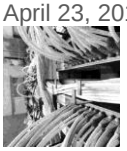
Dell Strikes First, Cuts Deep With Tomahawk Switches

April 23, 2015 Timothy Prickett Morgan



The next generation of higher bandwidth and lower cost Ethernet switching for the datacenter is beginning, and privateer Dell is coming out swinging with new open switches based on the "Tomahawk" network ASICs from Broadcom. Dell is pitting itself against rival Hewlett-Packard with its Accton partnership, also based on the Tomahawk chips, and if the rumors are right, impending switches from Arista Networks that are based on the XPliant family of chips now controlled by Cavium Networks. Interestingly, Dell is going to be very aggressive about pricing.

The Tomahawk ASICs were unveiled by Broadcom last summer, and embody an Ethernet standard called 25G that was pushed by hyperscalers Google and Microsoft, switch chip makers Broadcom, Mellanox Technologies, and XPliant (which was acquired by Cavium after dropping out of stealth), and upstart switch maker Arista Networks.

The current crop of Ethernet ASICs like the "Trident-II" from Broadcom have serializer/deserializer (SerDes) circuits that run at 10 GHz, and four lanes yield 40 Gb/sec of bandwidth and ten lanes yields 100 Gb/sec of bandwidth. With the 25G effort, vendors wanted the IEEE to adopt SerDes with 25 GHz lanes, allowing for a 100 Gb/sec link to be made with four links instead of ten. This would allow a 25 Gb/sec network adapter card for servers that are already swamping 10 Gb/sec adapters and for a switch to have downlinks running at 25 Gb/sec that have 2.5X the bandwidth of the 10 Gb/sec link with somewhere around half the power per port and around 1.5 times the cost per port, and therefore much lower cost per bit transferred.

The Tomahawk chips are but one implementation of the 25G spec, which has now been adopted by the IEEE after it was apparent that switch makers and switch customers were not going to sit around and wait for its approval. The Trident-II ASIC, which was launched in August 2012, had a total of 128 SerDes and had an aggregate switching bandwidth of 1.28 Tb/sec and a 7.6 MB packet buffer. As *The Platform* has already reported, Broadcom just this week unveiled the Trident-II+ follow-on, which has a slightly deeper packet buffer at 10 MB and some tweaks that make VXLAN layer 2 overlays for cloudy networks run a lot more efficiently and without putting a burden on hypervisors and virtual switches running on servers. But this is just a tweak to the existing Trident-II design to make it more compelling for enterprises that are largely stuck at Gigabit Ethernet on their networks to move on up to 10 Gb/sec speeds.

Large enterprises, hyperscalers, and some HPC shops that prefer Ethernet over InfiniBand need something more – and more cost effective per bit than 10 Gb/sec Ethernet – for their networks, and the Tomahawk ASICs were designed explicitly for this purpose. The top-end Tomahawk ASIC will have a whopping 680 Mb (85 MB) packet buffer and support 3.2 Tb/sec of aggregate switching bandwidth. It can drive 32 ports at 100 Gb/sec, 64 ports running at either 40 Gb/sec or 50 Gb/sec, or 128 ports running at 10 Gb/sec or 25 Gb/sec. The Tomahawks will deliver port-to-port hop in 400

nanoseconds or less, which is pretty fast; the chips support RDMA over Converged Ethernet (RoCE) direct memory access technology to get that low latency.

The Z9100-ON 100 Gb/sec switch from Dell

Bandwidth and low latency are important, but what is interesting – and expected – about Dell is its aggressiveness on the pricing of 100 Gb/sec switching.

Arpit Joshipura, vice president of product management and strategy at Dell Networking, tells *The Platform* that it costs around $5,000 per port for a 100 Gb/sec switch these days based on earlier generations of ASICs, and further that this price does not include the cost of the optical connectors and the cabling, which is not cheap. (In hyperscale and HPC datacenters, the cost of the cabling usually exceeds the cost of the switches, so this is not a trivial matter.) Joshipura says that the plan with its first Tomahawk-based switch, the Z9100-ON, is to get that price per port down below $2,000 per port *including optics and cabling*. Joshipura is not making promises yet – the switch is in beta now and won't ship until later in the second quarter – but that is the idea. Dell wants to push hard against existing Cisco Nexus 9504 and Juniper QFX10002 gear that is already in the field with 100 Gb/sec ports and get in position to counteract any moves by HP with its rebranded Accton switches (which are based on the Tomahawks) and Arista with whatever it ships (it could be XPliant machines or a mix of XPliant and Tomahawk machines).

There is a lot of talk about only hyperscalers and cloud builders wanting to move to 25G ASICs for their switching, but Dell is having none of that.

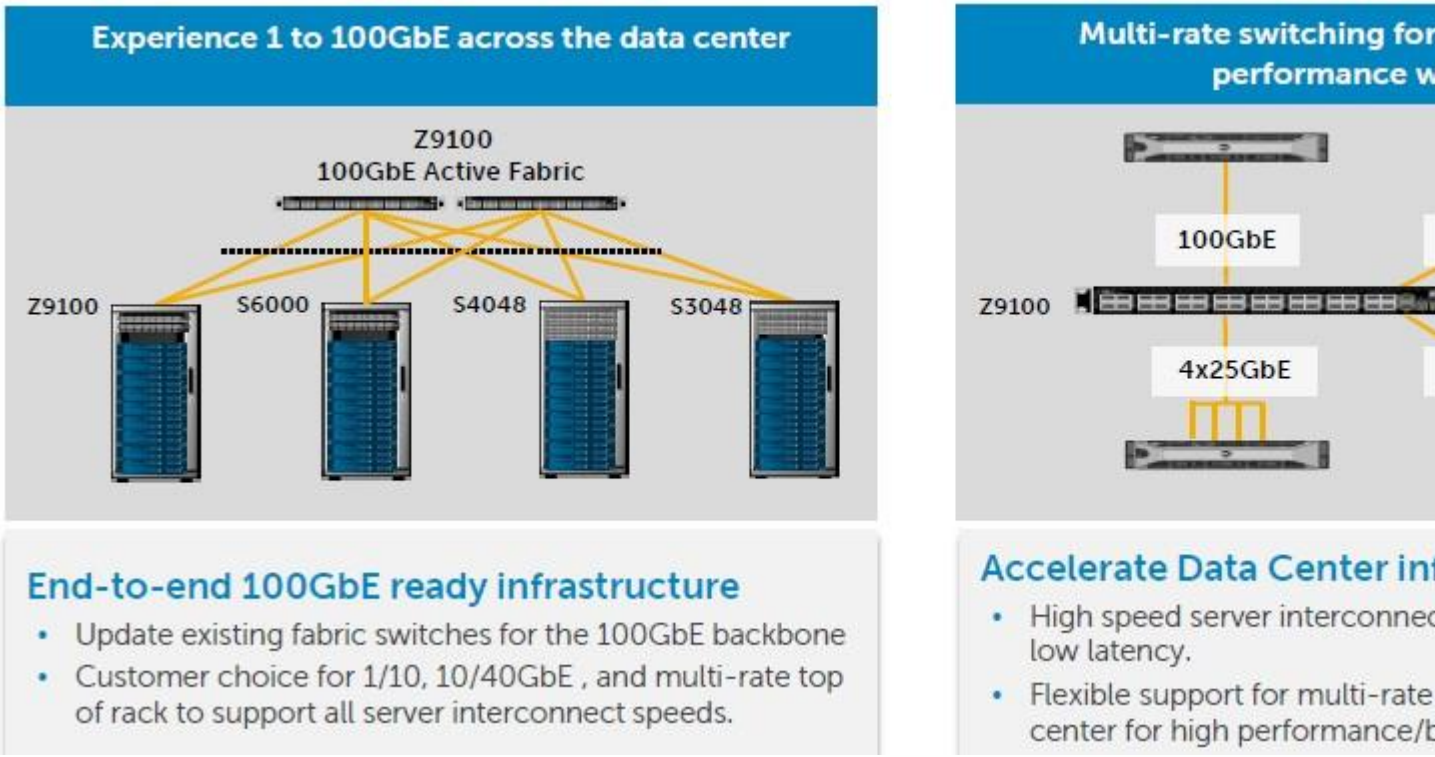
"Let our competitors keep on believing that," Ashley Gorakhpurwalla, vice president of Dell's Enterprise Solutions Group that oversees servers, switching, and storage at the IT supplier, tells*The Platform* with a smile.

Joshipura says that the Z9100-ON switch will be a core or spine switch for enterprises that need more bandwidth at that layer of their networks, and will be a top-of-racker for the hyperscalers who need even more bandwidth in their backbones.

"We are seeing demand for multi-rate switching not only from clouds and hyperscale companies, but also among enterprises that want to future proof," says Joshipura.

There are not multiple SKUs to create myriad variations of downlink and uplink speeds on the Z9100-ON. Rather, customers use breakout cables to split up the capacity on a hefty switch and can

take out the splitters later when they need more bandwidth per port. (You get 128 ports running at 25 Gb/sec by using a four-way breakout cable on a 32-port 100 Gb/sec switch, in fact, and a two-way breakout gives you 50 Gb/sec ports.) The multirate demand is not just within the Z9100-ON switch itself, which can mix and match different cables and speeds on the same box, but also across the network, thus:



The “ON” appended to the end of the switch names in the Dell line stand for “open networking,” and that means that the switch is not only capable of running Dell’s own OS9 network operating system (which came from its Force10 Networks acquisition a few years back) but also can be equipped with the ONIE open network operating system installer and can be used to load Cumulus Linux from Cumulus Networks and Switch Light from Big Switch Networks.

So who is looking for 100 Gb/sec switches at this point? HPC centers, for one.

“HPC is an area that we are seeing a lot of interest from,” says Joshipura. “While InfiniBand has its niche with the 30 percent to 40 percent of the workloads that are extremely latency sensitive, there are a lot more HPC workloads that can tolerate higher latencies. Moreover, the differences in the latencies are getting closer. 100 Gb/sec Ethernet is around 300 nanoseconds to 400 nanoseconds, and InfiniBand is like 200 nanoseconds to 250 nanoseconds. The differences are getting very minute, and it is almost noise, in my mind. With Ethernet, you get a whole ecosystem that supports it, rather than two vendors with InfiniBand.”

Hyperscalers, cloud builders, and universities are among the early tire kickers for the Z9100-ON switch, and are therefore an early indicator of who is looking for 25G-style switching. “The

hyperscale guys are clearly the first movers on anything. But when enterprises want to move to speeds higher than 10 Gb/sec because of the intensity of their east-west network traffic, the next question they will ask is should they go to 25 Gb/sec, 40 Gb/sec, or 100 Gb/sec? And I think what is happening is we are decoupling a vertically stacked solution and they might run different speeds on different ports."

Refreshing Slower Switches, Too

In addition to the top-end Z9100-ON switch, Dell also beefed up some slower-speed switches. The S4048-ON switch *is not based* on the just-announced Trident-II+ ASIC, but rather the Trident-II, and is a follow-on to the S4810 that used the older Trident ASIC several years ago. The S4048-ON comes in two configurations: one with 48 ports running at 10 Gb/sec with six uplinks running at 40 Gb/sec or one with 72 ports running at 10 Gb/sec. Joshipura says that it has half the port latency and half the power consumption of Cisco's Nexus 5672 top of racker. The S4048-ON is available now and costs $18,500.

At the very low-end, Dell is introducing the S3048-ON, which is a Gigabit Ethernet switch with 48 ports plus four uplinks running at 10 Gb/sec. It is based on an updated ASIC (presumably from Intel, Dell's other ASIC supplier) that draws half the power and has half the latency on a port hop than its predecessor. It costs $8,000 and is also available now.