Arista 10G switch: Fast and flexible

Linux-based core switch sets records with high density, blazing performance

By David Newman, Network World
March 26, 2012 12:07 AM ET

Packing 384 10G Ethernet ports into an 11-rack-unit form factor is only the beginning for Arista Networks' DCS-7508 data center core switch.

In this exclusive Clear Choice test, the 7508's performance set one high-water mark after another. It switched 5.7 billion frames per second, the highest throughput ever seen in a Network World test. It moved multicast traffic to more than 4,000 groups on all ports, another record for a modular switch. And it ran at wire speed in almost every case except when we deliberately congested the switch, and there it buffered up to 83MB per port.

Arista again looks to buck Cisco, others in data center

On top of its impressive performance stats, the 7508 also showed off multiple redundancy and load-balancing mechanisms and recovered quickly from failures. And it did all this running on Linux, with all the extensibility that comes with Unix-like operating systems.

For network managers wondering why they'd need this much port density: It might not happen this quarter or next, but 10G Ethernet is already well on its way to replacing gigabit as the pervasive data center transport.

The signs are all there: Intel is about to ship 10G-equipped server motherboards in quantity. A gaggle of storage vendors already send iSCSI traffic over converged 10G Ethernet backbones. And faster 40G and 100G Ethernet uplinks are starting to appear. Given the usual multi-year depreciation cycles for networking gear, high-density switches like Arista's 7508 are starting to make sense as data center workhorses.

A well-considered design

Beyond its high density, the 7508 offers some seriously nice hardware. Airflow is excellent, thanks to fans on each fabric card and a lattice inside the chassis. Power management allowed us to drive all 384 ports at full tilt using just two power supplies, instead of the standard four.

The design smarts extend to Arista's EOS software. Underneath a Cisco IOS-like command-line interface (CLI), EOS offers modularity and a complete Linux command set. Modularity, also seen in Cisco's NX-OS for Nexus switches, means the failure of any one process doesn't take down the entire system, as it would in monolithic designs like Cisco's mainline IOS. We verified this by intentionally killing EOS processes and watching them automatically respawn; there was no effect on other system functions.

But EOS's greatest strength is its extensibility. Because it's Linux under the hood, EOS is highly customizable. The vendor provides source code for its CLI and many other (though not all) system components and actively encourages customers to hack its code.

To demonstrate EOS extensibility, Arista recently gave a group of its system engineers, most of whom aren't programmers, 24 hours...
to get new projects running. The system engineers produced dozens of tools, ranging from useful (say you're on a Mac, and want Growl notifications when particular interfaces go up or down) to plain crazy (Pandora radio running on the switch, fed to external speakers via a $20 USB sound card). Essentially, any task that can run on Linux can probably run on EOS.

Also, a single EOS binary image runs on all Arista switches, both core boxes like the 7508 and various top-of-rack systems. Having one system image eliminates the feature and command mismatches sometimes seen across competitors' switch product lines.

**Wire speed all the time**

We assessed the Arista switch mainly in terms of performance, with a long battery of tests intended to determine the system's limits (see "How We Did It").

Describing the 7508's unicast throughput is easy: It always went at wire speed. With the Spirent TestCenter traffic generator/Analyzer blasting away in a fully meshed traffic pattern on all 384 10G Ethernet ports, the 7508 didn't drop a single frame in any of our unicast tests. At rates of up to 3.832 terabits per second, the 7508 was perfect, both in layer-2 and layer-3 configurations.

The 7508 is also non-blocking when handling multicast traffic, provided frame lengths are 70 bytes or longer. With minimum-length 64-byte frames, the system's throughput is equivalent to 92.588% of line rate. For every other frame size we used, the system again forwarded all traffic at wire speed without loss, both in layer-2 and layer-3 setups. (We've added 70-byte multicast tests to show the system will forward at line rate when frames are that long or longer.)

The layer-2 and layer-3 multicast tests also involved very high control-plane scalability. We ran the layer-2 tests with 383 receiver ports all subscribed to 4,095 multicast groups. That's much higher than previous Network World tests we've done involving modular core switches; typically those tests involved 1,024 or fewer groups.

In the layer-3 case, subscribers on 383 receiver ports joined "only" 512 multicast groups, but then again the system also ran a different PIM-SM multicast routing session on each of 384 ports.

Latency was generally low and consistent. Layer-2 and layer-3 delays were virtually identical. When handling unicast traffic, the 7508 delayed traffic, on average, by less than 9 microseconds with frame lengths of up to 1,518 bytes; with jumbo frames, average delay was around 13 microseconds.
One exception: Maximum latency was substantially higher with short and medium-length unicast frames than long ones, reversing the pattern often seen with Ethernet switches where delay increases with frame length. This was only seen in unicast tests.

In the multicast tests, both average and maximum latency were significantly lower than unicast, regardless of frame size (see Figure 2). This is important for the growing number of users who make heavy use of multicast in the data center (for example, many stock quote and trading applications used in the financial services industry).

Here, average delays were less than 5 microseconds for frame lengths of 1,518 bytes or shorter, and around 6 microseconds with jumbo frames. Again, there were no significant differences between layer-2 and layer-3 test cases. And unlike the unicast tests, maximum multicast latency was not significantly higher than average latency.

**Maximizing uptime**

While high performance is essential for core switches, high availability is at least as important. The 7508's highly redundant design extends to many components: There are six fabric cards, each with their own fans along with multiple power supplies and redundant supervisor modules.

To measure the time needed to recover from the loss of a redundant fabric module, we physically removed one of the fabric cards while offering unicast 64-byte unicast frames to all 384 ports. By dividing frame loss into frame rate, we determined that the system recovered in 31.84 microsec.

That’s not instantaneous, but it’s still pretty fast; performance of many enterprise applications, especially those running over TCP, won’t degrade until disruptions run up into the milliseconds. Arista says the 32-microsec figure represents only those frames that were “in flight” between transmit and receive ports at the time we pulled the fabric module.”
Power consumption is another key consideration, especially as data centers scale up to support hundreds or thousands of 10G Ethernet ports. We measured power usage in two modes: Fully loaded, with traffic from the Spirent test instrument offered to all 384 ports at line rate, and 50% loaded, with only half the line cards inserted (but still offering traffic at line rate to all those cards). In these and all other tests, the switch used direct-attached copper (DAC) cables and transceivers.

When fully loaded, the 7508 drew 4,358 watts, or about 11.3 watts per port. With only half the ports inserted, the system used 1,598 watts, or about 8.3 watts per port. The fully loaded number is a worst-case scenario, while the 50% case is more representative for many enterprises, especially those who don't populate all line cards on day one.

Buffering capacity

Arista requested that we measure the burst-handling characteristics of the 7508, specifically to verify Arista's claim that the system can buffer up to 50MB per port. Handling short, high-speed bursts of traffic is especially important in many high-performance computing applications, where multiple senders may present data to the same receiver at the same instant.

While many vendors talk about microbursts in marketing collateral, there isn't yet an industry-standard method of measuring burst handling. We used a couple of methods here: First, with a 2:1 oversubscription of steady-state traffic, where we offer traffic to 256 ports, destined to all the remaining 128 ports. That's a simple buffer test and should work regardless of burst length.

Second, to assess microburst buffering, we sent bursts of varying sizes at line rate from multiple sources to the same destination port at the same time. By experimenting with different burst lengths, we found the maximum microburst length the system could buffer without frame loss.

While the microburst method is arguably more interesting due to the dynamic nature of enterprise traffic, the first method produced a surprising result.

Faced with a 2:1 oversubscription, the switch initially dropped nearly 60% of traffic rather than the expected 50% or less, meaning it wasn't buffering at all. Arista attributed the loss to a combination way the 7508's virtual output queuing (VOQ) works and the totally nonrandom order of our test traffic. After setting the VOQ scheduling to a non-default setting ("petra voq tail-drop 2"), packet loss fell to 50% or less, as expected.

Another lesson learned, both in steady-state and microburst buffering tests, is that buffer capacity depends in part on the number of senders and receivers involved. When we ran the microburst test with 256 transmitter and 128 receiver ports, the 7508 buffered up to 83.49 megabytes on each receiver port with zero frame loss, well in excess of Arista's claim of 50MB/port. That's equivalent to around 56,300 1,518-byte frames.

However, if we ran the same test with 383 transmitters all aimed at one receiver, the largest amount of traffic that could be buffered without loss was much lower, around 6.85MB (or around 4,600 1,518-byte frames).

The results differ because of the 7508's VOQ and credit-based architecture. When frames enter the switch, it will allocate buffers and issue forwarding credits if, and only if, sufficient resources exist to forward the traffic. The higher the ratio of transmitters to receivers, the greater the imbalance between requested and available resources. In this light, Arista's 50MB claim is really a composite figure, one that assumes transmit and receive port counts are somewhere between the best- and worst-case scenarios.

Boosting bandwidth with MLAG

Mention Spanning Tree to any data center architect, and you're likely to be greeted with a scowl. Besides cutting bandwidth in half with its active/passive design (where 50% of links and ports sit idle), the protocol can be tricky to troubleshoot, especially when multiple VLANs are involved.

Many switch vendors, including Arista, have methods to eliminate spanning tree, in turn enabling larger, faster, flatter data center designs. While all the various approaches are proprietary, Arista's approach, called multi-switch link aggregation (MLAG), starts with the IEEE 802.3ad link aggregation specification.

With MLAG, each attached server or switch can use standards-based link aggregation to form a virtual pipe with two physical Arista switches, and see those switches as one logical entity. MLAG works with any device that uses the link aggregation control protocol (LACP). It doubles available bandwidth with its active/active design, while still preventing loops like spanning tree.

We verified MLAG functionality with two pairs of eight-port MLAG trunks, each split across two 7508 switches. First we verified MLAG could forward across all ports by offering bidirectional test traffic from 256 hosts emulated by the Spirent test instrument. MLAG perfectly distributed traffic from these hosts, with each MLAG port forwarding the exact same number of frames.
To test MLAG resiliency, we then rebooted one of the 7508s, forcing traffic onto the remaining ports in the MLAG trunk. By deriving cutover time from frame loss, we determined that it took 158.81 milliseconds for the system to resume forwarding all traffic without loss. In comparison, Rapid Spanning Tree typically takes 1 to 3 seconds to converge after a similar failure.

While MLAG represents an interesting approach in that it's based on a simple and well-understood standard, there's still a proprietary component: The two MLAG peers must be Arista switches, which share learning and state information using a proprietary protocol. For the devices attached to the peers, however, it's just standards-based LACP.

**OSPF Scalability**

We've already used multicast routing in the throughput and latency tests, but we also assessed unicast routing with tests of OSPF routing scalability and equal cost multipath (ECMP) capabilities.

To measure routing capacity, we configured the Spirent test instrument to advertise progressively larger numbers of networks over OSPF, and then determined whether the 7508 could forward traffic to all these networks without loss. The largest number of routes the system could install in its hardware forwarding tables was 13,500.

That's certainly more than enough for most data centers, though routing capacity is also a topic of some specsmanship among vendors. In its internal testing, Arista says the 7508 learned 42,500 routes when the advertised networks primarily used /24 and /32 prefix lengths (we used an Internet distribution in our tests, modeling the prefix lengths seen on Internet backbones).

We've previously tested other modular systems with much higher capacity; the record-holder is still the 3Com (now HP) S7906E, which learned more than 200,000 routes over OSPF. Further, Internet-facing routers running BGP will need far higher capacity; one view of a full BGP table is about 400,000 routes, and many sites will have multiple BGP feeds.

But this is a core data center switch, not a core router. Distinctions between 10,000 and 200,000 or 400,000 routes or more simply aren't meaningful for the vast majority of enterprises running OSPF; even in very large data centers, 10,000 routes is plenty.

Another part of the EOS routing feature set is the ability to load-balance using OSPF ECMP, distributing traffic on as many as 16 links. Most competitors support a maximum of eight-way links when doing ECMP.

In our tests, the 7508 spread traffic pretty evenly across all 16 OSPF sessions. We saw a variation of around 1.215% between the least and most used OSPF ports.

Our complaints with the 7508 are relatively minor. Unicast maximum latency is higher with short- and medium frames than longer frames. The 7508's hardware routing table doesn't scale as high as some other modular switches we've tested, but that's more a battle for bragging rights among vendors than a concern for most data centers. EOS doesn't yet support hitless failovers or IPv6, though Arista says it's working on both. And the 7508's buffering capabilities differ depending on traffic patterns, but that's true of all credit-based systems.

Speed and flexibility trump these complaints. As enterprises roll out more 10G Ethernet in the data center - and that will happen soon - they'll need high-density ways to tie everything together. The Arista 7508 is ready to do that, today.

**Thanks**

Network World gratefully acknowledges the assistance of Spirent Communications, which supplied its Spirent TestCenter traffic generator/analyzer and 10G Ethernet HyperMetrics dX modules for this project. Spirent's Jurrie van den Breekel and Timmons Player also provided technical support. Thanks also to Fluke Corp. for supplying a Fluke 335 clamp meter used in power measurement.

Newman is a member of the Network World Lab Alliance and president of Network Test, an independent test lab and engineering services consultancy. He can be reached at dnewman@networktest.com.

Read more about lans & wans in Network World's LANs & WANs section.
