

SUMMARY TEST REPORT

Core Ethernet Switches Buffering and Control Plane Performance Comparison:

Cisco Systems Catalyst 6500

Foundry Networks BigIron 8000

July 12, 2000

ABSTRACT: A review of all the existing test suites and RFCs on throughput testing (as described in RFC 1242, 1944 and 2285) revealed that there were no established guidelines for testing how a core Ethernet switch is actually stressed in a busy service-provider or enterprise network. In this environment, the Layer-3 switch must respond to heavy, bursty traffic loads, with thousands of addresses dynamically changing (including route flapping). Cisco Systems and MIER Communications worked in a joint field study and testing project involving first of its kind measurements to determine how a core switch would perform in such an environment—using tools from NetCom Systems and Shomiti Systems. This report includes results from tests applied on Cisco Catalyst 6500 and Foundry BigIron 8000 core Ethernet switches.

Background

Based on newly devised test scenarios, Cisco Systems sponsored a switch review, executed by MIERcomms Labs, on Layer-3 core Ethernet switches from Foundry Networks (BigIron 8000) and Cisco Systems (Catalyst 6500).

MIER Communications and Cisco Systems conferred to discuss network design issues and the state of network technologies involving core Ethernet switches for both service-provider and enterprise-customer applications. We shared our experiences regarding customer installations, product claims, test studies that have been conducted, and came to the mutual conclusion that there was still work to be done in the test market regarding fair, accurate and meaningful tests on core Ethernet switch products. In particular, we could find no lab-test methodologies, or existing testing guidance via RFCs, that addressed realistic methodologies to test switches in an environment that most closely replicated large enterprise and service-provider environments. We decided to investigate this further.

Customer field studies indicated – During a recent field-analysis study of core switch deployment, which included interviews with network design professionals, as well as our own experiences with clients from the service-provider and enterprise environments, we found that core switches are predominantly configured in a “many-to-one” relationship: that is, many 10/100 Ethernet links pass to a single, high-speed (typically Gigabit Ethernet) uplink at co-location points-of-presence (POPs), wiring closets, distribution points and server farms. The port density arrangement for these deployments was between 10 and 36 in our immediate consulting customer base. However, engineers from both Foundry and Cisco reported that they were aware of sites with a considerably higher ratio of 10/100 Ethernet links to Gigabit uplinks.

During the interviews and visits, we immediately recognized two major issues that occur in large service-provider and enterprise networks that deploy Layer-3 switches: First that the contention for the uplink(s) was causing Gigabit choke points in which the uplink would become overloaded for brief periods of time; second, that packet loss occurs due in networks that must contend with thousands of dynamically changing IP flows. This condition, which requires the route processor on the CPU to switch the first packet in the flow, is seen when new IP flows enter a switch or a route flap occurs.

Buffering

When analyzing the traffic at some of these customer sites, we found it very bursty in nature. Since the bulk of traffic in networks is TCP-based, traffic bursts occur as TCP sends a window of data, waits for a response, and then bursts again (a burst is defined as a series of packets sent with a minimum inter-packet gap). When multiple bursts are sent to a port, for instantaneous periods, the capacity of the uplink on the core switch can be exceeded. We conferred with the customers’ design engineers to discuss

methods to alleviate the bottlenecks. We discussed three possible solutions: 1) provide additional hardware and uplink capacity through trunking techniques and/or additional switch hardware; 2) enable a flow control or QoS mechanism in the switches; or 3) select a product with the most effective buffering capability.

Our conclusion on these alternatives is as follows. We discounted the first option due to the complexity and costs involved in adding Gigabit uplinks, which would not be practical in most cases. Additionally, the problem is not completely solved since now two ports need to have efficient buffers.

The second option, enabling flow control, could prevent the over-subscription from occurring, but only at the expense of pushing the congestion further back to the edge of the network. While this makes the switch in question not drop frames, congestion is merely pushed back to the upstream device, which may or may not have effective buffering. This results in the entire network slowing down. (Cisco offers a solution called Weighted Random Early Detection, which preemptively drops a packet from random TCP streams based on IP precedence. This allows small numbers of streams to back off at Layer 4 of the protocol stack, where some intelligence can be provided. At this time, Foundry does not offer this feature)

Our final decision was to recommend a product with the most effective buffering capability. The challenge, then, was to develop a test to measure the effectiveness of buffering on core Ethernet switches when they were handling loads consisting of bursty traffic of varying packet sizes (the typical service-provider environment).

After analyzing the architectures of switches used in service-provider and large enterprise applications, we found that they relied heavily on their buffering capabilities to handle the load that occurs when uplinks become over-subscribed for brief periods of time.

How effectively the switch buffers traffic is dependent upon the design of the buffering mechanism, which might include the following: the size of buffers, type of allocation, shared memory or fixed memory per port, and minimum block size that can be allocated per packet for buffering.

We theorized that the switch with superior buffer management would achieve less packet loss, and better overall throughput, in the “many-to-one” test scenarios. Therefore, we decided not to measure the aggregate size of the switch buffers themselves, nor the minimum allocation size, as neither would really allow us to rate overall effectiveness of buffering under the overload conditions described above. Rather, while exercising a switch’s buffering capability as it handled traffic, we measured packet loss, which is the source of end-user complaints when the excessive TCP retransmissions cause application slow down and time outs.

Control-Plane Performance

During the previously mentioned ISP and Enterprise studies, another noteworthy occurrence involving particular stress on switch architectures was encountered. The typical mid-size, service-provider environment processed many thousands of IP address flows per second (source-destination traffic pairs), which had to be learned and cached by the switch at great speed. In many cases, such as in an e-commerce network, many new flows are being set up as customers and content viewers access the server. In these types of environments, peak demand can surge to well over 8,000 new IP address conversation pairs per second.

It was further discovered that when the new address flow rate exceeded the ability of the switch to cache the new entries, packet loss would occur. This situation is exacerbated when, for example, route flapping occurs in which intermittent link-outage requires whole tables of addresses to be repopulated quickly.

In many architectures, such as the Catalyst 6500 Supervisor I and Foundry BigIron Management Module 3, the first packet of each new flow must go to the CPU, which performs a routing-table lookup, switches the packet in software, then creates an entry for this flow in the ASIC-based hardware. This architecture is commonly referred to as a “flow-based” architecture. Since the CPU is now involved in switching packets, the limiting factor is no longer the capability of the ASICs to switch at high speed, but how fast the CPU can switch frames and create hardware-based entries.

Due to potential network disruption, we did not have the opportunity to test or observe route flapping or peak IP address flow surges at a service-provider site, but we conferred with senior network engineers at network operation centers as to how to best replicate this type of situation in our lab.

The conclusion was to use test and measurement equipment to induce thousands of varying IP address flows to a switch that had its IP address cache cleared, and then measure how quickly (by virtue of lost routable packets over time) the switch could repopulate its IP address cache. These benchmark tests provide overall insight as to how robust the flow-control architectures are and how well the IP address caches of the switches function.

Test Results

Buffering Test (SmartFlow Test)

This test is designed to measure the switch's ability to buffer traffic during peak bursty periods of traffic. We used a NetCom SmartBits running the latest version (1.12) of SmartFlow. We applied 40 ports of load (40 flows of IP traffic). The SmartFlow Application conducts IP initialization dialog (ARP initialization) and then transmits bursts of traffic on inbound 10/100 Ethernet links unidirectional to a Gigabit uplink. Unidirectional flows were selected since the bulk of Internet traffic (over 90%) is more or less unidirectional in nature (e.g. upload or download transfers of data, or web inquiries one way with acknowledgement traffic coming in the other direction). The tests were done at 5% average utilization and with 15-packet bursts using packet sizes of 64, 512 and 1,518 bytes. (Note: results for 64-byte packets and 1,518-byte packets, but not 512 bytes, are shown in the chart below.) This load was applied to 15, 20, 25, 30 and 35 ports inducing peak demand up to twice the uplink capacity for some of these tests, but for only brief periods of time. During the brief period, in which the link was oversubscribed with packet bursts, the buffers must be utilized. (Throughput results for 15 to 30 ports are shown in the chart below. See Table 1, next page, for all results, expressed in terms of packet loss.) Each test lasted for 10 seconds.

The switch under test (SUT) was configured with one port per VLAN per IP subnet. VLANs were configured on each SUT with single subnets per port per VLAN. ARP tables within the switches were populated before the throughput tests were conducted and measurements were taken.

**Chart 1: Buffering Test Results - Percentage of Throughput*
(Results shown for 64- and 1,518-byte packets and 15, 25 and 30 ports.)**

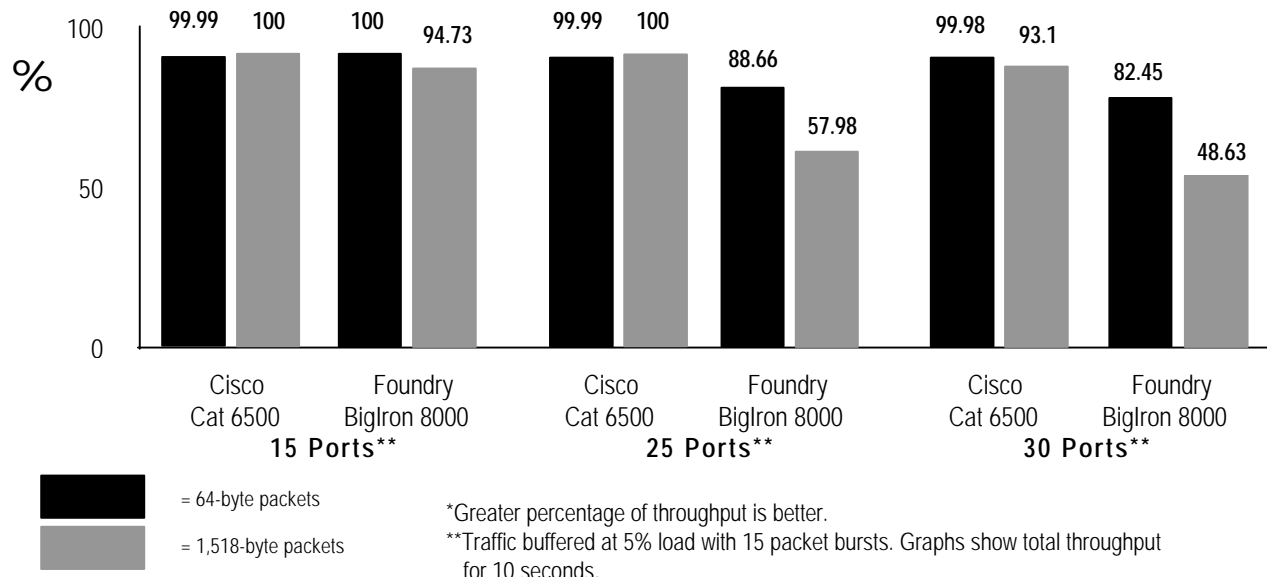


Table 1: Buffering Test Results - Percentage of Packet Loss		
	Cisco Systems Catalyst 6500	Foundry Networks Biglron 8000
Tested with 15 Ports	Percentage of Loss*	Percentage of Loss
64-byte packets	0.0012	0
512-byte packets	0	1.53
1,518-byte packets	0	5.27
Tested with 25 Ports	Percentage of Loss	Percentage of Loss
64-byte packets	0.0012	11.34
512-byte packets	0	37.68
1,518-byte packets	0	42.02
Tested with 30 Ports	Percentage of Loss	Percentage of Loss
64-byte packets	0.0013	17.55
512-byte packets	0	46.78
1,518-byte packets	6.87	51.37
Tested with 35 Ports	Percentage of Loss	Percentage of Loss
64-byte packets	0.0013	24.31
512-byte packets	0	53.79
1,518-byte packets	19.99	58.03

*All percentages of loss shown above are averaged over a period of 10 seconds.

There are two aspects of the buffering tests that readers should keep in mind when relating the results to their own network environment. They are, first, the ratio of 100-Mbps links contending for the same Gigabit uplink; and, secondly, the fact that the test system we used delivered all data-traffic surges—in bursts of 15 or 20 packets—on all input ports at the same time.

In a theoretical sense, no more than ten 100-Mbps links feed into the same Gigabit uplink, thereby eliminating congestion and packet loss. However, this network is unrealistically “over engineered”; many customer deployments push a more cost-effective ratio of 10/100 links-to-Gigabit Ethernet uplinks. Ratios of 24-to-1 are not uncommon in some ISP networks. In the buffering tests conducted here, we started with a 15-to-1 ratio and increased that to a 35-to-1 ratio.

The test demonstrates that, under the most common bursty traffic scenarios, the Catalyst 6500 delivers superior performance relative to the Biglron. This means that, in a real network scenario, the Catalyst 6500 would be able to handle congestion much better than the Biglron, (subject to the randomness of bursts, which we could not introduce in this test bed). Since we saw fewer packets dropped in the Catalyst, the applications on the user’s end station, whether in the Enterprise or a home-user shopping over the Internet, would witness faster response time and better performance.

The Flow Test (Learning Address Test)

We specifically tested the switches' ability to populate IP address cache tables and switch IP traffic to handle surges of address flows. This occurs when new IP flows need to be learned by the switch. In many architectures (including the Foundry BigIron and Cisco Catalyst 6500), high-speed switching can occur only after the first packet has been switched by the CPU. As new flows enter the switch, the CPU must switch them first before they can be switched in hardware. An extreme example of dynamically changing IP addresses is in the event of a route flap.

In the flow test, we used a NetCom SmartBits running the latest version available of SmartWindows (6.53.23). We first allowed the SmartBits to populate the MAC address table; each port loaded would apply a thousand flows. We incremented the number of ports from 5 to 10 to 15 and to 20, thus applying 5,000, 10,000, 15,000, and 20,000 flows (unique IP address pairs) per port. We applied 64- and 1,518-byte packet streams for 10 seconds, and measured how many of the flows were successfully added to the IP address cache. As shown in the graphs that follow, loss occurs disparately, based on delivery to system of up to 15,000 new addresses per port.

The flow tests we applied exercised an aspect of many Layer-3 switches, including the Cisco and Foundry models tested, where the first packet of a "flow"—a logical connection between a unique IP source and destination pair—is processed in software. This switching scheme is often referred to as a "flow cache" or an "exact-match lookup" implementation. Afterwards, subsequent packets in the flow are Layer-3 switched (in essence, IP routed) in ASIC-based hardware.

The flow tests apply 5,000 to 15,000 new address flows in a cyclic fashion at line rate for a period of 10 seconds. We chose a 10-second period to give each product a fair chance to learn the new address pairs and to run closer to a steady-state condition. In similar tests of one-second duration, both products exhibited packet loss, but a much wider gap in performance of the two products was observed. The results shown in the following charts (next page) represent the effect of packet loss while attempting to transmit traffic as the switch is learning new addresses. The rate at which the switch learns the new addresses greatly affects the amount of traffic throughput in this measurement.

Chart 2: Flow Test Results - Percentage of Throughput*
 (Results shown for 64- and 1,518-byte packets and 5K, 10K and 15K flows)

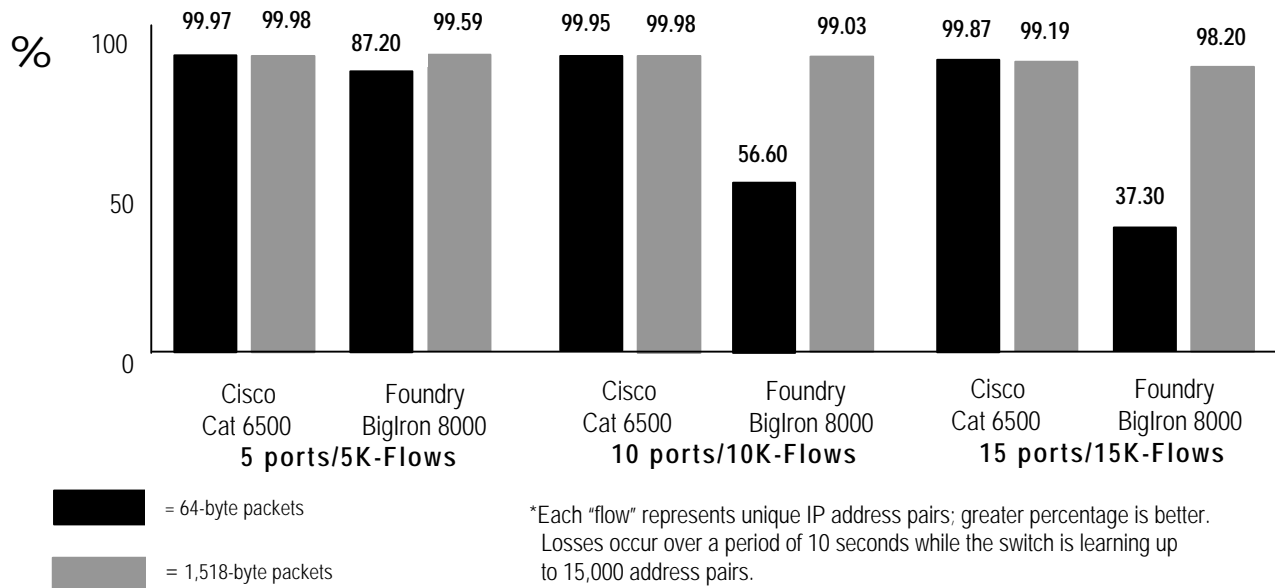


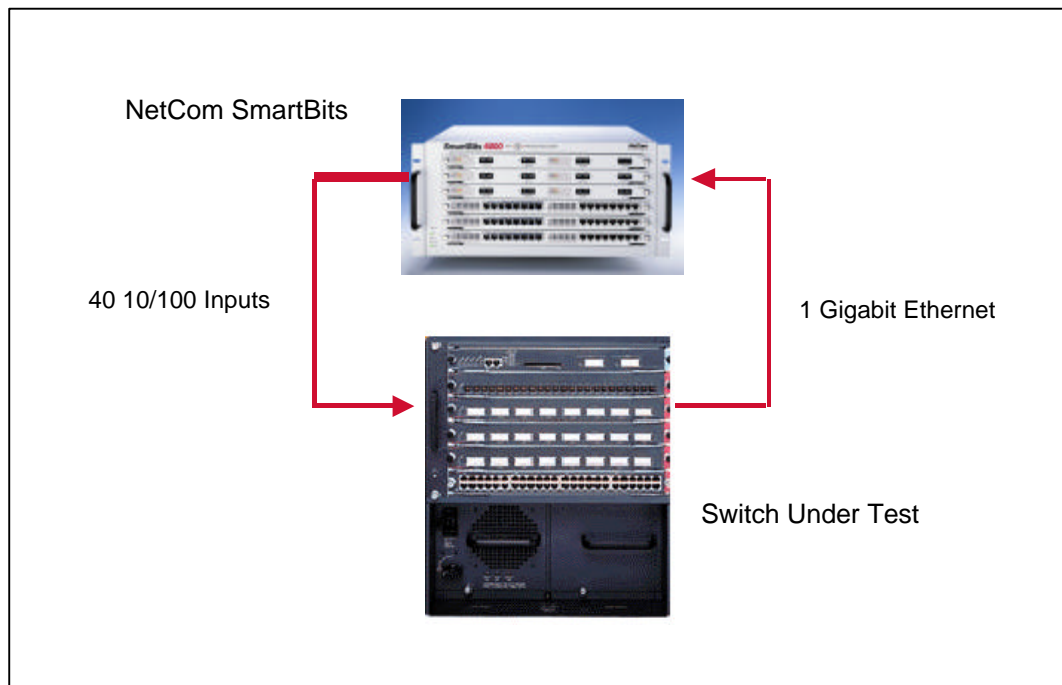
Table 2: Flow Test Results - Percentage of Packet Loss

	Cisco Systems Catalyst 6500	Foundry Networks BigIron 8000
Tested with 5K-Flows	Percentage of Loss	Percentage of Loss
64-byte packets	0.03	12.80
1,518-byte packets	0.02	0.41
Tested with 10K-Flows	Percentage of Loss	Percentage of Loss
64-byte packets	0.05	43.40
1,518-byte packets	0.02	0.97
Tested with 15K-Flows	Percentage of Loss	Percentage of Loss
64-byte packets	0.13	62.70
1,518-byte packets	0.81	1.8
Tested with 20K-Flows	Percentage of Loss	Percentage of Loss
64-byte packets	0.20	72.40
1,518-byte packets	0.63	2.48

On the Foundry BigIron 8000 switch, we had to increase the ARP table size and IP address cache size from the default values of 8,000 each to 32K each (although we needed only 20K at most).

The results of this test indicate that the Catalyst 6500 was able to software-switch frames with a minimum of packet loss relative to the Foundry BigIron. In an enterprise, this is important when, for example, many users log onto the network at 9 a.m., all of them hitting the network core within a short time window. In a service-provider network, thousands of new flows may need to be learned per second if the switch is connecting into e-commerce or application-hosting sites. If a device is not able to handle this overhead traffic, packets will be lost and the end-user will witness slow performance, regardless of what the maximum packet-per-second of the switch claims to be.

Test-Bed Description



Equipment used for testing included NetCom Systems SmartBits 6000 configured with 7710 Layer-3 cards running “SmartFlow” application tests. We also used a Shomiti Systems Century Analyzer to validate the test flows produced by the SmartBits.

MIERComms completely validated the test bed, including the test equipment and vendors’ equipment used in the review. We contacted all vendors whose products and equipment are mentioned in this report to ensure we had the latest versions of operating code, the proper configurations for the SUTs and test systems, and that what we were testing actually made sense in light of how customers have deployed these products. Each of the tests MIERComms conducted here are repeatable, at any lab using similar test equipment and the same products under test. All test-configuration files for systems under test and the test equipment have been saved for future reference, and are available upon request.

The Foundry Network BigIron 8000 switch tested was configured with an 8000 chassis SYSIF, version 21, populated with the latest shipping hardware. The chassis was populated with one B8GMR3 Management Module Three, the recently released management module that offers eight ports of Gigabit Ethernet; three 24-port 10/100 line cards; one 8-slot Gigabit Ethernet module; and three power supplies. We tested with firmware version 06.6.05T53 Enterprise Version. Through Foundry’s Web site and tech support, we validated that the firmware release we were using was the most current (generally available) shipping release.

Conclusion

The current RFCs that specify performance are interesting in evaluating the maximum capabilities of a switch and its architecture. However, they are not at all realistic in terms of how networks actually run. For this reason, MIERComms and Cisco devised a test methodology in order to test two critical performance-related items that correspond to real networks: buffering and control-plane performance. The methodology was developed after consultation with end-users (both enterprise and service providers). The results of this test indicate that the Catalyst 6500 demonstrates superior performance to the Foundry BigIron in the overload tests that we have run, which directly model what happens in real-world networks.