ENTERPRISETECH
INSIDE ADVANCED SCALE CHALLENGES

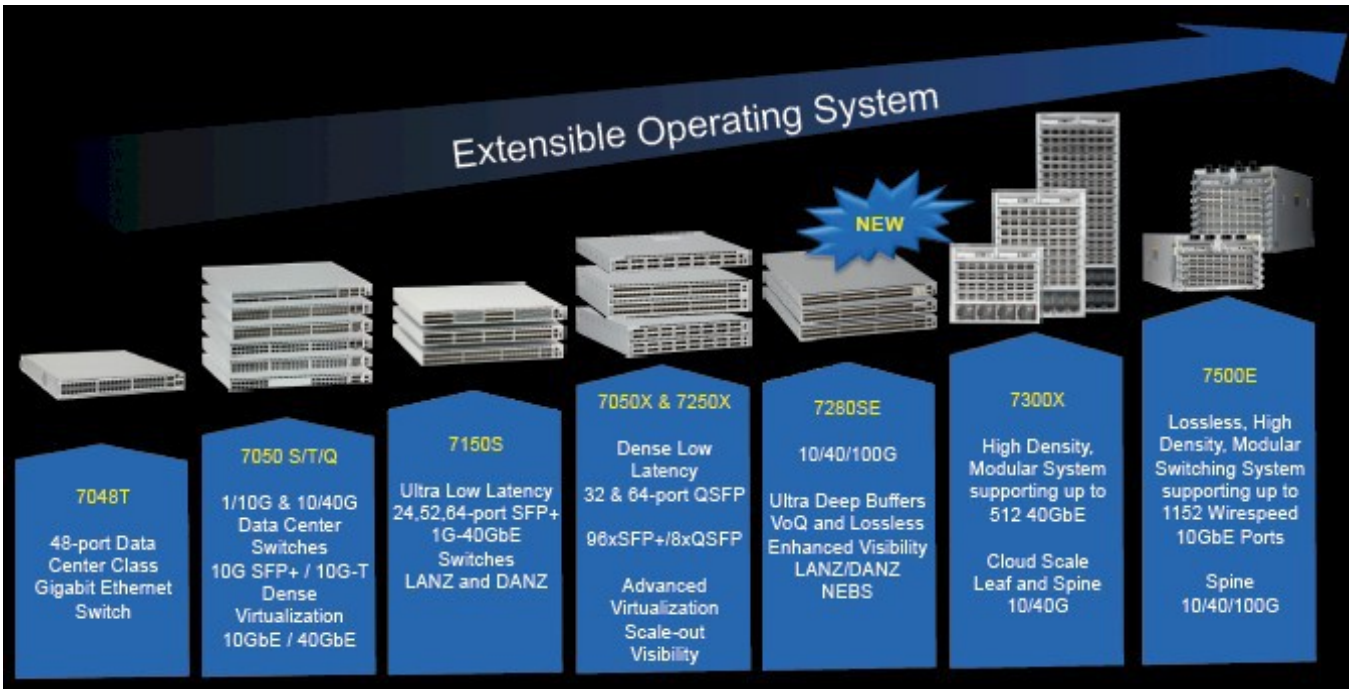
Arista Cranks Leaf Switches To 100G For Big Data, Storage

July 15, 2014 by Timothy Prickett Morgan

The emergence of big data, storage, and streaming media applications that have bursty behavior on the network and the prevalence of 10 Gb/sec network adapters on servers is choking the uplinks on top of rack switches. Switch makers are being pushed to offer more uplink bandwidth and much fatter packet buffers, and Arista Networks is jumping out ahead of the pack and offering 100 Gb/sec uplinks in a new family of switches.

The adoption of 100 Gb/sec protocols in routers and switches follows the traditional path, starting out in core routers first many years ago. These core routers, explains Anshul Sadana, senior vice president of customer engineering at Arista Networks, had a very high price tag and low port density but offered long distance links. More recently, spine switches for the network backbone (hence the name) have come out with datacenter-spanning distances for their links and higher density line cards that are also less costly than the router ports were. And now, top-of-rack switches are getting 100 Gb/sec uplinks so they can feed up to the spine at higher rates.

The shift to faster uplinks in top-of-rack switches is progressing faster than the shift to faster ports on servers and their companion ports in the switches they link to. Arista got out in front of the 10 Gb/sec wave when it launched in 2008 with its first switches. It delivered 10 Gb/sec switches with 40 Gb/sec uplinks in 2011, and now it is getting 10 Gb/sec switches out with much deeper packet buffers and a choice of 40 Gb/sec or 100 Gb/sec uplinks.



The new 7280SE series switches are based on Broadcom's "Dune" network ASICs, and in particular use the "Arad" variant of the chip. This ASIC can process 900 million packets per second, and its throughput ranges from 1.28 Tb/sec to 1.44 Tb/sec, depending on the mix of ports in the fixed port switch in the 7280SE family. The machines offer a port-to-port latency of 3.8 microseconds, which means they are not aimed for low latency work but are plenty zippy enough for cloudy infrastructure, data analytics, and compute clusters.

The switches are also equipped with a quad-core X86 processor for network functions pulled into the switch, and have 4 GB of system memory, 4 GB of flash storage memory, 120 GB of solid state drive capacity, and importantly 9 GB of packet buffer memory.

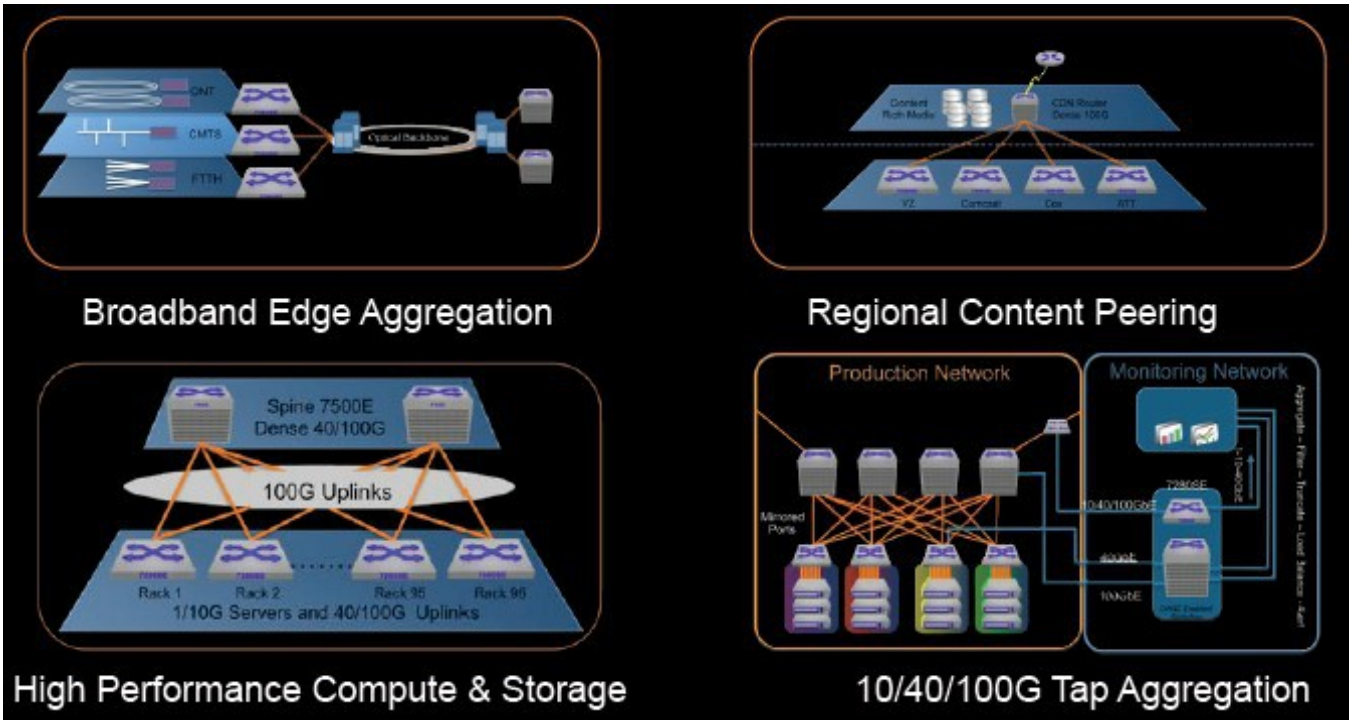
That is around 1000X the amount of packet buffering in a typical leaf switch these days, says Sadana, and Arista Networks is not doling this as some kind of engineering publicity stunt. The reason for the deep packet memory is because the Ethernet protocol underpins a lot of object and file storage in the datacenter, driving all manner of workloads, and if a packet gets lost, retransmission delays can cause huge latencies in applications because the whole packet needs to be resent. Keeping scads of packets in the cache means a server that didn't get its data the first time can often grab it from the switch's packet buffer the second time without having to traverse the network again.

Here is a benchmark test performed by Datasift, a social media data analytics service provider, comparing switches with deep packet buffering to those without running its workloads:

As you can see, there is a big drop in latency for this workload, which is centered around big data and storage. eBay was also trotted out in the announcement as an early customer for the 7280SE.

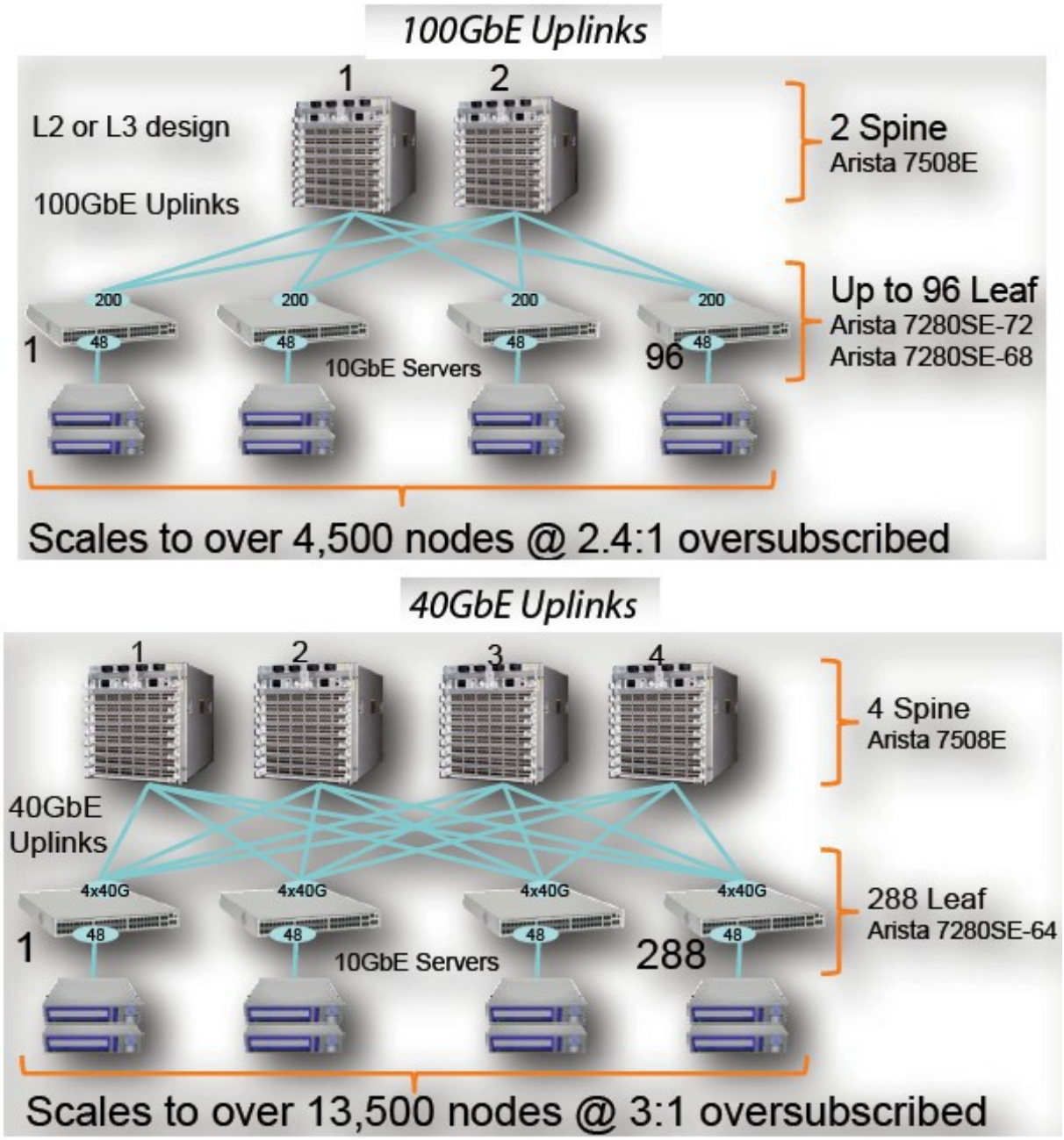
The new 7280E switches not only have deep packets, but also support the VXLAN Layer 3 overlay protocol for Layer 2 networks that VMware, Arista Networks, and others created to allow for cloud-scale networks and seamless live migration of virtual machines across what are physically distinct yet logically united networks. The 7280SE family also supports 36,000 entries on its access control list, compared to somewhere between 2,000 and 4,000 entries on the typical top of racker. With that size ACL, the 7280SE can function as a router with Layer 3 functions hanging off of the core router, doing some of its work. The new switches also support the OpenFlow 1.3 protocol, joining the 7050, 7050X, and 7500 families which, with the 7280SE represent about 75 percent of the switches sold by Arista Networks today.

While there are plenty of applications that still do not push 2 Gb/sec or even 3 Gb/sec data rates out of their servers, Sadana says that some storage and big data workloads can easily push 10 Gb/sec per port and that with only four such I/O-intensive servers, a 40 Gb/sec uplink on a top of rack switch can quickly become flooded. If you can get a near-lossless Ethernet network, you can also dump the Fibre Channel switches that link multiple servers to storage area networks and just use Ethernet for everything. In general, says Sadana, two ports of 100 Gb/sec is going deliver more efficient uplink flow than four ports at 40 Gb/sec because it is more difficult to choke one of the faster uplinks. You also get an extra 40 Gb/sec of aggregate bandwidth moving up to the core or spine switches, obviously, too. (Why switch makers didn't put five 40 Gb/sec uplinks in their machines if the bandwidth was there is a bit of a mystery. . . .)



There are three different models of the 7280SE switches. The 7280SE-64 has 48 SFP+ ports running at 10 Gb/sec and four QSFP+ ports running at 40 Gb/sec. The main difference between this box and other existing switches with 40 Gb/sec uplinks from Arista Networks is the deep packets, expanded ACL entries, and OpenFlow 1.3 support. This machine can handle 1.28 Tb/sec and will be available in the third quarter at a cost of $40,000.

The second machine is the 7280SE-68, and this one has 48 downlink ports at the 10 Gb/sec speed a server expects and two QSFP100 ports running at 100 Gb/sec. It has 1.36 Tb/sec of switching bandwidth across the Dune ASIC. The pricing for this box has not been set yet, since it will not be available until late in the fourth quarter when the transceivers are ready. The expectation is that this switch will cost somewhere around a few thousand dollars more than the 7280SE-64.



If you can't wait for 100 Gb/sec uplinks, then Arista Networks has forged the 7280SE-72 just for you. This machine has the base 48 SFP+ downlinks running at 10 Gb/sec, but has 1.44 Tb/sec of bandwidth and two 100 Gb/sec MXP ports with integrated transceivers. With the those two integrated transceivers, the switch costs $50,000. In other words, not too much more than the switch above that doesn't have the integrate transceivers.

In addition to the new switches, Arista Networks is announcing a new universal optics that allows for 40 Gb/sec Ethernet to run over a single pair of either single-mode or multi-mode fiber – the cables that existing in a lot of datacenters today. With cabling costing about 35 percent of the total networking bill, the ability to shift to 40 Gb/sec without having to recable with fatter and more expensive cables is a big deal. The single-mode version of the 40G-UNIV transceiver can reach distances of up to 500 meters, and the multi-model version can stretch to 150 meters. This, says Sadana, addresses about 90 percent of the distances that customers building leaf/spine networks and is

interoperable with 40GBASE-LR4 and LRL4 transceivers that also stretch to 500 meters.

Arista Networks is also expanding the use of its Smart System Upgrade feature, which allows for a single switch to be upgraded without taking it offline. Eventually. The SSU feature is being rolled out in phases, with the first phase requiring only 30 seconds of downtime. (This is good because a switch operating system upgrade usually takes 2 hours and can be plagued by human error.) In the fourth quarter, the second phase of SSU will be available and it will allow for a "hitless upgrade" of the underlying EOS operating system on the switch. The SSU feature will be available on the 7050X switches to start, which are commonly deployed by enterprises and cloud builders alike. The new 7280SE switches do not have SSU support, but Sadana says it can be added at some point in the future. It is just a matter of taking into account the differences in the chips in the boxes and their drivers.