Tolly/Force10 math computes that there are 56 x 20,993 = 1,175,608 ACL entries supported by the E1200

This interpretation does not conform to any known industry accepted standard for counting of ACLs. We believe that creating a counting practice outside the accepted norm misleads the reading audience.

Let's give an everyday example of what Tolly/Force10 would have us believe. In our example, the headmaster of a school draws up a list of ten rules that all pupils must abide by. This is the equivalent of the rules/statements in the ACL.

The headmaster photocopies the list; pins the list up at all twenty entrances into the school and asks the teaching and security staff to enforce them. This is the same as associating the ACL to each interface.

How many rules are to be enforced for the whole school? … TEN rules right?

Wrong … Tolly and Force10 would have you believe the headmaster just issued 10 rules x 20 entrances = 200 new rules for the pupils to obey.

```
F10-E1200(conf-if-te-5/0)#ip access-group million_ACL in
09:15:08: %EXW4PF:5 %ACL_AGENT-2-ACL_AGENT_ENTRY_ERROR: Unable to apply seq 61390 onwards of access-
list million_ACL applied on TenGigabitEthernet 5/0
09:15:08: %EXW4PF:5 %ACL_AGENT-2-ACL_AGENT_ENTRY_ERROR: Unable to apply seq 61480 onwards of access-
list million_ACL applied on TenGigabitEthernet 5/0
```

**Figure 7: Error messages while applying 20,993-entry ACL to a single interface**

```
F10-E1200#show ip accounting access-list million_ACL
Standard Ingress IP access list million_ACL on TenGigabitEthernet 5/0
Total cam count 12281
List is not in sync with running config
 seq 5 permit host 144.45.1.0 count (0 packets)
 seq 10 permit host 144.45.1.1 count (0 packets)
```

**Figure 8: Number of Installed ACL Entries in Run #2**

We believe the test results as reported by the Tolly Group lead the reader to believe the Force10 E-Series can support 1 Million ACLs, which we did not find.

Next, we investigated whether we could apply the Tolly Group ACL (20,993 entries) to all four interfaces on a ten Gigabit Ethernet line card, as would have been necessary to run the original Tolly Group 1-Million ACL test.

**Run #2.** Having received from the Tolly Group the exact configuration file used in the 1-Million ACL test, we tried to apply this to all four interfaces on a ten Gigabit Ethernet line card, just as Tolly Group must have done. We approached this task in stages, applying the ACL to one interface at a time.

The test was performed using FToS version 6.2.1.3 and proves that the use of Beta Version 4.4.2.107 allowed Force10 to configure the switch in ways not possible with current production software. A formal response letter was received from Force10 and admitted that they re-defined CAM partitions to succeed in this test, confirming our conclusions. The full letter is attached in the appendix.

*Figure 7: Error messages while applying 20,993-entry ACL to a single interface* contains an extract from the console log, which shows what happened when we tried to apply the 20,993-entry ACL (as used by Tolly) to the first 10-Gigabit Ethernet interface on the 4-port line card.

The console log shown in *Figure 8: Number of Installed ACL Entries in Run #2* confirmed that Force10's latest production software FtoS Version 6.2.1.3, can't load the ACL on a single interface.

Finally, the `show ip accounting access-list` command confirms that out of 20,993 ACL entries, only 12,281 could be successfully written to CAM. As the CAM must dedicate a further 20,993 entries to apply the ACL to a second interface, a further ((20,993 – 12281) + 20,993) = 29,705 extra CAM entries must be allocated to run this test (that's over double the number provided by default).

Either special line card hardware with larger CAM memory was used, or the CAM was specially partitioned to support this test, deleting some of the other partitions and dedicating their CAM space to this ACL. Either way, the Tolly Group overstates the test results about the performance and scalability of the Force10 E1200 switch.

We then repeated the test and tried to assign the Tolly Group ACL to multiple Gigabit Ethernet interfaces. The exact same results as above were recorded, we could not apply the ACL to even a single interface.

**Run #3.** Next we wanted to answer the following questions:

- How many Egress IPv4 ACL entries does the E1200 support?
- Does each port need its own copy of the ACL, as it did with Ingress ACLs?
- Does the E1200 create "partial" Egress ACLs on an interface when it runs out of CAM space?

To investigate this we created a much smaller ACL of just 150 entries. We calculated that if each interface needed its own copy of the ACL in CAM, just like Ingress ACLs, we would be able to apply the ACL to six interfaces without oversubscribing the CAM.

For the first six interfaces, gi0/0 through gi0/5, the ACL was applied without any problems. shows the

```
F10-E1200(conf-if-gi-0/6)#ip access-group TEST-150-ACL out
03:32:05: %E48TF:0 %ACL_AGENT-2-ACL_AGENT_ENTRY_ERROR: Unable to apply seq 455 onwards of access-list
TEST-150-ACL applied on GigabitEthernet 0/6
03:32:05: %E48TF:0 %ACL_AGENT-2-ACL_AGENT_ENTRY_ERROR: Unable to apply seq 545 onwards of access-list
TEST-150-ACL applied on GigabitEthernet 0/6
03:32:05: %E48TF:0 %ACL_AGENT-2-ACL_AGENT_ENTRY_ERROR: Unable to apply seq 635 onwards of access-list
TEST-150-ACL applied on GigabitEthernet 0/6
```

**Figure 9: Error messages while applying 150-entry ACL to a 7th interface**

```
F10-E1200#sho ip accounting access-list TEST-150-ACL | fi 0/6
Extended Egress IP access list TEST-150-ACL on GigabitEthernet 0/6
Total cam count 96
List is not in sync with running config
 seq 5 deny tcp any any eq 26024
 seq 10 deny tcp any any eq 28863
```

**Figure 10: Number of Installed ACL Entries on gi0/6 in Run #3**

error messages we received on the console when we applied the ACL list to the seventh interface, gi0/6.

Once again, the `show ip accounting access-list TEST-150-ACL` confirmed that interface gi0/6 had a partial access-list applied (see *Figure 10: Number of Installed ACL Entries on gi0/6 in Run #3*).

In summary, we saw exactly the same problems with the egress ACLs as we had with the Ingress ACLs. We confirmed that each interface needs a full copy of the ACL in CAM and that the CAM partition dedicated to Egress IPv4 ACLs is 1K entries in size.

This test confirms the CAM partition information displayed with the `show cam profile` command is accurate, even under FTOS Version 6.2.1.3.

### EANTC Analysis

First, we found that the original `show cam profile` command has been modified by Force10 and now does not show any details about the actual CAM scalability and partitioning.

EANTC can confirm that contrary to Force10's marketing claims, no CAM partition sizes can be configured by the user. Only sub-partitioning of the IPv4 CAM partition is allowed.

EANTC also found that Force10's claim to support 1-Million ACL entries to be unsubstantiated. Our tests discovered that the E1200 suffers from severe ACL scalability limitations caused by small CAM partitions and the need for each port to have its own "personal copy"of the ACL resident in CAM.

EANTC can confirm that Force10's 1-Million ACL test was actually carried out using a single Ingress ACL of 20,993 entries in size and that this would reflect an ACL scalability limit of 20K. Further investigations found this ACL could not even be applied to a single 10 Gigabit Ethernet interface without overloading the CAM. So we believe Force10 used either special hardware or software to pass this test.

### Test Configuration and Methodology

- Configure a 48-port L2 full-mesh test configuration.
- Create a 5,000 entry non-contiguous, non-repeating ingress ACL and apply this to interfaces, one interface at a time.
- Confirm the 5,000 entry ACL can be applied to all 48 interfaces on the line card and that the rules within the ACL are correctly enforced on each port.
- Note any problems with scalability when applying the ACL to the ports.
- Send full mesh traffic between all 48 ports in the test. Ensure the traffic does not match any DENY statements in the ACL, forcing the switch to match on the very last PERMIT IP ANY ANY statement.
- Record frame-loss and any problems.
- If the DUT passes this test, send traffic which matches the last DENY statement and confirm the full ACL is being enforced on each port.
- If the DUT passes the test, increase the size of the ACL until 1 million ACL entries are reached, as reported by the Tolly Group.

**48-port 10/100/1000 Base-T Ethernet Line Card**

Feature Highlights and Installation Instructions

Catalog Number: LC-EF-GE-48T

Feature Highlights

- 48 ports with RJ-45 connectors that support auto-negotiation or 10/100/1000 Base-T speed.
- The EF series line cards support a single 18M user configurable CAM with flexible partition assignments. The following max entries are:
  - 256K Layer 3 IP forward information base (FIB)
  - 256K Layer 2 FIB
  - 64K Layer 3 IPv6 FIB
  - 128K Layer 2 and Layer 3 access control list (ACL) entries
  - 24K IPv6 ACL entries
- Supports online insertion and removal (OIR) of line card.
- Supports ingress and egress Layer 2 and Layer 3 ACL processing across all ports.
- Requires FTOS 6.1.1.0 or higher

**4-Port 10-Gigabit Ethernet LAN/WAN PHY Line Card**

Feature Highlights and Installation Instructions

Catalog Number: LC-EF-10GE-4P

Feature Highlights

- Uses approved 10G Small-form Factor Pluggable (XFP) laser modules. Laser modules must comply with 21 CFR 1040 Class 1 requirements.
- Pluggable XFP optics providing support for SR, LR, ER, and ZR optical interfaces.
- The EF series line cards support a single 18M user configurable CAM with flexible partition assignments. The following max entries are:
  - 256K Layer 3 IP forward information base (FIB)
  - 256K Layer 2 FIB
  - 64K Layer 3 IPv6 FIB
  - 128K Layer 2 and Layer 3 access control list (ACL) entries
  - 24K IPv6 ACL entries
- - Supports online insertion and removal (OIR) of line card.
- - Supports ingress and egress Layer 2 and Layer 3 ACL processing across all ports.
- - Ports are user configurable, on a per port basis, to be either LAN PHY or WAN PHY.
- - Requires FTOS 6.1.1.0 or higher

**Figure 11: Excerpt from Force10 Manual: Line Card Feature Highlights**

# Executive Summary — High Availability Tests

In June 2005, Cisco Systems commissioned the European Advanced Networking Test Center (EANTC) to independently validate the performance, scalability, and availability of Force10 TeraScale E-Series switches.

See http://www.tolly.com/Search.aspx?VendorID=27

In this category of tests, we attempted to reproduce a number of the original Tolly Group high availability tests found in the Tolly Group Test Report 204148. We also added further tests to verify the high availability claims made in other Force10 marketing literature.

Tolly Group Test Report 204148 focuses on high availability features and route scalability.

High availability is key to the success of Force10's high port-density network designs. If extremely large numbers of users are served by a single device, continuous system availability even under hardware component failure is vitally important.

Unfortunately, our findings cast doubts on the TeraScale E-Series' reliability and device resilience features when tested in more realistic network topologies using common features and protocols.

In the Tolly Group tests, both Route Processor Module (RPM) and Switch Fabric Module (SFM) failures were simulated by the Tolly Group, who recorded zero packet loss to test traffic, implying the Force10 TeraScale E-Series supports "hitless" control-plane or data-plane failovers.

Closer inspection of Force10's "RPM Hitless Failover Test" reveals that this failure condition only affects the control plane of the switch; the data plane is totally unaffected and continues to forward traffic, possibly why the Tolly Group recorded zero loss on the data streams under these conditions.

### ➜ Tolly Group RPM Failover Result is Not Hitless:

EANTC believes Tolly Group's test methodology presents a limited view of how the TeraScale E-Series would cope with such an RPM failure in a real-world network.

We used a multi-switch network topology and ran various L2 and L3 control-plane protocols and resilience features to see if the E-Series would deliver zero-loss failover, now that the data-plane is controlled by and reliant upon the uninterrupted functioning of the protocols and resilience features running on the RPM.

This is a better test of RPM resilience and we found that in these more realistic network conditions, where the control plane is actively involved in controlling the data plane forwarding (something not present in the Tolly Group's tests), the RPM not only failed to deliver zero packet loss, it exhibited periods of unstable behavior caused by the Virtual Router Redundancy Protocol (VRRP), lost OSPF neighbor adjacencies, failed to protect multicast traffic, and disrupted SNMP management of the switch.

Our tests show that the only traffic unaffected by the RPM failover was the L2 Snake traffic, which was not dependent on the state of the control protocols running on the RPM.

Force10 E-Series provides guaranteed system management access by dedicating individual processors to specific tasks for additional fault tolerance and redundancy, as shown in Figure 2. This architecture isolates IP routing, Layer 2 tasks, and management functionality to three individual CPUs, respectively. This ensures that the administrator can always log on to the "management CPU" (Control CPU #3, shown in the diagram), receive precise knowledge as to the cause of the problems, and take corrective action thereby avoiding disruptive system reboots.

**Figure 1: Extract from Force10 Whitepaper, "Guaranteed Access to System Management Even Under CPU Overload"**

### ➜ SFM Failure Causes Switch Fabric Collapse:

Force10 employ an 8+1 switch fabric redundancy scheme, where eight out of nine switch fabric modules must be active to maintain proper switch operation. If two fabrics fail simultaneously, the whole switch dies and all interfaces are shut down.

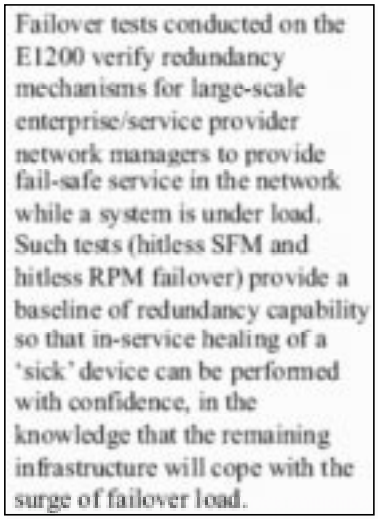

Failover tests conducted on the E1200 verify redundancy mechanisms for large-scale enterprise/service provider network managers to provide fail-safe service in the network while a system is under load. Such tests (hitless SFM and hitless RPM failover) provide a baseline of redundancy capability so that in-service healing of a 'sick' device can be performed with confidence, in the knowledge that the remaining infrastructure will cope with the surge of failover load.

**Figure 2: Extract From Tolly Report 204148**

We reproduced the L2 Snake configuration used in the Tolly Group test using Gigabit Ethernet rather than 10-Gigabit Ethernet interfaces due to lack of 10-Gigabit line cards. In this configuration, we found that we could not achieve zero-loss hitless failover recorded by the Tolly Group. As soon as an SFM was removed, the remaining eight SFMs had insufficient capacity to forward all traffic, introducing additional frame loss, which at certain frame sizes more than doubled.

We believe the Tolly Group's test methodology is limited, as it does not represent a typical network topology in which the switch would be deployed. EANTC assessed whether the E-Series could provide hitless failover in a more real-world multi-switch test topology, employing protocols and features commonly found in real networks.

In our pre-tests using FTOS Version 6.2.1.1 we found that when we removed an SFM (any SFM) it caused a collapse of the whole switch fabric and all forwarding ceased. See the individual test report for further details.

When we investigated further, we found that a unrelated software feature (sFlow Statistics) caused the Control Processor (CP) utilization to increase to 40% and was the main triggering factor in the collapse. Version 6.2.1.3 release notes showed no information relating to this problem, yet when we tested with Version 6.2.1.3 the problem appeared to be resolved.

In fact, we saw that Force10's workaround does not fix the underlying problem; it basically disables sFlow flow statistics.

### ➜ TeraScale E1200 Has Overheating Problems:

We found that removing a single line card and leaving the slot open for just ten minutes results in the whole switch fabric collapsing due to overheating, bringing down all 672 interfaces (or even 1,260 interfaces if 90-port cards had been used).

The two Switch Fabric Modules (SFMs) immediately below the open line card slot overheat and are automatically powered off to prevent permanent damage. With two SFMs shutdown, the switch fabric no longer has the mandatory eight active SFMs, and the switch is therefore totally disabled.

The Cisco Catalyst 6500s showed no such problems, even with all blanking panels removed, the Catalyst's ventilation system kept the line cards and supervisor modules cool for the whole period of the tests (well over a month).

### ➜ Force10's 3-CPU Control Plane Problems:

Force10 claims that when one of their three CPUs is overloaded, the customer will always be able to get precise information to pinpoint a problem and will still have full control over the switch to effect a remedy without having to reboot.

We found that the Control Processor (CP), responsible for all management of the switch, is reliant on the other two processors for switch status and control. Our tests proved that if one of the other processors is overloaded, the CP cannot gain access to information controlled by that processor, nor can it control those aspects of the switch controlled by the other processor.

When we tried, we had lengthy timeouts on the CLI where the switch would not allow the entry of any further commands while the previous command timed out due to no response. In one instance, we were effectively locked-out of the console for over 20 minutes.

### ➜ FToS Modular Operating System Claims Are Unverifiable:

In the product specification area of the Tolly Group report, an area Tolly Group disclaims verifying for authenticity, Force10 made the following claim described in *Figure 3: Excerpt from Tolly Report 204148*.

Our tests with Version 6.2.1.3 software, show that FToS is not modular. It is a single monolithic software image with no ability to provide software upgrades to individual software modules. It does not provide the ability to restart and control individual software processes and does not support zero-loss hitless software upgrades.

In separate tests, we put a new version of Cisco IOS offering Software Modularity on the Catalyst 6500 through its paces. We found that a number of components in IOS have been modularized. For more details see: *http://www.eantc.com/downloads/test_reports/2003-2005/EANTC-Summary-Report-Cisco-ION.final3.pdf*

Force10 Networks, Inc.
E-Series E1200
Product Specifications*

Resilient architecture

- Distributed control plane with separate processors and memory for Layer 2 switching, Layer 3 routing and system management
- Robust Force10 FTOS™ modular software delivers full functionality Layer 2 switching and Layer 3 routing while also supporting zero packet loss hitless restarts and software upgrades

**Figure 3: Excerpt from Tolly Report 204148**

## Summary

Our tests revealed reliability and resilience problems not exposed by Force10's Tolly Group tests.

The E1200's cooling system design needs to be investigated. It is almost unavoidable that a customer's network technician will leave a card slot open without a blanking panel, either through forgetfulness, or just because they cannot find a blanking panel and decide to "risk it."

EANTC believes that the overheating problem on the E1200 poses a significant business risk and that Force10 need to address this design fault.

When taken as a whole, the Force10 switch did not live up to the high availability claims made in Force10's marketing literature, showing problems with overheating, RPM resilience, SFM resilience, recovery from CPU overload conditions and unstable VRRP router resilience software.

Force10 TeraScale E-Series Competitive Test

# RPM Hitless Failover (Tolly Test Simulation)

## Test Objectives

In Tolly Group Test Report 204148, Force10 claims to provide hitless fail-over to the Standby RPM in the event the online RPM fails and that this action results in zero packet loss. Assess the validity of this claim using directly connected hosts simulating L2 unicast traffic in a 672-port "Snake" configuration, the same as in the Tolly Group Tests.

The objective of this test is to measure whether there is any additional packet loss for established flows when an RPM failure is simulated by removing the online RPM.

## EANTC Analysis

EANTC was able to reproduce the results of the RPM Hitless Failover test published by the Tolly Group (Published on http://www.tolly.com/Search.aspx?VendorID=27 Report 204148), using the same test topology.

The test conducted by the Tolly Group caused a failure of the control plane only, the data plane remained intact throughout the test, so it was not unexpected that the traffic passing through the switch's data plane was unaffected.

EANTC feels however, that the Tolly Group test did not assess the E1200's claim of "Hitless RPM Failover" in a realistic network setting, employing control plane protocols that one would expect to find in a real-world production network. EANTC therefore decided to introduce an additional test to investigate whether the RPM failover would still be hitless if a more realistic test scenario is used and multiple control plane protocols and traffic types were present. These additional tests can be found in the RPM Hitless Failover (Systems Test) report.

**Test Highlights**

- **EANTC confirmed that the Force10 TeraScale E1200 supports RPM hitless failover when tested in the same configuration as used by the Tolly Group. The RPM failover did not introduce any additional packet loss.**

## Test Configuration and Methodology

This test used the same 672-port L2 snake topology as the Tolly Group test.

Prior to the doing the failover test, we ran a baseline test to record the level of packet loss to expect. Contrary to the Tolly Group reports, the TeraScale E-Series is not wire-rate non-blocking at all frame sizes. Having determined normal baseline operation, we re-ran the test, this time removing the active RPM. Once again frame loss was recorded and compared to the baseline result to determine if any additional loss was introduced by the RPM failover.

Test traffic was run at 100 % load, using 64-byte frames.

Force10 TeraScale E-Series Competitive Test

# RPM Hitless Failover (Systems Test)

## Test Objectives

n Tolly Group test report 204148, Force10 claims to provide hitless fail-over to the Standby RPM in the event the online RPM fails and that this action results in zero packet loss.

In the Tolly Group test, the RPM failover tests were conducted in a single switch, with a limited test configuration, without any of the services or traffic types that might be expected in a true production network.

This RPM failover test designed by EANTC takes a systems approach, simulating part of a typical network design. Our test used three Force10 TeraScale E-Series switches. One was configured as a L2 wiring closet switch, dual-homed to two L3 core/distribution switches that employed VRRP for first hop router resilience. The test also adds OSPF routing to simulate an end-to-end enterprise network. In addition, IP multicast traffic is added to the traffic mix as multicast is an important component of many production networks, especially mission critical trading rooms of large financial services companies, where multicast availability is critical (See the separate IP Multicast test reports for further multicast specific resilience tests).

We also added various services typically found in a production network to assess the DUTs performance under these more realistic network conditions.

The objective of this test is to assess whether the Force10 switch is able to provide hitless RPM failover in these more typical network conditions, or confirm this capability is only achievable in a lab test where the test traffic is not dependent on the state of the control plane protocols. Moreover, the Tolly test only focused on what existing traffic flows experienced, our test also investigated the effect on the establishment of new flows.

**Test Highlights**

- ➜ **Unicast traffic in the "Tolly" L2 Snake was unaffected by the RPM Failover.**
- ➜ **The RPM Failover triggered multiple instances of VRRP flapping. The effect on the unicast traffic can be clearly seen on the Unicast (Existing Flows) Results graph.**
- ➜ **VRRP flapping also caused OSPF instability. OSPF adjacencies on the VRRP protected VLAN/Subnet went down multiple times during the test.**
- ➜ **As seen in other tests, we observed intermittent OSPF problems on multiple switches. OSPF neighbors were in "full" state, the OSPF database was correct, but the switch failed to populate its routing table.**
- ➜ **After the RPM failover event, multicast flows suffered two periods of disruption, the first affected all flows for 20 seconds, the second lasting 176 seconds, affecting 50% of the flows.**
- ➜ **After the RPM failover event, new flows were not established until 145 seconds had elapsed. During this period all frames from the new flows were flooded.**
- ➜ **SNMP management was unavailable for 45 seconds, and then suffered intermittent outages of 4–6 seconds duration.**

## Test Results

**Existing Unicast Flows.** *Figure 1: Unicast (Existing Flows) Rx Rate* shows the disruption to "existing" unicast traffic. The disruption (flapping) was caused by VRRP becoming unstable shortly after the RPM failover. The VRRP master flip/flopped between DUT(A) and DUT(B), changing ownership numerous times before finally settling down. Fortunately, there was an alternative path for the "existing" traffic via DUT(B).

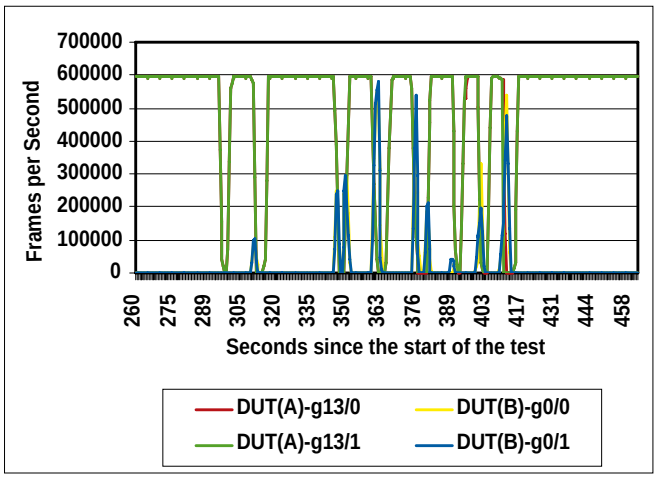



**Figure 1: Unicast (Existing Flows) Rx Rate**

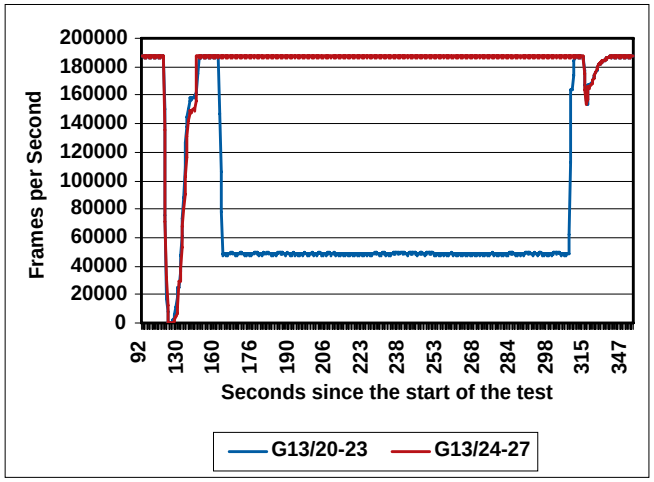



**Figure 2: Multicast Results**

*Figure 3: VRRP Flapping* contains an extract from the console log from DUT(B), which shows an example of the VRRP flapping causing unstable OSPF adjacencies.

**Multicast Flows.** The mission-critical multicast traffic fared worse than unicast. As can be seen from the *Figure 2: Multicast Results*, after initially recovering, half of the multicast receivers lost a significant portion of the traffic a short time later and did not recover for 176 seconds.

**New L2 Unicast Flows.** Immediately after RPM failover, we started the "new flows" and tried to establish communication using a full mesh traffic pattern. DUT(A) only needed to learn a total of 800 MAC addresses (100 hosts per port).

The traffic from these new flows was flooded for 145 seconds, causing over-subscription of all links in VLAN 998, in turn this resulted in 34 % packet loss.

**Results for "Tolly" L2 Snake (VLAN 2–313).** The traffic on the L2 snake remained unaffected by the failover and lost no additional traffic. This shows the Tolly Group tests only provided a limited view of the TeraScale E-Series' high availability.

```
4d1h8m: %RPM0-P:RP2 %VRRP-6-VRRP_MASTER: vrid-1 on Vlan 999 entering MASTER.
4d1h8m: %RPM0-P:RP2 %VRRP-6-VRRP_BACKUP: vrid-1 on Vlan 999 leaving MASTER.
4d1h9m: %RPM0-P:RP2 %VRRP-6-VRRP_MASTER: vrid-1 on Vlan 999 entering MASTER.
4d1h9m: %RPM0-P:RP2 %VRRP-6-VRRP_BACKUP: vrid-1 on Vlan 999 leaving MASTER.
4d1h9m: %RPM0-P:RP1 %OSPF-5-ADJCHG: OSPF Process 1, Nbr 7.7.7.7 on interface Vl 999 change state from
FULL to DOWN
4d1h9m: %RPM0-P:RP1 %OSPF-5-ADJCHG: OSPF Process 1, Nbr 7.7.7.7 on interface Vl 999 change state from
NULL to DOWN
4d1h9m: %RPM0-P:RP1 %OSPF-5-ADJCHG: OSPF Process 1, Nbr 7.7.7.7 on interface Vl 999 change state from
DOWN to INIT
4d1h9m: %RPM0-P:RP1 %OSPF-5-ADJCHG: OSPF Process 1, Nbr 7.7.7.7 on interface Vl 999 change state from
INIT to EXSTART
4d1h9m: %RPM0-P:RP1 %OSPF-5-ADJCHG: OSPF Process 1, Nbr 7.7.7.7 on interface Vl 999 change state from
EXSTART to EXCHANGE
4d1h9m: %RPM0-P:RP1 %OSPF-5-ADJCHG: OSPF Process 1, Nbr 7.7.7.7 on interface Vl 999 change state from
EXCHANGE to LOADING
4d1h9m: %RPM0-P:RP2 %VRRP-6-VRRP_MASTER: vrid-1 on Vlan 999 entering MASTER.
4d1h9m: %RPM0-P:RP2 %VRRP-6-VRRP_BACKUP: vrid-1 on Vlan 999 leaving MASTER.
```

**Figure 3: VRRP Flapping**

## EANTC Analysis

This test clearly shows that Force10's claim to support RPM Hitless Failover is not possible if the test is conducted in a real-world scenario. All configured services, except the existing directly connected L2 snake flows, which were not dependent on the stability of the control plane protocols, experienced forwarding problems after the active RPM was removed.

These test results are not reassuring when one considers that Force10 promotes the idea of "collapsing" thousands of hosts/servers into fewer, but hugely denser, core switches supporting up to 1,260 ports. To make this feasible, the switches must provide absolutely "rock solid" device resilience.

The Tolly Group's test methodology gave the illusion that RPM Hitless Failover actually worked. Our tests however, employing a real-world network design, requiring the control plane to failover whilst maintaining the state of control plane protocols, showed that the Force10 switch failed to provide a hitless failover in the event of an online RPM failure.

## Test Configuration and Methodology

The topology shown in *Figure 4: Test Topology* was used to simulate part of a typical network design. We used the "Multi-User" feature on the SmartBits to split the SmartBits into four logical pieces of test equipment.

- SMB-1a — Emulates a multicast distribution tree using PIM Sparse Mode and IGMP Version 2.
- SMB-1b — Emulates the traffic between a L2 wiring closet and the network, this represents our "existing" unicast flows, already established before the RPM failure.
- SMB-1c — Emulates 800 directly connected hosts. These hosts will try to establish communication in a full mesh traffic pattern immediately after the RPM failure. These represent clients/servers seeking to establish new sessions with each other.
- SMB-1d — Provides background traffic load (same as in the Tolly Group tests). This is a 624-port snake, leaving the 48-port line card in slot 13 to provide the other interfaces in the test.

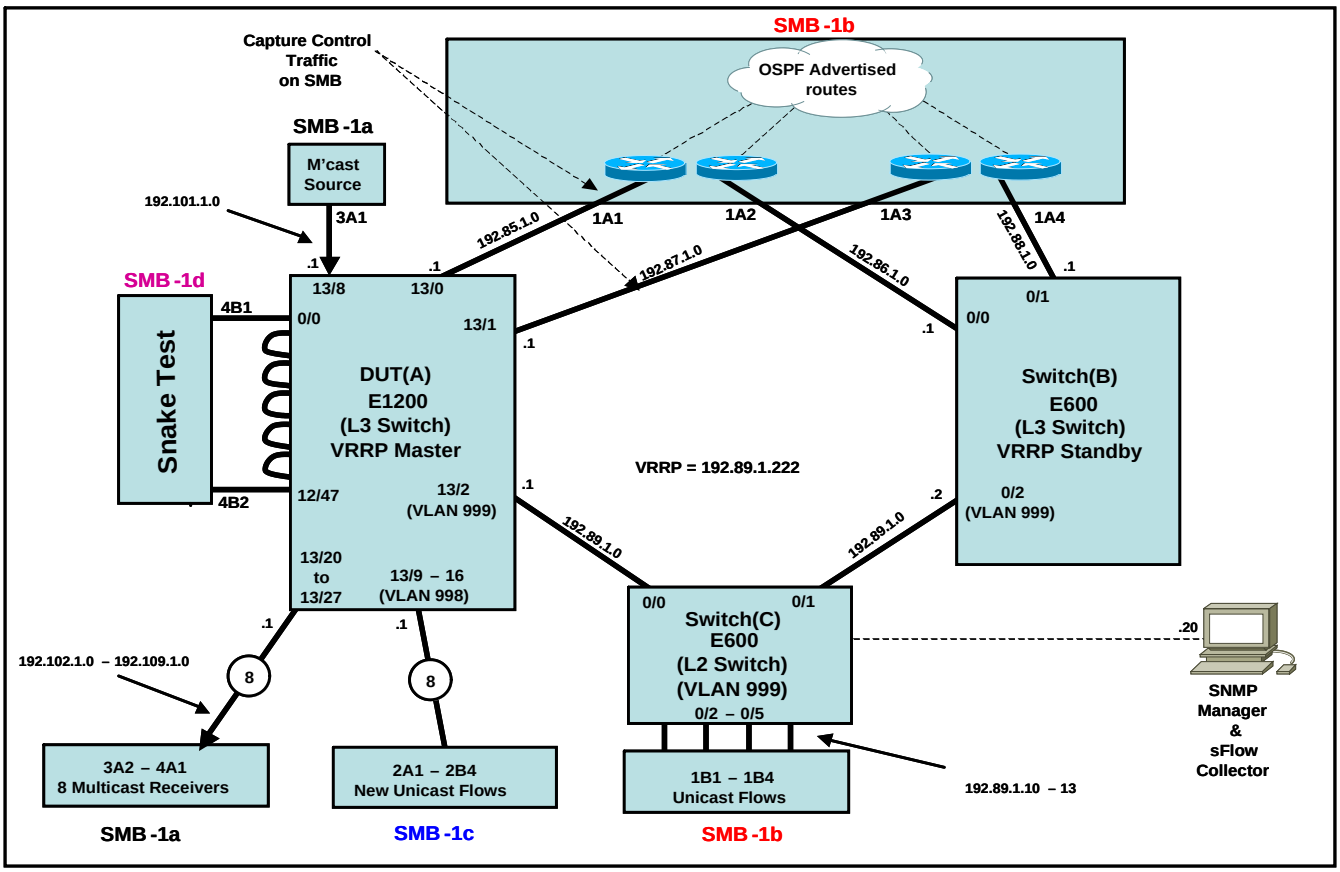



**Figure 4: Test Topology**

- DUT(C) — Was configured with a single VLAN (VLAN 999) and acted as our dual-homed wiring closet switch in the test topology. All ports were untagged and no spanning tree was running.
- DUT(A) — The Force10 TeraScale E1200 was configured with redundant RPMs and a full complement of 9 switch fabric modules, providing 8+1 switch fabric redundancy and represents the device under test (DUT). The switch was configured as a L3 core/distribution switch running OSPF and PIM-SM routing protocols. In a real network this switch would serve multiple wiring closets, with potentially hundreds or even thousands of users.
  On VLAN 999, the switch ran VRRP first hop router resilience protocol and was configured to become the VRRP master, responsible for forwarding all traffic received from the protected VLAN.
  DUT(A) also functioned as the PIM-SM Rendezvous Point (RP), which was statically configured and served 8 multicast receivers attached to ports 13/20 - 27; only DUT(A) supported multicast traffic in this test. A single multicast source sent traffic to 100 groups, creating a total of 100 S,G mroutes.
- DUT(B) — Was also configured as a L3 core/distribution switch and also ran OSPF. DUT(B) has no multicast receivers and does not take part in the multicast portion of the test.
  DUT(B) runs VRRP on VLAN 999 and is configured to become the VRRP standby router.

In addition to the above, we also configured a high powered PC to act as an SNMP management station and sFlow Statistics Collector, receiving management traffic from DUT(A). The SNMP management station polled an IP address on DUT(A) associated with a loopback interface as this should remain available throughout the test.

Finally, we used the packet capture capability on the SmartBits to capture the OSPF control plane traffic from DUT(A), this was to provide further analysis of what transpired from a routing protocol perspective when the active RPM in DUT(A) failed.

All switches in the test were TeraScale E-Series running the latest production software, FTOS Version 6.2.1.3 we repeated the test multiple to ensure repeatability.

Test traffic was started in the following sequence:

4. OSPF routing and existing unicast flows between DUT(C) via the VRRP master, DUT(A) and the emulated routers on the SmartBits at the top of the diagram. All OSPF neighbor states and the route table were checked as correct before the unicast "existing flows" traffic was started. Traffic was sent at 20% of line rate.
5. PIM-SM and multicast traffic. Prior to starting the multicast traffic we checked the correct IGMP and PIM-SM multicast routing state was present. Only after we confirmed the correct state did we start the multicast traffic.
6. Next we started the L2 background traffic on the snake. We also started SNMP polling, sFlow collection (configured on less than 50% of the interfaces) and packet capture on the SmartBits.

Once all traffic (with the exception of the "new flows") was started, we waited a short while for the network to stabilize, checked the VRRP state on switches A & B, confirming DUT(A) was the master and DUT(B) the standby. We only proceeded further when we were satisfied the network was stable and in its initial acquiescent state.

Next we failed the active RPM in DUT(A). Immediately after the RPM was removed we started the "new flows", trying to establish communication during the recovery period of the RPM.

Once all traffic returned to normal and the new flows were fully established we stopped the test and gathered the results.

**Force10 TeraScale E-Series Competitive Test**

# SFM Hitless Failover (Tolly Test Simulation)

## Test Objectives

The Force10 TeraScale Architecture employs a multi-card switch fabric design comprising nine switch fabric modules (SFMs).

In the TeraScale E1200, all nine modules are in active use by default. In the event that one module fails, Force10 claims the remaining eight modules have sufficient capacity to continue forwarding all traffic without loss, thereby providing a "Hitless SFM Failover."

Force10 commissioned the Tolly Group to test this high availability feature and their findings are documented in Tolly Group Test Report 204148.

The objective of this test is to assess the validity of this claim using the same 672-port "L2 Snake" configuration used in the Tolly Group Tests. The test measures whether any additional loss for established flows is introduced when an SFM fails.

Test traffic will use the same frame size and transmit rate as used in the Tolly Group tests.

**Test Highlights**

- ➜ **EANTC was unable to reproduce the SFM hitless failover results measured by the Tolly Group. When one out of 9 SFMs is removed, the received frame rate drops from 1,486,247 fps to 1,398,014 fps per port.**
- ➜ **EANTC tests show that contrary to Force10 claims, the capacity of the remaining 8 SFMs is insufficient to maintain the full traffic load from 672 Gigabit Ethernet ports.**
- ➜ **Additional EANTC tests showed that with only 8 out of 9 SFMs active, packet loss increased to 40 % worst case, confirming the SFM failure has a detrimental effect on throughput at all frame sizes and is certainly not "Hitless."**

## Test Results

Contrary to Tolly Groups Test Reports 204147 and 204148, the Force10 TeraScale E-Series is not wire-rate non-blocking at all frame-sizes.

The failure of a single SFM module does not result in zero packet loss. Indeed, as soon as we removed an SFM to simulate failure, we noticed a drop in the rate of the traffic being received on the SmartBits ports. Replacing the SFM in the chassis saw the receive rate restored.

To investigate this further, we constructed additional performance tests to explore the frame loss that can be expected if the E1200 is running with the minimum eight SFMs. These tests were run at all frame sizes and compared with the results from tests where all nine SFMs are active. This report is titled “Frame Loss Under Switch Fabric Module (SFM) Failure Conditions.”

## EANTC Analysis

Normally, all nine SFMs in the TeraScale E1200 are used to forward the traffic. During our tests to confirm the Tolly Group results, we found that the removal of one out of the nine SFMs results in packet loss. Once again, EANTC believes this test does not accurately record the true impact of an SFM failure on a real-world network. As with the RPM Failover test, we also re-ran this test using a more realistic multi-switch test topology, including multiple traffic types and services.

## Test Configuration and Methodology

This test used the same 672-port L2 snake topology as the Tolly Group test. Prior to the execution of the failover test, we ran a baseline test to record the level of packet loss to be expected. Contrary to the Tolly Group reports, the TeraScale E-Series is not wire-rate non-blocking at all frame sizes. Having determined normal baseline operation, we re-ran the test, this time removing one out of nine SFMs. Once again frame loss was recorded and compared to the baseline result to determine if any additional loss was introduced by the SFM failover. Test traffic was run at 100 % load, using 64-byte frames.

Force10 TeraScale E-Series Competitive Test

# SFM Hitless Failover (Systems Test)

## Test Objectives

In Tolly Group Test Report 204148, Force10 claims to provide "SFM Hitless Failover" and claims this action results in zero packet loss.

In the Tolly Group test however, the SFM failover tests were conducted in a single switch, limited lab configuration, without many of the services or traffic types that might be expected to be operational in a production network.

This SFM failover test designed by EANTC takes a systems approach, simulating part of a typical network design. Our test used three Force10 TeraScale E-Series switches. One was configured as a L2 wiring closet switch, dual-homed to two L3 core/distribution switches that employed VRRP for first hop router resilience. The test also added OSPF routing to simulate an end-to-end enterprise network. In addition, IP multicast traffic was added to the traffic mix as multicast is an important component of many production networks, especially mission critical trading rooms of large financial services companies, where multicast availability is critical. (See the separate IP Multicast test reports for further multicast specific resilience tests)

We also added various services typically found in a production network to assess the DUTs performance under these more realistic network conditions.

The objective of this test is to assess whether the TeraScale E1200 is still able to provide hitless SFM failover in these more realistic network conditions.

**Test Highlights**

- **When any out of the 9 SFMs was removed, the complete switch fabric collapsed and all traffic through the switch ceased. The test was repeated multiple times, each time with a different SFM being failed. (FtoS Version 6.2.1.1) — See video clips at *http://www.eantc.com/video/F10-SFM*.**
- **The total switch fabric collapse was related to Control Processor (CP) utilization of 40 % caused by the sFlow statistics process on the RPM.**
- **The same test, run with FToS 6.2.1.3 saw no switch fabric collapse, but introduced further problems relating to Force10's bug fix. (See detailed test results)**
- **EANTC confirmed a single SFM failure is not "hitless," as claimed by Force10. All streams see lower forwarding rates and increased packet loss when an SFM fails. (FToS 6.2.1.1 & 6.2.1.3)**

## Test Results

**SFM Hitless Failover Test – Version 6.2.1.1.** In preparing our more sophisticated test scenario for this feature earlier in the year, we developed the tests on a network using FToS Version 6.2.1.1, the latest production software available from Force10 at that time.

As part of this effort, we chose a multi-switch topology that simulated part of a typical network design, and added multiple traffic types and control plane protocols to the test. We also configured a number of "services" such as SNMPv2 management and IP Flow Statistics using sFlow.

To generate the traffic load for the test, we used a L2 Snake of 624-ports, just the same as Tolly Group did in the original test. It was felt that this would give us a direct comparison, in the same switch, between how the original L2 Snake flows were affected, compared to the other flows and traffic types we used in our extended test scenario.

When we ran the test, the action of removing one out of the nine SFMs resulted in the immediate and total collapse of the whole switch fabric. This was unexpected, and we ran the test numerous times, each time with the same result, regardless of which SFM we removed.

We were logging the console port of the switch at the time, and the capture log shows what was seen on the switch CLI.

```
00:21:55: %RPM0-P:CP %TSM-2-SFM_SW_FAB_DIAG_MULTIPLE: Multiple SFMs failed SW FAB port pipe diags
00:21:55: %RPM0-P:CP %TSM-6-SFM_SWITCHFAB_STATE: Switch Fabric: DOWN
00:21:55: %RPM0-P:CP %IFMGR-5-OSTATE_DN: Changed interface state to down: Gi 1/0
00:21:55: %RPM0-P:CP %IFMGR-5-OSTATE_DN: Changed interface state to down: Gi 2/0
00:21:55: %RPM0-P:CP %IFMGR-5-OSTATE_DN: Changed interface state to down: Gi 3/0
[...]
```

**Figure 1: All interfaces down**

We also took video footage as a further record of this significant failure. Please see www.eantc.com/video/F10-SFM.

**SFM Hitless Failover Test – Version 6.2.1.3.** After pre-staging many other related tests in this extensive series of test reports, covering a wide range of categories, EANTC and Cisco were finally ready to run the formal tests and fully record the results. Just before starting however, Force10 released FToS Version 6.2.1.3, so all of the officially validated tests were run with this latest production software.

Using exactly the same switch and test equipment configuration as in our pre-staging tests, we found that the Force10 TeraScale E1200 no longer exhibited the same behavior. The switch fabric did not collapse, but once again we did see throughput degrade at the same amount as in the previous test case.

Clearly something had changed between the two tests. We scrutinized the Release Notes for Version 6.2.1.3, but could not find any reference to these fault symptoms even though this must have been seen as a critical Severity #1 issue for Force10.

We therefore decided to investigate the cause of this failure ourselves.

## EANTC Investigates

We reverted the switches in the test to Version 6.2.1.1 and began our investigation, by reducing the number of protocols, traffic types, and features running on the DUT. We started with only the L2 Snake flows in the test and gradually added each protocol or feature one at a time.

The SFM failover remained stable in all such tests until we re-introduced sFlow Statistics. As soon as sFlow was active and processing sFlow samples, the act of removing any of the SFMs resulted in immediate switch fabric collapse. We had successfully isolated the root cause of this problem as being some form of negative interaction between sFlow and the hardware switch fabrics of the TeraScale E-Series.

We could not understand how a seemingly unrelated software process could cause such a major hardware failure.

**SFlow Caused Switch Fabric Collapse.** sFlow is comprised of two parts, the first is a small software agent that runs on the line module processor on each line card and takes the packet samples that it forwards to the second part that is a central process running on the RPM. This process running on the RPM then generates a pack in software for export to sFlow collector (Management Station).

We monitored the behavior of the sFlow agent on the line cards under each software version, and found their behavior to be about the same. Next, we looked at the main sFlow process on the RPM, and noted that the average load on the Control Processor (CP) increased to a constant 39 % due to sFlow. When we looked at the CP processor's utilization under Version 6.2.13 however, we observed a constant utilization of 0–1 %, contrary to the behavior we saw under Version 6.2.1.1

We noted that traffic passing through the switch in each test run was exactly the same and should therefore have generated the same number of sFlow samples. At the line card level, the number of samples sent to the RPM was constant in each test.

**SFlow is restricted by Force10's workaround in FToS6.2.1.3.** As sFlow is a sampling technology, based on capturing one out of every N packets per interface, it stands to reason that the more interfaces that are being monitored or the greater the traffic rate, the greater the number of samples that will be generated and sent to the sFlow Collector for processing.

In their attempt to prevent sFlow from completely bringing down the whole switch fabric, it appears that Force10 has significantly rate-limited all sFlow traffic to the CPU, effectively throwing away the majority of samples gathered by the hardware on the line cards.

To confirm this theory, we monitored the volume of sFlow traffic being sent to the sFlow Collector under each software version. The results were significantly different.

In each test we incremented the transmit rates through the L2 Snake, and monitored the volume of sFlow traffic received by the sFlow collector.

As can be seen in *Figure 2: Volume Of Traffic To sFlow Collector*, for Version 6.2.1.1 the volume of sFlow samples increased with the traffic rate, but remained absolutely constant in Version 6.2.1.3.

For further details on sFlow visit:
http://www.sflow.org/sFlowOverview.pdf.

## EANTC Analysis

Two problems were seen in this test. First, there was insufficient switch fabric capacity to provide full performance when an SFM failed. See the "Gigabit Ethernet L2 Frame-Loss Under Switch Fabric Module (SFM) Failure Conditions" test report for full details.

Second, with Version 6.2.1.1 we witnessed a total collapse of the switch fabric. Force10's workaround in Version 6.2.1.3 is unsatisfactory. It rate limits sFlow so aggressively that it will interfere with the management application's ability to accurately record the traffic levels in the network. This will have a detrimental effect on the accuracy of network capacity planning, billing or any other management tasks that rely on accurate IP flow statistics.

Note that for FToS Version 6.2.1.1, as the traffic rate increases in steps from 1 % up to 100 %, the numbers of samples sent to the sFlow collector also increase proportionally, until a plateau is reached at 50 % traffic rate. In contrast, the plot for Version 6.2.1.3 is flat, regardless of traffic rate, something that is impossible for a sample-based flow statistics technique such as sFlow.

Note also, how the volume of sFlow traffic forwarded to the sFlow Collector is roughly the same volume as is seen when the network is passing traffic at 1% load.

In fixing one problem, Force10 has created another and effectively disabled sFlow altogether. We believe that for most customers this will prove an unsatisfactory workaround as flow statistics are a vital tool in day-to-day traffic management.

It is unclear whether Force10's workaround for the original switch fabric collapse was fixed in the newest firmware version or the restricting of sFlow just hides the problem.

Furthermore, it is unclear to us how a unrelated software process such as sFlow, which took the control processor (CP) to 39 % utilization, can result in the total failure of the switch.

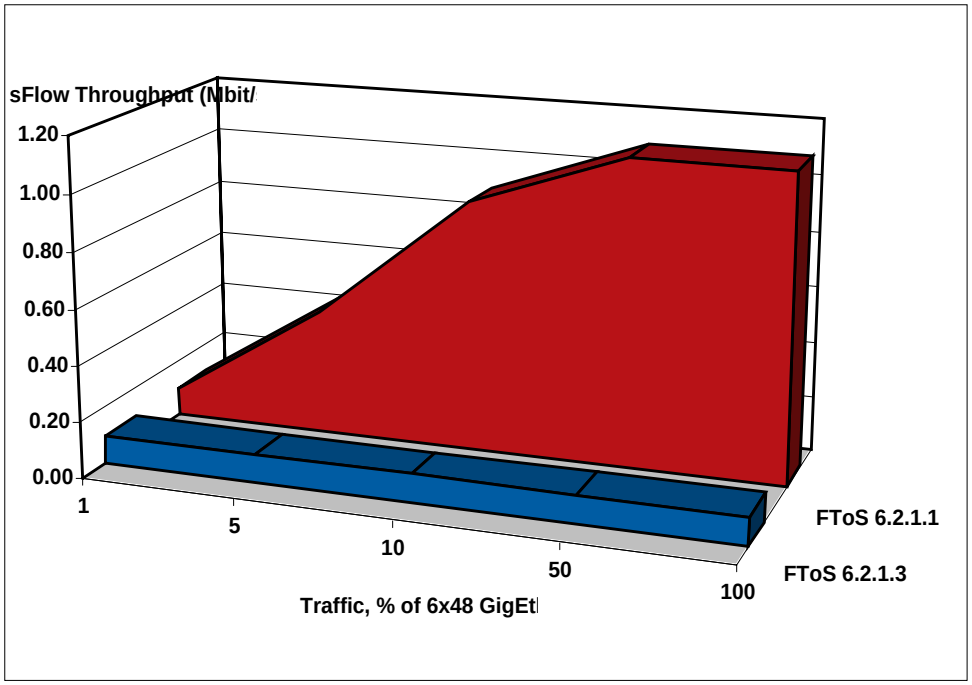



**Figure 2: Volume Of Traffic To sFlow Collector**

## Test Configuration and Methodology

The topology in *Figure 3: Test Topology* was used to simulate part of a typical network design. We used the "Multi-User" feature on the SmartBits to split the SmartBits into four logical pieces of test equipment.

- SMB-1a — Emulates a multicast distribution tree using PIM Sparse Mode and IGMP Version 2.
- SMB-1b — Emulates the traffic between a L2 wiring closet and the network, this represents our "existing" unicast flows, already established before the RPM failure.
- SMB-1c — Emulates 800 directly connected hosts. These hosts will try to establish communication in a full mesh traffic pattern immediately after the RPM failure. These represent clients/servers seeking to establish new sessions with each other.
- SMB-1d — Provides background traffic load (same as in the Tolly Group tests). This is a 624-port snake, leaving the 48-port line card in slot 13 to provide the other interfaces in the test.
- DUT(C) — Was configured with a single VLAN (VLAN 999) and acted as our dual-homed wiring closet switch in the test topology. All ports were untagged and no spanning tree was running.
- DUT(A) — The Force10 TeraScale E1200 was configured with redundant RPMs and a full complement of nine switch fabric modules, providing 8+1 switch fabric redundancy and represents the device under test (DUT). The switch was configured as a L3 core/distribution switch running OSPF and PIM-SM routing protocols. In a real network, this switch would serve multiple wiring closets, with potentially hundreds or even thousands of users.
  On VLAN 999, the switch ran VRRP first hop router resilience protocol and was configured to become the VRRP master, responsible for forwarding all traffic received from the protected VLAN.
  DUT(A) also functioned as the PIM-SM Rendezvous Point (RP), which was statically configured and served eight multicast receivers attached to ports 13/20 - 27; only DUT(A) supported multicast traffic in this test. A single multicast source sent traffic to 100 groups, creating a total of 100 S,G mroutes.

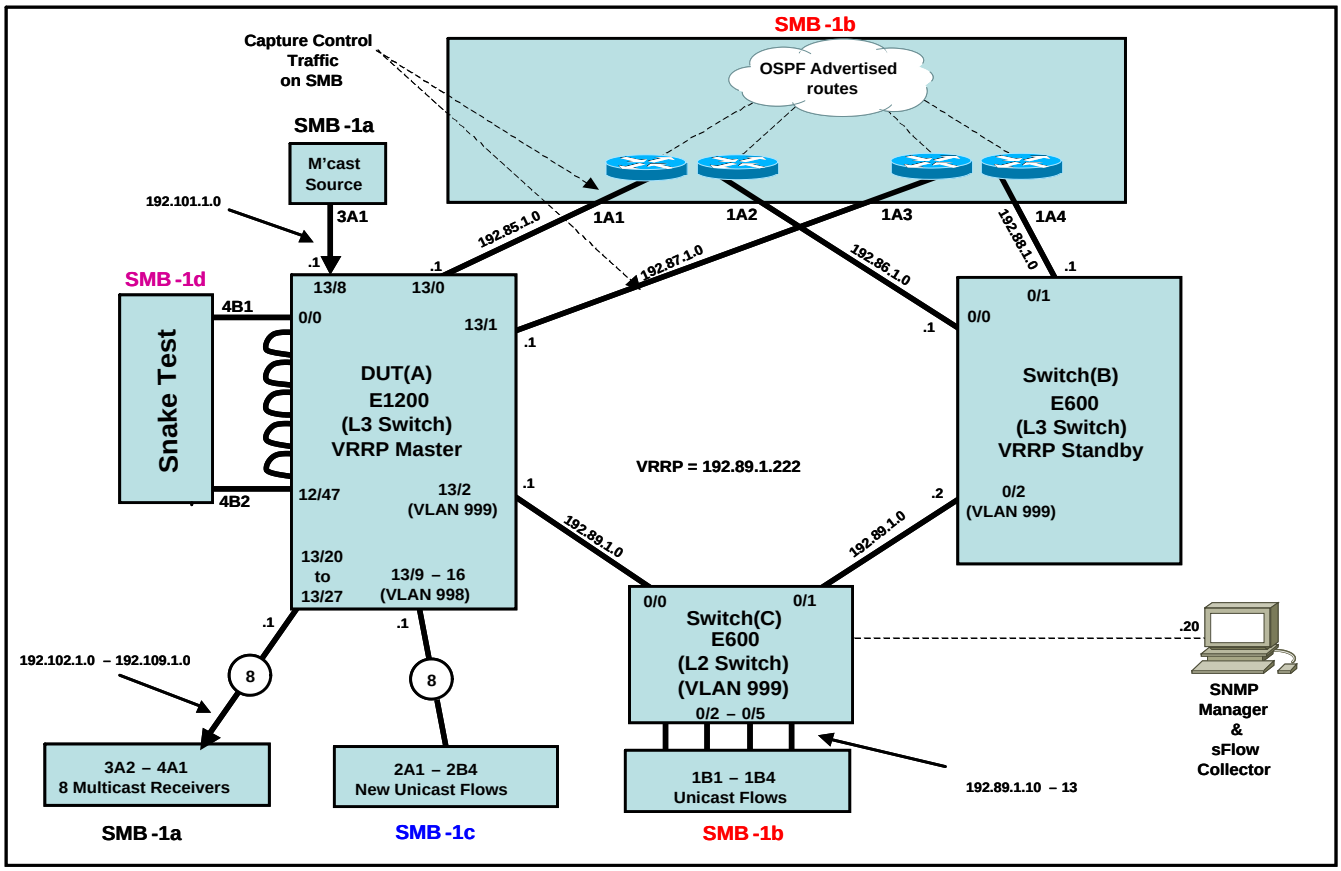



**Figure 3: Test Topology**

- DUT(B) — Was also configured as a L3 core/distribution switch and also ran OSPF. DUT(B) has no multicast receivers and doesn't take part in the multicast portion of the test.
  DUT(B) runs VRRP on VLAN 999 and is configured to become the VRRP standby router.

In addition to the above, we also configured a high powered PC to act as an SNMP management station and sFlow Statistics Collector, receiving management traffic from DUT(A). The SNMP management station polled an IP address on DUT(A) associated with a loopback interface as this should remain available throughout the test.

Finally, we used the packet capture capability on the SmartBits to capture the OSPF control plane traffic from DUT(A). This was to provide further analysis of what transpired from a routing protocol perspective when one out of nine SFM in DUT(A) failed.

All switches in the test were TeraScale E-Series running the latest production software, FtoS Version 6.2.1.3. We repeated the test multiple to ensure repeatability.

Test traffic was started in the following sequence:

7. OSPF routing and existing unicast flows between DUT(C) via the VRRP master, DUT(A) and the emulated routers on the SmartBits at the top of the diagram. All OSPF neighbor states and the route table were checked as correct before the unicast "existing flows" traffic was started. Traffic was sent at 20% of line rate.
8. PIM-SM and multicast traffic. Prior to starting the multicast traffic we checked the correct IGMP and PIM-SM multicast routing state was present. Only after we confirmed the correct state did we start the multicast traffic.
9. Next we started the L2 background traffic on the snake. We also started SNMP polling, sFlow collection (configured on less than 50% of the interfaces) and packet capture on the SmartBits.

Once all traffic (with the exception of the "new flows") was started, we waited a short while for the network to stabilize, checked the VRRP state on switches A & B, confirming DUT(A) was the master and DUT(B) the standby. We only proceeded further when we were satisfied the network was stable and in its initial acquiescent state.

Next we caused one out of nine SFMs in DUT(A) to fail. Immediately after the SFM was removed we started the "new flows," trying to establish communication during the recovery period of the SFM.

Once all traffic returned to normal and the new flows were fully established we stopped the test and gathered the results.

Force10 TeraScale E-Series Competitive Test

# Hitless Software Upgrade

## Test Objectives Hitless Software Upgrade

Tolly Group test report 204148 states that, *"Force10's FToS Modular Operating System supports zero packet loss hitless restarts and software upgrades"*.

The objective for this test is to investigate whether these claims are true. All tests were run with Version 6.2.1.3 software.

## Test Results Hitless Software Upgrade – Version 6.2.1.1 to 6.2.1.3

We followed the procedure documented in the FToS configuration manual titled "Streamlined Upgrade of the Software Image" (See FtoS Configuration Guide Version 6.2.1.1 – March 2005, p. 476).

To confirm which RPM module was active and which was in Standby state we issued `show redundancy` command.

The console log shown in *Figure 3: Ouptut of "show redundancy" Command* confirmed that RPM0 was the active RPM, with RPM1 the standby. Both RPMs were running FToS Version 6.2.1.1 at this time.

EANTC's understanding is that hitless software upgrade allows the standby RPM to be upgraded to a new version of software while the online RPM continues to manage the switch and forward traffic as normal. To initiate transition to the new software, the online RPM is then commanded to failover to the Standby RPM, which will be running the new software image.

This premise is based on Force10/Tolly Groups claim that the E-Series switches support "Hitless RPM Failover," and that no packet loss will occur during the controlled failover to the Standby RPM, which is running the new image. There should be no reason to reboot the switch as this would interrupt normal service and cause major packet loss.

To provide background traffic and some control plane routing protocol activity, Smartbits was configured with eight ports emulating OSPF neighbor routers, advertising 14,040 OSPF external routes. An additional eight SmartBits ports were used to send traffic to all advertised routes, simulating existing traffic flows at the time of the hitless failover and software upgrade.

**Test Highlights**

- **EANTC confirm FToS is not a modular operating system as claimed by Force10/Tolly Group. FToS is a single monolithic image with no way to support in-service software upgrades to individual software modules.**
- **EANTC tested Cisco's IOS with Software Modularity, a new feature in the core infrastructure of IOS on the Catalyst 6500. See EANTC's report at http://www.eantc.com/downloads/test_reports/2003-2005/EANTC-Summary-Report-Cisco-ION.final3.pdf.**
- **EANTC's tests confirms that contrary to Force10's claims, it is not possible to perform a hitless software upgrade on the TeraScale E-Series**
- **EANTC confirms there are no user commands under FToS for stopping/restarting any of the individual processes.**

As a first step we ran a baseline test, checked all OSPF neighbors were in full adjacency and that the route table was correctly populated. Traffic was generated for all advertised OSPF routes, and we verified that no packet loss occurred before starting the Hitless Software Upgrade test. Any packet loss would be attributable to the Hitless Software Update.

**Download new image to active RPM and synchronize this with the standby RPM.**
Following instructions in the FToS Configuration Guide, we downloaded the new image file FTOS-EF-

6.2.1.3.bin from the FTP server and copied it to the active RPM using the following command:

**`copy ftp://172.26.196.185 flash:/FTOS-EF-6.2.1.3.bin boot-image synchronize-rpm`**

After the download was completed to the active RPM's flash memory, the active RPM automatically copied the image to the standby RPM (output from F10: *Copying the image file to peer RPM*). As is depicted in Figure 1, the old software version is still the active one at this point.

**Reset standby RPM (RPM1) to boot up using the new image.** Next, we issued the following command: `reset rpm 1`. This should reload the Standby RPM and boot it up with the new image. This clearly worked because we received the following error message on the console of the active RPM, (RPM0).

```
*********************************************
*
*     Warning !!!  Warning !!!  Warning !!!
*
* -------------------------------------------
*
*       Different SW Version detected !!
*
*       This RPM -> 6.2.1.1
*       Peer RPM -> 6.2.1.3
*
*********************************************
```

**Figure 1: Different SW Version Error Message**

**Force a "Hitless RPM Failover" to complete the upgrade.** According to the vendor's manual the next command to use should be failover.

The CLI didn't recognize this syntax, but did accept the redundancy force-failover rpm command which clearly did the same job. Before the active RPM0 allowed us to force a failover to the standby RPM1, it asked us to save the configuration, which we did as can be seen below:

```
System configuration has been modified. Save? [yes/no]: yes
Synchronizing data to peer RPM
!!!!!!!
Proceed with RPM fail over [confirm yes/no]:yes
```

**Figure 2: RPM Failover**

*Figure 4: Output messages on the standby RPM* is an extract from the console log, which shows the messages we observed on the standby RPM as it tried to become the active RPM during the RPM failover process.

**Standby RPM1 cannot successfully become the active RPM.** The message in Figure 4 suggests that the Standby RPM (RPM1) was experiencing difficulty in coming online with Version 6.2.1.3, unable to determine the card slot numbers in the chassis.

At this point, it was not exactly clear whether the failover had been successful or not. So we checked the running-configuration on RPM1 (the new active RPM) by issuing the following command: `show running-config`.

Unexpectedly, we found all configuration commands were gone (the running configuration was completely empty). Further investigations showed that for some reason the startup-config file was not loaded during the failover to the standby RPM.

The lack of any configuration commands caused OSPF failure and 100% packet loss on all streams.

**Attempted manual recovery of the startup-config file.** In the next step we tried to load the configuration manually by using the following command: `copy startup-config to running-config` on RPM1. The command `show running-config` showed that all configurations were copied successfully. However, the command `show ip ospf` unexpectedly showed that no OSPF areas were configured on this device. So we executed the command `copy startup-config to running-config` again. No improvement was observed.

After the third execution of the command `copy startup-config to running-config` all OSPF neighbors suddenly became active.

Finally, to fully recover in preparation for the next test, we power-cycled the DUT. This time the startup configuration was loaded successfully.

### Test Objectives Modular Operating System with Zero-Loss Process Restart

In Force10's Tolly Group Test Reports 204147 and 204148, in the section titled "Force10 Networks, Inc. E-Series E1200 Product Specification, Force10 claims FToS is a modular operating system supporting "Zero Packet Loss Hitless Restarts…."

The Tolly Group printed a disclaimer at the bottom of the section, stating that none of these claims had been verified by Tolly Group. Since Tolly included this disclaimer, we cannot be sure why this feature was included in the Tolly Test Report.

Our goal with this test was to discover exactly what processes in FToS were modular and which processes could be manually stopped and restarted using the CLI.

### Test Results Modular Operating System with Zero-Loss Process Restart

EANTC could find no evidence to support Force10's claim that FToS is a modular operating system. No CLI commands were available to stop or restart an individual software process.

In a separate test on the Catalyst 6500 with Software Modularity, we were able to perform process restarts and apply patches to individual subsystems during runtime without any packet loss. A variety of other tasks one would associate with a true modular operating system could also successfully be verified.

### EANTC Analysis

EANTC's results disprove all Force10's claims regarding FToS being a modular operating system, supporting "zero packet loss hitless restarts and software upgrades". We were not able to verify these claims.

```
Force10-E1200#sho redundancy

--  RPM Status  --
-----------------------------------------------
 RPM Slot ID:               0
 RPM Redundancy Role:       Primary
 RPM State:                 Active
 Link to Peer:              Up

--  PEER RPM Status  --
-----------------------------------------------
 RPM State:                 Standby

--  RPM Redundancy Configuration  --
-----------------------------------------------
 Primary RPM:               rpm0
 Auto Data Sync:            Full
 Failover Type:             Hot Failover
 Auto reboot RPM:           Enabled
 Auto failover limit:       3 times in 60 minutes

--  RPM Failover Record  --
-----------------------------------------------
 Failover Count:            0
 Last failover timestamp:   None
 Last failover Reason:      None

--  Last Data Block Sync Record:  --
-----------------------------------------------
  Line Card Config:         succeeded  Jun 26 2005 21:26:37
   Start-up Config:         succeeded  Jun 26 2005 21:26:37
 Runtime Event Log:         succeeded  Jun 26 2005 21:26:37
    Running Config:         succeeded  Jun 26 2005 21:26:38
```

**Figure 3: Ouptut of "show redundancy" Command**

```
Force10-E1200(standby)>00:04:55: %RPM1-S:CP %RAM-6-FAILOVER_REQ: RPM failover request from active peer:
User request.
00:04:55: %RPM1-S:CP %RAM-6-ELECTION_ROLE: RPM1 is transitioning to Primary RPM.
00:04:55: %RPM1-P:CP %TSM-6-SFM_SWITCHFAB_STATE: Switch Fabric: UP
00:04:55: %RPM1-P:CP %IFMGR-1-DEL_PORT: Removed port: Ma 0/0
00:05:32: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 5 sucessfully switched to master RPM
00:05:33: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 1 sucessfully switched to master RPM
00:05:33: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 11 sucessfully switched to master RPM
00:05:33: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 8 sucessfully switched to master RPM
00:05:33: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 7 sucessfully switched to master RPM
00:05:33: %E48TF:5 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:33: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 13 sucessfully switched to master RPM
00:05:33: %E48TF:7 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:33: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 12 sucessfully switched to master RPM
00:05:34: %E48TF:1 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:34: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 0 sucessfully switched to master RPM
00:05:34: %E48TF:12 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:34: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 2 sucessfully switched to master RPM
00:05:34: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 6 sucessfully switched to master RPM
00:05:34: %E48TF:11 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:34: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 10 sucessfully switched to master RPM
00:05:34: %EXW4PF:2 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:34: %RPM1-P:CP %CHMGR-2-SWITCH_MASTER: Line card 4 sucessfully switched to master RPM
00:05:35: %E48TF:13 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:35: %E48TF:10 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:36: %E48TF:6 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:36: %E48TF:8 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:36: %E48TF:4 %IFAGT-3-BAD_SLOT: Unable to determine slotId: 0
00:05:37: %RPM1-P:CP %RAM-5-HOT_FAILOVER: RPM Failover Completed.
```

**Figure 4: Output messages on the standby RPM**

Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test

# Force10 E1200 — Cooling & Ventilation Problem

## Test Objectives

Cisco Systems asked EANTC to document what happens if a single line card is removed from the Force10 TeraScale E1200 chassis for a period exceeding ten minutes. (Without a blanking panel being inserted in the open card slot).

Cisco claims that Force10 has serious cooling and ventilation problems with the E1200 that can bring the whole switch down.

Cisco also claims the Catalyst 6500 does not suffer from such cooling problems and offered to run all tests to be validated by EANTC without blanking panels in any unoccupied card slots. We agreed that one switch should be left without blanking panels, but suggested that the rest should have them installed.

**Test Highlights**

- **10 minutes after a single line card was removed, two SFMs were deactivated by the E1200 due to critical overheating. The whole switching fabric collapsed and all 672 ports were disabled. See the video at *http://tools.cisco.com/cmn/jsp/index.jsp?id=46123***
- **It was not possible to reset the SFMs by using CLI (only line cards and RPMs can be reset from CLI).**
- **Catalyst 6500s in the same lab operated with no blanking panels in the majority of their card slots and never suffered overheating.**

## Background:

The diagram below shows the ventilation and cooling system in the Force10 TeraScale E1200, 14-slot chassis and Force10's flagship product.

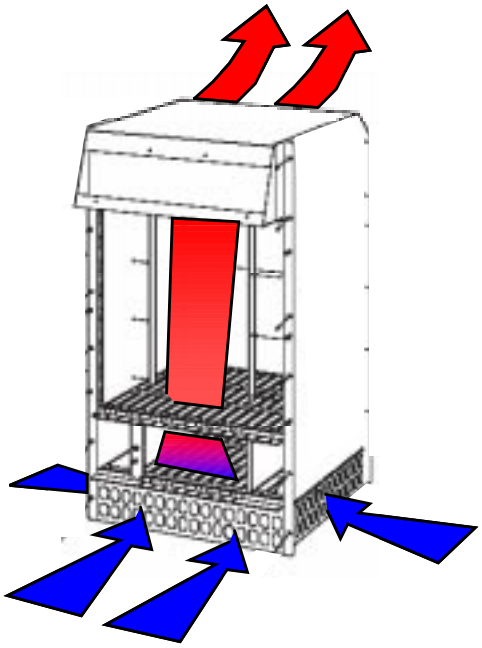


**Figure 1: Force10 TeraScale E1200**

Force10's TeraScale E1200 can support up to 1,260 Gigabit Ethernet, or 56 Ten Gigabit Ethernet ports, so the chassis needs absolutely rock solid high availability.

Cisco believes that poor air circulation in the E1200 poses a threat to the E1200's availability and states the E1200's cooling and ventilation system represent a single point of failure for the whole chassis.

The E1200 has a bank of eighteen fans, in six groups of three, located at the top rear of the chassis. These fans draw cool air in from open grids on the bottom of the chassis, up past the Switch Fabric Modules (SFMs), up past the Route processor Module and/or Line Cards, to be vented out the top rear of the chassis.

**Cisco claims the TeraScale E1200 is vulnerable to overheating.** Cisco explained that even if a single line card is removed, perhaps as a result of failure, and the card slot is left open while a new card is retrieved from the store room; it only takes approximately ten minutes before the two Switch Fabric modules immediately below the open card slot overheat and are shut down by the system.

```
Force10-E1200#
Force10-E1200#!!!!!!!!!!!!!!!!!!!! Remove Line Card in Slot 9 !!!!!!!!!!!!!!!!
Force10-E1200#
02:38:58: %RPM0-P:CP %CHMGR-2-CARD_DOWN: Line card 9 down - card removed
[?]
02:44:44: %RPM0-P:CP %CHMGR-2-MINOR_TEMP: Minor alarm: chassis temperature (SFM temperature reaches
minor threshold of 65C)
[?]
02:50:01: %RPM0-P:CP %CHMGR-2-MINOR_SFM: Minor alarm: only eight working SFM
02:50:01: %RPM0-P:CP %TSM-2-SFM_OVER_TEMP: SFM 6 powered off due to over temperature
[?]
Force10-E1200#02:50:13: %RPM0-P:CP %TSM-6-SFM_SWITCHFAB_STATE: Switch Fabric: DOWN
02:50:13: %RPM0-P:CP %IFMGR-5-OSTATE_DN: Changed interface state to down: Gi 0/0
02:50:13: %RPM0-P:CP %IFMGR-5-OSTATE_DN: Changed interface state to down: Gi 0/1
02:50:13: %RPM0-P:CP %IFMGR-5-OSTATE_DN: Changed interface state to down: Gi 0/2
```

**Figure 2: Force10 Console Log**

As the E1200 employs an 8+1 switch fabric redundancy scheme, the simultaneous loss of two SFMs renders the switch totally inoperative.

The positioning of the fans on the top rear of the chassis mean it is easier for them to suck cool air in through the open line card slot, than it is for the air to be dragged up from the bottom of the chassis, past the SFMs; the incoming cool air taking the shortest path to the fan assembly.

## Force10 Test Results

*Figure 2: Force10 Console Log* shows the effect of removing the line card in Slot 9 without inserting a blanking panel in its place. Take note of the timestamps on the left hand side of the console messages.

During the test we used the `show sfm environment` command to keep track of the SFM temperatures at regular intervals. From this data, we produced *Figure 3: Force10 E1200 SFM temperatures* to depict exactly how the temperature rose for the two SFMs immediately below the open line card slot (*http://tools.cisco.com/cmn/jsp/index.jsp?id=46123*).

## Cisco Test Results

The photograph in *Figure 4: Catalyst 6509 Used in the Tests*, taken after the completion of all the tests validated by EANTC shows the main Catalyst 6509 used in the tests.

The Catalyst 6509 ran in this condition, without overheating, or issuing an overtemperature warning message for the whole period of the tests, which lasted over four weeks.

This proves that the ventilation and cooling systems in the Catalyst 6500 operate properly even when half of the line card slots are left empty and without blanking panels. EANTC staff also noted other Catalyst 6500s in the same lab where only a single line card was installed and where all others slots were open, and yet these Catalyst 6500's also continued to function without problem.

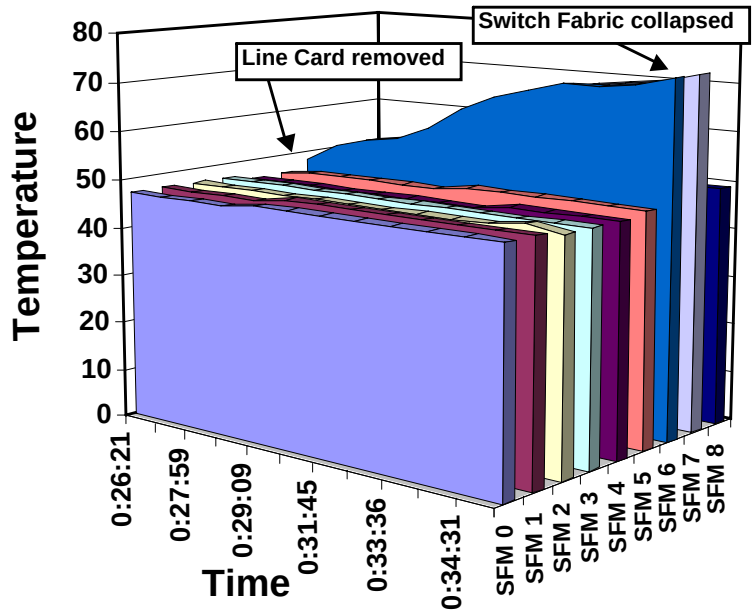



**Figure 3: Force10 E1200 SFM temperatures**

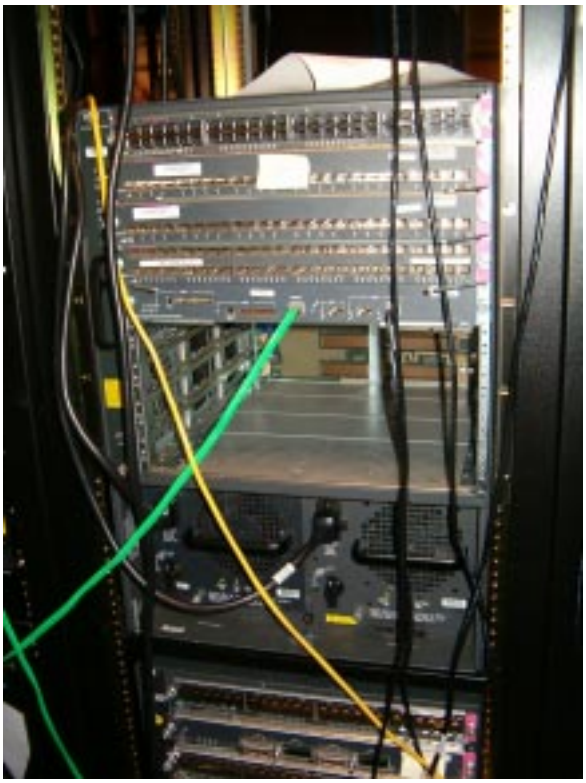


**Figure 4: Catalyst 6509 Used in the Tests**

### EANTC Analysis

EANTC confirm the Force10 TeraScale E1200 exhibits an unacceptably high risk of overheating even when only a single line card or RPM management module is removed from the chassis (and not immediately replaced with a blanking panel). EANTC can also confirm that the Cisco lab in which the tests were conducted was adequately air-conditioned.

The speed at which the temperature of the SFMs adjacent to the open slot rose from 47°C to 74°C in ten minutes once the line card was removed was verifiable. Also, EANTC re-ran this test numerous times to better understand this problem.

There is also no warning notice on the front of the chassis alerting operations staff that all the slots must be kept closed in order to avoid a total crash of the switching fabric.

**Force10 TeraScale E-Series Competitive Test**

# Gigabit Ethernet Layer 2 Frame Loss Under Switch Fabric Module (SFM) Failure Conditions

## Test Objectives

The Force10 TeraScale E-Series supports an 8+1 switch fabric redundancy scheme. Force10's claim to support "Hitless SFM Failover", implies that after a switch fabric module (SFM) failure the remaining eight SFMs can continue to support wire-rate, zero-loss performance at all frame-sizes on all ports.

This test compares the frame-loss recorded when all nine SFMs are active with the frame-loss seen when only eight SFMs are functioning following an SFM failure. The test is run at each switch fabric Epoch setting, as this also has a impact on latency and frame-loss.

**Test Highlights**

- **Under normal operation, all 9 SFMs are used to forward traffic. The failure of one out of nine SFMs results in additional frame loss.**
- **Force10's claim that the TeraScale E-Series provides *"zero packet loss hitless failover of all components for non-stop applications"* could not be verified with regard to SFM failover.**

## EANTC Analysis

All nine SFMs are used to forward the traffic under normal conditions. Although the system continues to work with only eight SFMs, it operates at a reduced capacity, and does not meet Force10's claim of being hitless.

## Test Configuration and Methodology

The DUT was configured with 336 two-port VLANs; VLAN 2-337. VLANs were interconnected with external crossover cables to form a 672-Port snake. L2 forwarding-table aging time was disabled for all VLANs. All control plane protocols were disabled to ensure no control protocol frames are being sent by the DUT as this might interfere with the test. One out of nine SFMs was physically removed.

This test conforms to the methodology for measuring frame loss as specified in RFC 2544 and conforms to the definition of frame-loss specified in RFC 1242.

Bidirectional traffic was transmitted between the two SmartBits ports at 100% wire-rate for each of the frame-sizes. For each frame size, frame loss was recorded as a percentage of the number of frames transmitted. As the frames pass along the snake, they are effectively multiplied and are switched again in the next VLAN, until they exit the final VLAN in the snake. This exercises the switch as if all 672 ports were connected to separate traffic sources/sinks.

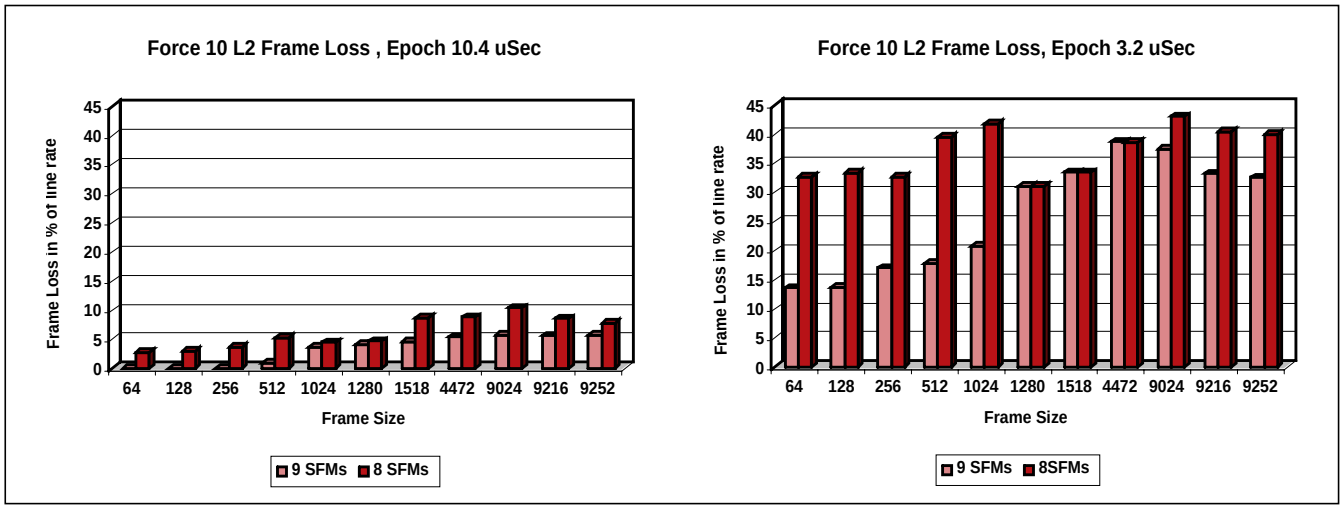



**Figure 1: Force10 Layer 2 Frame Loss With Eight (8) and Nine (9) SFMs**

Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test

# Gigabit Ethernet L2 Address Learning Rate Tests

## Test Objectives

The objective of this test is to compare the L2 address learning-rate of the Force10 and Cisco switches and look at the effects on learning rate caused by:

- Number of learning ports in the test
- Aggregate number of MAC-addresses to be learned

## Test Results

**Multi-Port Learning Rate Tests.** We tested the switches with 64-byte frames, and offered a total of 5,040 addresses to be learned (5,040 addresses were chosen, as this number is divisible by the numbers of ports used in each test) and representative of a large L2 network. For instance, in the 12-port test, each port is asked to learn 5,040/12 = 420 addresses per port.

In all tests the Catalyst 6500 learned all addresses at wire-rate (1,488,095 fps).

We found the Force10 switch uses software learning of MAC-addresses. Returning results of 75 fps when learning on two ports; dropping to 32 fps when asked to learn on 48 ports.

**Test Highlights**

- ➜ **Force10's L2 address learning rate drops to 57 frames per second, when the total number of MAC-addresses to be learned exceeds 1,026.**
- ➜ **Force10 E1200 supports L2 address learning in software. The measured learning rate (32 -75 fps per port) is insufficient for a core component.**
- ➜ **Catalyst 6500 learns MAC addresses 35,430 times faster than Force10 E1200.**
- ➜ **Catalyst 6500 learns MAC addresses wire-rate in hardware.**

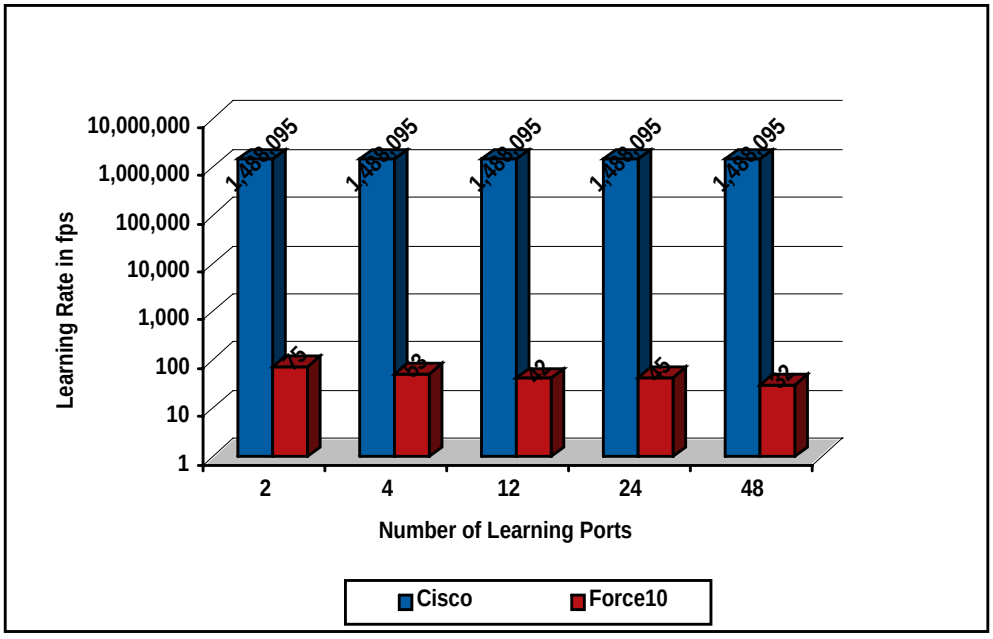



**Figure 1: L2 Address Learning-Rate Per Port**

On the Force10 switch, the function of address learning is shared between two processors. A small processor located on each line card and shared by all ports on the card, plus the central L2 processor (RP2) located on the RPM card. (See architecture section at the end of this report for more details)

**Single Port Learning Rate Tests.** To investigate Force10's learning-rate further, we reduced the test topology to a single learning port.

We started by investi-gating whether the total

number of MAC-addresses the switch had to learn had an effect on its learning rate. We offered the Force10 switch increasing numbers of MAC-addresses and recorded the learning rate.

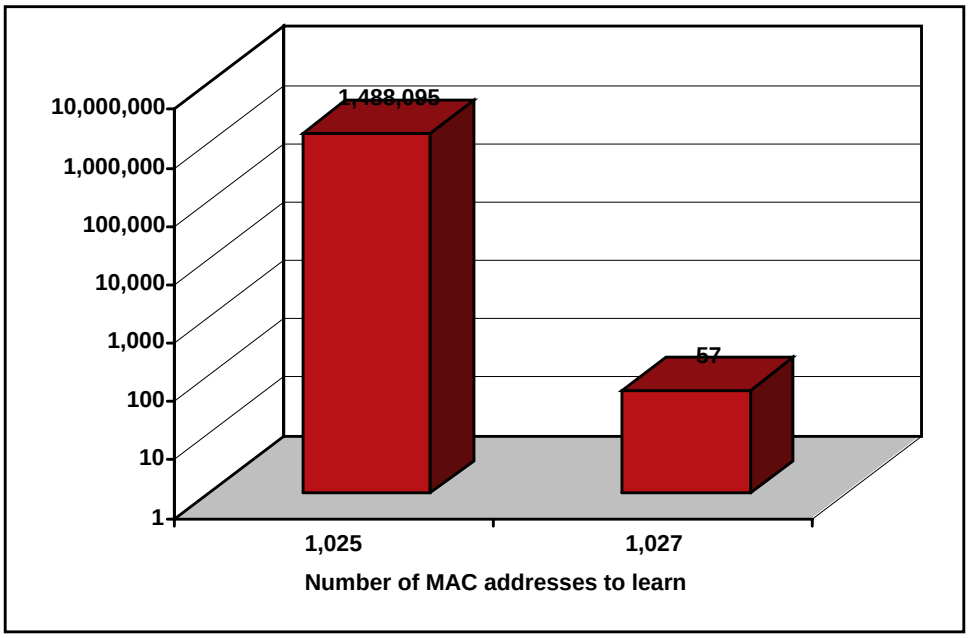



**Figure 2: L2 Address Learning on Single Port**

In tests where the number of addresses to be learned was 1,026 or less, the Force10 switch recorded wire-rate address learning. However, once the total number of MAC-addresses exceeded 1,026 the learning rate decreased to 57fps, repeatable across numerous test runs.

This is very confusing, because if tests were run with a small number of MAC-addresses to be learned, the Force10 switch would appear to learn at wire-rate. In fact when we kept the total number of mac-addresses below 1,026 and shared this number across multiple ports, as we did in the multi-port tests, the E-Series appeared to learn at wire-rate.

## EANTC Analysis

In today's Triple Play, Voice, Video, and Data networks, there is a need for a campus-wide VLAN to support the IP phones. This could potentially mean thousands of MAC addresses in the same VLAN, supporting what is considered a mission critical 24/7 service. The Catalyst 6500 and Force10 E-Series are deployed as core switches, or used in high-density wiring closets. On the uplink ports linking such wiring closets to the core, there is the need to learn large numbers of MAC addresses, especially in response to any topology changes caused by link or node failure.

Our tests confirm the Force10 TeraScale E-Series supports L2 address learning in software. The measured learning rate (45 - 70 fps per port) is low for such a high density core switch.

For example, if the uplink on a switch fails, and the alternate uplink to the core must re-learn 10,000 addresses, the Force10 TeraScale E-Series would take:

10,000 addresses / 57fps learning-rate = 133 seconds to learn all addresses

In the meantime, the traffic from all these hosts would be flooded throughout the VLAN, oversubscribing the links thereby causing further network problems.

Catalyst 6500 learns MAC addresses 35,430 times faster than Force10 E1200. In the example above, the Catalyst 6500 would take six milliseconds to learn all addresses.

## Test Configuration and Methodology

All control plane protocols were disabled to ensure no control protocol frames are being sent by the DUT as this might interfere with the test. SmartBits Advanced Switch Tests (AST) application was used to test both switches.

The address learning rate tests were conducted in conformance with RFC 2889.

**Background.** We have included this additional section in this test report to aid the reader in understanding Force10's mixed results. This section gives a brief overview of the Force10 switching architecture in relation to address learning and offers an explanation of why some tests appear to support wire-rate address learning.

The address learning-rate test offers a fixed number of MAC-addresses to be learned, on one or more learning ports and confirms the addresses have been learned by sending packets from the second port to each of the mac-addresses. If the switch failed to learn all addresses at the offered rate, they are flooded and

picked up by a monitor port. This constitutes a test FAIL.

In response, the SmartBits uses a binary search algorithm to increase or decrease the rate at which the addresses are offered, finally homing-in until the exact address learning-rate of the switch is found. This methodology is slightly flawed, as there is a small delay between offering the addresses to the switch and generating the traffic that confirms they've been learned.

**Force10 Address Learning Architecture.** The small processor responsible for L2 address learning is located on each line card and is shared by all 48 ports on the card.

As with most processors, it has a "hold queue" that temporarily stores packets that cannot be immediately processed by the CPU. When the number of addresses to be learned is less than the depth of the processor hold queue, the Force10 switch can give the impression that it can learn at wire-rate.

What is actually happening though is the addresses arrive at wire-rate, but not processed at wire-speed. Instead they are held in the processor's hold queue and written to CAM at a much slower pace. Because there is a delay introduced by the SmartBits at this point, this delay allows the Force10 switch to slowly process the packets in its hold queue and appear to have processed these at wire-rate by the time the actual test traffic begins.

Our tests showed that even one packet over the size of the hold queue, means that one packet is flooded and the binary search reduces the offered rate until the true learning rate of the switch is found: somewhere in the region of 32–75 frames per second per port.

# Executive Summary — L2 Scalability Tests

In June 2005, Cisco Systems commissioned the European Advanced Networking Test Center (EANTC) to independently validate the performance, scalability, and availability of Force10 TeraScale E-Series switches.

In this category of tests, we evaluated the L2 MAC address scalability and address learning rate of the Cisco Catalyst 6500 and Force10 TeraScale E-Series. We compared our results with the various L2 scalability claims made in vendor marketing literature.

## Selection of Force10's MAC Address Scalability Claims

*Figure 1: Extract from Force10 E-Series Data Sheet* and *Figure 2: Extract from Force10 48-port Line Card Installation Manual* are examples from various Force10 documentations pertaining to scalability.

**FToS Key Features**

**L2 Switching**

- 4,096 VLANs
- 16M VLANs with VLAN stacking
- Up to 1.4 Million MAC addresses per system
- Link aggregation
- 802.1p prioritization
- FVRP VLAN redundancy
- MSTP (802.1s)/RTSP (802.1w)

**Figure 1: Extract from Force10 E-Series Data Sheet**

**48-port 10/100/1000 Base-T Ethernet Line Card**

Feature Highlights and Installation Instructions

Catalog Number: LC-EF-GE-48T

Feature Highlights

- 48 ports with RJ-45 connectors that support auto-negotiation or 10/100/1000 Base-T speed.
- The EF series line cards support a single 18M user configurable CAM with flexible partition assignments. The following max entries are:
  - 256K Layer 3 IP forward information base (FIB)
  - 256K Layer 2 FIB
  - 64K Layer 3 IPv6 FIB
  - 128K Layer 2 and Layer 3 access control list (ACL) entries
  - 24K IPv6 ACL entries
- Supports online insertion and removal (OIR) of line card.
- Supports ingress and egress Layer 2 and Layer 3 ACL processing across all ports.
- Requires FTOS 6.1.1.0 or higher

**Figure 2: Extract from Force10 48-port Line Card Installation Manual**

## Gigabit Ethernet MAC Address Scalability & Learning Rate — EANTC Findings

**L2 Address Learning-Rate Tests.** EANTC's extensive tests in this area found that the Cisco Catalyst 6500 exhibited wire-rate MAC address learning, regardless of the number of learning ports in the test.

The Force10 E-Series learned MAC addresses only at a speed of 50-70 addresses per second per line card. Force10's address learning is software-based and the processor can buffer a small number of addresses (up to 1,026) so that it appears to show wire-rate performance from SmartBits' point of view (due to restrictions in the SmartBits test methodology). In fact MAC addresses are always written to hardware memory (CAM) at the low speed we measured.

In contrast, the Cisco Catalyst 6500 showed true wire-speed MAC address learning on the Gigabit Ethernet links, learning all 5,000 addresses in the test.

**Mac Address Scalability Tests.** Force10 claims to support 256K MAC addresses per line card, and a total of 1.4 Million MAC addresses on the 14-slot E1200 chassis. Our tests were confined to testing the MAC address scalability offered by a single line card, so we expected the TeraScale E-Series to support around 256K addresses.

Our tests revealed that the E-Series only managed to cache 27,792 addresses, a little over 10 % of the number claimed by Force10. Further investigation showed that it is not possible for an E-Series running production software to support the scalability claimed by Force10 as there simply is not enough memory space in the Gigabit Ethernet line card CAM to support it.

The Catalyst 6500 claims to support 64K entries and managed to cache 54,816 addresses, representing 83.6 % of the number claimed by Cisco.

**Summary.** Force10's MAC address learning rate is slow for a core switch, especially considering the number of interfaces these switches are expected to support and the large number of MAC addresses this will represent in Layer 2 configurations where Ethernet aggregation networks are attached.

Such a slow address-learning rate means that the E-Series will be slow to adapt to Layer 2 topology changes where the switch must learn large quantities of addresses on a new interface. During this period, excessive amounts of traffic will be flooded, causing congestion on previously uncongested links.

The relatively small MAC address table size on the E-Series line cards is different from Force10's data sheets and means that this literature does not present a realistic view of the switch's capabilities.

Cisco's Catalyst did not quite match its claimed scalability of 64K either — we reached 54K instead.

However, the Catalyst learned all addresses at wire-speed. Even when we asked it to learn at wire-rate on all 48 ports on the card, the Catalyst learned all addresses instantaneously. This is an impressive performance meeting and exceeding large-scale Enterprise requirements.

**Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test**

# OSPF Route Scalability Test

## Test Objectives

This test assessed the OSPF scalability up to a limit of 50,000 routes — which we believe is a safe upper limit in real-world networks using OSPF.

The primary metric of an OSPF implementation is its ability to scale to large networks with many routes. For a backbone switch, OSPF scalability is always a critical factor because it has to process all routes present in the backbone area (Area 0) plus all other areas network wide.

## EANTC Analysis

In EANTC's opinion, both devices supported a reasonable number of External LSA OSPF routes. However, the Catalyst 6500 was able to learn more routes and managed to bring up all the OSPF adjacencies faster than Force 10. The Catalyst took six minutes to accept all 50,000 routes; whereas the Force10 E1200 required nine minutes to learn 48,000.

In addition to confirming the software routing table contained all advertised routes, we used a full-mesh traffic pattern to confirm that each network prefix was correctly written to the hardware forwarding ASICs in the data plane of each switch. By comparing the average latency seen in this test with expected averages from previous latency tests, we were able to confirm both switches had correctly populated their hardware forwarding engines; the E-Series displaying a latency of 44 µSec, while the Catalyst came in with just 11 µSec.

**Test Highlights**

- ➜ **Catalyst 6500 easily scales to at least 50,000 OSPF routes. (Test was capped at 50,000 routes).**
- ➜ **Force10 E1200 failed the 50,000 routes test, reaching a maximum of 48,000 routes.**
- ➜ **Catalyst 6500 brought up OSPF adjacencies and learned all routes 30% faster than TeraScale E-Series.**

## Test Configuration and Methodology

50,000 non-contiguous routes (external LSAs) were distributed over 24 ports, with all ports on the same line card.

The Spirent SmartBits emulated a single OSPF neighbor per port, with all ports in Area 0.

Routes employing an "Internet Mix" of various prefix lengths were distributed evenly across all 24 emulated OSPF neighbors. Full-mesh traffic was sent over each advertised route to confirm full functionality and hardware forwarding. We sent bidirectional full-mesh traffic with 64-byte packets through each port.

We then transmitted full-mesh traffic over all routes, and confirmed traffic was forwarded in hardware. We did this by measuring the latency introduced by each switch to confirm hardware-based forwarding. We compared the latencies seen in this test with the latencies recorded in earlier latency tests to confirm all traffic streams were being forwarded in hardware.

# Executive Summary — OSPF Route Scalability Tests

In June 2005, Cisco Systems commissioned the European Advanced Networking Test Center (EANTC) to independently validate the performance, scalability, and availability of Force10 TeraScale E-Series and Cisco Catalyst 6500 switches.

In this category of tests, we attempted to reproduce a number of the original Tolly Group route scalability tests found in the Tolly Group Test Report 204148 and added further tests to fully understand the scalability of the switches under test.

Like the Tolly Group, our tests were constrained by the availability of test equipment and the quantity of each vendor's switching equipment. In general, our tests focused on Gigabit Ethernet, as Force10 acknowledges their sales of Gigabit Ethernet far outstrip those of 10-Gigabit.

The Tolly Group's reports state the Force10 tests were run using Force10 Beta Version 4.4.2.107, Tolly Group state *"According to Force10 Networks, this Beta release was replaced by Version 6.1.1.0"*. There is documented *evidence to suggest this beta software was actually a special engineering image*, customized to achieve these test results. (See the letter in the Appendix dated 9/28/2005 from Force10 to EANTC.)

## Tolly Group OSPF Route and Neighbor Scalability Claims

The Tolly Group states the E-Series supported over 500 peers (OSPF Neighbors) and was tested with 25,000 OSPF routes. Our OSPF route scalability tests were capped at 50,000 routes, as this is a safe upper limit for real OSPF networks.

For performance tests, The Tolly Group tested a Force10 Networks Inc. E-Series E1200 Resilient Switch/Router Software Version 4.4.2.107. According to Force10 Networks, this Beta release was replaced by Version 6.1.1.0.

Tests also determined that the E1200 can scale to support communication with over 1,500 peers and 6 million paths with background traffic using the BGP protocol. With OSPF as the protocol supported, the E1200 can scale to support communication with over 500 peers.

**Figure 1: Extracts from Tolly Group Test Report 204148**

## OSPF Route Scalability

The Catalyst 6500 learned all 50,000 routes, populating its software and hardware forwarding tables in six minutes, and forwarding traffic on all advertised routes with an average latency of 11 µSecs.

The Force10 E-Series failed to learn all 50,000 routes, never fully converging. When offered 48,000 routes, the E-Series took nine minutes to fully converge all routes, and forwarded traffic with an average latency of 60 µSecs.

## OSPF Neighbor Scalability

OSPF neighbor scalability saw a significant difference between the two switches.

To pass the test, the switch had to establish adjacencies with all neighbors and forward traffic to all advertised routes with no loss. Each neighbor advertised just ten routes and traffic was sent at less than wire-rate.

The Force10 TeraScale E-Series reached its limit at 48 OSPF neighbors. When tested with 72 neighbors, it learned the routes but recorded 6 - 9 % packet loss, even though the traffic rate was quite low. At 96 adjacencies, the E-Series was able to maintain 33 out of the 96 neighbors and never converged. This was unexpected?> for a switch that claims up to 1,260 interfaces.

The Catalyst 6500 scaled to 360 OSPF neighbors and passed traffic on all routes without loss. That is over seven times more OSPF neighbors than the Force10 switch.

## OSPF Continuous Route Flap Test

In this test we created a scenario which exercised all the major elements of an OSPF routing and L3 switching implementation.

- We chose to employ continuous route flaps as this mimics real network conditions, particularly experienced

with frame-relay networks. Continuous route flaps often find routing protocol implementation problems such as memory leaks that single flap tests, such as the Tolly Group test, never discover.

- We created route flaps that flapped 2,000 routes at a time. This should be within the capabilities of any core switch. In fact the Tolly Group tests claims the TeraScale E-Series handled OSPF route flaps of 25,000 routes.
- We advertised routes with overlapping network address ranges, forcing the switch under test to correctly apply the "longest prefix match" algorithm with each route flap
- We aged out 2,000 routes, all advertising the same prefix. This has the effect of thoroughly exercising the read/write management of the line card memory (CAM) and exercises the critical need to keep RIB/FIB and CAM totally synchronized at all times.

The Catalyst handled continuous route flaps for 15 minutes. No problems were encountered.

The TeraScale E-Series failed after just three runs. The route processor went into overload and no matter what we did, it remained there until we rebooted the switch. As soon as the RP1 processor became overloaded, the switch failed to respond to further route flaps. In addition, we experienced periods where the command-line interface was unresponsive for extended periods after entering commands. At one point we were locked out of the CLI for over 20 minutes.

## Summary

Both switches learned roughly the same number of routes, but the Catalyst processed the routes and populated the route table faster than the E-Series.

In other tests where OSPF was configured, we had numerous instances where the OSPF routing process in the E-Series switches (this problem occurred on multiple switches) would bring up all its adjacencies, correctly populate the link state database, but never actually populate the routing table.

Resetting the OSPF process made no difference, and we had to resort to rebooting the switch before it would recover.

The route flap test resulted in the total collapse of the TeraScale E-Series, no matter what we did we did not manage to recover, forcing us to abandon the test and reboot the switch. This test result was repeatable.

The Catalyst's OSPF implementation was more stable and we had no recourse to reset the OSPF process during any of the tests. The Catalyst brought up neighbors faster, learned routes faster and responded to all route flaps we threw at it with ease.

**Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test**

# OSPF Neighbor Scalability Test

## Test Objectives

This test explores the OSPF neighbor scalability offered by the Force10 E1200 and Cisco Catalyst 6500. The test reflects the need for a backbone, or aggregation switch to support a large number of OSPF neighbors.

In the Force10 Tolly Group Test Report 204148, Force10/Tolly Group claimed that the E1200 supported 500 OSPF peers (neighbors). This test is designed to confirm that claim.

With a maximum port density of 1,260 Gigabit Ethernet ports on the Force10 E1200 (90-port cards with mini-RJ21 interfaces), it is likely that many customers will expect the switch to support very significant quantities of OSPF neighbors, potentially greater than even the 500 neighbors recorded by the Tolly Group.

Our tests employed the 48-port RJ-45 Gigabit Ethernet cards from each vendor. With these cards, the Force10 E1200 can scale to 672 Gigabit ports, while the Catalyst 6509 scales to a maximum of 384.

## EANTC Analysis

The OSPF implementations of the Cisco Catalyst 6509 and the Force10 E1200 showed significantly different behavior and capabilities. The Catalyst showed the performance expected from its Cisco IOS heritage, setting up adjacencies with 360 neighbors and learning all routes at a rate of 180 neighbors per minute.

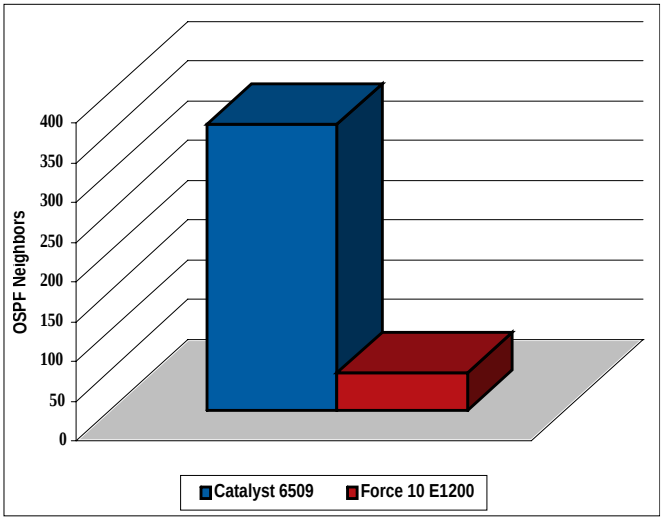



**Figure 1: OSPF Neighbor Scalability**

### Test Highlights

- ➜ **EANTC was unable to reproduce the Tolly Group's claim that the Force10 E-Series can support up to 500 OSPF neighbors.**
- ➜ **Force10's maximum OSPF neighbor scalability was 48 neighbors. That's just 3.8 % of the maximum number of interfaces supported by the switch.**
- ➜ **Network growth to 96 OSPF neighbors resulted in only 33 adjacencies being maintained by the Force10 E1200.**
- ➜ **Cisco's Catalyst 6500 supported up to 360 OSPF neighbors; over seven times the scalability shown by the E1200. In addition, the Catalyst processed OSPF at a rate of 180 adjacencies per minute, whereas the Force10 processed only EIGHT per minute.**
- ➜ **In the overload test, when offered a 432 OSPF neighbors, the Cisco Catalyst 6500 failed to converge and the CPU remained at 100 % load.**
- ➜ **Although the Catalyst 6500 showed 100 % CPU load during route learning, it remained responsive to CLI commands, (the same as the Force10, with its 3-CPU control plane). Once routes had been fully learned, the Catalyst's dual CPUs returned to 2–3 % utilization.**

Force10's OSPF implementation only managed eight adjacencies per minute, considerably slower than the Catalyst. In a real network, this would mean slower recovery from failure by the Force10 switch.

Beyond their respective maximum OSPF neighbor scalabilities, both switches failed to fully converge their routing tables. For example, when asked to support just 72 neighbors, the E1200 managed to bring all neighbors to full adjacency, but failed to process the OSPF database and correctly update the route table. Out of 5,976 external routes advertised by SmartBits, the E1200 only managed to learn 5,626; some 350 routes short.

| Device Under Test | OSPF Neighbors | | Packet Loss At 3,600 pps Per Port | Verdict |
|---|---|---|---|---|
| | Total Configured | Established By Switch | | |
| Force10 E1200 | 600 | 25 | Never converged | FAIL |
| | 312 | 41 | Never converged | FAIL |
| | 96 | 33 | Never converged | FAIL |
| | 48 | 48 | 0% | PASS |
| | 72 | 72 | 7.95% | FAIL |
| Cisco Catalyst 6500 | 600 | N/A | Never converged | FAIL |
| | 312 | 312 | 0% | PASS |
| | 432 | N/A | Never converged | FAIL |
| | 360 | 360 | 0 | PASS |

EANTC believes that Force10's OSPF implementation failed to scale sufficiently to support the very high-density single-switch network designs proposed by Force10 and certainly did not match the findings reported by the Tolly Group.

The Catalyst was not perfect either, when pushed beyond its scalability limit on the failed 432 and 600 neighbor tests, the Catalyst displayed out-of-memory and CPU utilization problems, but remained responsive to the CLI.

In previous tests of the Catalyst 6500 conducted by EANTC, the Catalyst 6513 (the larger brother of the Catalyst 6509) using the same line cards and Supervisor 720 showed it could support up to 410 OSPF neighbors, one neighbor per physical interface. See, see http://www.eantc.com/press/pressreleases/sep03/EANTC-Summary-Report-Cisco-GigE-Catalyst6500-Supervisor720.pdf. This represented the ability to support OSPF neighbors up to the maximum physical scalability of the switch.

This test shows that the Force10 switch could not support the scalability claimed in the Tolly Group reports, supporting less than 10 % of the claimed 500 OSPF neighbors. Of the two switches, the Catalyst 6500 was the winner in this test.

## Test Configuration and Methodology

As before, 24 ports were used in the test, with all ports on the same line card. To scale the test, we configured 25 VLANs and virtual router interfaces per physical port, allowing up to 600 OSPF neighbors to be emulated.

Both Cisco and Force10 process the OSPF routing protocol on a centralized route processing module, so the fact that OSPF neighbors were not distributed across multiple cards did not affect the test results, which primarily exercised the switch's control plane.

The switch under test was configured as an OSPF area border router (ABR); with the SmartBits emulating neighbor routers evenly distributed across 24 areas, including Area 0.

We kept the total number of advertised routes constant, as close to 6,000 as possible to ensure that OSPF route scalability issues did not affect the neighbor scalability test results.

After all routes had been learned, we sent unidirectional background traffic to all routes advertised by each OSPF neighbor at a low rate of 3,600 frames/second per port, measuring packet loss and latency to verify that all packets were forwarded in hardware.

We started with 600 neighbors and employed a manual binary search to find the maximum number of neighbors supported by the switch under test. The switch passed the test if all neighbors were brought to full adjacency, the routing table was populated with the correct numbers of external routes, plus the switch forwarded all traffic in hardware exhibiting zero loss and low latency.

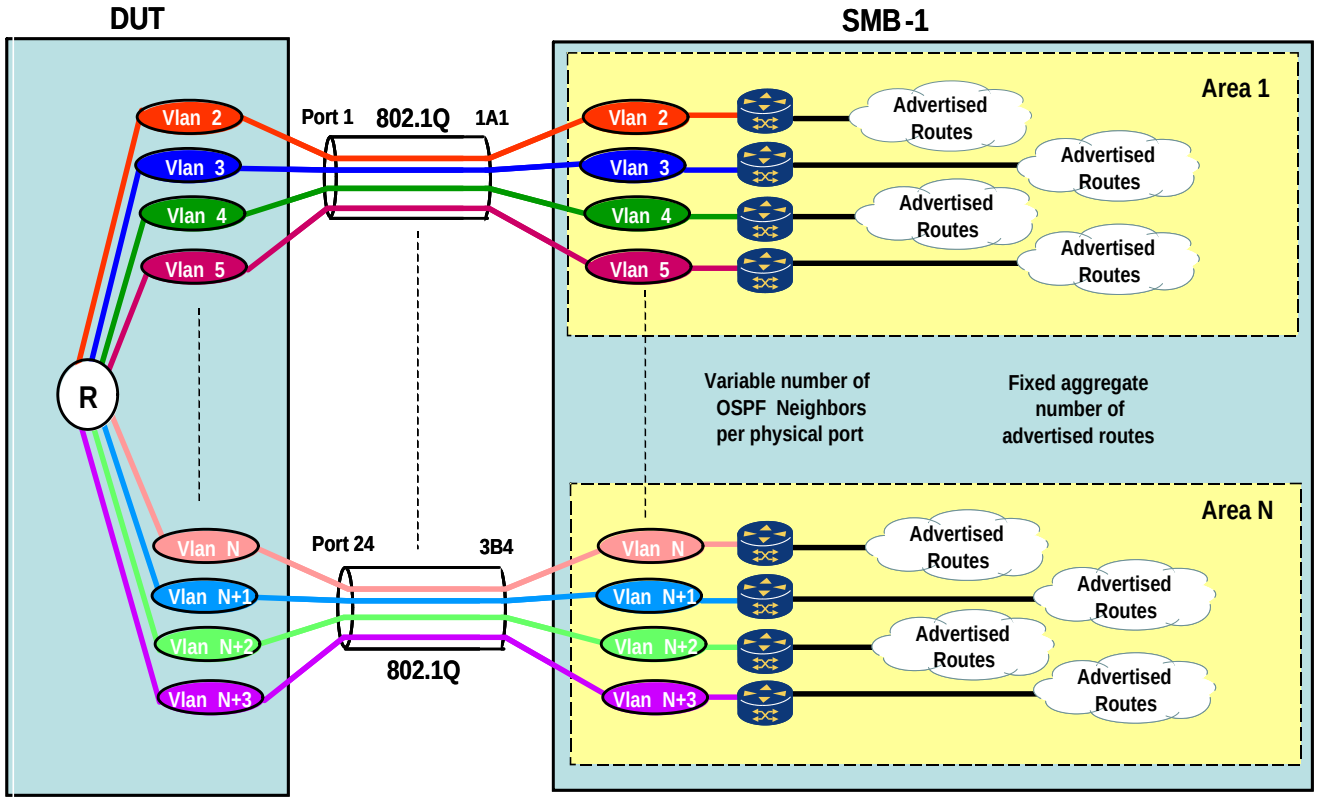



**Figure 2: Test Topology**

Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test

# OSPF Equal-Cost Multipath Test

## Test Objectives

OSPF ECMP is the ability to load-share traffic across multiple routes of equal cost. As load-sharing decisions must be taken at aggregated data rates that are up to 160Gbps (16 x 10Gbps) wire-rate, the load-sharing algorithms must be burned into the hardware switching ASICs if the switches are to provide bandwidth aggregation services at wire-speed.

In addition, to maintain correct end-to-end packet sequencing, the algorithms must ensure that all traffic associated with a particular flow follows the same path and uses the same link for all packets. (A flow is identified by its IP Source Address, IP Destination Address, L4 Protocol-Type, L4 Source Port & L4 Destination Port - this is also known as a 5-Tuple flow.)

As the load-sharing is based on "hashing" the flow information in the packet header using a sophisticated load-sharing algorithm, it is difficult to provide "perfect" sharing across all links. The load-sharing is statistical in nature and depends entirely on the hash result derived from the IP/Ethernet headers, so it is expected that some links will be more heavily utilized than others.

Some vendors recognize this fact and offer multiple load-sharing algorithms, allowing the customer to match the best algorithm to his exact traffic flows, thereby approximating the best load-sharing distribution.

To pass this test, the switch must make use of all available links at each hop in the topology, and must load-share the traffic across these links as evenly as possible. Where possible, we tested the load-sharing using all available hashing algorithms available from the vendor and recorded the results for comparison.

## Test Results

**Force10.** The switches (A) & (C) (see *Figure 3: Test Topology*) did a reasonably good job of sharing the traffic evenly across all outgoing ports. The level of traffic being forwarded to Switch (B) was such that if shared correctly across all eight outgoing ports none of the ports would be oversubscribed and no traffic would be lost.

**Test Highlights**

- ➜ **For Force10, half of the ECMP links remained idle while others were oversubscribed.**
- ➜ **Force10's solution showed a packet loss of 17.5% of all traffic. Even though half the links remained available for use, the Force10 solution was incapable of employing them.**
- ➜ **EANTC tested all 16 hashing algorithms offered by Force10 – none of them cured the problem; in each case 50% of the links remained unused, the algorithms merely changed which links remained idle.**
- ➜ **Force10's OSPF implementation showed problems throughout many of the tests run by EANTC. In two of the OSPF ECMP test runs, the Force10 switches did not install any routes in the routing table, forcing the test run to be aborted. In each case, the OSPF process required a manual "Reset" via the CLI before it would learn the routes and populate the routing table.**
- ➜ **Cisco's Catalyst 6500 distributed traffic over all available links; however, some links were oversubscribed, while others saw 51–65% utilization.**
- ➜ **Cisco's OSPF implementation remained stable in all test runs.**
- ➜ **Overall the Catalyst moved more traffic and at just 5.95% overall loss, it suffered 60% lower loss than the Force10 solution.**

The Force10 switches exhibited a problem on the second hop in the network. On Switch (B), the Force10 default hashing algorithm always forwarded traffic on ***four out of eight*** outgoing ports, ***zero*** packets were forwarded by the other ports.

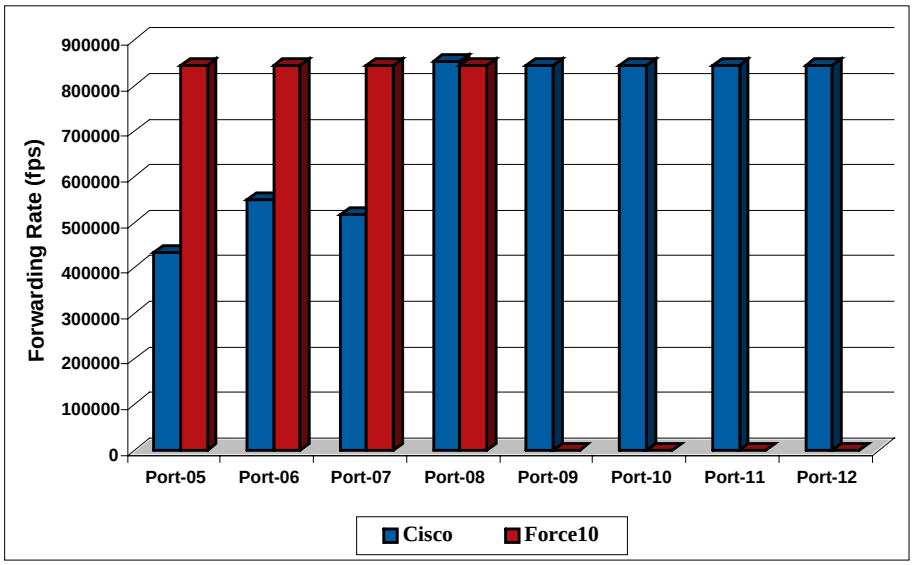



**Figure 1: Traffic Distribution on DUT(B) with default hashing algorithm**

Next, as Force10 offers a choice of 16 alternative hashing algorithms, we investigated whether the problem we had uncovered was nothing more than a badly matched hashing algorithm. We repeated the test a further 15 times, using a different hashing algorithm on each run. The alternative hashing algorithms merely changed which four out of the eight outgoing interfaces forwarded traffic. *None of the hashing algorithms managed to forward traffic over more than four outgoing interfaces.* (see *Figure 2: Force10 DUT(B) - effect of different hashing algorithms*)

A further problem was encountered with the Force10. Prior to the start of each test run, we executed various `show` commands to ensure that all OSPF adjacencies were established and the routing tables were correctly populated.

On two of the test runs, we found that although the links between the Force10 switches showed full OSPF adjacencies, and the link state database was correctly populated, no routes were written to the route table. As tests progressed, this phenomenon re-appeared in numerous different tests. We had to manually reset the OSPF routing process in the switch to get it to populate the routing table.

**Cisco.** Overall the Cisco Catalyst 6500 used the available bandwidth much better than the Force10 E-Series, with the Catalyst forwarding traffic on all eight outgoing interfaces. However, Catalysts (A) & (C) didn't do quite as good a job as Force10 of sharing the traffic evenly across all links. Both Cisco switches (A) & (C) sent a higher percentage of their traffic to Catalyst (B). In fact, due to the imbalance at Catalysts (A) & (C), Catalyst (B) was presented with over 22% more traffic than Force10.

Catalyst (B) correctly forwarded traffic on all eight interfaces; however, the default hashing algorithm on the Catalyst did not manage a perfect distribution across all links. Five of them were oversubscribed, while the others were loaded between 51-65% each.

## EANTC Analysis

The objective of this test was for all switches under test to load-share the traffic as evenly as possible across ALL available links. Neither switch achieved a perfect distribution of traffic, however the Catalyst 6500 came the closest to this goal.

The Force10 E-Series switches do not seem to be designed with multi-hop link aggregation in mind; the fact that half of the links remained idle while the others were oversubscribed will present network managers with a challenge on capacity planning.

| DUT(B) Forwarding Rate per Port | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| F10 Hash Algo-rithm | Port-05 | Port-06 | Port-07 | Port-08 | Port-09 | Port-10 | Port-11 | Port-12 |
| Default | 844,590 | 844,588 | 844,579 | 844,592 | 0 | 0 | 0 | 0 |
| Hash-01 | 844,590 | 844,588 | 0 | 0 | 844,590 | 844,588 | 0 | 0 |
| Hash-02 | 844,589 | 0 | 844,588 | 0 | 844,594 | 0 | 844,591 | 0 |
| Hash-03 | 844,588 | 844,587 | 844,586 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-04 | 844,577 | 844,575 | 0 | 0 | 844,592 | 844,578 | 0 | 0 |
| Hash-05 | 844,591 | 0 | 844,579 | 0 | 844,584 | 0 | 844,581 | 0 |
| Hash-06 | 844,590 | 844,588 | 844,588 | 844,653 | 0 | 0 | 0 | 0 |
| Hash-07 | 844,590 | 844,587 | 0 | 0 | 844,584 | 844,591 | 0 | 0 |
| Hash-08 | 844,590 | 0 | 844,588 | 0 | 844,593 | 0 | 844,570 | 0 |
| Hash-09 | 844,590 | 844,589 | 844,588 | 844,588 | 0 | 0 | 0 | 0 |
| Hash-10 | 844,586 | 844,589 | 0 | 0 | 844,592 | 844,587 | 0 | 0 |
| Hash-11 | 844,590 | 0 | 844,588 | 0 | 844,580 | 0 | 844,593 | 0 |
| Hash-12 | 844,569 | 0 | 844,588 | 0 | 844,593 | 0 | 844,591 | 0 |
| Hash-13 | 844,581 | 0 | 844,588 | 0 | 844,592 | 0 | 844,589 | 0 |
| Hash-14 | 844,589 | 0 | 844,588 | 0 | 844,592 | 0 | 844,591 | 0 |
| Hash-15 | 844,589 | 0 | 844,588 | 0 | 844,572 | 0 | 844,592 | 0 |

**Figure 2: Force10 DUT(B) - effect of different hashing algorithms**

## Test Configuration and Methodology

As explained in the introduction to this group of tests, our test topology is designed to simulate part of a typical enterprise network design. DUT(A) & (C) represent building aggregation switches, aggregating traffic from multiple wiring closet switches, emulated by SmartBits SMB-1.

Each building aggregation switch is dual-homed to two backbone L3 switches. DUT(B) represents one of these backbone switches, with the other being emulated by the eight SmartBits ports 2B1–2B4 and 4B1 - 4B4.

To ensure all switches saw eight equal cost paths, the OSPF cost was adjusted on Ports 13–16 on DUT(A) and DUT(C), compensating for the fact there was one less hop on this path.

All switches were configured in OSPF Area 0 and a total of 4,800 external routes advertised by SmartBits. The same routes were advertised on each SmartBits port, thereby creating the conditions necessary for equal cost paths to be seen by each DUT. Unidirectional traffic (128 Byte packet size, 750 Mbit/s per port) was sent from the emulated users on the left of the diagram to the emulated servers and external routes on the right of the diagram.

We ran each test for 100 seconds and recorded the number of packets forwarded on each interface of each switch so we could assess the traffic distribution.

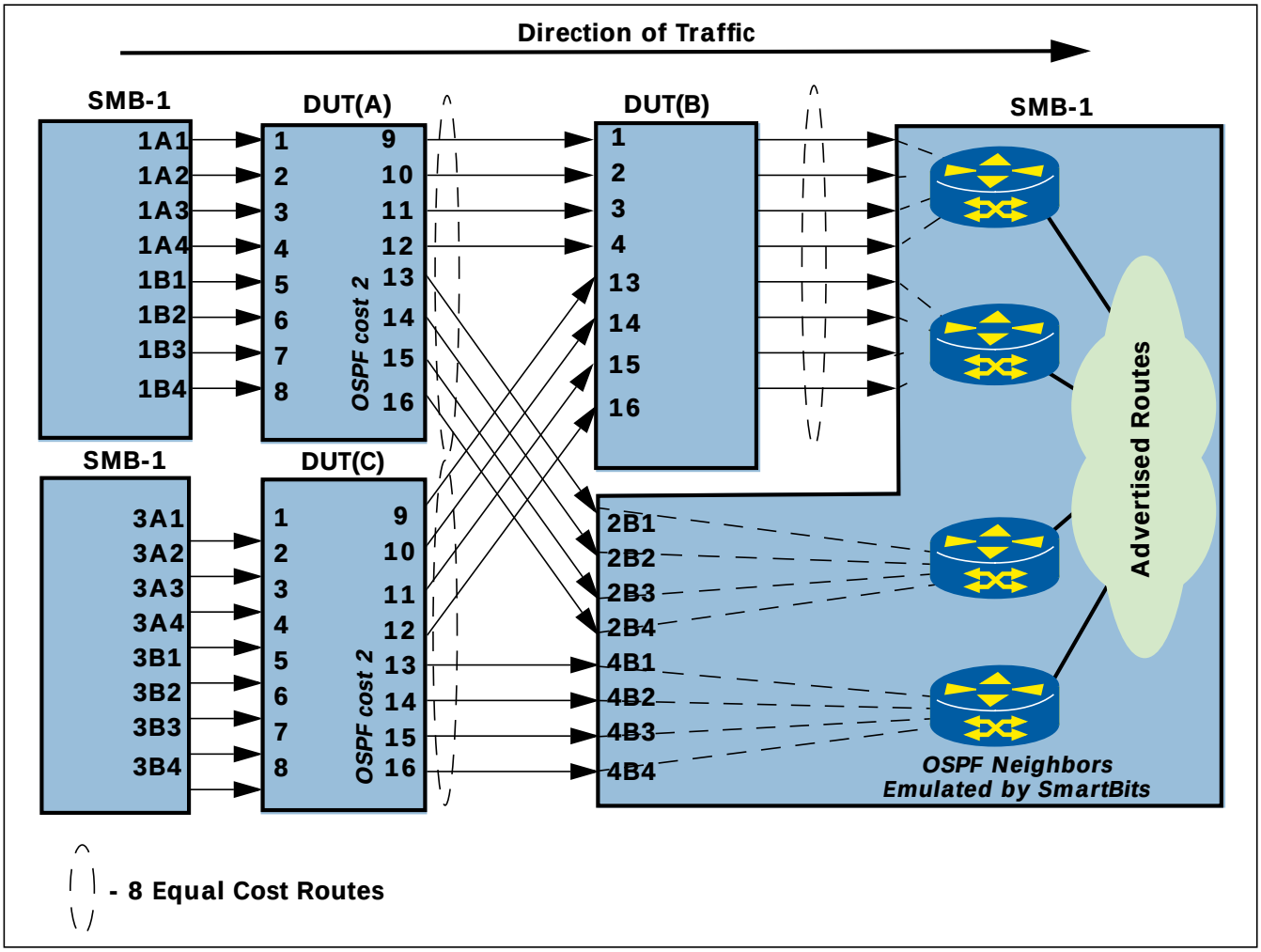



**Figure 3: Test Topology**

# Executive Summary — Bandwidth Aggregation Group

In June 2005, Cisco Systems commissioned the European Advanced Networking Test Center (EANTC) to independently validate the performance, scalability, and availability of Force10 TeraScale E-Series and Cisco Catalyst 6500 switches.

Our tests in the Bandwidth Aggregation Group examine the switch's ability to provide high volume data paths between switches. Backbone switches such as the Force10 E-Series and the Cisco Catalyst 6500 need this ability to provide high capacity inter-switch data paths for traffic that exceeds the capacity of a single Gigabit or 10-Gigabit link. This test report covers Gigabit Ethernet.

There are two primary bandwidth aggregation schemes available.

- OSPF equal cost multi-path routing (ECMP)
- Link aggregation, also known as Trunk-Groups or EtherChannel

Often, network designs incorporate an element of each scheme in the same network (see the right hand diagram below). In such circumstances, it is important that both OSPF ECMP and Link-Aggregation work together to provide the most efficient use of the available link bandwidth end-to-end.

Our tests were designed to simulate part of a typical enterprise network design, with the rest of the network being emulated by the SmartBits test equipment.

In *Figure 1: EANTC bandwidth aggregation tests simulated typical enterprise network designs*, the three

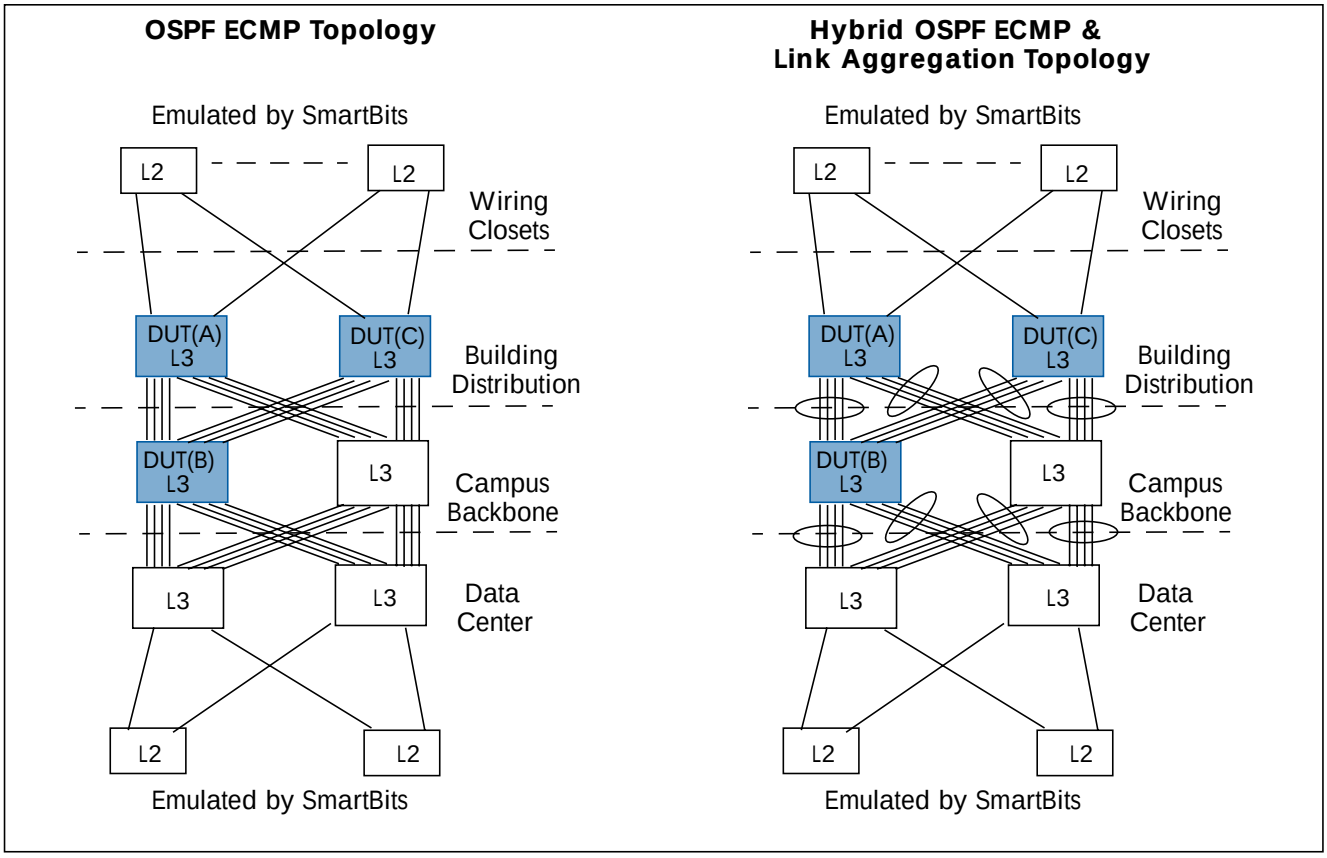



**Figure 1: EANTC bandwidth aggregation tests simulated typical enterprise network designs**

switches with bold outlines were the switches under test and represent where these switches would sit in a typical network. The switches with broken outlines were emulated by SmartBits.

The devices under test (DUTs) were configured so that as far as the OSPF routing protocol is concerned, there are eight equal cost routes for any destination at each point in the network.

In the OSPF ECMP topology, each switch sees eight equal cost paths, whereas in the hybrid network, each aggregated link is seen by OSPF as a single logical link with high bandwidth. In the hybrid network, OSPF ECMP balances the traffic between each link-aggregation group, while link-aggregation is responsible for load-sharing traffic across the individual links comprising the group.

Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test

# OSPF ECMP & Link Aggregation Test

## Test Objectives

Many enterprise networks employ different techniques of load sharing in parallel. For example, link aggregation enables the use of multiple links between switches, presenting these multiple links as a single logical link of high capacity to higher level protocols such as OSPF. If two link-aggregation groups offer equal cost paths to a destination network, OSPF will load-share the traffic equally between the two paths.

This test assessed the ability of the devices under test to evenly load-share traffic across all available equal cost paths, and make use of all available links in each Link-Aggregation Group (LAG). This topology represents an alternative to the use of OSPF ECMP alone.

Once again our test topology was designed to simulate a real-world network design, with DUTs (A) & (C) representing building aggregation switches, servicing numerous wiring closets spread over multiple floors. While DUT (B) represents a backbone switch, dual-homed to two other backbone switches, the rest of the topology is emulated by SmartBits test equipment. As before, OSPF costs applied to Ports 13 - 16 of DUT(A) & (C) compensate for the fact that this path is one hop less than the path via DUT(B), thereby providing DUT(A) & (C) with equal cost paths once more.

## Test Results

**Force10.** Please refer to the topology diagram Figure 2, Page 3. Switches (A) and (C) only forwarded on two out of the four links available within each link aggregation group (LAG). These links were oversubscribed and DUT(A) & (C) both dropped frames.

When the traffic reached Switch (B), one of the two LAGs remained unused, whilst on the second LAG, only two out of the four links forwarded traffic. Of course, these two links were heavily oversubscribed and resulted in packet loss.

In total, end-to-end only six out of 16 possible links were used by the Force10 switches to forward traffic. This reduced the end-to-end aggregate capacity of the network by 62.5 % and resulted in 51 % traffic loss. We also re-ran the test a further 15 times to check out all of Force10's alternative hashing algorithms on DUT(B). None of the algorithms made a difference as can be seen in *Figure 2: Force10 Forwarding Rate DUT(B) With Different Hash Algorithms.*

**Test Highlights**

- ➜ **Force10 TeraScale E-Series performed poorly with just OSPF ECMP alone in a multi-hop configuration.**
- ➜ **DUTs (A) & (C) only used *four out of eight* available links**
- ➜ **DUT(B) used only *two out of eight* available links**
- ➜ **Force10's solution showed a packet loss of 51% even though many links remained idle throughout the test.**
- ➜ **Force10's solution reduced available end-to-end bandwidth by over 63% to just six Gbit/s.**
- ➜ **EANTC tested all 16 of Force10's hash algorithms, but these had no effect on the number of ports used. DUT(A) & (C) never used more than four out of eight available links, while DUT(B) never used more than two.**
- ➜ **Cisco Catalyst 6500 distributed traffic over all links and achieved a near-even traffic distribution.**
- ➜ **Catalyst 6500 forwarded all traffic without packet loss when using a Cisco recommended hash algorithm and incurred only 1.1% traffic loss with the default hash algorithm.**

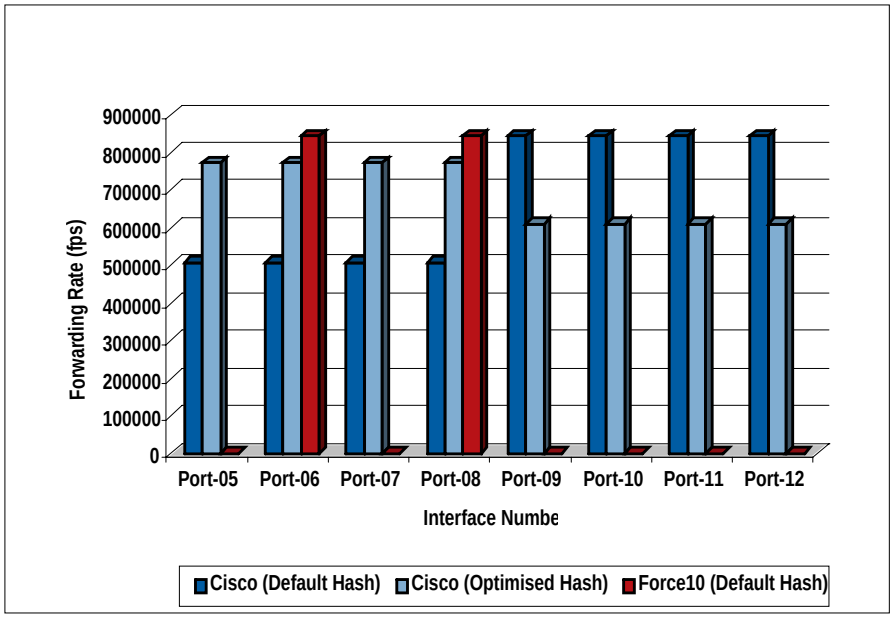



**Figure 1: OSPF/ECMP Traffic Distribution Across Ports**

**Cisco.** Cisco's Catalyst 6500 showed almost even traffic distribution across all available links with all three DUTs. Though we observed 1.16 % packet loss when we used the default hashing algorithm, optimizing the hash algorithm helped to forward all traffic without any packet loss.

## EANTC Analysis

Many test labs only test link aggregation in a single hop environment. However, this is rarely how it is employed in real networks. As discussed in the executive summary for this test section, there are typically either multiple hops of link aggregation or OSPF equal-cost multi-path and link aggregation used in unison. Our test results show that single-hop link aggregation tests can hide significant problems with the vendors' bandwidth aggregation features.

Due to the very large volumes of traffic, the hashing algorithm must be implemented in hardware to provide wire-speed bandwidth aggregation. The problems seen in this test are hardware related and therefore not easily fixed by software.

| Hash Algorithm | Port 05 | Port 06 | Port 07 | Port 08 | Port 09 | Port 10 | Port 11 | Port 12 |
|---|---|---|---|---|---|---|---|---|
| Default | 0 | 844,544 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-01 | 0 | 844,544 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-02 | 0 | 844,544 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-03 | 0 | 844,545 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-04 | 0 | 844,544 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-05 | 0 | 844,543 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-06 | 0 | 844,543 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-07 | 0 | 844,543 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-08 | 0 | 844,544 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-09 | 0 | 844,544 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-10 | 0 | 844,544 | 0 | 844,586 | 0 | 0 | 0 | 0 |
| Hash-11 | 0 | 844,543 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-12 | 0 | 844,543 | 0 | 844,586 | 0 | 0 | 0 | 0 |
| Hash-13 | 0 | 844,543 | 0 | 844,586 | 0 | 0 | 0 | 0 |
| Hash-14 | 0 | 844,543 | 0 | 844,587 | 0 | 0 | 0 | 0 |
| Hash-15 | 0 | 844,543 | 0 | 844,586 | 0 | 0 | 0 | 0 |

**Figure 2: Force10 Forwarding Rate DUT(B) With Different Hash Algorithms**

## Test Configuration and Methodology

Link aggregation is a feature that only applies to outgoing traffic. The group of physical ports that comprise a Link-Aggregation Group (LAG) is seen by the higher level routing protocols in the DUT as a single logical interface that represents the whole LAG.

In real-time this logical interface applies a hashing algorithm to the traffic being forwarded across the LAG and selects one of the "N" links available in the LAG to forward the traffic. The hashing algorithm is built into the hardware switching ASICs of the switch as it must cope with hashing N x wire-rate traffic streams simultaneously. The hashing algorithm uses the address fields present in each packet as its input and derives a single hash result that represents one of the "N" links in the LAG; this is the link that will be used to forward all packets in this particular 5-tuple flow.

This address-based hashing technique ensures that all packets matching the same addresses will always derive the same hash result and will therefore always be forwarded down the same link in the LAG. This guarantees end-to-end packet sequencing is preserved.

By default, SmartBits cannot emulate a LAG. However, we devised a configuration without the 802.3ad Link Aggregation Control Protocol (LACP) to enable SmartBits to emulate an adjacent node in a LAG and fool the DUT into thinking it is attached in the logical topology shown in the Test Topology diagram. LAGs in the DUTs were configured statically rather than being automatically configured through LACP.

Test traffic was sent at 750Mbps per port, leaving sufficient room for all traffic to be delivered should the traffic distribution be less than perfect. When conceiving the test however, we did not expect the Force10 switch to only forward traffic over such a limited subset of its available interfaces.

The packet counts from each DUT and from the SmartBits were recorded, as were the Port Receive Rates on the SmartBits as the test proceeded. This gave us multiple data points from which to analyses exactly what was happening during the test, which interfaces were forwarding and at what rate and why we got the results we did.

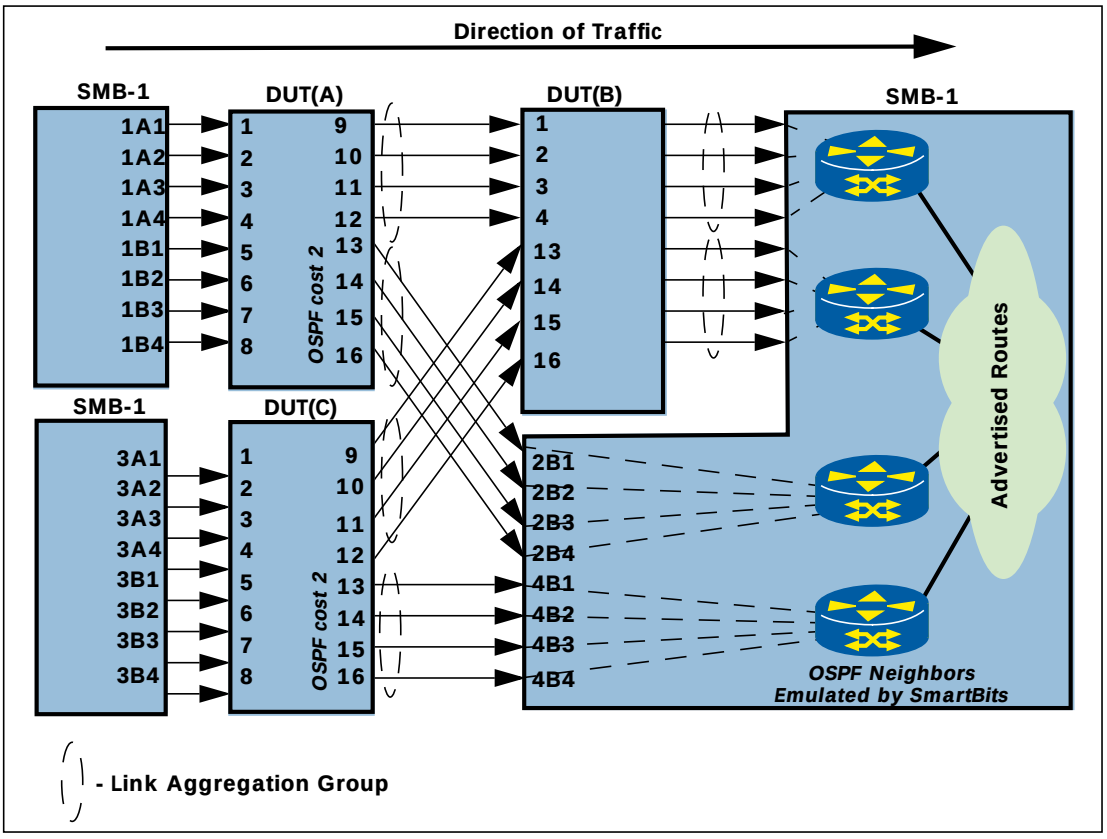



**Figure 3: Test Topology**

# OSPF Continuous Route Flap Test

## Test Objectives

This test assesses the ability of the switch to respond to continuous rather than single one-off route flaps. Continuous route flaps are common in networks and can expose potential weaknesses in the OSPF protocol implementation; memory management processes and RIB/FIB/CAM table synchronization that single-flap tests do not uncover.

Glossary:

- RIB = Routing Information Base, a.k.a the Routing Table, this is a software table held in memory on the management card.
- FIB = Forwarding Information Base, this is built from network prefixes taken from the RIB with additional adjacency information derived from the ARP-Cache. It is a software table stored on the management card and used as the master L3 forwarding table for the switch. A copy of the FIB is downloaded into the CAM on each line card.
- CAM = Content Addressable Memory, this is the hardware-based table, present on each line card that is used by the hardware switching engine to make wire-rate forwarding decisions

Our objectives for the test were to assess whether any problems would develop when the switch was subjected to continuous route flaps for a period of up to 15 minutes.

**Test Highlights**

- ➜ **Force10's FIB (software) and CAM (hardware) became desynchronized, with each table containing different numbers of entries. In a real network this could cause serious problems.**
- ➜ **Flapped prefixes were not completely removed from the Force10's CAM**
- ➜ **Force10's route processor CPU (RP1) went to 100 % overload and remained there even after OSPF was terminated on the test equipment. All manual recovery actions failed, the system had to be rebooted.**
- ➜ **Force10's 3-CPU control plane "locked up" for all commands related to OSPF. At one point we were locked out of the switch console for over 20 minutes.**
- ➜ **Cisco's Catalyst 6500 worked its way through 15 minutes of continuous route flaps, remaining stable throughout the test.**

## Test Results

In this test it did not take long before we observed routing problems on the Force10 E-Series. During the 3rd round of route-flaps, error messages were displayed on the Force10 console, indicating that a route the switch was trying to delete did not exist in the CAM. This indicates that on the previous round of route flaps, the RIB/FIB/CAM contents were not restored to their original state (see Figure 1).

We noted that this error condition caused CPU (RP1), which is responsible for L3 routing to go into an overload condition, remained there for the rest of the test.

```
03:45:03: %RPM0-P:RP1 %RTM-3-RT NOT FOUND: RTM: rtmDelRouteIPv4:1:  No such route cid 21 (10.2.107.0, 0x1a, O)
03:45:03: %RPM0-P:RP1 %RTM-3-RT NOT FOUND: RTM: rtmDelRouteIPv4:1:  No such route cid 21 (10.2.113.0, 0x1a, O)
03:45:03: %RPM0-P:RP1 %RTM-3-RT NOT FOUND: RTM: rtmDelRouteIPv4:1:  No such route cid 21 (10.2.57.0, 0x1a, O)
03:45:03: %RPM0-P:RP1 %RTM-3-RT NOT FOUND: RTM: rtmDelRouteIPv4:1:  No such route cid 21 (10.2.71.0, 0x1a, O)
[...]
```

**Figure 1: Force10 RIB/FIB and CAM Become Desynchronized**

```
Force10-E1200#sho proc cpu sum

CPU utilization        5Sec    1Min    5Min
-------------------------------------------
CP                       0%      0%      0%
RP1                    100%    100%     99%
RP2                      0%      0%      0%
```

**Figure 2: CPU RP1 Overload Lasted Greater Than 5 Minutes**

**Force10's 3-CPU Control Plane.** The above situation presented EANTC with an opportunity to investigate another claim made by Force10 in their white-paper entitled, "Guaranteed Access to System Management even during Processor Overload".

Force10 claims, "*This (3-CPU control plane) architecture isolates IP routing, Layer 2 tasks, and management functionality to three individual CPUs, respectively. This ensures that the administrator can always log on to the "management CPU", receive precise knowledge as to the cause of the problems, and take corrective action thereby avoiding disruptive system reboots*".

We could see the process holding the CPU at 100% load was OSPF, so we tried to recover from this situation using console commands.

**Recovery Attempts Without Rebooting.** Firstly we issued a `show ip ospf neighbor` command to make sure all our OSPF neighbors were in "Full" adjacency state. Once this command was issued, the console "locked up" for a period of two minutes, during which time no further commands could be entered.

After the two minutes passed, the console prompt once again returned, but the information we had asked for was not displayed. It appears the show command had timed out. If we were trying to use such commands to gain "*precise knowledge*" of what was happening, we would not have been successful.

Next to avoid rebooting the switch, we issued a `clear ip ospf process` command, in the hope that we could clear the OSPF process, teardown and reestablish adjacencies, and rebuild the OSPF database without the need to do a full reboot. (Of course, in a real network this action would be highly disruptive, but we wanted to see what would happen). *Figure 3: Attempting To Clear The Force10's OSPF Routing Process* is an extract from the console log, which shows the output of `clear ip ospf process` command.

Once again the console locked up for 120 seconds, allowing no further commands to be entered. When the CLI prompt returned, it was accompanied by an error message stating that our request to clear the OSPF process had timed out.

Even after we had stopped OSPF on the SmartBits traffic generator and it closed all its OSPF adjacencies, the RP1 CPU continued to show 100% overload.

In a final attempt to recover the situation without rebooting the switch, we issued a `shutdown` command to all the interfaces used in the test. The immediate effect of this action was to lock-up the console for approximately 20 minutes.

We believe the reason for this 20 minute lock out was that each interface needed to tell the OSPF process that it had been shutdown, so the OSPF process could remove those routes from its database. Because RP1 was still overloaded, these attempts timed out and it was only after the sum of these timeouts that we gained access to the CLI prompt once more.

To summarize, when RP1 became overloaded, neither show nor configuration commands relating to OSPF worked on the Force10 switch indicating that the facilities provided by the control processor (CP) are dependent on it being able to access and communicate with the other two processors on the RPM module.

The only way to recover from this CPU overload was to reboot the switch.

**Cisco Test Results.** In contrast to the Force10's performance, the Cisco Catalyst 6500 remained stable for the whole 15 minute duration of the test and correctly handled all route-flaps. The average CPU load showed peaks of only 10%, and the console remained responsive throughout.

The Catalyst 6500 recovered completely from the test, and there was no need to reboot or clear OSPF processes during or afterwards. The Catalyst showed no signs of memory management problems, and no memory leaks were discovered.

```
Force10-E1200#clear ip ospf proc
Reset ALL OSPF processes? [y/n]: y
03:50:45: %RPM0-P:CP %MIB-6-SRC_TO: Task CLI send ReqType 3 ObjId 10148. Request times out.
03:50:45: %RPM0-P:CP %MIB-6-TIMEOUT: No response from IFMGR after 120 seconds. Timing out....
```

**Figure 3: Attempting To Clear The Force10's OSPF Routing Process**

## EANTC Analysis

The Force10 E1200 did not pass the test. Once the CPU (RP1) became overloaded (by a route flap of 2,000 routes each time), it refused to respond to further route flaps. It also lost synchronization between the RIB, FIB and CAM, leaving the system in an unusable state from where it could not recover without a manual reboot.

In EANTC's opinion, Force10's 3-CPU architecture showed no real advantage over a single CPU architecture, failing to deliver on Force10's promises about unaffected management operation under CPU overload conditions. Indeed, the separate control processor (CP), which was not overloaded continued to pass heartbeats to the standby RPM (management module) thereby preventing the switch from initiating a failover; something that might have happened in a single CPU architecture and that might have recovered the situation without operator intervention.

Although we are not absolutely sure about the root cause for Force10's failure, we know it is critical that the hardware and software forwarding tables never desynchronize in a FIB-based switch. Once they did, the E1200 was not able to continue correct operation, and in a real network could have been responsible for routing loops or for forwarding traffic into black holes.

## Test Configuration and Methodology

In this test we configured 16 ports on the DUT. Ports 1 through 8 were attached to traffic sources, while Ports 9–16 advertised routes and were the destination of the traffic.

Each port on the right of the diagram advertised a range of networks; each with a single prefix length as shown in the diagram. A percentage of those networks were also advertised by the next port, using a shorter prefix length, thereby forcing the DUT to choose the best route using a longest-prefix-match routing algorithm.

Networks are often advertised with multiple prefix lengths. For instance an OSPF Area Border Router (ABR) will learn networks advertised within an area with a certain set of prefix lengths and will aggregate these multiple routes together, re-advertising them as a single aggregate inter-area route with a shorter prefix length. Such OSPF border routers must be able to efficiently apply the longest prefix match algorithm when making forwarding decisions, something this test exercises.

Each port advertised 2,000 routes, each at a specific prefix length as shown in the diagram. The only exception to this were the routes advertised with a /8 prefix. It wasn't possible to advertise 2,000 of these routes as we'd run out of IPv4 address space, so instead we settled for just 40 routes with a /8 prefix. In total, this meant that 14,040 routes of different prefix lengths were used in the test

Sequentially for each prefix length, 2,000 routes were withdrawn and then restored after 10 seconds (40 routes for the interface with /8 prefix length). In effect, the DUT had to reroute traffic to whatever route represented the longest prefix match.

When the route flap happened, a subset of the routes were completely unreachable, while the rest were still reachable via routes advertised with a less specific, shorter prefix length. The DUT should continue to forward traffic to all available routes and should discard packets destined to the withdrawn routes that have no alternate path.

Every time routes were withdrawn or restored, by design of the test, the DUT had to update its forwarding information base (FIB) and content-addressable memory (CAM) on each of its line cards.

Test traffic was sent over the routes at a low rate of 5,000 frames per second per port with 64 Byte frame length to confirm the routes were being withdrawn and restored correctly.

The test records whether the DUT maintains the correct forwarding of the test traffic, the correct management of the RIB/FIB/CAM tables, which must remain synchronized at all times, whether the longest-prefix-match algorithm is applied correctly and whether any memory management problems manifest themselves during the extended test period.

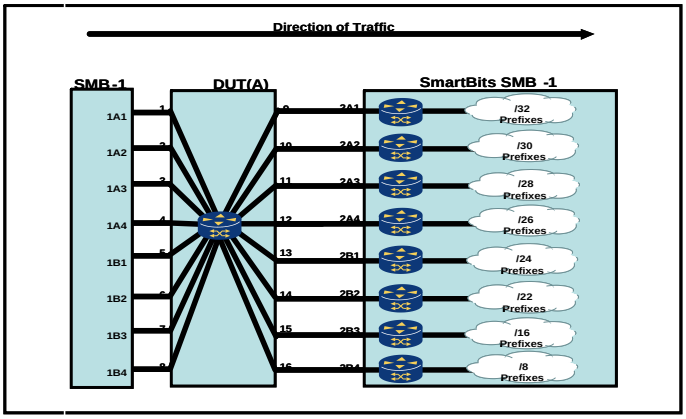



**Figure 4: Test Topology**

# BGP Route Scalability Test

## Test Objectives

This test assessed the maximum BGP Route scalability and recorded what happens when the device under test exceeds its BGP capacity. Unlike Force10's Tolly Group tests, we chose not to use BGP Multipath (equal cost routes), but instead to focus on the true BGP route scalability of the DUT, not the BGP Path scalability.

## Test Results Force10

**250,000 Non-Contiguous BGP Routes – Single Line Card.** Based on our previous test results, where we tried to recreate the Tolly Group's 6-million BGP path test (and failed), we started with a 250,000 route configuration. (We used prefix sizes of /18 to /32 evenly distributed from 1.0.0.0 to 33.140.192.0.) It took Force10 5:09 minutes to fully populate the software-based FIB. The CAM on the line cards was never populated with the correct number of entries (only 214,978 entries were in the CAM). *Figure 4: Force10 Inter-Process Communication (IPC) Problems* is an extract from the console log, which shows the warning messages we observed repeatedly during the test.

The software-based FIB (forwarding information base), which acts as a master table from which the individual CAMs (content-addressable memory) on each line card are populated were now de-synchronized.

In a FIB-based switch such as the Force10 TeraScale E-Series, it is imperative that the FIB and CAM remain synchronized, otherwise the switch's control plane loses all visibility of exactly what is in the CAM and what is not.

We saw this exact same problem in the OSPF continuous route flap tests we ran; where the desynchronization of the RIB/FIB/CAM caused the route processor RP1 to go into an unknown state, never to recover. In this situation, the switch failed to respond to route-flaps, potentially putting the network in danger of routing loops.

**Test Highlights**

- ➜ **Catalyst 6500 scaled to 500,000 BGP routes maintained in the Routing and Forwarding Information Bases and the Content Addressable Memory (RIB/FIB/CAM). Force10's RIB/FIB/CAM managed 200,000 BGP routes.**
- ➜ **Catalyst 6500 fully populated its hardware with all 500,000 routes in less time (1min 51secs) than it took Force10 to populate just 200,000 routes (3min 2secs).**
- ➜ **At 250,000 BGP routes and above, Force10 ran out of FIB/CAM on the line cards and couldn't forward traffic to all routes in hardware.**
- ➜ **Force10 displayed error messages showing congestion on the route processor module (RPM) to line card internal data path. Inter-process messages were lost. This resulted in the RIB/FIB/CAM becoming de-synchronized, a potentially dangerous situation.**

From the results above, it appears the same could happen with BGP and the loss of IPC messages could be rooted in architectural problem in the E-Series.

When traffic was sent, we observed between 7–14 % packet loss for different test runs.

```
Force10#show ip fib line 0 summary
 Total Number of Routes in the FIB database is 250102
 Total Number of Routes in the CAM is 232381
 Total Number of Routes which can be entered in CAM is 262144
```

**Figure 1: FIB, CAM Summary**

**50,000 Non-Contiguous BGP Routes — 4 Line Cards.** As the test evaluated not only the control plane's ability to learn the routes, but also the ability of the DUT to provide hardware-based forwarding of traffic to each network, the above tests were deemed to have failed. We therefore reduced the number of routes to be advertised to 200,000. (We used prefixes /16 to /32 evenly distributed from 1.0.0.0 to 92.231.0.0.)

At this network size, no packet loss was observed, so the Force10 switch passed the test.

| Total Number of Advertised Routes[a] | All BGP Peers Established? | Convergence Time | Packet Loss (%) | Average Latency (µSec) | Verdict |
|---|---|---|---|---|---|
| 500,000 | Yes | 7m 5s | 47.6[b] | 35 | FAIL |
| 200,000 | Yes | 3m 2s | 0 | 34 | PASS |
| 250,000 | Yes | 4m 6s | 7[b] | 35 | FAIL |
| 250,000 | Yes | 3m 58s | 13.4[b] | 40 | FAIL |

a. 16 ports were distributed over 4 line cards
b. Packet loss (in percent of total traffic) is provided as an indication of whether RIB/FIB/CAM were fully populated with all routes only; it was not a frame loss test.

**Figure 2: Force10 Route Scalability Results**

### Tests Results Cisco

During the test with 1,000,000 routes, we noticed that though all tables (FIB/CAM) contained all 1,000,000 entries, some streams were forwarded in software (see *Figure 5: Cisco FIB TCAM Exception*).

| Total Number of Advertised Routes | All BGP Peers Established? | Convergence Time | Packet Loss (%) | Average Latency (µSec)[a] | Verdict |
|---|---|---|---|---|---|
| 250,000 | Yes | N/A | 0 | 9 | PASS |
| 500,000 | Yes | 1m 25 s | 0 | 9 | PASS |
| 1,000,000 | Yes | 4m 51s | 47.8[b] | 2912 | FAIL |
| 900,000 | Yes | 3m 53s | 40.5[2] | 2606 | FAIL |
| 700,000 | Yes | 2m 47s | 23.9[2] | 2036 | FAIL |
| 600,000 | Yes | 2m 31s | 11.4[2] | 1637 | FAIL |

a. Cisco's latency is skewed by the fact that certain flows were forwarded in software
b. Packet loss (in percent of total traffic) is provided as an indication of whether RIB/FIB/CAM were fully populated with all routes only; it was not a frame loss test.

**Figure 3: Cisco Route Scalability Results**

## EANTC Analysis

For a platform to offer high levels of scalability, it is necessary to look beyond simple port density and packet forwarding performance alone as the key criteria for selecting a switch and consider whether the switch will provide a stable platform that can be relied upon for the transport of mission critical traffic.

In this test we saw a difference between the Cisco Catalyst 6509 and the Force10 TeraScale E-Series when tested with exactly the same number of ports. Throughout the tests, the Catalyst impressed us with its rock-solid scalability, fast route learning, and overall routing protocol stability.

Catalyst 6500 learned more than twice as many routes (500,000) as Force 10's E1200 (200,000), completing the task in just one minute 51 seconds. During the Force10 test runs with 250,000 routes or more, we observed serious internal communications problems between RPM (management card) and line card(s). Some internal messages were dropped leading to inconsistent state between FIB and CAM. We also noticed that this problem got worse if ports were distributed over many line cards because the number of CAMs that need to be synchronized with the FIB, increased. In a multi-card configuration we also noted a doubling of the packet loss, even though the number of routes and the traffic rate remained the same.

## Test Configuration and Methodology

32 ports were used in the test. 16 ports were connected to eBGP routers emulated by SmartBits, each in a different Autonomous System (AS). A manual search was used to find the maximum BGP route scalability. Traffic at 64 bytes frame size was sent to all routes advertised by each BGP Peer at 75% of the line rate to confirm full functionality and hardware forwarding of all traffic. SmartBits advertised an even distribution of prefixes with different prefix lengths. An Internet Mix of prefix sizes was not possible in this test due to the very large number of routes to be generated.

```
22:19:04: %RPM0-P:RP1 %SWP-2-IPC SEND FAILURE: SWP:swpSend: queue 27, rtm0 to fibAgt6, failed to send
out an IPC buffer
22:19:04: %RPM0-P:RP1 %SWP-3-IPC HDR DUMP: SWP: IPC Header Dump: SWP IPC header dump: mlen = 1804
source.c
22:19:04: %RPM0-P:RP1 %RTM-3-SWP DEQUEUE FAILURE: RTM: rtmMessageSwitch MT_SWP_ACK: Failed at SWP
dequeue
```

**Figure 4: Force10 Inter-Process Communication (IPC) Problems**

```
2d17h: %MLSCEF-SP-7-FIB_EXCEPTION: FIB TCAM exception, Some entries will be software switched
```

**Figure 5: Cisco FIB TCAM Exception**

**Force10 TeraScale E-Series Competitive Test**

# Force10-specific 6 Million Paths BGP Route Scalability

## Test Objectives

Force 10's Tolly Group Test Report (204148) states that the E1200 scales to 6-million BGP Paths while passing "background traffic."

Firstly, it is important to understand exactly what Force10 is claiming. BGP paths are in fact very different from BGP routes:

*BGP Routes* x *Equal Cost Paths* = *Total BGP Paths*

The objective of this test was to exactly reproduce the Tolly Group test configuration and validate Force10's results.

## EANTC Expectations

In order to pass this test, we decided on the following PASS criteria. All routes and paths should be learned and correctly written to the software and hardware forwarding tables in the switch enabling traffic to all destinations to be forwarded in hardware.

## Test Results

First, we aimed to reproduce the Tolly test configuration exactly. This attempt failed because we mis-calculated that Tolly Group would have used 375,000 routes, learned over the maximum supported 16 equal cost routes.

In our test we found the DUT's software routing table managed to learn all 375,000 routes offered over the 16 equal cost paths as can be seen in the SHOW IP BGP SUMMARY command below. The route table also confirms that 6,000,000 paths are not equal to 6,000,000 BGP routes — note that only 375,000 routes are in the software routing table (RIB).

Although the software route table demonstrated it was able to hold 375,000 BGP routes, the same cannot be said of the CAM serving the hardware forwarding engines on each line card. Based on Force10 documentation, the CAM can hold 256K (262,144) IPv4 FIB entries and as such we expected to see CAM overload and calculated the expected packet loss as follows:

((375,000-262,144)/375,000)*100 = 30.09% loss

**Test Highlights**

- **Force10's 6-million BGP paths result was achieved using just two switch interfaces, with 260,000 advertised BGP routes, learned over 25 equal cost paths.**
- **EANTC was unable to exactly reproduce Force10's 6-million BGP path result with the Tolly Group, because FTOS production versions support 16 instead of 25 equal cost paths. Contrary to the Tolly Group report, the software used for the Tolly test has not been publicly released as a production version.**
- **EANTC's attempt to replicate the Tolly Group test using 16 equal-cost paths failed. Clearly, the resulting 375,000 unique routes exceeded Force10's BGP scalability from both the control and data plane hardware perspective.**

**During the 375,000 route test, the Force10 TeraScale architecture exhibited the following fatal issues:**

- **Severe IPC (Inter-Process Communication) problems resulting in "very slow" BGP route/path learning.**
- **A task crashing on a single line card caused the switch to power-cycle all line cards in the chassis.**
- **Low memory as the number of routes and paths increased, causing already learned BGP peer sessions to be lost.**
- **Force10 switch ran out of hardware switching engine FIB/CAM space. This means any additional routes would not be hardware switched, only learned by the control plane. This caused packet loss even at 100 packets/sec.**

**Route Table Convergence.** We noted that Force10's BGP route table converged very quickly; however, updating of the FIB/CAM on each line card took a long time and the switch issued large numbers of error messages indicating the internal switch architecture was under stress (see *Figure 1: RP1 running low on memory*).

*Figure 1: RP1 running low on memory* contains an extract from the console log, which shows the switch complaining about lack of memory. Note that BGP peers were lost, as the E1200 could not allocate sufficient memory to sustain the connection.

The console log shown in *Figure 2: Out of IPC buffer* confirmed that the route table management process (RTM) was experiencing difficulty in updating the line cards. This appears to be due to congestion on the 100Mbps internal control plane to line card management path, indicated by the fact that IPC (inter-process communication) messages are being queued for transmission.

It is clear from the error messages above, (which were repeated numerous times) that inter-process messages were lost and the system waited for acknowledgements from the line cards that never arrived.

All line cards were affected by this problem, slowing down the test progress drastically because the switch had to send the same messages to line cards repeatedly in an attempt to maintain RIB/FIB/CAM synchronization, which had clearly been lost at this stage.

As the test progressed, the DUT reported that the FIB/CAM on each line card ran out of entries and could not be written to learn all routing table entries (see *Figure 3: RADIX insertion failure*).

This message confirmed the test's 375,000 BGP routes exceeded the DUT's CAM scalability limit, and that it is not possible that the E1200 could learn 6,000,000 unique BGP routes.

The system was stalled for a further 45 minutes, in a state where only 173,681 entries were successfully written to the FIB/CAM on each line card — showing that Force10 did not implement any working counter measures against this system overload.

Next we started the test traffic and recorded 75% packet loss — greater than we expected, it seems the switch did not know where to send the majority of packets generated by SmartBits.

At the end of the test, SmartBits brought down all BGP peers, and the Force 10 switch cleared the routing table. Due to the earlier problems with internal communication; however, the DUT found it impossible to remove the obsolete network prefixes from the

```
1d1h23m: %RPM0-P:RP1 %KERN-5-INT: RP1 running low on memory, available memory 181157888 bytes
1d1h32m: %RPM0-P:RP1 %MEMMGR-2-LOWWATERMARK: ALLOC - size: 32946, Memory requested exceeded low water
mark, Current Usage: 94
00:24:38: %RPM0-P:RP1 %BGP-5-ADJCHANGE: Connection with neighbor 192.93.1.3 closed. Cannot allocate
memory
[...]
```

**Figure 1: RP1 running low on memory**

```
1d1h32m: %RPM0-P:RP1 %SWP-2-IPC SEND FAILURE: SWP:swpSend: queue 72, rtm0 to fibAgt11, failed to send
out an IPC buffer
1d1h32m: %RPM0-P:RP1 %RTM-3-SWP DEQUEUE FAILURE: RTM: rtmMessageSwitch MT_SWP_ACK: Failed at SWP
dequeue
1d1h32m: %RPM0-P:RP1 %SWP-2-IPC SEND FAILURE: SWP: swpDequeue: queue 47, rtm0 to fibAgt0, failed to
send out an IPC buffer
[...]
```

**Figure 2: Out of IPC buffer**

```
1d1h35m: %E48TF:5 %FIBAGT-6-IPFIBT2_RDX_INSERT_FAILED: IPFIBT25, RADIX INSERTION FAILED, LOCATION:
1[...]
```

**Figure 3: RADIX insertion failure**

forwarding hardware (CAM). In a real network, this could lead to traffic being forwarded into a black hole, or in a worse case scenario being sent into a routing loop.

We repeated the same test three times with smaller amounts of background traffic (100 packets/s per port). All test runs showed unstable behavior. At one point, a task crashed on line card eight, and in an attempt to recover, the Force10 switch automatically powered off every line card.

- Route processor module (RPM) to line card communication problems.
- Inconsistent state between route table (RIB), FIB and CAM.
- A task crash on a single line card that triggered the switch to power off ALL line cards.

We never succeeded in forwarding traffic to all 375,000 routes (only 173,681 entries were ever entered in the CAM) and at 100 packet/sec

```
00:27:39: %E48TF:8 %TME-2-TASK SUSPENDED: SUSPENDED - svce:170 - inst:8 - task:fat2Mover
00:27:39: %RPM0-P:CP %CHMGR-2-CARD_DOWN: Major alarm: Line card 8 down - task crash
00:27:39: %RPM0-P:CP %IFMGR-1-DEL_PORT: Removed port: Gi 8/0-47
00:27:39: %RPM0-P:CP %CHMGR-5-CARD_RETRY_RESET: line card 8 power-cycled to attempt bring-up
00:27:46: %RPM0-P:CP %CHMGR-5-CARD_RETRY_RESET: line card 12 power-cycled to attempt bring-up
00:27:53: %RPM0-P:CP %CHMGR-5-CARD_RETRY_RESET: line card 6 power-cycled to attempt bring-up
00:28:01: %RPM0-P:CP %CHMGR-5-CARD_RETRY_RESET: line card 0 power-cycled to attempt bring-up
[? repeated for all line cards present in the system ?]
```

**Figure 4: Task crash**

A single process, initially failing on a single line card caused a "domino effect" in which all line cards in the chassis were power-cycled. In a real network, users would be badly affected by such a failure.

Once the line cards came online again, the system was still unstable, getting 15 out of the 16 BGP peers to work and losing more than 48% of all packets, even at 100 packets/sec.

## EANTC Analysis

With the co-operation of the Tolly Group, we were able to see exactly how Force10 set up the test with both a limited configuration and specially adapted software that allowed up to 25 equal cost paths, reducing the number of routes necessary to 260,000; inside the scalability limits of the FIB/CAM on each line card.

We could not recreate the test configuration exactly, as Force 10's production software allows a maximum of 16 equal cost paths.

The important point is that Force10's Tolly Group test did not scale to 6,000,000 BGP routes as it might first appear; it scaled to 260,000 routes. In our recreation of the tests using 375,000 BGP routes over the maximum 16 equal cost paths supported in FtoS Version 6.2.1.3, we observed the following issues:

- Memory management problems.

forwarding rate, we incurred substantial packet loss.

We found the BGP scalability published in the Force10 Tolly Group test report is not attainable with production hardware and software.

## Test Configuration and Methodology

When we started our tests, we were unaware of exactly how the Tolly Group had configured the test, there not being enough details in the Tolly Group report to allow the test to be accurately reproduced. We therefore proceeded with our best estimation of how the 6-million BGP paths were achieved.

Later, after contacting the Tolly Group, we found the Tolly Group test was limited to two interfaces. We feel that such a two-port test configuration is not representative of any enterprise or service provider network architecture.

EANTC's BGP route scalability test discussed here used a more realistic 32 Gigabit Ethernet port test configuration to provide 16 traffic sources transmitting to the 16 equal cost paths used in the test. We started our tests by advertising the exact same 375,000 non-contiguous eBGP routes per port using an "Internet Mix" of network prefix sizes between /17 and /32 bits in size, thereby simulating a real-world network.

To test that all routes/paths were learned and written to the hardware forwarding engines on each line card, we then transmitted uni-directional 64-byte traffic at 650,000 packets/sec per port as background

traffic and measured the packet loss and average latency. Just in case we misunderstood the Tolly Group's meaning of "background traffic", we also ran the test with each port configured to transmit at 100 packet/sec. Finally, we also noted any problems we came across during the test.

EANTC recorded the time taken to establish a stable state routing state with all BGP routes/paths learned, the time taken to reflect this state in the CAM of a single line card, plus packet loss and average latency. In addition, we recorded any problems we encountered.

# Executive Summary — BGP Routing Scalability Tests

In June 2005, Cisco Systems commissioned the European Advanced Networking Test Center (EANTC) to independently validate the performance, scalability, and availability of Force10 TeraScale E-Series and Cisco Catalyst 6500 switches.

We evaluated two scenarios in these tests.

10. Recreate the Tolly Group 6-Million BGP Path Test and confirm results.
11. Record BGP route (not path) scalability for each switch.

## Tolly Group 6-Million BGP Path Scalability Test

This was a Force10-only test in which we tried to recreate the Tolly Group 6-Million BGP Path scalability test using the latest Force10 production software, Version 6.2.1.3.

We later learned from the Tolly Group that the test used two switch ports and comprised 260,000 BGP routes learned over 25 equal cost paths (BGP peer sessions), thereby creating 6.5 million BGP *paths*. BGP paths should not be confused with BGP routes — there were only 260,000 BGP routes in the Tolly Group tests.

Force10 was able to achieve their Tolly Group results through the use of "special" engineering software customized to allow 25 equal cost paths, while Force10's production software limits the number of paths to 16. When we attempted the test using 375,000 BGP routes learned over 16 equal cost paths:

375,000 x 16 = 6,000,000 BGP paths

EANTC test engineers could not achieve the same zero-loss test result seen by the Tolly Group. In fact, at a traffic rate of 650,000 pps per port, packet loss on the TeraScale E-Series was 75 %. When we dropped the transmit rate per port to100 pps, the E-Series continued to dropped 48 % of all packets.

Force10 and the Tolly Group did not explain to readers exactly how the test was configured, failing to explain the difference between BGP Paths and BGP Routes, plus misinforming the reader that this "special" engineering software would be released as FToS Version 6.1.1.0. The specific software used in this test has not been made publicly available.

## BGP Peer Scalability

In the Tolly Group reports, Force10 claims to support 1,500 BGP Peers. Each peer represents a TCP session that must be established and maintained by the switch under test. The Tolly Group's test was limited to two ports and did not stress the switch very heavily.

We used 24 ports in our test which stepped up the stress on the switch, by forcing the routing protocol to converse on multiple ports simultaneously. EANTC found the Force10 switch could not scale beyond 1,008 peers, above this number the route processor (RP1) went into overload condition and did not recover.

The Cisco Catalyst 6500 scaled to 2,496 BGP peers and established them as twice as fast as the TeraScale E-series. Beyond this number of peers, the Catalyst failed to reliably maintain peer sessions.

## BGP Route Scalability

EANTC believes Force10's Tolly Group test does not give a realistic picture of the TeraScale E-Series BGP protocol scalability. Therefore, we conducted our own BGP route scalability tests, a more realistic scalability metric than BGP paths as tested by the Tolly Group.

Our BGP route scalability test used 32 ports: 16 ports advertising routes, the other 16 ports acting as traffic sources, sending traffic in a unidirectional partial-mesh to all advertised routes.

The Catalyst 6500 scaled to 500,000 BGP routes and passed traffic with zero-loss on all routes. Above this number the RIB/FIB/CAM showed correct numbers of entries (even at 1,000,000 routes) but the traffic experienced packet loss, meaning the test was classed as a fail.

The Force10 TeraScale E-Series could manage 200,000 BGP routes. At 250,000 routes and above we recorded serious Inter-Process Command (IPC) problems; apparently related to congestion on the internal control-plane to line card data path resulting in loss of significant numbers of IPC packets.

The critical synchronization between Force10's software-based Master FIB Table on the RPM module and the Slave FIB Tables resident in the line card memory (CAM) was lost — this could potentially result in incorrect forwarding of traffic, black holes, or in the worst case scenario, routing loops.

### Summary

We found that as with other tests performed by the Tolly Group, Force10 utilized "special" engineering software images and used switch configurations that cannot be replicated using production versions of FToS, even though Force10 stated the beta image used in the Tolly Group tests would be released as Version 6.1.1.0.

Force10's route scalability proved far less capable than the Cisco Catalyst 6500 and we noted that the task of processing routes was noticeably slower on the E-Series.

EANTC's numerous route scalability tests showed some significant scalability problems with the TeraScale E-Series. The most critical of which was the loss of IPC messages and the resulting RIB/FIB/CAM desynchronization that occurred.

The Catalyst 6500 processed BGP peer sessions and routes more quickly and scaled to more than double the number of routes and peers offered by the Force10 switch.

Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test

# BGP Peer Scalability

## Test Objectives

In Force10's Tolly Group Test Report 204148, it is claimed that the Force10 TeraScale E1200 was able to maintain up to 1,500 BGP Peers.

This test uses a realistic 24-port test configuration, with each port emulating a number of BGP peer routers. During the test, the number of BGP routes was kept as constant as possible, only the number of BGP peers was varied to assess maximum scalability.

## EANTC Analysis

Test results show that the BGP implementation on Catalyst 6500 scales much better and proved to be more stable than the BGP implementation on the Force10 TeraScale E1200.

The Catalyst 6500 scaled to 2,496 BGP peers. The Force10 E1200 scaled to 1008 BGP peers.

## Test Results Force10

| Total Number of Peers | All Peers Established? | All Peers Stable? | Traffic Passed? | Time To Converge (min) | Verdict |
|---|---|---|---|---|---|
| 696 | Yes | Yes | Yes | 4:09 | PASS |
| 1,560 | Yes | No | N/A | N/A | FAIL |
| 1,500 | No | No | N/A | N/A | FAIL |
| 1,412 | Yes | No | N/A | N/A | FAIL |
| 1,200 | Yes | No | N/A | N/A | FAIL |
| 1,008 | Yes | Yes | Yes | 5:39 | PASS |

**Figure 1: Force10 BGP Peer Scalability with 200,000 routes**

**Test Highlights**

- ➜ **The TeraScale E-Series scaled to a maximum of 1,008 BGP Peers, not greater than 1,500 as claimed in the Tolly Group Report.**
- ➜ **Beyond 1,008 peers, the Force10 switch saw sustained 100 % CPU utilization on the route processor RP1.**
- ➜ **At 1,200 peers and above, the Force10 E1200 continuously cycled large numbers of BGP peers, first losing, then restoring peer sessions, never achieving a stable state.**
- ➜ **The Catalyst 6500 scaled to 2,496 BGP Peers, 2.5 times Force10's maximum scalability and at this network size remained totally stable.**
- ➜ **The Catalyst also established BGP Peer sessions nearly twice as fast as the Force10 switch.**

## Test Results Cisco

Though the number of BGP peers established was 2,496, Cisco informed us their best practice and design guides do not recommend using such a high number of BGP peers in real networks unless absolutely critical. (Cisco normal guideline: 1000 BGP peers).

The convergence time (time to start BGP peers plus time to populate Forward Information Base/Content Addressable Memory) was calculated by monitoring the start-up of the BGP peers on the device under test.

| Total Number of Peers | All Peers Established? | All Peers Stable? | Traffic Passed? | Time To Converge (min) | Verdict |
|---|---|---|---|---|---|
| 1,008 | Yes | Yes | Yes | 3:29 | PASS |
| 1,560 | Yes | Yes | Yes | 4:23 | PASS |
| 2,496 | Yes | Yes | Yes | 4:00 | PASS |

**Figure 2: Cisco BGP Peer Scalability with 200,000 routes**

## Test Configuration and Methodology

- This test uses a number of pre-defined TRT configurations. Each configuration emulates a different number of eBGP Peers per port. As the test progresses, the number of BGP peers is increased, but the number of routes advertised as a whole remains approximately constant (within the constraints of the test topology) and easily within the route scalability of the DUT.
- TRT advertises non-contiguous route blocks so that the DUT cannot aggregate these into a single RIB/FIB/CAM entry.
- TRT uses a full-mesh traffic pattern between all BGP Peers in the test and exercises every advertised route, thereby confirming the switch is able to not only learn the routes, but also forward traffic on all routes in hardware.
- After setting up BGP Peer Sessions and advertising routes to the DUT, TRT will pause before sending traffic.
- Use the DUT's console to confirm all peer sessions have been established, all routes have been learned and that the Route Table (RIB), Forwarding Information Base (FIB) and the hardware CAMs have been correctly programmed.
- If all peer sessions, routes/prefixes appear correct, start the test traffic.
- Record the number of peers successfully established.
- Record frame-loss and average latency.
- If there is no frame-loss and the average latencies indicate all flows were hardware switched, increase the number of BGP peers and repeat the test.

# Executive Summary — IP Multicast Tests

In June 2005, Cisco Systems commissioned the European Advanced Networking Test Center (EANTC) to independently validate the performance, scalability, and availability of Force10 TeraScale E-Series switches and Cisco Catalyst 6500 switches.

In this category of tests, we evaluated the IP Multicast scalability, stability, and resilience offered by the two solutions.

This is a relatively large test category with many results; this executive overview only discusses the top few findings.

## IP Multicast Scalability, Stability, & Resilience — EANTC Findings

There was a clear difference between the Catalyst 6500 and the Force10 E-Series in this test category.

**PIM Sparse Mode Neighbor Scalability.** The TeraScale E-Series imposes a limitation in IP Multicast scalability. Force10 limits the number of interfaces that can support IP Multicast (PIM-SM) to 32 for the whole switch.

In contrast the Catalyst 6500 scaled to at least 552 PIM neighbors and even went beyond the test equipment's ability to scale the test.

**PIM Sparse Mode S,G Multicast Route Scalability.** When we tested multicast route scalability, the Force10 E-Series showed inconsistent results, each new test topology that had IP multicast as an element showed different levels of scalability. Force10 claims to support a maximum of 15,000 S,G mroutes. In a simple single switch test, we successfully recorded 14,400 S,G mroutes for Force10. In the same test, the Catalyst 6500 exceeded 100,000 mroutes. But these levels of scalability are misleading, as a single switch test does not put the switch under the same stress as it would experience in a real network.

To highlight this fact, and to provide a more realistic picture, we re-ran the multicast route scalability tests in a 3-switch test topology. This new topology forced each switch to communicate with each other, maintaining routing protocol state machines, timers, tables, and coping with the additional stress of PIM Register packets (multicast packets encapsulated inside unicast packets, present during route establishment and very CPU intensive). PIM Register handling is not present in single switch test topologies.

In this topology, the Catalyst supported 49,000 mroutes, establishing them all in just 163 seconds, the E-Series on the other hand could not scale beyond 10,240 mroutes and even with this relatively small number took almost ten minutes (599 seconds) to establish all mroutes.

When we moved onto other tests, where the test topology was a little more complex than the 3-switch topology used above, we found the E-Series could not learn even 3,000 mroutes. See the other multicast test reports for more details.

**PIM Sparse Mode Rendezvous Point Resilience.** PIM Sparse Mode (PIM-SM) is the most widely deployed multicast routing protocol in use today. The protocol provides routing services to dynamically build a multicast tree between the multicast receivers (the leaves) and the multicast source (the trunk of the tree). A critical element in the protocol is a single router that has additional responsibilities to allow multicast receivers and sources to meet at a single point in the network. This switch is called the Rendezvous Point (RP). Unfortunately, the PIM-SM protocol only permits a single RP per multicast group in the whole routing domain. If the RP fails, no new multicast routes can be established for these groups.

To overcome this single point of failure, a combination of supplementary protocols and network design techniques are combined to provide multiple and RPs. RP redundancy is achieved should one of them fail. RP Redundancy is the focus of this test.

The Catalyst 6500 performed as expected and recovered from RP failure in 13–21 seconds.

The Force10 TeraScale E-Series exhibited the following problems.

- We had to reduce the traffic levels in the test when we found that the E-Series couldn't load-share the

multicast traffic over multiple links. All multicast traffic flowed down a single link.

- E-Series exhibited slow multicast route learning and could not scale to the same levels we had recorded in the multicast route scalability tests.
- Force10 never recovered from the RP failure. Even when we reduced the number of multicast routes by 60 %.

When we restored the failed RP to operation, the TeraScale E-Series was not able to restore the network to its original state.

The Force10 failed this critical test of high availability.

**IP Multicast Protocol Interaction with Router Resilience Protocols.** The Catalyst 6500 performed as expected and no adverse interaction between Cisco's IP multicast features and the chosen router resilience protocol was observed.

Force10 also showed no adverse interaction between multicast and VRRP, with the exception that we had to use statically defined IGMP Snooping entries to help the E-Series understand where the PIM-SM multicast routers were attached.

We found that although the Force10 solution worked as expected, in one of the two topologies tested, the Force10 solution saw failover times of 505 seconds compared to the Catalyst's 93 second failover time in the same test.

**IP Multicast Protocol Interaction with 802.1Q VLAN Trunks.** This test is designed to check whether the switch under test correctly implements IGMP Snooping when applied to 802.1Q VLAN trunks. VLAN trunking is used to allow a single physical link to transport the traffic from multiple VLANs and is a critical element in the deployment of "Triple Play", Voice, Video, and Data networks.

The test assesses whether the switch correctly controls the forwarding of multicast traffic at the logical VLAN level across the trunks, or whether it is only capable of controlling the traffic at the physical interface level. If the latter, then users in one VLAN leaving a multicast group, would cause the switch to block the multicast group at the physical trunk level, thereby also terminating the multicast traffic to the other VLANs sharing the trunk.

We tested the Catalyst 6500 in a multi-switch environment and ran through a series of IGMP Joins/Leaves to test the switch, the Catalyst performed as expected, controlling the traffic at the logical VLAN level.

Next we tested the TeraScale E-Series. Running through the same series of IGMP Joins/Leaves we experienced the total collapse of all multicast groups. The collapse even affected ports on other switches and terminated traffic not only on the VLAN where the IGMP Joins/Leaves were executed, but also on all VLANs. After a short delay, multicast traffic to all multicast groups was unexpectedly terminated. Our attempts at sending new IGMP Joins to try and re-establish the groups failed.

**Summary.** In EANTC's opinion, the Force10 TeraScale E-Series does not provide the level of scalability or resilience needed to support today's mission critical multicast networks.

Multicast scalability is limited by Force10's decision to limit the number of IP multicast interfaces to 32 for both routed interfaces and virtual router interfaces servicing VLANs. Even if this could be overcome, the lack of line card memory (CAM) dedicated to IP multicast would impose its own limitations as detailed in our test reports.

In addition, the inability to load-share multicast traffic across multiple inter-switch links, means that end-to-end bandwidth is restricted.

The fact that the Force10 solution couldn't provide the levels of high availability expected of today's networks and failed to recover in response to an RP failure also gives cause for concern.

The fact the Catalyst 6500 scaled better, learned faster, recovered from RP failure in just 13-21 seconds (even with default settings) means it is far more capable in supporting IP multicast than its Force10 counterpart.

We also found that the Catalyst 6500 supports other important multicast related features such as IGMPv3, Source-Specific Multicast, and Bidirectional PIM.

**Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test**

# Multicast PIM Scalability Test

## Test Objectives

This test explores the Protocol Independent Multicast Sparse Mode (PIM-SM) neighbor scalability offered by the Force10 E1200 and the Cisco Catalyst 6500.

Since guidelines and limitations are not discussed in either vendors' software documentation, the tests focused on PIM-SM neighbor requirements as experienced by EANTC in previous, real world, service provider tests.

Our tests employed the 48-port RJ-45 Gigabit Ethernet cards from each vendor. With these cards, the Force10 E1200 can scale to 672 Gigabit ports, while the Catalyst 6500 scales to 384.

In order to be able to scale the number of PIM-SM neighbors beyond the physical port limitation of our test setup, and to provide a configuration closer to a real world "triple play" voice, video, and data application, we tested both routed ports and virtual router ports (VLAN Interfaces) serving VLANs equipped with 802.1Q trunks for added scalability.

### Test Highlights

- ➜ **Out of 672 physical interfaces on the Force10 TeraScale E1200, EANTC found that 31 ports could be configured for PIM-SM.**
- ➜ **When virtual router ports (VLAN interfaces) are tested, the same restriction applies. Out of 4,096 VLANs supported, 31 VLAN interfaces can be configured for PIM-SM.**
- ➜ **The PIM-SM limitation appeared to be hardwired into the system. In both test configurations, once this limit was exceeded, error messages indicated that the limit for PIM-SM interfaces had been reached.**
- ➜ **The Catalyst 6500 was limited by the test equipment hardware, which could not allow greater than 552 PIM neighbors on a single port. Even with the limit of 552 PIM-SM neighbors, the Catalyst still supported more than 17 times the number of PIM neighbors supported by Force10.**

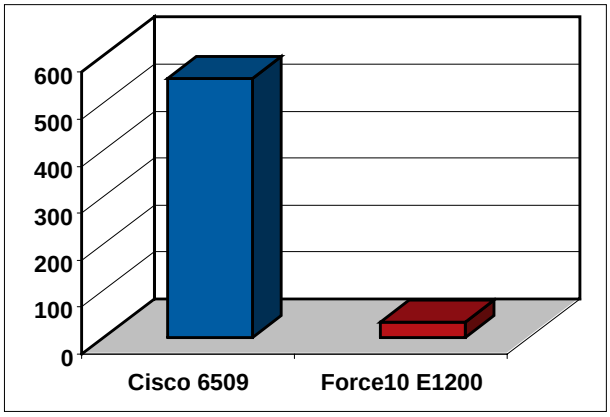



**Figure 1: Number of PIM-SM Neighbors**

## Test Results Force10

The Force10 switch exhibited scalability issues, which seem to be directly related to its architecture. It appears to EANTC that Force10 is imposing a limit on the number of interfaces that can be configured to support IP multicast. When we attempted to configure the 32nd interface in the test with PIM-SM, we got an error message on the console indicating the maximum number of PIM interfaces had been reached (see *Figure 2: Force10's 31 Interface PIM-SM Limit With Routed Interfaces*).

```
E1200(A)(conf)#interface range g0/0 - 11
E1200(A)(conf-if-range-gi-0/0-11)#ip pim sparse-mode
E1200(A)(conf-if-range-gi-0/0-11)#interface range g1/0 - 11
E1200(A)(conf-if-range-gi-1/0-11)#ip pim sparse-mode
E1200(A)(conf-if-range-gi-1/0-11)#interface range g2/0 - 11
E1200(A)(conf-if-range-gi-2/0-11)#ip pim sparse-mode
% Error: Cannot configure pim sparse mode, maximum pim interface limit reached.
% Error: Cannot configure pim sparse mode, maximum pim interface limit reached.
[?]
```

**Figure 2: Force10's 31 Interface PIM-SM Limit With Routed Interfaces**

```
E1200(A)(conf)#interface range vlan 2 - 49
E1200(A)(conf-if-range-vl-2-49)#ip pim sparse-mode
% Error: Cannot configure pim sparse mode, maximum pim interface limit reached.
% Error: Cannot configure pim sparse mode, maximum pim interface limit reached.
[?]
```

**Figure 3: Force10's 31 interface PIM-SM limit with VLAN interfaces**

## Cisco Results

The Catalyst exhibited greater PIM-SM neighbor scalability, and was limited by the test equipment's ability to emulate more than 552 PIM neighbors on a single multicast source port.

## EANTC Analysis

The PIM-SM implementations of the Cisco Catalyst 6500 and the Force10 E1200 showed different behavior and capabilities.

Force10's operating system (FToS) notified us that it could not configure further PIM-SM interfaces when the 32nd interface was reached. This limitation could affect anybody considering the E1200 for its very high port density.

The Cisco Catalyst 6500 was able to outperform the test equipment used for the test. In all, 552 PIM-SM interfaces were configured and neighbor relationships established without any problems. The Catalyst also had no problems in routing traffic to all its PIM-SM neighbors without loss.

## Test Configuration and Methodology

Two test configurations existed: Routed ports and virtual routed ports (VLAN interfaces).

**Routed Ports Test.** In the routed ports configuration, 12 ports were configured on each of four line cards, making a total of 48 physical router ports in the test. On each of these ports, the SmartBits was configured to emulate a directly attached PIM-SM neighbor.

**Virtual Router Ports (VLAN Interfaces).** 577 VLANs were configured on the device under test (DUT) and a virtual router port (VLAN Interface) configured to serve each VLAN. The VLANs comprised 47 tagged member ports, all configured as 802.1Q trunks. Each VLAN interface was assigned a unique IP address with OSPF as the unicast routing protocol.

One interface on the DUT was directly connected to the multicast source emulated by SmartBits. This port was configured as a routed port rather than a logical VLAN Interface.

We started by configuring unicast on the DUT as described above. Next we configured the multicast routing protocol PIM Sparse Mode (PIM-SM) and enabled PIM-SM on each interface in turn.

After all ports were configured, PIM-SM neighbor relationships were established and traffic was passed to all neighbors.

The tests recorded correct IP Multicast state, packet loss and multicast latency.

Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test

# PIM-Sparse Mode Multicast Route Scalability Test

## Test Objectives

This test assesses the number of S,G (Source, Group) multicast routing table entries the device under test (DUT) can support. Multicast routing tables are constructed from source and group pairs, each of which has a number of outgoing interfaces. These are held in a table called the Outgoing Interface List (OIL) describing the downstream interfaces through which the multicast traffic should be forwarded.

In busy multicast networks, such as financial trading rooms or various military simulations, every host can serve both as a source and a subscriber on numerous multicast groups.

At the control-plane level of the switch, IP multicast routing is a processor intensive activity requiring each route table entry to maintain a number of timers, finite state machines, and software tables. In addition the switch must write multicast forwarding entries in hardware on each line card and must maintain synchronization between software and hardware tables. In general, all these elements are referred to as "multicast state."

In large multicast networks described above, the Rendezvous Point (RP) plays a critical role, and often needs to maintain "multicast state" for very large numbers of sources and groups.

The objective of this test case is to explore the limitations of the multicast S,G route scalability in the DUT, both at the control plane and data-plane levels. To verify the data-plane hardware has been correctly programmed, we verify correct operation of all routes by sending traffic from multicast sources to all groups and onward to all receivers.

## Test Results

Force10's release notes for software version 6.2.1.3 document a maximum of 15,000 multicast entries supported by FToS.

**Single Switch Topology Results.** The Force10 switch was able to scale to 14,440 S,G mroutes almost reaching the 15,000 multicast entries documented in the release notes. In the same single switch topology, the Cisco Catalyst 6500 scaled to 100,000 S,G mroutes.

This single switch topology is unrealistic however, and doesn't place the switch under the same amount of stress as a real network.

**Test Highlights**

- ➜ **In a realistic 3-switch network topology, where the DUT acted as the PIM-SM Rendezvous Point, Force10's TeraScale was unable to scale above 10,240 S,G multicast entries.**
- ➜ **Cisco's Catalyst 6500 reached 49,000 S,G multicast routes in the same 3-switch network topology.**
- ➜ **Force10's TeraScale E1200 exhibited a slow multicast route-processing rate, requiring ten minutes to create all 10,240 S,G mroutes. For the same number of entries the Catalyst 6500 required only two minutes and 43 seconds, more than three times faster than Force10.**

## Why use a multi-switch topology?

Single switch multicast tests are not stressful because:

- The switch is its own Rendezvous Point (RP) and is statically defined.
- The switch doesn't have to communicate and process unicast and multicast routing protocols, it only needs to populate IGMP Joins into the outgoing interface list for each mroute.
- The switch doesn't have to process PIM Register packets. These are sent during the setup stage of an S,G mroute between RP and multicast source. PIM Register packets encapsulate multicast frames inside unicast frames sent to the RP until a multicast route

exists. The RP must process these PIM Registers, create the S,G mroute entry, set up the various timers, de-capsulate the multicast and forward it down the *,G shared tree, as well as sending Register Stop messages to the first hop router and PIM JOINs down the shortest path to the source. Finally it must write the correct information needed by the hardware forwarding engines.

### Multi-switch Test Results:

To create a more realistic test topology we used three switches (see test topology diagram).

In this topology, Force10's solution was able to scale to 10,240 S,G mroutes, taking 599 seconds to populate its tables. When offered 11,560 mroutes, the network failed to maintain correct multicast state end-to-end. The Cisco Catalyst 6500 was able to scale to the same number of mroutes in 163 seconds, over 3.5 times faster than the Force10.

The maximum number of S,G mroutes reached by the Catalyst was 49,000.

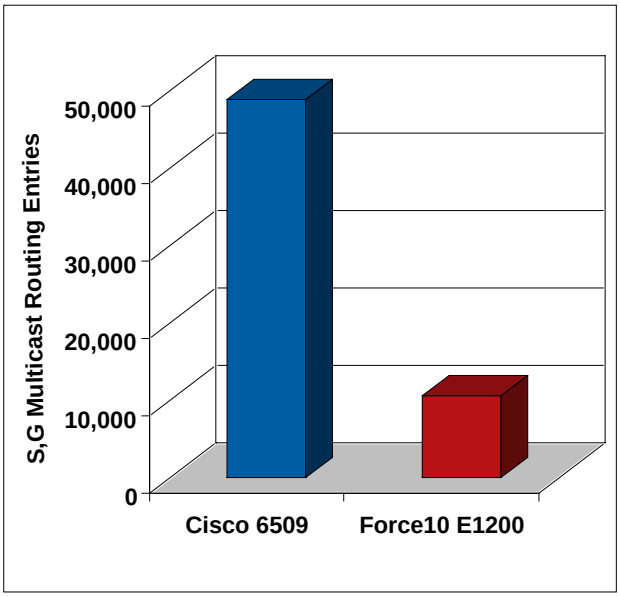



**Figure 1: Multi-Switch Test Results**

### Test Observations:

On the Force10 switch, we received error messages from the CLI stating that the IPv4Flow CAM partition was full. Force10 FTOS release 6.2.1.3 allows for a sub-partitioning of this CAM partition at the expense of other features sharing the same CAM-space (QoS, System-Flows, Policy-Based Routing and Trace-Lists), but contrary to Force10's claims, the rest of the CAM partitions are of fixed size.

We increased the multicast area of this sub-partition to the maximum possible size available to multicasts (17K entries) and observed that the messages disappeared; however, this did not improve the mroute scalability of the switch, it only prevented the error messages from being displayed. All tests were conducted with this large CAM partition.

Average end-to-end switching latency for the Force10 topology never stabilized. Even when all multicast routes entries were confirmed to exist in the hardware and software multicast routing tables, the network latency continued to oscillate between 138-150 µSec.

In contrast, the Catalyst solution displayed a stable switching latency of 57-60 µSec once all multicast routes were populated.

### EANTC Analysis

The multicast routing scalability test exposed two problems on the Force10 switches that have a impact on real world network design.

12. Only 31 out of the possible 1,260 interfaces on the Force10 switch can be configured to run multicast routing.
13. The limited size of the multicast routing table and Content Addressable Memory (CAM) on each line card constrains multicast scalability.

The multicast scalability limitations we discovered in these tests do not appear in Force10's marketing literature or data sheets; they are only documented in release notes.

The Catalyst scaled three and a half times better than the Force10 and remained stable throughout the whole test.

## Test Configuration and Methodology

Two test topologies were used.

The first is a single DUT test with a number of emulated PIM-SM neighbors and multicast hosts attached. The DUT serves as the PIM RP while a SmartBits tester is emulating the peers and hosts.

The second test employs a topology closer to a real world deployment in which three switches participate. The DUT is the center switch, acting as the Rendezvous Point (RP) in the simple 3-switch network. The first switch functions as a "first hop router" directly attached to multicast sources emulated by SmartBits.

The second, center switch is the device under test (DUT) that acts as the RP; while the third switch is a "last hop router" serving downstream IGMP hosts emulated by SmartBits.

The three switches topology is illustrated below. Note that Switches A and C were used for facilitate the topology; however the device under test (DUT) was switch B that was serving as the RP.

NOTE: All multicast routing scalability tests were conducted with the maximum CAM size configuration to solicit the best result from the Force10 switch.

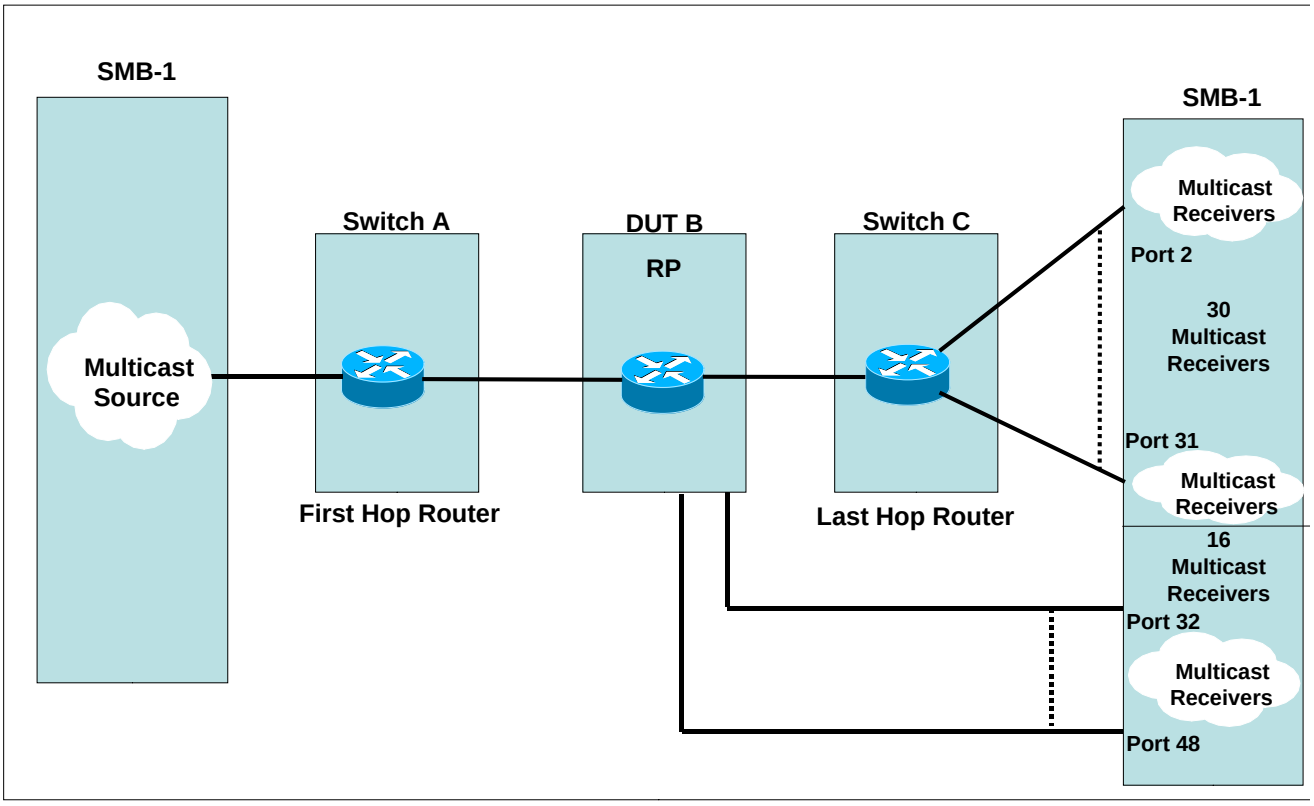



**Figure 2: Three-Switches-Topology**

Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test

# Multicast PIM-SM Rendezvous Point Failover Test

## Test Objectives

The PIM Sparse Mode (PIM-SM) Rendezvous Point (RP) Failover Test focuses on multicast resiliency. Every PIM-SM network must be configured with at least one RP. The RP acts as a meeting point for multicast sources and receivers; which by default do not know each other's location. Unfortunately, the RP also represents a single point of failure. If the RP fails, no new multicast trees can be established.

The IETF has defined various mechanisms and protocols to facilitate RP resilience and fast recovery. Since there can be only one RP for any single group in a PIM domain, a network wide algorithm is used to divide the multicast groups between the active RPs.

RP resiliency is facilitated by the use of the following:

- Anycast-RP provides the mechanism to facilitate fast convergence in a PIM domain where two RPs are provided for resilience. The failure of one of the RPs causes the multicast groups to be re-established via the alternate RP once the underlying unicast routing protocol has converged.
- MSDP (Multicast Source Discovery Protocol) provides rapid RP failover and recovery, among other tasks. It provides the mechanism where each RP learns of the multicast sources being serviced by its partner RP. After an RP fails, this aids the surviving RP to rapidly establish its multicast state and forward the affected multicast groups.

The objective of this test was to fully exercise the above protocols in a realistic network topology, measuring the time taken for the network to restore multicast forwarding to the multicast groups affected by the RP failure.

**Test Highlights**

- ➜ **RP Failover: EANTC's tests show that Force10 failed to recover from an RP failure. Cisco recovered all affected multicast groups within 13-21 seconds.**
- ➜ **Anycast-RP: Although Force10 claims to support Anycast-RP, EANTC found that Force10's implementation results in no load-sharing of multicast groups and no recovery. In contrast, Cisco's Catalyst 6500 correctly implements the Anycast-RP mechanism and shared multicast groups evenly between both RPs.**
- ➜ **Link-Aggregation: Force10's TeraScale E-Series cannot load-balance multicast traffic, using only one out of four available links. Force10 effectively reduce inter-switch multicast bandwidth by 75 %. Cisco's Catalyst 6500 evenly distributed multicast traffic across all links in the link-aggregation groups, making full use of all available bandwidth.**
- ➜ **Line Card Reset: During this test, one of the Force10 switches experienced a line card reset; this appeared to be triggered by a L3 Content Addressable Memory (CAM) overload condition, even though the total number of multicast routes were within the scalability limits of the Force10 switch.**

## Test Results Force10

**Inconsistent Performance.** We configured the Force10 switch to use the maximum amount of CAM possible for IP multicast, a total of 17,000 entries. We ensured the 9,000 S,G multicast routes used in the test were within the multicast route scalability demonstrated by the Force10 E-Series in earlier EANTC multicast route scalability tests where it learned 10,240 S,G mroutes.

Once again, the Force10 switch's performance proved inconsistent, failing to recover all 9,000 mroutes. Less than one third of the mroutes converged after failover. We had to reduce the number of multicast routes to just 2,890 (170 groups, with 17 sources per group) in an attempt to achieve network recovery. Even after such reduction in the number of multicast routes in the test, the Force10 switches never converged after failover.

**Slow Multicast Convergence.** As in previous tests, we observed that Force10 required a considerable time to establish all multicast routes compared to the Cisco Catalyst.

**Anycast-RP Problems.** Force10's marketing materials and configuration guide led us to believe the TeraScale E-Series switches fully supported Anycast RP, a vital component in multicast resilience.

We were able to configure both DUTs as candidate RPs; however, Force10's implementation of Anycast RP did not function correctly and no load-sharing of multicast groups between the two RPs was possible. Using the CLI we observed that DUT (B) was hosting all multicast groups in the network, while DUT(C) was effectively idle, serving zero groups.

**RP Failover Never Recovers.** We simulated the failure of DUT(B), forcing an RP failover to DUT(C). After waiting more than 60 minutes, the Force10 solution still did not converge its multicast routing tables and fully restore service.

We restored DUT(B) to full operation and expected the network to return to its original state. We found that DUT(B) was unable to converge its multicast routing table, even though it had been the active RP and had supported this number of mroutes prior to the RP failover.

**CAM Overloads.** Several times during the execution of these tests we received messages from Switch(D) complaining about lack of CAM space; these were displayed even though the total number of multicast routes was within the capabilities of the DUT. We can assume that the Force10 switch makes inefficient use of its available CAM, something we had noticed in the ACL and OSPF scalability tests.

### Test Results Cisco

The Cisco Catalyst tests resulted in a very different experience. We were able to configure Anycast RP and record multicast load sharing between the two active RPs. The test was repeated three times with the full 9,000 S,G mroutes, and failover times were recorded.

Recovery from failover varied between 13 to 21 seconds.

Restoring the original network state (when both RPs are active) took 57 to 61 seconds.

Both multicast traffic and unicast traffic were evenly load-balanced across all members of the link-aggregation groups used in the test.

When the Cisco solution was tested using the same 2,890 mroutes as the Force10, it converged and restored service to all multicast groups in just 8.74 seconds.

### EANTC Analysis

**Cisco.** The Catalyst's ability to correctly support Anycast-RP and MSDP as well as perform load-balancing of multicast traffic makes it well positioned for mission critical multicast topologies. The multicast section of the tests proved Cisco's major strengths in this area.

In our tests with 9,000 mroutes, the RP failover took on average 16 seconds and worked consistently on each test run.

**Force10.** In all the failover tests, even with a greatly reduced number of mroutes, the Force10 switches consistently failed to recover from an RP failure. Multicast routes were still cycling in and out of the multicast route table 60 minutes after the failover.

The E-Series only used one quarter of the available link bandwidth created by link-aggregation protocol for multicast traffic, severely restricting multicast scalability.

EANTC tests proved that the Force10 E-Series needs to improve in these areas to be able to handle mission critical multicast applications requiring high availability.

## Test Configuration and Methodology

The test employed four switches to create a resilient multicast topology, employing dual RPs for redundancy. RP (A) and RP (B) were configured for Anycast RP and also MSDP to ensure the fastest possible recovery from RP failure.

All switches were inter-connected using 4-port link-aggregation groups, and we expected the multicast traffic to be load-shared across all available links, thereby providing maximum inter-switch capacity for multicasts. Unfortunately, Force10 could not load-share multicast traffic, so we were forced to reduce the unicast rate to just 10%, so that when 80% multicast load was added to the single link used by Force10 it would not be oversubscribed. Of course, this level of traffic engineering would not be possible in a real network, so both unicast and multicast loss would be inevitable.

Unicast traffic comprised 4,800 OSPF routes with bidirectional 64-byte traffic at 10% load.

Multicasts used 300 groups and 30 sources per group to generate a total of 9,000 S,G mroutes. A single multicast source port generated multicast traffic with 512-byte frames at 80% load; this was forwarded through the network and distributed to eight multicast receivers on DUT (D).

(Note: We had to reduce the number of multicast routes to 2,890 to accommodate Force10)

After establishing all unicast routes and confirming they had been correctly learned, we started the unicast traffic. Next we established all multicast routes, confirming that all mroutes were learned in each switch before starting the multicast traffic.

We then conducted a baseline test to confirm there was no packet loss under normal network conditions.

After we were confident that the network topology was working as expected, we simulated a failure of DUT (B) and used the real-time graphs available on SmartBits' TeraRouting as well as the DUTs' consoles to confirm the recovery of all mroutes.

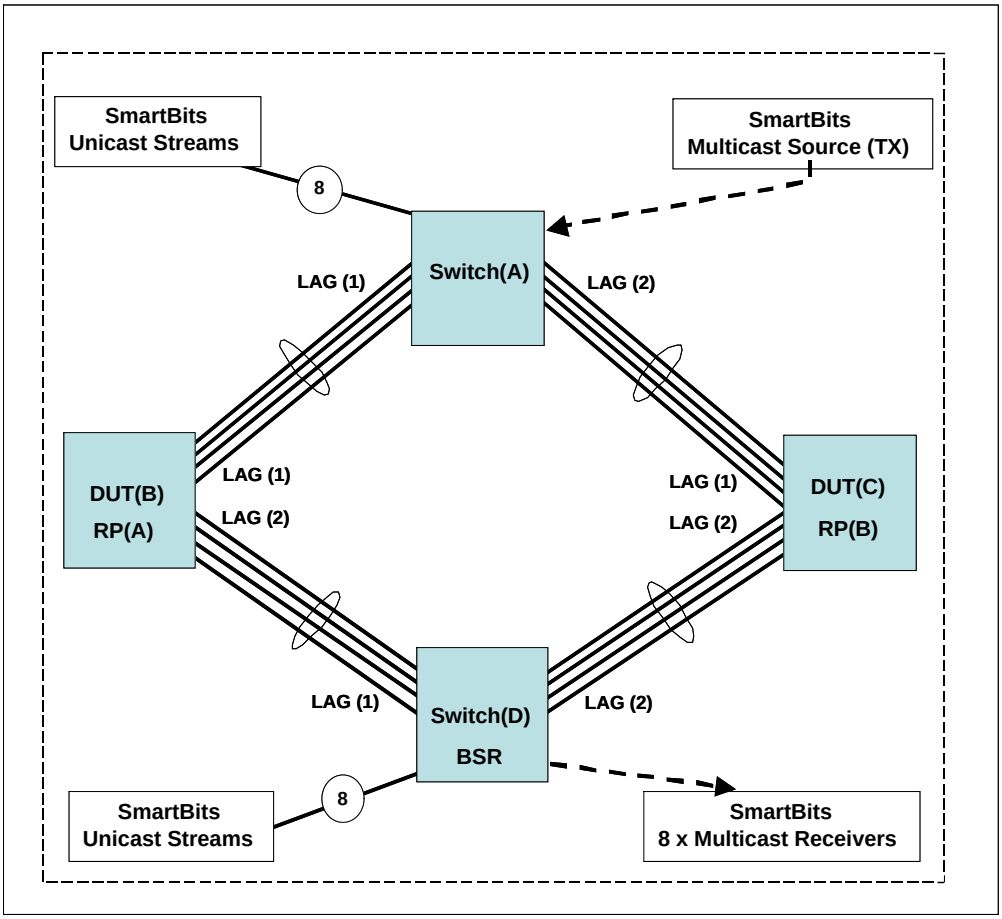



**Figure 1: Test Configuration**

**Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test**

# IGMP JOIN/LEAVE over 802.1Q VLAN Trunks Test

## Test Objectives

Today's "Triple Play", Voice, Video, and Data networks often share a common cabling infrastructure, with all three services transported over the same physical links. Triple play networks pose their own unique design challenges. For instance, to make IP phones plug-and-play and accommodate their adds, moves and changes in as simple a manner as possible, most vendors recommend placing the IP phones in their own unique campus-wide VLAN. To support this topology, all switches in the network must use 802.1Q trunks to maintain VLAN isolation between voice and other services.

IP telephony is regarded as a mission critical service that must be available 24/7, so to meet this requirement, networks are designed to be highly resilient and offer alternate paths around link or node failures. This adds a further requirement to the network design, alternate paths must be available, but Layer 2 loops must be avoided, as these can lead to broadcast storms and network meltdown.

The objective of this test is to verify the correct operation of IGMP in the presence of 802.1Q trunks and spanning tree. Each switch in a network supporting 802.1Q trunks should control multicast traffic at the logical VLAN level and not at the physical trunk level. For instance, if an 802.1Q trunk serves three VLANs and all the users in one VLAN leave the multicast group, other group subscribers in other VLANs should not be affected.

This test investigated two areas:

14. **IGMP/Spanning-Tree Interaction**: Often protocols such as spanning tree and IGMP are tested in isolation, with little regard for the fact that they are often deployed in unison. These tests take an holistic approach and test these features in the way they would be deployed in a real network.
15. **IGMP/802.1Q Trunk Interaction**: We test the correct operation of IGMPv2 (Internet Group Management Protocol Version 2) when applied to 802.1Q trunks. IGMPv2 is crucial for the reception of multicast traffic to end users and in the maintenance of multicast streams.

**Test Highlights**

**The Force10 switches did not pass this test:**

- ➜ **The simple action of users "leaving" a multicast group in one VLAN resulted in all other subscribers within the same VLAN having their multicast traffic terminated even if they were connected to other switches.**
- ➜ **After a short time, the "leaves" in one VLAN triggered the failure of all multicast traffic to all VLANs.**
- ➜ **EANTC confirmed the Force10 solution resulted in total collapse of all multicast streams to all virtual LANs.**
- ➜ **This failure occurred regardless of the spanning tree protocol used.**

**In the same test, the Cisco Catalyst 6500 solution worked as expected, even when subjected to multiple IGMP JOINs/LEAVEs.**

These tests use a 3-switch loop topology and "standards-based" loop avoidance protocols. We tested multicast control while also running standard IEEE 802.1d spanning-tree or the more modern alternative IEEE 802.1w Rapid Spanning Tree (RSTP).

## Test Results

### Force10 Results:

Please refer to the topology diagram. Regardless of which Layer 2 loop-avoidance protocol was used (STP or RSTP), Force10's solution displayed multicast control problems.

The simple operation of leaving a multicast group by hosts in VLAN 2 attached to Switch(B), caused hosts

connected to the same VLAN, even on Switch(A) and Switch(C), to lose their multicast streams.

At this stage, all multicasts ceased in VLAN 2, but continued to be delivered to the other two VLANs. After a short while however, the traffic from ALL multicast groups to all VLANs failed. In effect we observed a complete collapse of the multicast network, triggered by the simple actions of users in one VLAN.

When we attempted to recover the failed condition by sending further IGMP JOINs from hosts in VLAN 2 on DUT(B), we found that the operation failed to re-establish group membership. In fact, in the whole network, no further multicast group memberships could be established.

The test was repeated twice to confirm the results were reproducible.

**Cisco Results:**

The Cisco solution worked as expected. Hosts attached to the Catalyst switches were able to leave and join multicast groups in one VLAN without affecting other hosts in the same VLAN or hosts in other VLANs still requiring the multicast group.

## EANTC Analysis

The correct interaction between Layer 2 loop-avoidance protocols and Layer 3 multicast traffic is an important aspect of enterprise networks. We followed a common network design in which the DUT is only required to apply loop avoidance protocols, distinguish between physical and logical ports and to correctly implement IGMP snooping.

The operation of leaving a multicast group on a single VLAN should not, in any way, affect other multicast subscribers. In addition to the results from the previous multicast tests, these join/leave test results indicate problems with Force10's multicast implementation.

The Cisco Catalyst 6500 performed as expected.

## Test Configuration and Methodology

The receiver ports were configured as three IGMP hosts (one per port) and a monitor port thus making four ports together.

Three switches were used for the test and were connected in a loop topology, providing an alternate path for traffic in case of link or node failure. Three VLANs were then configured on each switch, and four untagged member ports were configured in each VLAN on DUTs (B) and (C). DUTs (B) and (C) were then linked to DUT(A) using 802.1Q trunks.

The multicast traffic source was connected to a routed port on DUT(A), so that we also exercised multicast routing, as well as IGMP Snooping capabilities.

Multicast clients were emulated using SmartBits on ports attached to DUT(B) and (C), while SmartBits emulated a multicast source attached to DUT(A).

The test was repeated three times; once with 802.1d Spanning Tree Protocol, and twice with 802.1w Rapid Spanning Tree. The configuration for both protocols was similar, with the Root Bridge being DUT(A) and the direct link between DUT(B) and DUT(C) blocked by the loop avoidance protocol.

SmartBits' TeraRouting test application was used to send IGMP JOINs and LEAVEs to multicast groups in VLAN 2, while the same software was used to verify the continuous reception of multicast streams to the other VLANs.

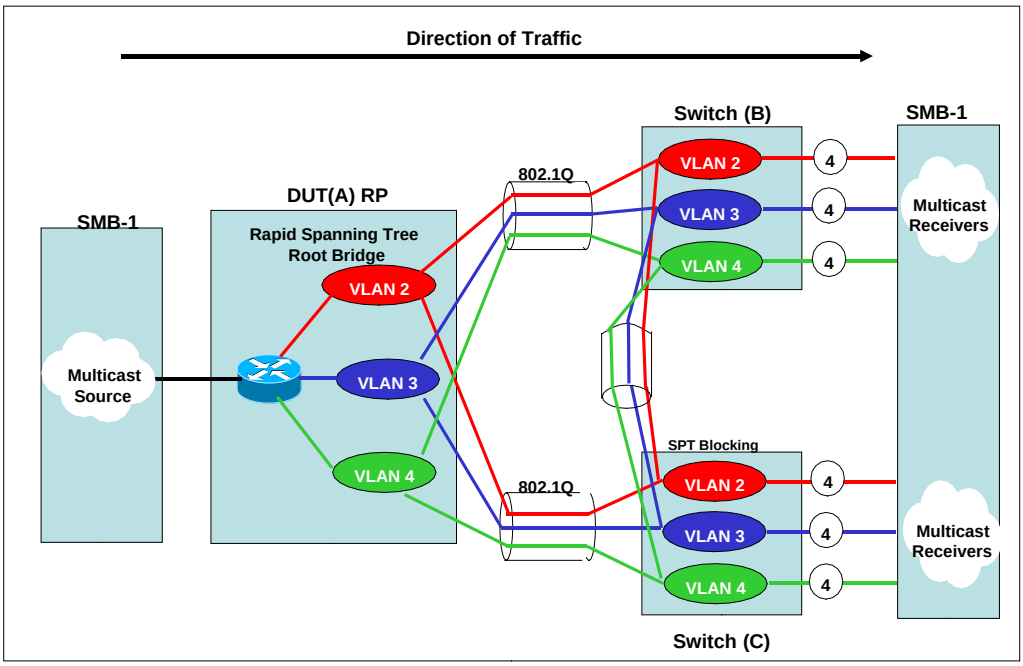



**Figure 1: Test Configuration**

Force10 TeraScale E-Series — Cisco Catalyst 6500 Competitive Test

# IP Multicast and Router Resiliency Test

## Test Objectives

In designing high redundancy networks, it is desirable to deploy a router resilience protocol. Three "Router Resilience" protocols exist for this purpose; standards-based VRRP, plus HSRP, and GLBP, which are unique to Cisco Systems. (Force10 only supports VRRP.)

VRRP and HSRP allow a subnet to be served by two routers; a master router (active forwarder), and a backup router that only becomes active when the master router fails. Using the above protocols, only an active forwarder is allowed to forward traffic from the protected VLAN/Subnet. The backup router listens to hello messages from the active forwarder and waits to take over that role should the active forwarder fail.

To maintain the master/backup state, the protocols use Hello Messages. If the backup router fails to see hello messages from the master router, it assumes the virtual MAC address of the master router and becomes the active forwarder.

Only one of the two available routers is actively forwarding at any one time and this means the two uplinks to/from a wiring closet cannot both be used; the one to the backup router remains inactive. Both VRRP and HSRP employ an active/standby paradigm. Cisco's GLBP (Gateway Load-Balancing Protocol) on the other hand, allows both routers to be active forwarders simultaneously.

To load-balance the hosts between each router, a master GLBP router is elected and is responsible for responding to all ARP requests. It alternately replies to the ARPs with either its own virtual MAC address, or that of the other router. This means that 50 % of the hosts in the protected subnet will use one router, while the remainder will use the other. Should either router fail, only 50 % of the active flows are affected by the failover.

Often, tests are conducted on each protocol in isolation. Testing router resilience as one test and multicast as another. In real networks, both these features are active simultaneously and must not interfere with one another.

### Test Highlights

- **Force10's E1200 switch did not load-balance multicast traffic across Link Aggregation Groups, making use of only one out of four links. The Catalyst 6500 load-shared the multicasts across all links.**
- **When configured as the VRRP Master, and PIM-SM Designated Router (DR), Force10's E1200 required over 8 minutes to recover from the failure of the master router.**
- **Unlike the Cisco Catalyst 6500, Force10's wiring closet switch failed to learn the location of the multicast routers automatically. Manual configuration of IGMP Snooping "mrouter location" was required for correct operation.**
- **Upon restoring the original network state, the Force10 switch caused further loss of multicast traffic streams over more than 60 seconds while the Cisco Catalyst switch introduced negligible disruption and continued in correct operation after only five seconds.**

The objective of this test is to verify the correct interaction of a router resilience protocol and IP Multicast and to measure failover and recovery times.

## Test Results

The tests were run in two configurations. In the first, the master router and the multicast designated router were configured on the same switch. In the second configuration, they were configured on separate switches.

| Test Topology | Cisco Failover Time[a] | Force10 Failover Time | Cisco Restoration Time[b] | Force10 Restoration Time |
|---|---|---|---|---|
| **Master Router = DUT(B)[c] Multicast DR = DUT(B)** | U = 12.44<br>M = 93.74 | U = 11.88<br>**M = 505.37** | U = 20.23<br>M = 4.24[d] | U = 11.64<br>**M = 66.95** |
| **Master Router = DUT(B) Multicast DR = DUT(C)** | U = 10.71<br>M = 8.71 | U = 12.61<br>M = 10.51 | U = 19.47<br>M = NI | U = 9.54<br>M = NI |

a. Failover time is the time taken for all multicast traffic to be forwarded via alternate route
b. Restoration time is the time taken to restore the original network state when DUT(B) recovers
c. DUT(B) is failed in each test run, DUT(C) is never failed
d. Please see Figure 1 for an explanation of this result
U = Unicast failover or restoration time
M = Multicast failover or restoration time
NI = No discernible interruption to multicast traffic streams

**Force10 Results:**

Force10 showed mixed performance in these tests. In the test case where the VRRP Master and PIM-SM Designated Router were hosted on different switches, the Force10 switch delivered similar results to the Cisco solution.

The problems came when the VRRP Master and PIM-SM Designated Router were hosted on the same switch, where failure of the switch requires both unicast and multicast traffic to failover. Under these conditions, both switches were affected by the new topology; however, the Force10 solution delivered high failover times.

In addition to the performance above, we noted several problems, initially found in previous tests. Multicast traffic was not load-balanced across the links in link aggregation groups and under certain conditions even unicast traffic was not load-balanced across all links in each aggregation group. This behavior was also seen in Force10 bandwidth aggregation tests previously conducted by EANTC.

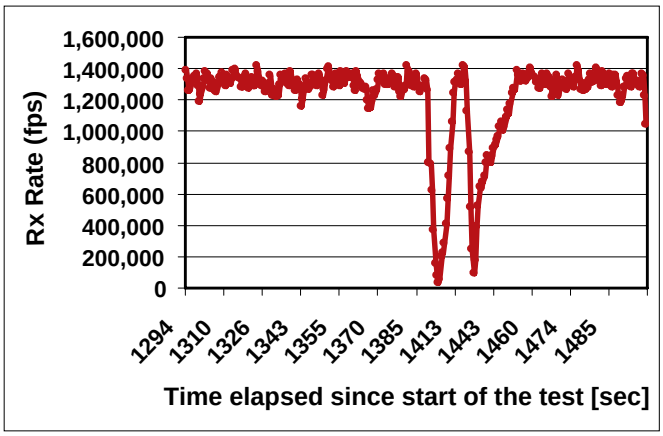



**Figure 2: Force10 Results**

In the test case where the multicast DR and VRRP master were configured on the same switch, the Force10 solution took 67 seconds to restore to the original state.

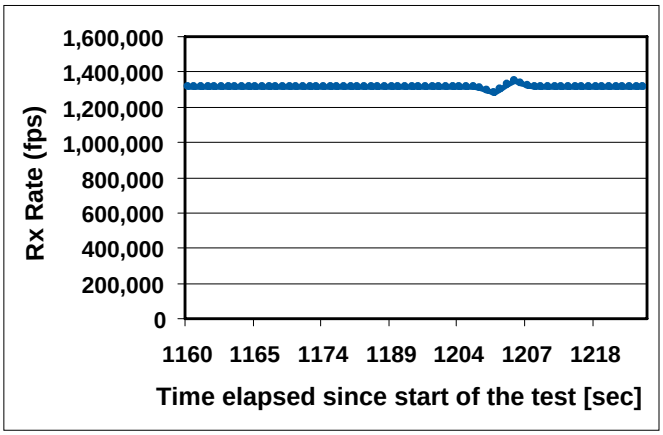



**Figure 1: Cisco Results**

**Cisco Results:**

In the test case where the multicast DR and the GLBP master were configured on the same switch, we saw a multicast recovery time of 93 seconds. Since this test case involved the failure of an element crucial to both unicast and multicast topologies, an increased recovery time was to be expected. Once restoration was complete, the network returned to normal state.

In the test configuration where the GLBP master and the multicast DR are hosted on different switches, we saw no restoration time required for multicast traffic.

## EANTC Analysis

While VRRP and GLBP both provided similar unicast failover and restoration results for both vendors, we were impressed by Cisco's GLBP load-balancing capabilities and the fact that only 50% of the unicast flows were affected by the node failure.

The length of time taken to restore multicast states in the Force10 solution is a cause of concern as is the TeraScale E-Series' inconsistent results, with multicast failover times ranging from 10 seconds to 505 seconds even though the same number of S,G mroutes were present in each test.

Although the Cisco solution also saw an increase in multicast failover times with the second test topology, going from eight seconds to 93 seconds, this was significantly less than Force10.

## Test Configuration and Methodology

Four switches were used for the test topology. Three of them were configured as multicast routers. DUT(D) was configured as a simple L2 switch with a single VLAN and as such was not regarded as an active element in the test.

Multicast and unicast sources were attached using a SmartBits to DUT(A) and received on DUT(D). Multicast Rendezvous Point (RP) was also configured on DUT(A).

DUTs (B) and (C) were configured with the optimal router redundancy protocols offered by each vendor. Force10 testing was conducted using VRRP while Cisco testing was done using Cisco's GLBP.

All links in the network were configured as Link Aggregation Groups (Port-Channels), each creating a logical 4-Gbit/s links between switches. Two alternative data paths were created between DUT(A) and DUT(D).

Unicast and multicast traffic were partitioned in such a way that Port-Channels were never over-subscribed, having discovered in previous tests that Force10's Link Aggregation implementation does not support multicast load sharing. We therefore kept the combined unicast and multicast traffic load below the capacity of a single link in the link-aggregation group.

We selected 170 multicast groups and 17 sources per group to generate a total of 2,890 S,G mroutes, which according to previous test results, both vendors easily supported.

In both vendors' switch configurations, the various clocks and timers associated with the router redundancy protocols were set to identical values.

Two possibilities exist for the position of the multicast designated router (DR) in the test topology. The DR can be installed on the VRRP/GLBP master or on the standby router. For this reason the test was repeated for each scenario.

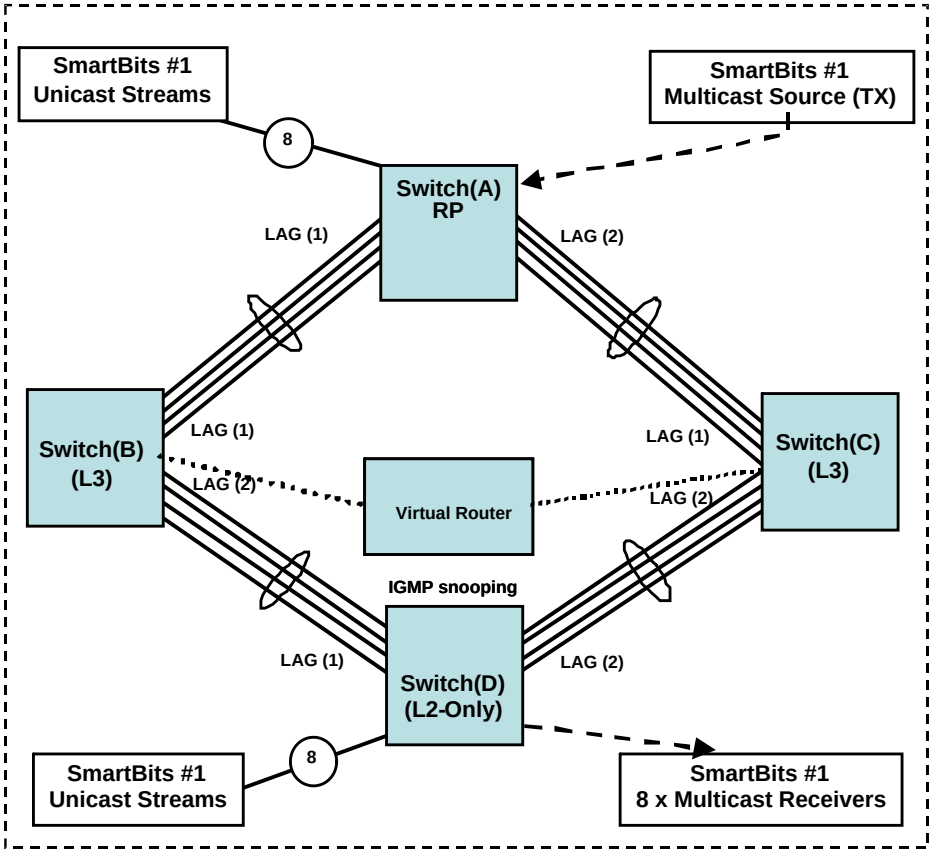



**Figure 3: Test Configuration**

The first test simulated failure of the master router that was also hosting the multicast DR, affecting both multicast and unicast topologies.

The second configuration saw the multicast DR located on the standby router. In each test, we always failed the master router. In this test configuration, the multicast topology remained unaffected.

Each test measured the time taken by the vendor's solution to restore all multicast traffic, as well as noting any problems.

# Appendix

## Hardware and Software Versions of the Devices under Test

| | Force10 E1200/E600 | Cisco Catalyst 6500 |
|---|---|---|
| **Route Processor** | IBM PowerPC 750FX (Rev D2.2) | Supervisor Engine 720 WS-SUP720-3BXL |
| **Gigabit Ethernet Line Card** | LC-EF-GE-48T | CEF720 48 port 10/100/1000 Ethernet WS-X6748-GE-TX with WS-F6700-DFC3BXL Distributed Cisco Express Forwarding (dCEF) daughter cards |
| **Software Version** | FTOS 6.2.1.3 | s72033-ipservicek9-mz.122-18.SXE2 |

**Figure 1: Force10 E1200**

**Figure 2: Force10 E600**

**Figure 3: Catalyst Product Family**

## Spirent SmartBits 6000C

The testing environment included Spirent's SmartBits load generator in combination with TeraRouting Tester software.

The TeraRouting Tester (TRT) application provides the first integrated control and data plane routing test that includes system-level configuration and analysis of routed networks. TRT is an easy-to-use application that allows network service providers and network equipment manufacturers to precisely determine the performance of a router under a variety of realistic and worst-case scenarios. TRT exposes the true performance of a router by providing stress testing of the routing software, the data forwarding hardware, and the overall system architecture under both static and dynamic routing conditions. A key strength of the application is its use of a deterministic and repeatable test methodology.

**Figure 4: Spirent SmartBits 6000C**

TRT includes a routing stack that allows each TeraMetrics port to act as one or more peer routers to the device under test (DUT). These peer routers establish adjacencies with the DUT to exchange routing information, and in the process, determine basic routing protocol performance. Test data is then transmitted to the advertised routes at user-specified rates to determine the forwarding performance of the hardware.

By simulating various network events (such as route, session, and link flaps), TRT then attempts to characterize the DUT behavior under various real-life stress conditions. With TRT, you can determine convergence time by observing the receive rate and the maximum latency over time for both the flapped routes and the stable routes in the router.

| Spirent SmartBits Hardware | Version |
|---|---|
| Card Type | LAN-3325A |
| Card Core Firmware | 5.30.036 |
| Routing Firmware | 5.00.024 |

### Test Topology Example

The four test areas explored various topologies and configurations. Each test report includes an in-depth description of test topology. As an example, *Figure 5: Force10 E1200 672-Port Snake Test Topology* shows the physical test topology used for Force10's frame loss test in a snake topology. In this example the SmartBits tester (the device on the left side of the picture) is cabled with only two ports to the DUT.

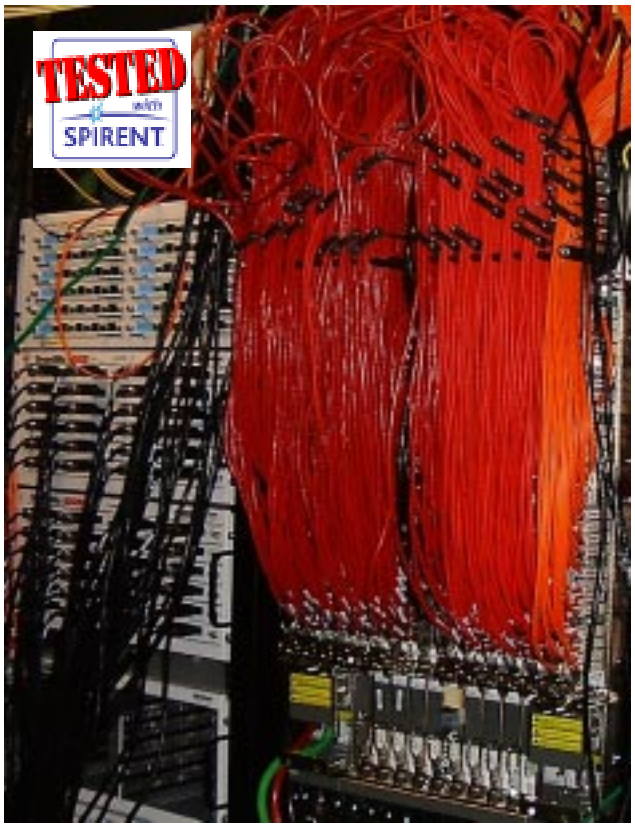



**Figure 5: Force10 E1200 672-Port Snake Test Topology**

EANTC AG
Einsteinufer 17, 10587 Berlin, Germany

Phone: +49.30.3180595-0
E-Mail: info@eantc.com
Web: http://www.eantc.com/

# About EANTC

EANTC (European Advanced Networking Test Center) is an internationally recognized test lab based in Berlin, Germany. We offer vendor neutral network test facilities for manufacturers, service providers and enterprise customers. Our business areas include:

## ➜ Test and Certification Services for Network Equipment Manufacturers

EANTC provides external quality assurance by testing conformance, performance and robustness of single systems as well as the interoperability of multiple components. More than 14 years of testing experience, the use of state of the art analysis equipment and active participation in industry forums enable us to conduct tests even for the most advanced customer requirements.

EANTC tests cover the full range of Carrier Ethernet, IP, and MPLS networks, IP services including firewalls and VPN gateways, ATM, and voice over packet networks.

## ➜ Test Services for Service Providers and Enterprises

EANTC supports network operators during all stages of the infrastructure life cycle: From network design and RfP creation through prototype testing during the tender process and acceptance testing up to trouble-shooting and future capacity planning. Our tests ensure network performance, availability and resilience. They reduce the risk of network failure and cost of operation.

## ➜ Philosophy

In all our activities — whether we conduct tests, provide consultancy or training — only the highest quality serves our purpose. EANTC invests much in research, education and development of new test methods to meet the demands of our customers. Our active participation in industry forums is a must to stay up to date.

We often find ourselves mediating between network operators, vendors and systems integrators in a cross-cultural environment. EANTC test services result in neutral, unbiased facts. Our detailed test reports provide the basis for rational decisions and actions. We ensure confidentiality in any required way.

## ➜ History

In 1991, EANTC was founded at the Technical University of Berlin. EANTC soon started to test and certify network components for high speed network technologies, and continued the close cooperation with the university after its spin-off in 1999.