Technical Validation

Broadcom Trident 3 Platform Performance Analysis

Achieving Predictably High Performance for Real-world Data Center Workloads

By Alex Arcilla, Validation Analyst; and Tony Palmer, Senior Validation Analyst

May 2019

This ESG Technical Validation was commissioned by Broadcom and is distributed under license from ESG.

Contents

ESG Technical Validations

The goal of ESG Technical Validations is to educate IT professionals about information technology solutions for companies of all types and sizes. ESG Technical Validations are not meant to replace the evaluation process that should be conducted before making purchasing decisions, but rather to provide insight into these emerging technologies. Our objectives are to explore some of the more valuable features and functions of IT solutions, show how they can be used to solve real customer problems, and identify any areas needing improvement. The ESG Validation Team's expert third-party perspective is based on our own hands-on testing as well as on interviews with customers who use these products in production environments.

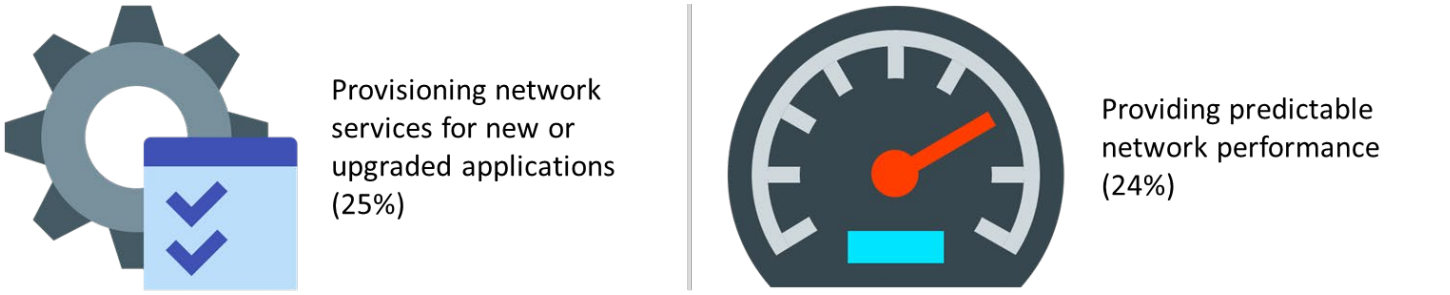
Introduction

This ESG Technical Validation documents hands-on testing of the Broadcom StrataXGS Trident 3 Switch Series, focusing on how the Trident 3 silicon architecture delivers consistent, predictably high performance in data center, enterprise, and cloud provider network environments. This report will validate multiple data center workloads and network configurations to illustrate how Trident 3 performs in real-world server-to-server and server-to-storage scenarios common in enterprise and cloud provider networks, while contrasting observed performance results with those of a competing switch silicon/system vendor. We conducted detailed tests exercising the switches' capabilities in terms of burst absorption, performance under congestion, security policy enforcement, and lossless storage traffic handling—all under in-network loading conditions. The goal of testing is to validate Trident 3's architectural approach and provide decision makers clear insight into the differentiated attributes of the Trident 3 solution versus potential alternatives when deployed in a live, cloud-scale network in terms of switch throughput, drop rate, latency, port fairness, and overall job completion time.

Challenges

IT is increasingly expected to improve application agility and responsiveness for users, but infrastructure, specifically network infrastructure, often holds them back. What are the biggest challenges facing networking teams as they look to maintain what's already in place while simultaneously evolving the network with new technologies? ESG's research indicates that provisioning network services for new or upgraded applications (25%) and providing predictable network performance (24%) are both high on the list of challenges.[1]

Figure 1. Challenges Facing Network Teams



Source: Enterprise Strategy Group

Broadcom Trident 3 Architecture

In 2017, Broadcom released the Trident 3 Switch Series, a programmable, multilayer Ethernet switch silicon family whose flagship device features 3.2Tb/sec of switching performance, 128 lanes of 25 Gb/sec serializer/deserializer (SerDes) interfaces, and high-density support for 10/25/40/50/100 Gigabit Ethernet ports. Designed to protect customers' networking equipment investments with a field-upgradeable packet processing data plane and 100% backward compatibility with the existing StrataXGS install base, Trident 3 technology is integrated into a broad range of OEM/ODM switching platforms, is supported by all popular network operating systems in the industry, and is deployed in many data center, enterprise, and service provider networks.

Building on Broadcom's well-established StrataXGS Trident and Tomahawk switch product lines, Trident 3 takes a balanced architectural approach designed for enterprise/cloud data center network operators that aims to provide high, deterministic throughput invariant with features; low latency at high offered load; Layer-2 through Layer-4 feature integration; large-scale reconfigurable switching and routing database; static and dynamic load balancing; programmable overlay and forwarding protocol support; fully shared buffering with RDMA support and congestion management; network

[1] Source: ESG Master Survey Results, *Trends In Network Modernization*, November 2017.

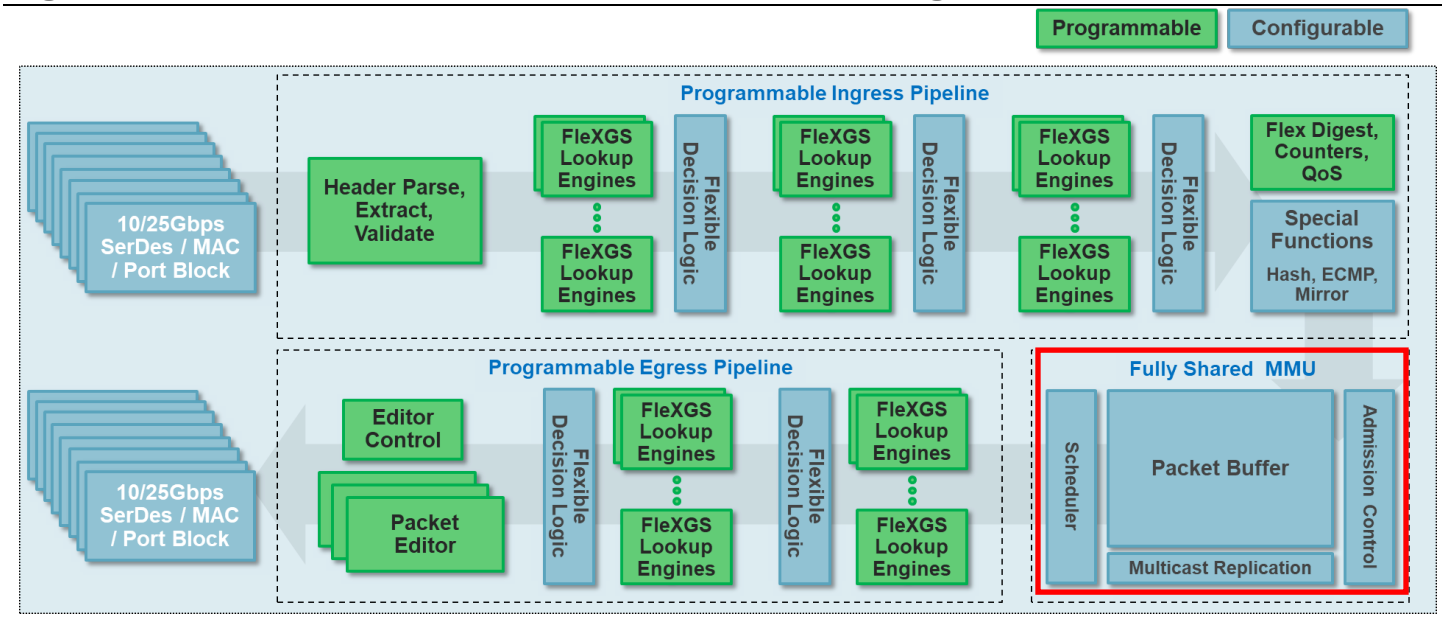
visibility augmented by telemetry sources and programmable processing; low power consumption scaling with performance; and cost-efficiency per Gbps for networks migrating from 10/40GbE to 25/100GbE.

To optimize the attributes of primary concern to network operators, Broadcom explicitly chose to avoid over-engineering on attributes that do not matter in real-world enterprise and cloud computing network environments. For example, Trident 3 does not optimize for small packet (e.g., 64B) benchmark performance at 100% load—a traffic pattern benchmark that is not relevant to compute/storage networks at scale—instead optimizing for significant efficiencies in power and cost per Gbps, in addition to increased table, buffer, and programmable resource scale. Broadcom also did not optimize the Trident 3 architecture for the lowest port-to-port *unloaded* latency—i.e., the measured latency for a single packet flow when there is no other traffic present, instead choosing to optimize for the generally accepted principle in distributed computing that the measure of application performance is in overall flow completion time (FCT) or query completion time (QCT) with live network traffic flowing, i.e., under load. In these real-world conditions, unloaded latency benchmarks are not relevant and FCT/QCT can be treated as the metric for application performance over the network.

This approach enables Trident 3 to take on the additional levels of processing required for a comprehensive L2-L4 switching and routing feature set with data plane programmability and provide a fully centralized buffering architecture that absorbs bursty traffic from all ports without loss and minimizes FCT/QCT in many-to-one burst scenarios, regardless of what features are activated in the switch.

Upon examining the switch architecture in further detail, it became clear that Trident 3 is engineered with price-performance in mind, leveraging parallelized packet processing engines with multiple lookups per clock and centralized, shared databases. Trident 3 is built on a new FleXGS pipeline architecture (see Figure 2) that is software API and feature compatible to current Trident, Trident 2, and Trident 2+ switch products, which are widely deployed in the field. Trident 3 adds programmable parsing, lookup, and editing engines with associated reconfigurable databases, through which new switching and instrumentation features can be integrated via verified, in-field upgrades—just as easily as users might update the software. Broadcom dimensioned and arrayed the packet processing engines with the goal of maximizing parallelism, performance, functional capacity, and area/power efficiency.

Figure 2. Broadcom Trident 3 Switch Silicon Internal Architecture Diagram



Source: Enterprise Strategy Group

Of note in Figure 2 is the fully shared Memory Management Unit (MMU) that serves as the centralized buffering resource in the Trident 3 architecture between the ingress and egress switch pipelines. In the enterprise/cloud data center,

workloads are moving vastly increasing amounts of data between compute and storage nodes. Organizations want to achieve higher utilization of network resources while maintaining responsive performance of their applications—in other words, trying to minimize end-to-end FCT/QCT under high offered load on the individual switches. To facilitate this, the Trident3 MMU is designed with 32MB of integrated, fully shared buffer memory that can respond to large concurrent bursts of activity from multiple server or storage endpoints. The full scale of the 32MB buffer is shared across all 128 ports of the device, while dynamic thresholding and active queue management in the Trident 3 MMU enables the buffer pool to be steered towards all ports that need the buffer at a given time. The architecture is designed to provide low overall FCT/QCT for application latency, with advanced congestion management mechanisms to provide a responsive feedback loop to the server and storage endpoints that are driving the network congestion.

The above examples illustrate the important tradeoffs made at design time in a switch architecture; In the following pages, ESG validates how these choices can have a significant impact on real-world network behavior.

ESG Technical Validation

ESG evaluated and tested the Trident 3 switch platform against a competitive switch platform at Broadcom's headquarters in San Jose, CA. Testing was designed to evaluate the performance of switches using industry-standard tools and methodologies, while emulating real-world data center workloads and use patterns.

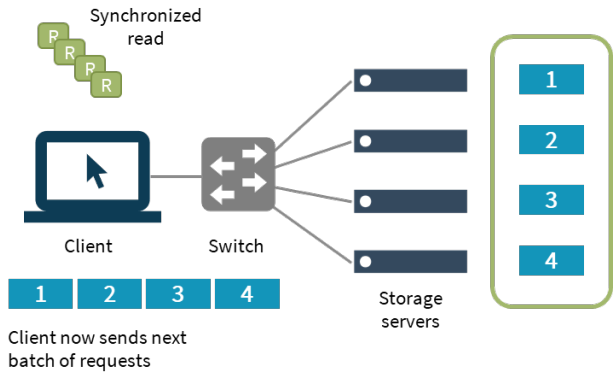
The test bed consisted of two top-of-rack 32-port 100GbE switches, an Arista 7050CX3 built on the 3.2Tbps Trident 3 switch platform, and an alternative 32-port 100GbE switch built on a competitor's 3.2Tbps platform. ESG ran a third-party operating system configured specifically for the competitor's switch platform. Additionally, we configured the competitor's switch according to the competitor's guidance via public documentation and operating system defaults. We then connected 30 Dell EMC PowerEdge R640 servers to each switch with 30 100GbE links. For tests related to TCP and TCP+RoCEv2 performance, we configured each switch with a RoCEv2-enabled, 2-port 100GbE network interface card (NIC) from the same manufacturer as the alternative switch (see Figure 3).

TCP Performance

TCP performance is critical for real-time applications such as file transfers, virtual desktop infrastructure (VDI), and database apps such as customer relationship management (CRM). Performance is especially critical with big data applications running in large distributed storage and computing environments such as Hadoop and Spark. End-users expect consistent, high performance from these applications, so they can perform their jobs efficiently.

A common issue that occurs in these large distributed storage and computing environments is TCP incast.[2] In this many-to-one-communication scenario typical between servers or between servers and storage in data centers, application performance can easily degrade, whether the number of servers transmitting data increases, or the amount of traffic increases from each server. TCP incast can result in congestion at the network port of the receiving server. In turn, the end user of the application subsequently experiences application latency and performance degradation, as multiple servers either wait until all application packets have been successfully transmitted before the next end-user data request is sent.

[2] TCP incast refers to the simultaneous transfer of traffic flows originating from multiple "child" servers to a single "parent" server within a scale-out distributed storage or computing environment



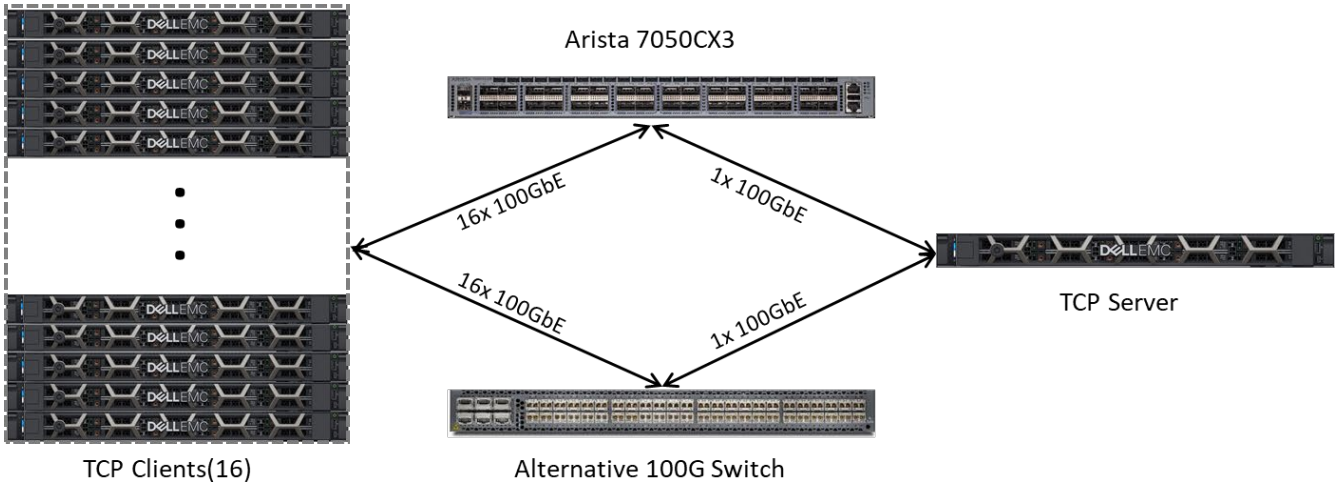
Recovering data on a failed server in a Hadoop cluster is a typical incast scenario. Hadoop is structured such that multiple copies of data are parsed out to the individual servers to guard against data loss in case one server fails.[3] Once the server is back online, it will rebuild its data by requesting copies from the other cluster servers, mimicking a many-to-one transfer scenario. As the performance of a Hadoop cluster can be gauged by its ability to rebuild a failed server, the cluster must effectively manage potential traffic congestion as data copies arrive from the other servers. Another scenario in which TCP incast can affect application performance is the rebuilding of failed storage nodes in a cluster-based storage system. When an end-user requests data, that request is distributed amongst multiple servers. Servers that have the relevant data stream that data to the client. As these multiple traffic streams converge onto a single port, congestion may arise as the client attempts to aggregate the data.

ESG Testing

ESG began by measuring TCP workload performance, under typical server-to-storage incast scenarios, of the Broadcom Trident 3-based switch against one using a competitive chipset. We emulated a single client transfer by attaching a single Dell R640 server, acting as a TCP server, to each switch using 100GbE links (see Figure 3). We used iPerf, an industry-standard network performance benchmark tool, to generate TCP traffic.

All Dell servers were equipped with dual 100GbE NICs manufactured by the competitor. We used Broadcom's system and SDK default settings for the MMU. In addition, we employed the competing vendor's prescribed guidelines for maximizing shared buffer allocation to a single service pool and dynamic buffer sharing between queues enabled, with a consistent amount of minimum reserved buffer space across ingress and egress ports and priority groups[4] on both switch platforms.

Figure 3. TCP Test Bed



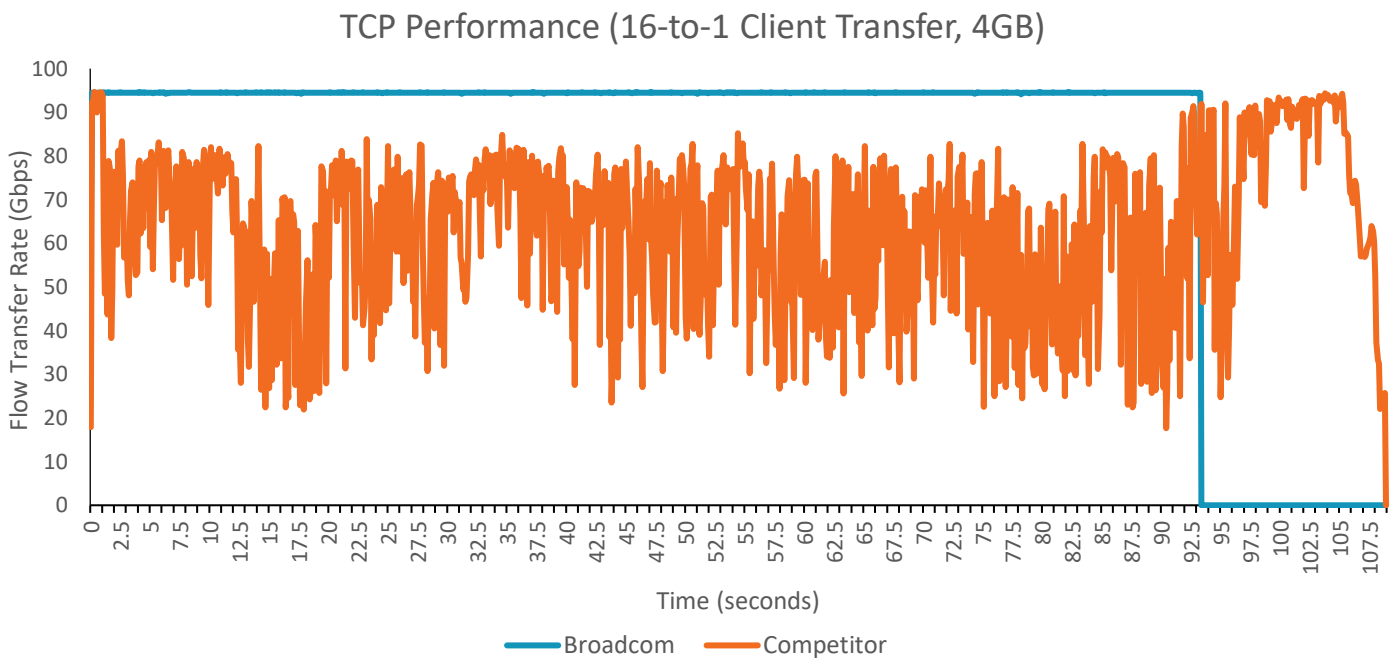
Source: Enterprise Strategy Group

We first performed a file transfer originating from sixteen servers simultaneously to the single TCP server, first using the Trident 3 switch, and then separately using the competitor switch. Each server transmitted 16 flows, with each flow

[3] In a Hadoop cluster, common practice dictates that three copies of the data are stored amongst multiple servers to tolerate server failures.
[4] A priority group is a group of output traffic queues assigned a forwarding class.

executing 4GB transfers at 256KB TCP window size. We measured the flow transfer rate at both the client and the TCP server, as well as the overall file transfer completion time; aggregate results at the server side are shown in Figure 4.

Figure 4. TCP Performance, 16-to-1 Client Transfer, 16 Flows per Client, File Transfer Size = 4 GB



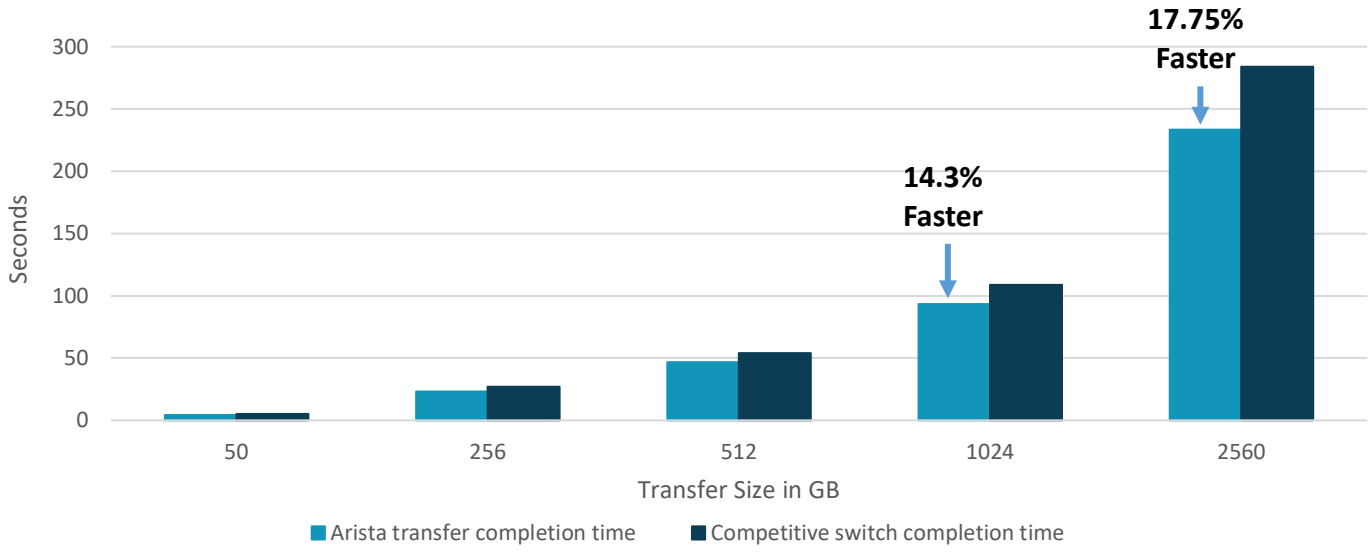
Source: Enterprise Strategy Group

ESG observed that the Trident 3 platform delivered a higher and steady flow transfer rate in aggregate over the entire file transfer job as opposed to the competitor's switch chipset. We also observed that the Trident 3 transfer almost achieved the maximum link bandwidth utilization. While we saw that the aggregate transfer rate measured for the Trident 3 was relatively constant between 93.9 and 94.6 Gbps for the test period, the transfer rate for the competitor chipset fluctuated throughout the test, dropping to as low as 22 Gbps on a 100 Gbps link. We also observed that all transfers from the sixteen clients through the Trident 3 Switch completed within 93.3 seconds. The same file transfers on the competitor's platform completed at 108.8 seconds, an additional 15.5 seconds. ESG concluded that the competitor chipset dropped more packets throughout the test, triggering frequent TCP retransmissions. The retransmissions caused additional congestion, increasing the transfer completion time by almost 17%.

ESG conducted this same test with per-flow transfer sizes of 200MB, 1GB, 2GB, and 10GB. We also aggregated all flows described and measured flow transfer rate and transfer completion time. We observed results similar to our first test with respect to flow transfer rate and completion time. As observed in our initial test, the aggregate per-port transfer rate on the competitor's platform fluctuated throughout all tests below 100Gbps. We also observed that all file transfer times on the competitor's platform exceeded those observed on the Trident 3 by 12 - 18% in total elapsed job completion time.

With 16 flows per client and 16 clients, the total emulated file transfer sizes were 50, 256, 512, 1024, and 2560 GB through the Trident 3 and the competitive switch chipset. As shown in Figure 5, we observed that the Trident 3 consistently completed transfers faster than the competitor's chipset in all cases. Completion times on the Trident 3 were up to 17% faster over all test cases.

Figure 5. TCP Performance Comparison-File Transfers



Source: Enterprise Strategy Group

The screenshot of the iPerf log in Figure 6 (Trident 3 results on the left) shows the results of a single client's file transfer across 16 server-to-storage TCP flows. We calculated an average 21% throughput advantage for Trident 3 over the competitor's chipset, both on a per-flow basis and in aggregate, demonstrating the sustained throughput advantage.

Figure 6. TCP Performance-iPerf Log

```
------------------------------------------------------------
Server listening on TCP port 10007
Binding to local address 172.16.19.142
TCP window size:  416 KByte (WARNING: requested  256 KByte)
------------------------------------------------------------
[ ID]     Interval          Transfer      Bandwidth
[ 19]   0.0-228.6 sec  10.0 GBytes   376 Mbits/sec
[ 12]   0.0-230.6 sec  10.0 GBytes   373 Mbits/sec
[ 18]   0.0-230.7 sec  10.0 GBytes   372 Mbits/sec
[ 11]   0.0-231.6 sec  10.0 GBytes   371 Mbits/sec
[ 14]   0.0-232.8 sec  10.0 GBytes   369 Mbits/sec
[  9]   0.0-233.2 sec  10.0 GBytes   368 Mbits/sec
[ 10]   0.0-233.3 sec  10.0 GBytes   368 Mbits/sec
[ 13]   0.0-233.4 sec  10.0 GBytes   368 Mbits/sec
[ 17]   0.0-233.4 sec  10.0 GBytes   368 Mbits/sec
[  7]   0.0-233.4 sec  10.0 GBytes   368 Mbits/sec
[ 16]   0.0-233.4 sec  10.0 GBytes   368 Mbits/sec
[  5]   0.0-233.5 sec  10.0 GBytes   368 Mbits/sec
[  6]   0.0-233.6 sec  10.0 GBytes   368 Mbits/sec
[  8]   0.0-233.6 sec  10.0 GBytes   368 Mbits/sec
[  4]   0.0-233.6 sec  10.0 GBytes   368 Mbits/sec
[ 15]   0.0-233.7 sec  10.0 GBytes   368 Mbits/sec
[SUM]   0.0-233.7 sec   160 GBytes  5.88 Gbits/sec
```

```
------------------------------------------------------------
Server listening on TCP port 20145
Binding to local address 192.168.49.142
TCP window size:  416 KByte (WARNING: requested  256 KByte)
------------------------------------------------------------
[ ID]     Interval          Transfer      Bandwidth
[  8]   0.0-201.7 sec  10.0 GBytes   426 Mbits/sec
[ 12]   0.0-260.9 sec  10.0 GBytes   329 Mbits/sec
[ 15]   0.0-261.9 sec  10.0 GBytes   328 Mbits/sec
[  7]   0.0-263.4 sec  10.0 GBytes   326 Mbits/sec
[  6]   0.0-264.5 sec  10.0 GBytes   325 Mbits/sec
[  4]   0.0-264.8 sec  10.0 GBytes   324 Mbits/sec
[ 18]   0.0-264.9 sec  10.0 GBytes   324 Mbits/sec
[ 17]   0.0-265.0 sec  10.0 GBytes   324 Mbits/sec
[  9]   0.0-265.5 sec  10.0 GBytes   324 Mbits/sec
[ 14]   0.0-265.9 sec  10.0 GBytes   323 Mbits/sec
[ 19]   0.0-266.0 sec  10.0 GBytes   323 Mbits/sec
[ 11]   0.0-267.0 sec  10.0 GBytes   322 Mbits/sec
[  5]   0.0-268.3 sec  10.0 GBytes   320 Mbits/sec
[ 10]   0.0-271.4 sec  10.0 GBytes   317 Mbits/sec
[ 13]   0.0-283.3 sec  10.0 GBytes   303 Mbits/sec
[ 16]   0.0-284.0 sec  10.0 GBytes   302 Mbits/sec
[SUM]   0.0-284.0 sec   160 GBytes  4.84 Gbits/sec
```

Source: Enterprise Strategy Group

Why This Matters

TCP performance is a critical factor to address when architecting a data center switch to ensure consistent performance in light of network congestion. Consistently high performance in a data center, especially when multiple machine-to-machine workloads coincide over the network, is critical to high utilization of network infrastructure and enterprise and cloud application performance and resiliency, enabling organizations to meet end-user SLAs.

ESG validated that the Trident 3-based switch can handle TCP incast scenarios with consistent and predictable performance compared with the competitor's switch chipset. Unlike what we saw with the competitive switch platform, we observed that TCP incast scenarios over multiple aggregate file transfer sizes (up to 2560 GB) on the Trident 3 platform completed simultaneously and quickly, primarily due to the switch device's shared buffer capacity. On the other hand, traffic flows on the competitive switch platform exhibited inconsistent flow transfer rates and packet drops, leading to congestion, higher transfer times, and degraded application performance.

TCP+RoCEv2 Performance

When designing a data center network infrastructure to support enterprise and cloud applications, IT usually accounts for the support of RDMA over converged Ethernet, version 2 (RoCEv2 traffic).[5] In these data centers, TCP and RoCEv2 workloads typically exist simultaneously to support the transmission of lossy and lossless traffic, respectively. Thus, any individual top-of-rack, leaf, or spine switch must concurrently ensure optimal TCP and RoCEv2 performance by configuring its internal memory management and congestion control to ensure both lossless transmission of RoCEv2 traffic as well as application performance for lossy and bursty applications (i.e., preventing excessive drops/retransmission).

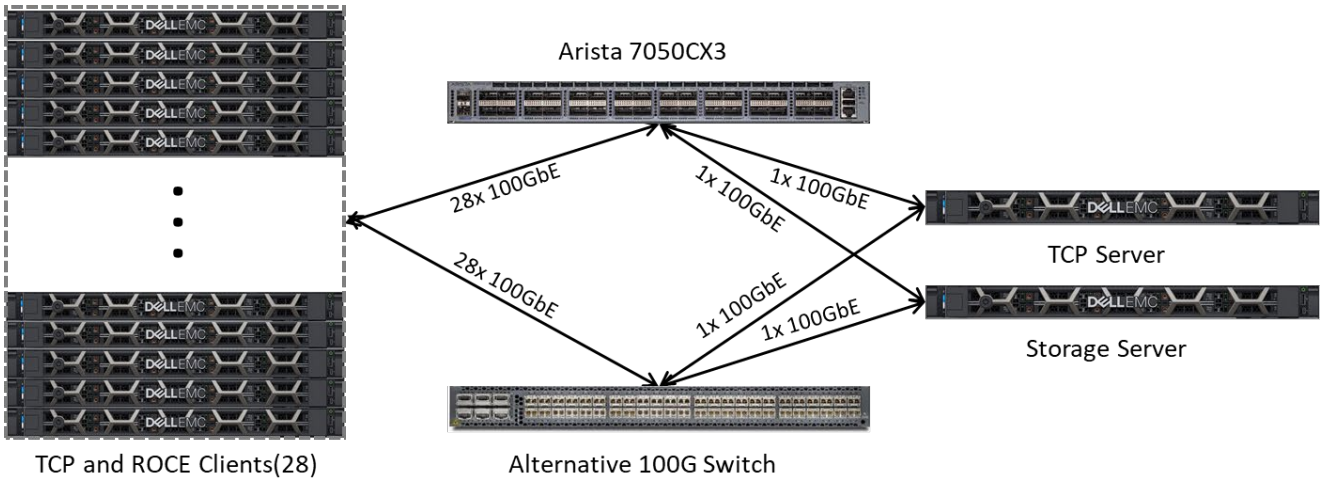
ESG Testing

To measure the combined TCP and RoCEv2 workload performance under typical server-to-storage traffic patterns, we modified the test bed introduced in Figure 3. First, we added a storage server equipped with a RoCEv2 enabled 2-port 100GbE NIC from the competitor. We then connected the Trident-3 based switch and the competitive switch to 28 TCP/RoCEv2 clients with 100GbE links (see Figure 7).

For each switch, ESG configured 14 clients each to transfer data to the TCP server and storage server simultaneously. Furthermore, we configured Explicit Congestion Notification (ECN) and Priority Flow Control (PFC) on each switch and the RoCEv2-enabled NICs according to the competitive vendor's configuration guidelines. Using iPerf and InfiniBand write bandwidth (IB_write_bw) tools, we measured job completion time of the combined TCP+RoCEv2 workload for each switch.

[5] RDMA over Converged Ethernet, version 2 (RoCEv2) is a network communication protocol used in data centers that can reduce latency in time-sensitive apps. The protocol enables faster transfer of packets between servers or between servers and storage via direct memory transfer between hosts without involving the hosts' CPUs, decreasing latency and increasing network utilization. RoCEv2 is typically applied in clustered storage applications.

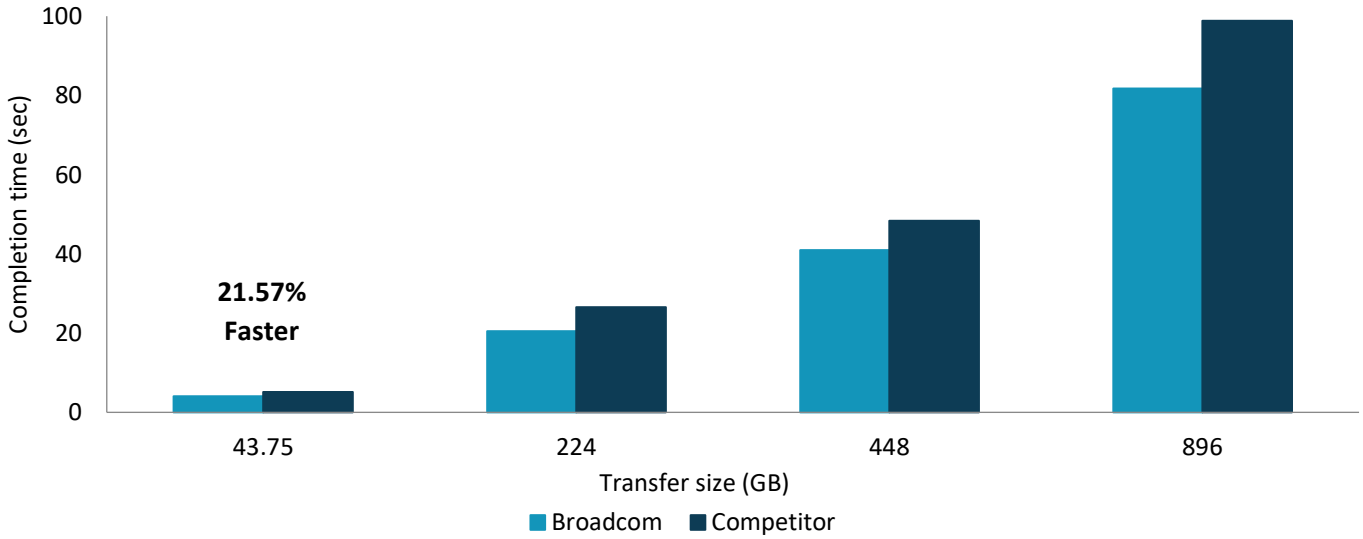
Figure 7. The TCP+RoCEv2 Test Bed



Source: Enterprise Strategy Group

ESG emulated TCP traffic by generating 16 flows per TCP client, with file transfer sizes of 200MB, 1GB, 2GB, 4GB, and 10GB, using a TCP window size of 256KB. We simultaneously emulated RoCEv2 traffic by configuring 100 queue pairs for each RoCEv2 client with a maximum transmission unit (MTU) of 2048 Bytes. We varied the total number of exchanges—100,000, 200,000, 400,000, 1,000,000—per client. Figure 8 shows throughput and completion time results for the combined workloads.

Figure 8. Combined Performance Broadcom versus Competitor—TCP+RoCEv2



Source: Enterprise Strategy Group

ESG observed across the combined TCP + RoCEv2 traffic patterns that the Trident 3 platform consistently completed TCP transfers faster than the competitive switch platform. For RoCE transfers on both platforms, we observed similar completion time and no packet drops with both PFC and ECN enabled. However, for the TCP portion of the combined workload, we noted Trident 3 had 21.57% faster completion time for the 43.75GB transfer and a 19.21% faster completion time on average across all TCP transfers between the servers and clients. While the lossless traffic was serviced

appropriately by both switches, the Trident 3 switch simultaneously maintained good performance on lossy TCP traffic, while the competitive switch platform did not.

Why This Matters

Organizations that support time- and loss-sensitive applications, such as search and storage, must architect their data center networks to reduce latency and increase throughput. RDMA over Converged Ethernet (RoCEv2) is ideal for these application types, as it reduces latency better than TCP while guaranteeing traffic delivery without host retransmission. However, data center network infrastructure that supports enterprise and cloud applications must manage both TCP and RoCEv2 traffic without optimizing one over the other.

ESG validated that the Broadcom Trident 3 switch platform did not lose any packets when emulating bursty data transfers from multiple storage clients to one storage server over a range of frame sizes. Trident 3 also maintained high throughput between all the TCP clients and the TCP server for the duration of the test, exercising nearly all of its link capacity. On the other hand, the competitive switch platform, while matching on lossless RoCEv2 performance, dropped a higher percentage of TCP packets regardless of file transfer size, increasing job completion time by up to 22.6%- for 224 GB transfers - and subsequently, increasing application latency. The Broadcom Trident 3 is well-suited to maximize the performance of lossless applications, such as storage services over RoCEv2 and TCP-based applications simultaneously.

Burst Absorption

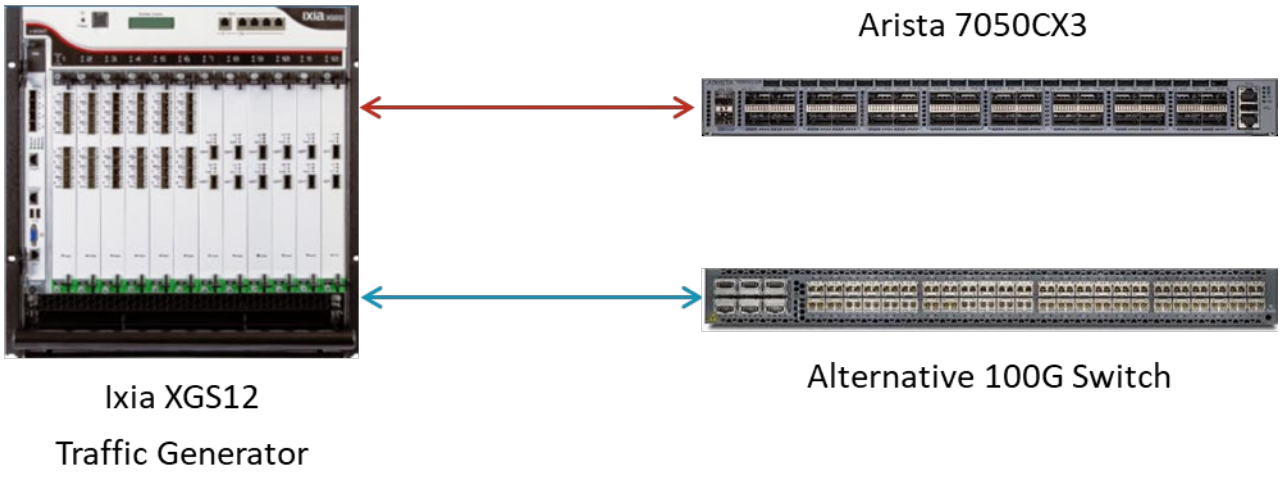
In “many-to-one” environments of today’s data center networks, switches must provide sufficient burst absorption capability to ensure that all application data transfers can simultaneously share buffer space. Should packets drop, app performance can degrade. This is especially relevant when multiple apps exhibiting bursty traffic patterns, such as Hadoop and MapReduce, contend simultaneously for one output switch port. The Trident 3 platform provides burst absorption capability to ensure the performance of multiple applications is maintained over a wide range of frame sizes.

It is important to note that TCP uses a “slow start” to probe and find the optimal window size. TCP connection reuse across applications is done extensively in cloud networks to reduce application start time. This can cause a burst of window size traffic at the start of the session, placing more burden on the network to absorb the bursts without drops. Many distributed application sessions can converge on a requestor, causing much larger bursts than generated by a single application session.

ESG Testing

ESG connected the Trident 3-based switch and the competitive switch with an IXIA XGS12 100GbE traffic generator (Figure 9). We simulated two-to-one and four-to-one traffic flow scenarios, with frame sizes ranging from 192 B to 9,216 B, on each switch.

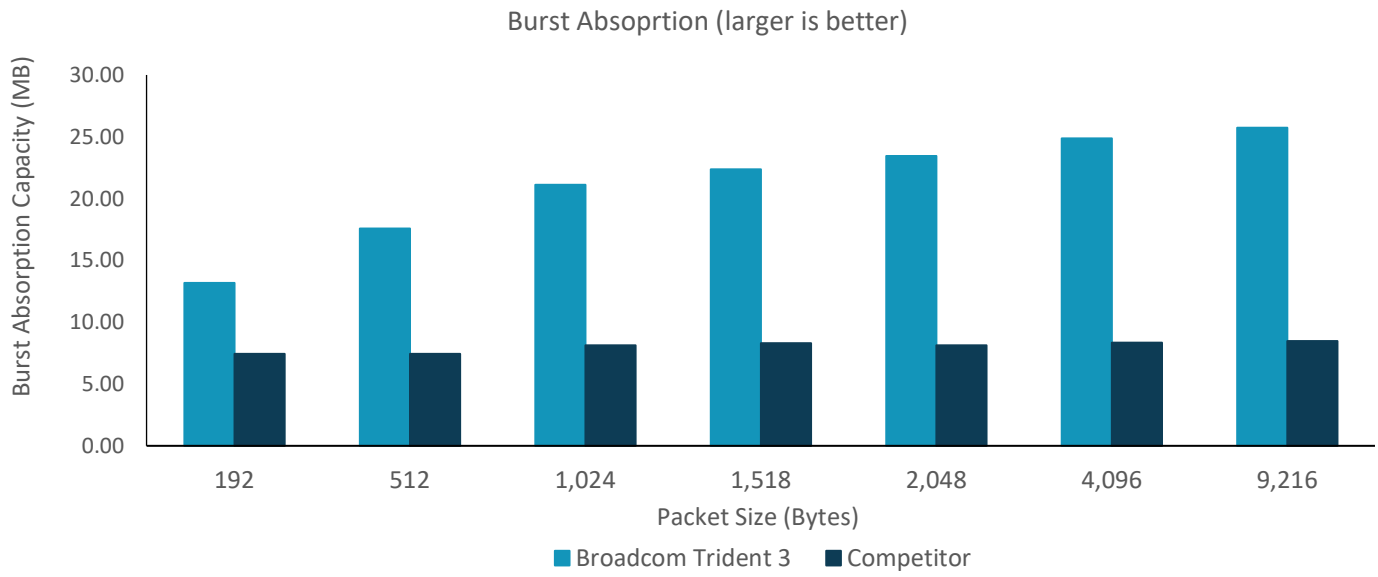
Figure 9. The Burst Absorption Test Bed



Source: Enterprise Strategy Group

For the results in the four-to-one scenario (Figure 10), ESG observed that the Trident 3 platform's burst absorption capacity exceeded that of the competitor. We observed similar results in the two-to-one scenario. The burst absorption capacity of the Trident 3 platform in the four-to-one scenario exceeded 77 – 204% over that of the competitive platform over all packet sizes.

Figure 10. Four to One Burst Absorption – Broadcom Trident 3 versus Competitor



Source: Enterprise Strategy Group

Why This Matters

Organizations that support enterprise-scale and cloud workloads exhibiting bursty traffic patterns, such as Hadoop and MapReduce, must design their data center networks with adequate switch buffer capacity to maintain acceptable and predictable application performance. Ideally, the switch buffer capacity will strike the appropriate balance between application performance and buffer capacity for high burst absorption.

ESG verified that the Trident 3 platform provides burst absorption capability for multiple traffic flows directed at one switch output port. Not only did the Trident 3 platform provide sufficient buffer capacity, but the burst absorption increased as we increased the packet size in both traffic scenarios. This confirmed that the Trident 3 platform is able to share its centralized buffer across all switch ports and has a much more robust burst absorption capability, which translates to predictable performance for unpredictable workloads.

Performance and Latency with Packet Processing Features Enabled

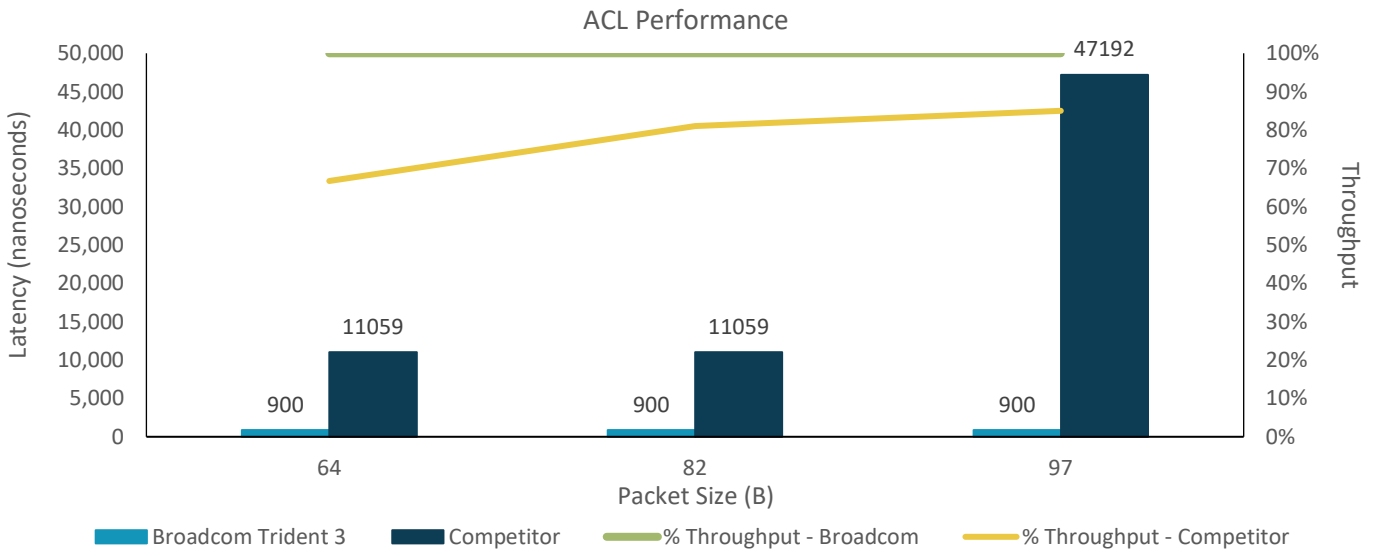
RFC benchmark tests traditionally measure Layer-2 switching and Layer-3 routing performance with no additional requirement for any switch or router functionality beyond the basic forwarding action. In real-world data center networks, significant packet processing operations are performed on any given packet, including VLAN lookups, tunnel encapsulation or de-encapsulation(e.g., VXLAN or IP-in-IP headers), ECMP hashing/load balancing, QoS processing/queue assignment, differentiated services code point (DSCP), and ingress and egress security policy enforcement via Access Control Lists (ACLs). Many of these functions require both inspection and extraction of fields from ingress packet headers, as well as editing of egress packet headers. For example, ACL rules provide a basic level of network security and traffic prioritization that is required in data center networks, where a lookup of packet header fields results in a permit or deny decision for packet forwarding on the destination egress port.

On some switch platforms, these functions are performed in software or in hardware at sub-line rate (using auxiliary processing paths), which can impact performance when these rules are used to permit and restrict data flowing through network interfaces. In ESG's opinion, an optimal switch architecture would perform all packet processing in hardware at line rate. Less than line rate performance when processing ACLs is a well-known attack vector for denial-of-service (DoS) attacks, and a typical response to a DoS attack is to drop the offending traffic. Organizations should architect their data center networks with switches that can minimize the impact on the performance of commonly deployed packet processing feature sets, such as ACL processing and DSCP modification.

ESG Testing

ESG proceeded to evaluate switching performance and latency under congestion with basic packet processing features, such as ACLs and DSCP modification enabled. ESG first populated a total of 24 ACL rules on both the Trident 3 and the competitor switches—where 12 of those ACL rules applied to all ports (permit/deny) while the other 12 rules were specific to select ports (modify DSCP of the packet). We then randomly chose six 100GbE adjacent port pairs (e.g. ports 1 and 2, or 31 and 32) between which we would stream test traffic. We generated traffic streams containing 64, 82, and 96 B packets, and then measured latency and throughput (see Figure 11).

Figure 11. ACL Performance for Traffic Streamed between Port Pairs



Source: Enterprise Strategy Group

As seen in Figure 11, the Trident 3 platform maintained a latency of 900 nanoseconds and line rate throughput between all port pairs across all packet sizes down to 64 bytes. The competitive platform's latency was 12x than that which Trident 3 achieved in the 64- and 82-byte packet size and 52x at a 97-byte packet size. Surprisingly, the competitive switch did not achieve line rate throughput in any test where basic ACL and DSCP rules were applied to a small subset of the device's ports as described in this test. The competitor's switch sustained lower than line rate performance when ACL rules or DSCP modifications were being executed. ESG observed that this ultimately caused non-offending traffic packets to be dropped inside the switch when the internal switch buffers filled up—which occurred silently with no notification to the user. The Trident 3 platform demonstrated the same line rate characteristics invariant with the number and type of packet processing rules applied within the switch.

Why This Matters

Managing data center network security with ACL rules is essential, yet organizations risk delivering unpredictable application performance when enabling rules on data center switches that do not design for feature-invariant packet processing and buffering. Organizations must architect their data center networks with a switch that can sustain the same predictable high performance regardless of how many features are activated at a given time on the switch.

ESG validated that the Broadcom Trident 3 switch platform can achieve line rate throughput on 100GbE links when ACL rules are enabled, DSCP modifications are executed, or any other feature is implemented in a single pass of the switch pipeline. We observed this behavior when generating incast traffic, containing traffic of increasing packet size, from multiple ports to a single output port. This is critically important as Trident 3's programmable architecture enables in-field upgrades of packet processing features, and by design any such feature upgrade will execute at the same deterministic line rate performance. The Trident 3 also maintained consistent, bounded port-to-port latency of 900 nanoseconds across all packet sizes regardless of features enabled. On the other hand, the competitive switch platform achieved substantially less than line rate performance in the presence of a very basic unit set of ACL rules for filtering and DSCP modification, with an offered load as low as two 100GbE ports running on the switch with much higher latency. Organizations can provide consistent application performance without sacrificing network functionality using the Trident 3 platform.

Bandwidth Fairness

When a data center switch handles traffic in a "many-to-one" environment, oversubscription at the output is a common occurrence. Organizations are correct to expect that switches will allocate available bandwidth of the output port fairly and evenly among the incast application streams, assuming that they all have the same priority. Bandwidth fairness maintains individual application performance and SLA guarantees when the switch is oversubscribed.

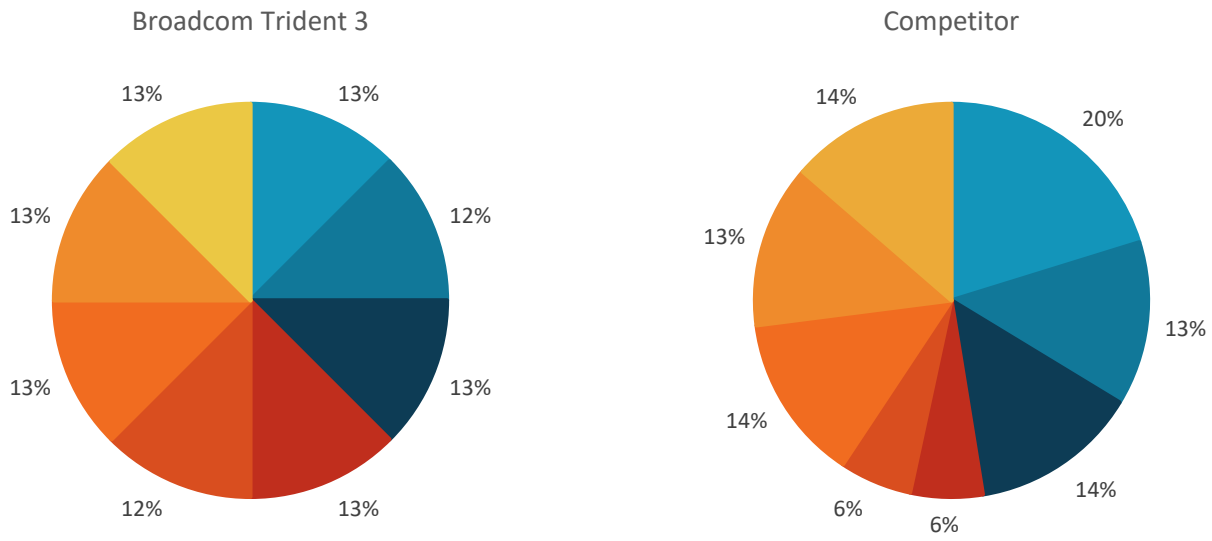
ESG Testing

For this test, ESG measured port bandwidth fairness under congestion while ACL rules were enabled on the switches. We set up ACL rules to modify DSCP based on QoS classification rules and drop a small percentage of the traffic. We chose random source ports from which traffic was sent and measured the bandwidth distribution, varying both the number of source ports and frame size of the packets. We generated traffic under the following scenarios:

- Six source ports to one destination port (6-to-1), frame size = 64 B and 1,280 B.
- Eight source ports to one destination port (8-to-1), frame size = 64 B and 1,280 B.
- 16 source ports to one destination port (16-to-1), frame size = 64 B, 97 B, and 1,280 B.

In all scenarios, we found that the Trident 3 platform distributed the available bandwidth across all source ports evenly regardless of the order or grouping of the ports, while the bandwidth distribution varied for the competitor in every scenario. Figure 12 shows the results of testing the 8-to-1 scenario, using 64B frames. While the Trident 3 platform distributed bandwidth equally among the source ports (13%), the competitor's switch varied its bandwidth distribution from 6-20% Conversely, the competitor's switch platform allocated unequal bandwidth across all ports. We observed similar results when we changed the frame size to 1,280 B.

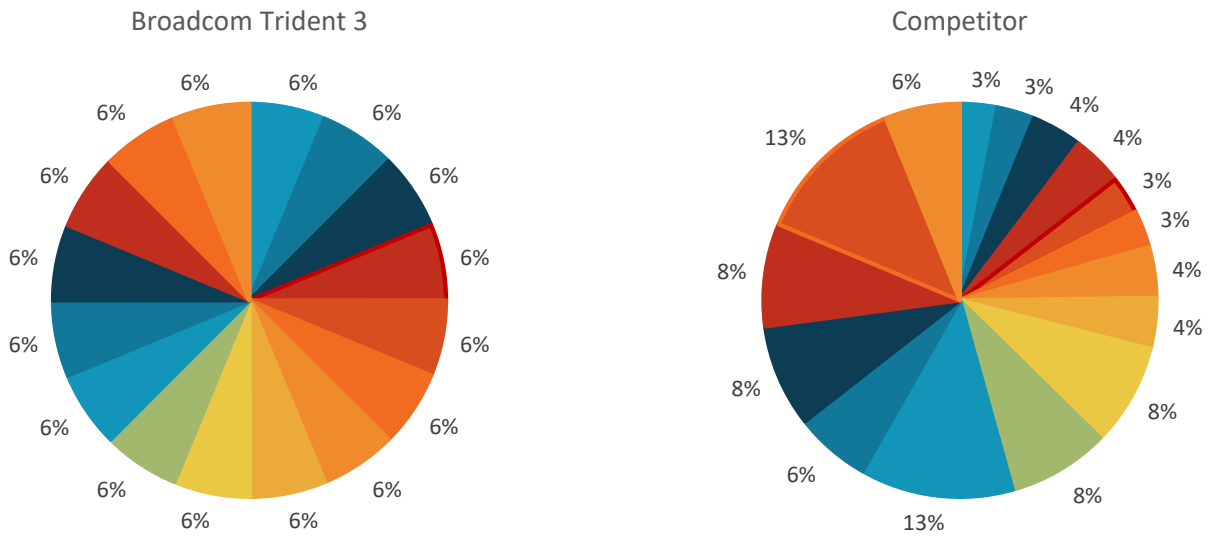
Figure 12. Bandwidth Fairness, 8-to-1 Scenario, Frame Size = 64 B



Source: Enterprise Strategy Group

Figure 13 shows the results of testing the 16-to-1 scenario, using 64B frames. Again, the Trident 3 platform distributed the bandwidth evenly (6%) among the source ports, while the competitor distributed allocated bandwidth ranging from 3% to 13%. As in the 8-to-1 scenario, we observed similar results when we changed the frame size to 97 B and 1,280 B.

Figure 13. Bandwidth Fairness, 16-to-1 Scenario, Frame Size = 64 B



Source: Enterprise Strategy Group

ESG particularly noted that Trident 3's even distribution pattern occurred regardless of the combination of source ports in all traffic scenarios. Conversely, the competitor's distribution varied repeatedly for all runs of all test traffic scenarios. We see this as further confirmation that the Trident 3 switch, with its fully shared and physically centralized buffer and fully in-order feed-forward packet processing pipeline, is architected for fairness.

Why This Matters

Streaming traffic in data center networks from multiple input switch ports to one output port can adversely affect individual application performance, as the multiple streams battle for available bandwidth resources. Organizations can alleviate this oversubscription by employing a switch that allocates available bandwidth of the output port fairly among the incast application streams, especially when applying ACLs for security and QoS classification.

ESG verified that the Broadcom Trident 3 platform allocated output port bandwidth evenly across multiple "many-to-one" traffic test scenarios and packet sizes, while the competitive switch did not divide bandwidth evenly in any test scenario. The fair-allocation approach of Trident 3 helps organizations to maintain predictable performance of multiple applications, making it easier for organizations to provide a better user/customer experience.

The Bigger Truth

While organizations are embracing the use of hybrid clouds to enhance their agility and responsiveness to their users and customers, they continue to be challenged by managing and orchestrating resources to extract the most value. ESG asked organizations to name challenges being faced by their networking teams. Provisioning network services for new or upgraded applications (25%) and providing predictable network performance (24%) are both in the top five most-cited challenges.[6]

Broadcom's Trident 3 platform is designed to address these challenges while protecting customers' switch investments with data plane programmability. Trident 3 technology is available to data center, enterprise, and service provider networks transitioning to high-density 10/25/100G Ethernet.

ESG validated that switches built on the Broadcom Trident 3 platform can easily service demanding "many-to-one" environments found in enterprises and cloud service providers with outstanding burst absorption, efficient congestion handling, zero performance or latency impact when deploying ACLs or other packet processing features, and bandwidth fairness under real-world congestion and load, compared with a current-generation switch based on another vendor's silicon. The Trident 3 based Arista 7050CX3 Switch we tested outperformed the competitive switch in every test and every scenario.

Broadcom continues to address the challenges of maintaining predictable application performance in large, complex network environments. Trident 3 is the latest in a line of highly successful switch products that are deployed in cloud service providers and enterprises across the globe. If your organization is looking to achieve efficient, scalable, predictable performance and ultimately a lower TCO from its network environment, it would be smart to take a close look at switches based on Broadcom's Trident 3 platform.

[6] Source: ESG Master Survey Results, *Trends In Network Modernization*, November 2017.

Appendix

Table 1. TCP Test Configuration

Configuration	
Arista 7050CX3	**Competitor**
Egress alpha has been changed from default value 2 to 8.	The entire shared buffer has been allocated to one service pool-0; Dynamic buffer has been enabled; Ingress and egress dynamic quota has been programmed to 10.
Max buffer usage during the TCP transfer was 27.6MB. (Total buffer 32MB; 4.4MB reserved for ingress min, PG headroom & egress min)	Max buffer usage during the TCP transfer was 9.4MB. (Reserved the same 4.4MB for ingress min, PG headroom & egress min)
Total input reserved cells: 8370 • Ingress generic frames: 3060 • Ingress control frames: 510 • In-flight frames (PG headroom): 4800	
Total output reserved cells: 9587 • Unicast queues: 2176 • Multicast queues: 2176 • Control-frame queues: 510 • High-priority CPU queues: 2880 normal-priority CPU queues: 1845	
Cell size = 256Byte	
Reserved Space: (Ingress reserved + egress reserved)*256 = 17957*256/1024 = 4.4MB	

Source: Enterprise Strategy Group

TCP + RoCEv2 Test Configuration:

Arista Test Configuration:

PG1 configured as lossy & PG3 configured as lossless; RoCEv2 data traffic has been assigned to priority 3, and CNP messages are assigned to strict priority queue 7, random-detect ECN minimum-threshold programmed to 153600 bytes maximum-threshold programmed to 1536000 bytes. ECN & PFC enabled on priority 3.

Competitor Test Configuration:

Pool-0 configured as lossy; pool-1 configured as lossless; RoCEv2 data traffic has been assigned to priority 3 and CNP messages are assigned to strict priority queue 7, random-detect ECN minimum-threshold programmed to 153600 bytes maximum-threshold programmed to 1536000 bytes. ECN & PFC enabled on priority 3.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides market intelligence and actionable insight to the global IT community.

 www.esg-global.com

 contact@esg-global.com

 P. 508.482.0188