Tomahawk 4 Switch First to 25.6Tbps

Broadcom Doubles 400Gbps Ports With Unprecedented 512 Serdes

December 16, 2019

PDF Version

By Bob Wheeler, The Linley Group, https://linleygroup.com/mpr/article.php?id=12237

Doubling the performance of a leading data-center chip usually requires breaking some limit, and that's what Broadcom did in developing its Tomahawk 4 switch. The new chip integrates an unprecedented 512 serdes, doubling the port density of its predecessor. Like Tomahawk 3, it uses 50Gbps PAM4 interfaces for compatibility with available 400Gbps Ethernet (400GbE) optical modules. The result is the industry's first 25.6Tbps switch chip, sampling almost exactly two years after its 12.8Tbps predecessor.

To help achieve this feat, Broadcom was still able to take advantage of Moore's Law, as Tomahawk 3 employs 16nm technology. By moving to a 7nm process, it delivered Tomahawk 4 within the power budget required to maintain an air-cooled system design. For maximum efficiency, the new chip also remains monolithic, avoiding the power penalty of chiplet-based architectures. Although it doubles port density, Tomahawk 4 uses the same top-level architecture as the prior generation, as Figure 1 shows.

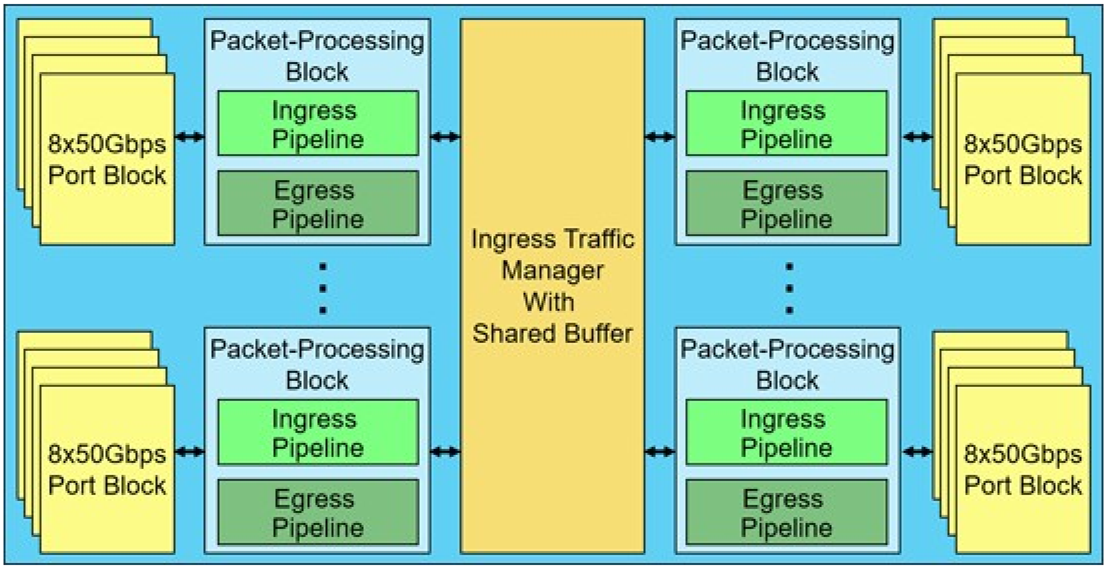


Figure 1. Tomahawk 4 switch chip The top-level architecture carries over from Tomahawk 3, but the new chip instantiates 64 port blocks, each with eight serdes.

The new chip's port count and speeds are, as expected, double those of Tomahawk 3: the BCM56990 implements up to 64x400GbE, 128x200GbE, and 256x100GbE ports. The lattermost figure is important to hyperscale-data-center operators, as it represents the maximum radix available for their tiered-network fabrics. As with other 400GbE chips, the serdes handle 25Gbps NRZ for backward compatibility with four-lane 100GbE as well as 25GbE. Owing to its narrow target of hyperscale customers, Broadcom is guarded about feature-enhancement details, but we view them as evolutionary.

Tomahawk 4 isn't Broadcom's first 7nm switch chip—that honor goes to Trident 4, which sampled in 2Q19 (see MPR 6/24/19, "Broadcom Samples Trident 4 Switch"). Although Trident 4 offers half the bandwidth of Tomahawk 4, it uses the same PAM4 serdes core. Reusing that serdes gives the company confidence it can qualify its newest chip for production in 2020. That's bad news for competitors, only one of which is shipping a 12.8Tbps switch chip. Barring unforeseen design or manufacturing problems, Broadcom is positioned to hold its share at the world's largest switch customers.

A Bigger and Sharper Weapon

In addition to keeping the bandwidth lead, the company is serving data-center customers with multiple product lines. The Tomahawk line delivers the greatest bandwidth and radix for hyperscale networks using a feature set optimized for that market. To address broader applications, the Trident line provides customer programmability and greater table flexibility than Tomahawk. Finally, the Jericho line enables modular chassis and deep buffers (see MPR 3/12/18, "Broadcom Router Chips Aim at ASICs"). Although Trident and Jericho address multiple markets, hyperscalers can use them in specialized network nodes with requirements that outstrip Tomahawk. By offering these alternatives, Broadcom can streamline Tomahawk 4 enough to double the port density of the 7nm Trident 4.

For the bulk of their network fabric, hyperscalers want to minimize the number of optical links and network layers. In a nonblocking leaf-spine architecture, Tomahawk 3 provides 64x100GbE access ports (down) and 64x100GbE ports to the next layer (up). Boasting twice that radix, Tomahawk 4 collapses two layers using six chips into a single 25.6Tbps switch. In addition to reducing cost and power, fewer network hops reduces end-to-end latency. That metric is particularly important for internal cloud services including disaggregated storage and distributed machine learning, both of which may use RDMA (RoCE).

Compared with the prior generation, Tomahawk 4 delivers the same streaming frame size—the packet size at which it can continuously stream data. To achieve that performance, Broadcom doubled the number of packet-processing pipelines. In other words, each pipeline handles the same number of port blocks as in Tomahawk 3. This approach avoids raising pipeline clock speed and thus potentially hurting power efficiency and latency. The downside of instantiating more pipelines is it usually means table-memory duplication, so memory area doubles without a corresponding increase in effective entries. Another design challenge is that more pipelines likely equates to more ports into the shared buffer memory. The company enhanced its traffic manager to improve burst absorption for high-priority traffic.

Congestion-control mechanisms can reduce the network's reliance on switch buffers, the size of which is constrained by die area. End-to-end traffic pacing requires support in both network adapters and switch chips. Congestion-management algorithms are proprietary, but Broadcom revealed some details of its NetXtreme Congestion Control (see MPR 9/17/18, "Broadcom First With 200Gbps NIC"). It also made undisclosed improvements to its load-balancing scheme, an important tool to maximize link utilization in hyperscale-data-center fabrics.

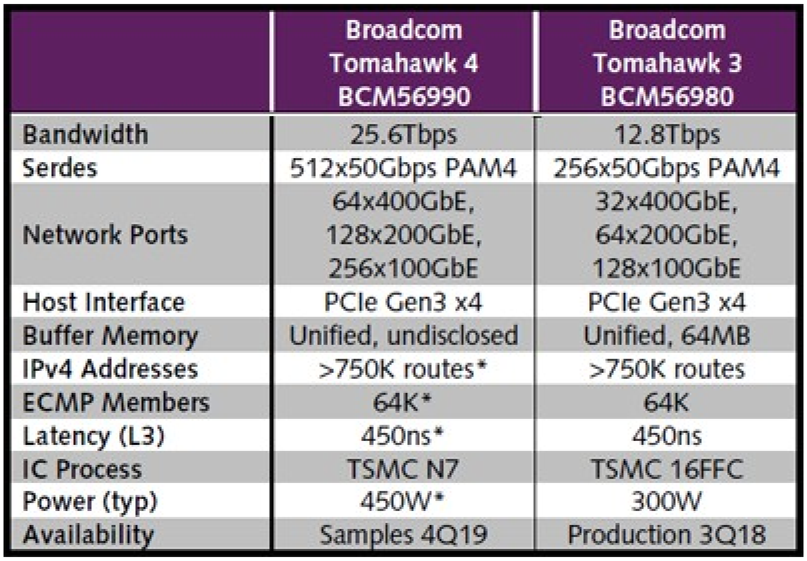
Another capability hyperscalers care about is instrumentation: it enables network automation, performance monitoring, and fault isolation. Carried over from its predecessor is in-band network telemetry (INT), which inserts telemetry metadata into network packets. Like that chip, Tomahawk 4 can handle various telemetry protocols including the IETF's Inband Flow Analyzer 2.0 (IFA 2.0), Cisco's IOAM, and UDP active probing. For out-of-band-telemetry processing, it has four 1.0GHz Arm Cortex-R5 CPUs. A unique new feature is real-time link-quality meters for accessing serdes-monitor statistics. These meters can identify problems with optical modules and direct-attach cabling (DAC).

More Ports, More Heat

When Broadcom sampled Tomahawk 3, we expected Tomahawk 4 would adopt next-generation 100Gbps PAM4 serdes. Such a move would enable a 25.6Tbps switch using the same number of serdes as 12.8Tbps designs. It has taken longer than expected, however, for 400GbE optics to reach high-volume availability from multiple suppliers. About 15 vendors have announced 400GbE optical modules, and production is finally ramping. Thus, to preserve compatibility with available optics, Broadcom decided to stay with 50Gbps serdes in Tomahawk 4. Another factor is DAC support, which will have less reach in the 100Gbps PAM4 generation (see MPR 12/10/18, "PAM4 Drives Serdes to 100Gbps").

In addition to module compatibility, staying with 50Gbps serdes enabled the company to reuse its port blocks from Trident 4. Sampled in May, that 7nm chip instantiates "Blackhawk7" serdes intellectual property (IP), which supports DAC and backplanes as well as optical modules. The serdes is functionally equivalent to the 16nm design in Tomahawk 3, so customers' qualified modules and cables should work seamlessly with the new switch. Doubling the number of 400GbE ports, however, means Tomahawk 4 designs with QSFP-DD or OSFP modules won't fit in one rack unit (1U). It's still possible to build a single-PCB switch in a 2U chassis by mounting module cages on both sides of the board.

As Table 1 shows, the BCM56990 also handles up to 256x100GbE ports, with each port using 2x50Gbps lanes. That capability allows customers to break out QSFP-DD ports as 4x100GbE without restrictions—that is, they can break out any or all ports. Broadcom withholds many details of its switch chips, although customers ultimately reveal some of them in their system specifications. We expect Tomahawk 4 offers table sizes like those of its predecessor, with Internet Protocol routes being most important for the Layer 3 fabrics that hyperscalers favor. For load balancing, equal-cost-multipath (ECMP) groups define a set of paths (members) to a single destination.



	Broadcom Tomahawk 4 BCM56990	Broadcom Tomahawk 3 BCM56980
Bandwidth	25.6Tbps	12.8Tbps
Serdes	512x50Gbps PAM4	256x50Gbps PAM4
Network Ports	64x400GbE, 128x200GbE, 256x100GbE	32x400GbE, 64x200GbE, 128x100GbE
Host Interface	PCIe Gen3 x4	PCIe Gen3 x4
Buffer Memory	Unified, undisclosed	Unified, 64MB
IPv4 Addresses	>750K routes*	>750K routes
ECMP Members	64K*	64K
Latency (L3)	450ns*	450ns
IC Process	TSMC N7	TSMC 16FFC
Power (typ)	450W*	300W
Availability	Samples 4Q19	Production 3Q18

Table 1. Comparison of Tomahawk switch generations. Externally, Tomahawk 4 is nearly identical to its predecessor but has twice the port count. (Source: Broadcom, except *The Linley Group estimate)

Broadcom made no mention of Tomahawk 4's buffer size, and SRAM is a major contributor to power dissipation. We therefore believe the new chip has less than double the 64MB of Tomahawk 3: likely 80–100MB of buffer memory. Because each packet-processing pipeline is similar to Tomahawk 3's, latency should also be similar at around 450ns. Note that Broadcom's minimum-latency figure excludes forward error correction (FEC), which adds about 150ns; customers' FEC requirements depend on their choice of optics and cabling. Tomahawk 3 dissipates 300W (typical) at 100% load, so doubling bandwidth in 16nm is impractical. By moving to 7nm technology and moderating memory growth, we estimate Broadcom fit Tomahawk 4 in a power envelope of around 450W, a 50% increase from its predecessor. The new chip requires a larger package, likely helping thermal management.

Feeding the Cloud

As the industry's first 25.6Tbps switch, Tomahawk 4 faces no direct competitors—an equivalent nonblocking configuration requires six 12.8Tbps chips. Although the field of switch-chip competitors has expanded, only Innovium has brought its 12.8Tbps switch chip to production (see MPR 10/21/19, "Innovium Moves Down Market"). That company's Teralynx 7 is also the only monolithic 12.8Tbps design aside from Tomahawk 3, and it's likewise built in 16nm technology. Other competitors use a mix of 7nm and 16nm chiplets in their respective 12.8Tbps designs, so we expect a mix of approaches in the 25.6Tbps generation. Latecomers may skip 512x50Gbps products in an attempt to time the 100Gbps serdes transition.

Tomahawk 4 is at once aggressive and conservative. Its design makes big changes in aggregate performance, I/O scale, process, and package. For customers, however, the chip maintains support for available optics without requiring external gearbox PHYs, easing upgrades. It's an engineering tour de force that smaller competitors will have difficulty matching. Unrelenting, Broadcom is keeping its foot on the gas to meet hyperscalers' needs as they build out new data centers using 400GbE.