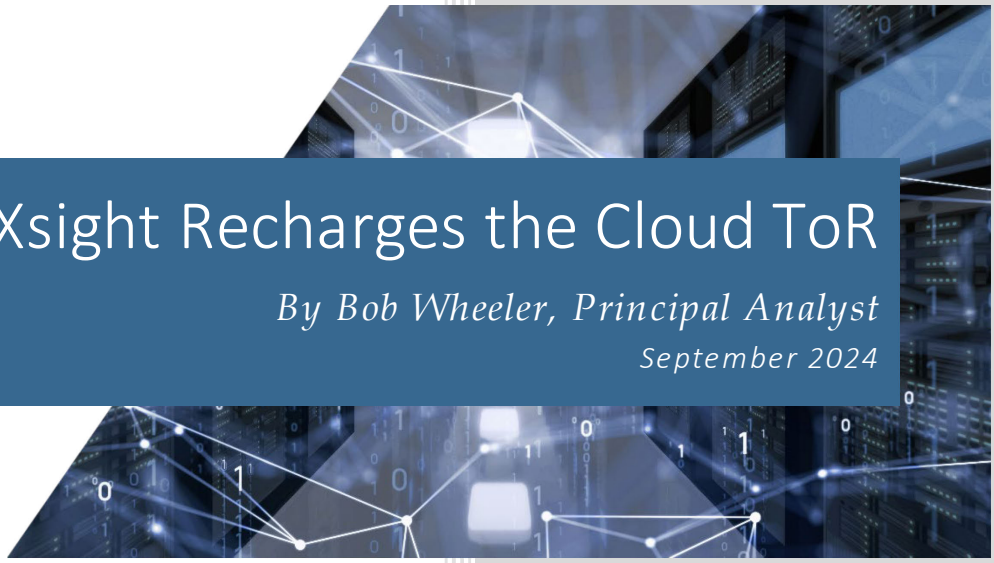

Xsight Recharges the Cloud ToR

By Bob Wheeler, Principal Analyst

September 2024

Wheeler's NETWORK

www.wheelersnetwork.com

Cloud-datacenter operators are driving rapid adoption of 800Gbps optical modules while also upgrading compute-server NICs to 400Gbps speeds. The 51.2Tbps switch chips designed for these network fabrics, however, deliver too much capacity for top-of-rack switch systems. With its X2, Xsight Labs developed a unique chip aimed at optimizing compute racks by enabling 100Gbps-per-lane server links and 800Gbps uplink optics. Xsight Labs sponsored the creation of this white paper, but the opinions and analysis are those of the author.

Cloud Servers' Missing Link

In the datacenter-switch market, the push for more bandwidth is relentless. Every two to three years, new Ethernet-switch chips enable a doubling of bandwidth, culminating in the first 51.2Tbps chips reaching production in 2023. Hyperscale datacenters must handle massive East-West traffic volumes, leading operators to deploy network fabrics with little or no oversubscription. To simplify congestion management, operators have even, in some cases, overprovisioned portions of their fabric hierarchy. Hyperscalers are beginning volume deployments of 51.2Tbps switch systems in 2024, and shipments should grow rapidly over the next three years. By adopting 800Gbps optical modules, system designers can pack 51.2Tbps of bandwidth into a 2U form factor. Because the available 51.2Tbps switch chips employ 5nm process technology, these compact systems remain air cooled.

The front panels of 51.2Tbps switch systems incorporate 64 optical modules using either QSFP-DD800 or OSFP800 form factors. Because end customers have different preferences, most OEMs and ODMs offer systems with both module types. Internally, these 800Gbps optical modules employ the same PAM4 DSPs and optoelectronic components, which dictate power and manufacturing cost. Several vendors have delivered second-generation 800Gbps DSPs built in 5nm technology that reduce module power to as little as 11W. These new DSPs also reduce cost by integrating drivers and transimpedance amplifiers (TIAs), which were previously discrete devices built in specialty process technologies. The combination of increased demand and competition should drive substantial price declines as these second-generation 800Gbps modules ramp to high volumes.

The optical standards dominating 800Gbps module shipments for the foreseeable future are DR8 and SR8, which both use eight parallel fibers to deliver 800Gbps in aggregate. Most hyperscale operators favor these standards because they enable flexible breakout of lower-speed ports, from 2x400Gbps down to 8x100Gbps. For the fabric leaf and spine nodes, a 51.2Tbps switch can be configured as 128x400Gbps Ethernet ports, for example. As server adapters (NICs) move to 400G Ethernet, the top-of-rack (ToR) switch can deliver two server connections per 800Gbps module. Whereas AI back-end networks will be first to adopt 800Gbps NICs, front-end networks will instead employ various smart NICs and DPUs at speeds up to 400Gbps in the near term. Products shipping next year should include AMD's Pensando "Salina" DPU and Intel's "Mount Morgan" IPU (for Google) as well as Broadcom's Thor2 standard NIC and internal (ASIC-based) designs at Amazon and Microsoft.

Next-generation 400G Ethernet server NICs are adopting 100Gbps serdes to drive passive-copper (DAC) and active-electrical (AEC) cables to the top-of-rack switch. Moving from 50G/lane to 100G/lane interfaces allows 400G Ethernet ports to operate over half the number of conductors used in first-generation (8x50Gbps) designs. There are problems, however, with existing ToR-switch designs. Available 12.8Tbps systems employ 50G/lane serdes, which are incompatible with 100G/lane DAC. AECs can "gearbox" from 8x50G to 4x100G lanes, but this adds power, cost, and latency. Alternatively, 25.6Tbps switches with 100G/lane serdes can serve in the ToR, but these systems are overkill for most general-purpose compute racks. Given that the 12.8Tbps ToR switch will serve compute infrastructure for years to come, the ideal solution is a refreshed design built around 100G/lane serdes.

Speed Matching from Fabric to Server

The reason hyperscalers want maximum-density switches in the fabric leaf/spine nodes is to flatten the network to the fewest possible levels. The switches' number of ports, or radix, determine how many network-hierarchy levels are required for a given number of servers. Because additional switching levels add power, cost, and latency, it is easy for hyperscalers to justify investing in leading-edge switches for the network fabric. The available 51.2Tbps switch systems have 512x100G serdes and can handle up to 256x200GbE or 128x400GbE ports. Two chip vendors' forthcoming products will offer the maximum-possible 512x100GbE ports. As Figure 1 shows, a two-level fabric based on 512-radix switches can handle more than 130,000 servers.

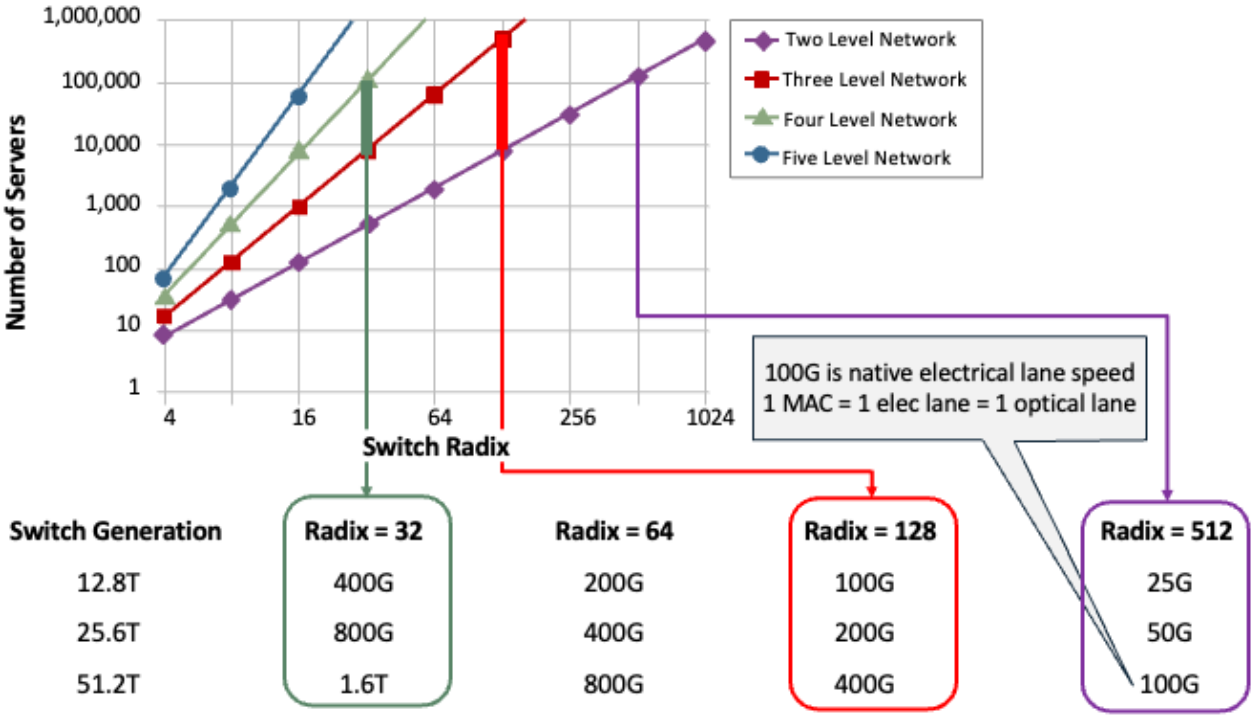

FIGURE 1. 100G/LANE FABRIC SCALE
(Source: Wheeler's Network)

The ultimate architecture would extend 100G/lane serdes through the ToR all the way to the server NIC. This approach eliminates gearboxing from 50G/lane to 100G/lane, whether implemented at the ToR or in an AEC. For example, connecting a next-generation 400GbE NIC to an existing 12.8Tbps ToR system would require an AEC with 4x100G lanes on the server end and 8x50G lanes on the switch end. Compared with DAC, such a design would add about 10 watts per cable as well as the cost of two PAM4 DSPs. For the uplink to the leaf switch, the ToR would use an optical transceiver that integrates the 8x50G-to-4x100G gearbox function, adding power per 400G port compared with newer 800G optics.

It is also beneficial to use the same Ethernet port (MAC) speeds throughout the fabric, as this enables the switches to perform cut-through packet forwarding. Cut-through operation enables the switch to begin sending packet data to the egress port before the entire packet has been received at ingress, minimizing port-to-port latency. Generally, cut-through forwarding is required to achieve sub-microsecond latencies, whereas store-and-forward latencies exceed one microsecond and increase with packet length. In addition

to minimizing latency, cut-through operation minimizes packet-buffer occupancy, leaving more buffer capacity for congestion events.

Although NICs can implement 100G, 200G, and 400G Ethernet ports using 100G/lane serdes, hyperscalers appear poised to adopt 400GbE NICs in the near term. Because their hardware roadmaps are usually opaque, it is instructive to look at what is already available. Amazon's 200Gbps Nitro System, for example, has been available in the AWS EC2 Elastic Fabric Adapter (EFAv2) since 2022. Also in 2022, Intel shipped the Mount Evans IPU to Google, its development partner. In July 2023, Microsoft announced the preview of Azure Boost, which delivers 200Gbps networking using the FPGA-based Microsoft Azure Network Adapter (MANA). By the end of 2024, Google should begin deploying 400G IPUs, and Amazon could be close behind with next-generation Nitro cards.

X2 Modernizes 12.8T Cloud Switch Silicon

Responding to customer needs for a modern ToR, Xsight Labs developed the X2 switch chip, a 12.8Tbps device with 128x100G serdes lanes and built in 5nm process technology. Although it's a new chip design, the X2 architecture is derived from the company's silicon-proven X1 switch chip. Xsight sampled the X1 to customers in late 2020 but decided not to take that 7nm chip to production. Because the X1 is fully functional, however, customers have used it as a software development platform in preparation for the X2. Xsight delivered X2 samples to lead customers in July 2024.

Figure 2 shows an X2 functional block diagram, which differs from the physical design. Internally, the X2 employs a modular architecture that allows more granular performance scaling than traditional pipeline-based switch architectures. The programmable packet processors enable multiple switch profiles that can trade off, for example, table scale versus latency. The architecture delivers deterministic latency, however, which can be as low as 450ns and around 700ns for a typical ToR profile. Programmability is an important underlying capability for many of the X2's advanced features, including queue management, congestion control, and telemetry. At the same time, the chip is designed for line-rate operation with 256-byte packets.

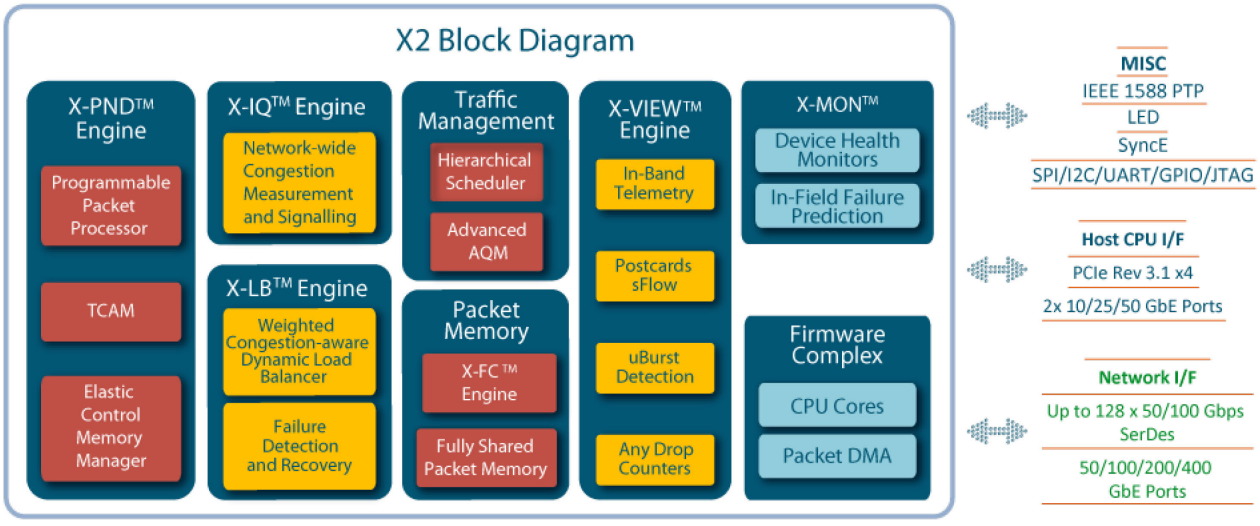

FIGURE 2.X2 FUNCTIONAL BLOCK DIAGRAM
(Source: Xsight Labs)

The X2 can implement congestion-aware routing, which comprises measuring and signaling congestion as well as load balancing across the network fabric. For a given fabric link, the chip can measure flow rate, latency, and queue level. When congestion is detected, it can signal using Explicit Congestion Notification (ECN), Congestion Signaling (CSIG), or other customer-defined mechanisms. The X2 can handle flowlet-based load balancing and packet spraying across fabric links as well as traditional flow-based approaches like ECMP. The chip's programmability means it will handle new notification mechanisms defined after Xsight finalized its silicon design. The company is confident that the X2 will handle advanced features currently being specified as a part of the Ultra Ethernet Transport (UET) protocol, such as packet trimming and back-to-sender signaling.

Owing to the massive scale of hyperscalers' networks, network visibility has a large impact on operational costs. These customers need to be able to identify network-performance problems as well as outright failures. Traditional out-of-band network TAPs are inadequate for the scale of large data centers. Instead, in-band network telemetry enables real-time collection of statistics, events, and timing information. Through its programmability, the X2 can handle various in-band telemetry standards such as P4-INT, IOAM, and Postcard-Based Telemetry (PBT). Another important real-time feature is microburst detection, which provides granular feedback to the active queue management (AQM) function.

Externally, the X2 handles up to 128x100G, 64x200G, and 32x400G Ethernet ports. Its long-reach 100G PAM4 serdes are designed to drive DAC cables up to 4m in length and to directly connect with 800G optical modules. The serdes also handle 50G PAM4 and 25G NRZ rates, which provide backward compatibility with existing server NICs. Separately, the chip provides dual 50G Ethernet management ports as well as a PCI Express v3.1 x4 host-processor interface. Thanks to its 5nm design, the X2 dissipates only 180W (typical) under full load, and Xsight expects to rate maximum power at 216W. The chip is packaged in a mainstream 55mm BGA with 1mm pitch, easing PCB routing.

Futureproofing While Saving

A typical ToR for a hyperscale customer might configure the X2 as 24x400G downlinks (using QSFP112 cages) and 4x800G uplinks using QSFP-DD800 or OSFP800 modules. Whereas hyperscalers need a 12.8T ToR to handle 400G-enabled servers, however, smaller cloud datacenters will typically employ 100G NICs for next-generation servers. To enable OEM customers to build a broad product line with common software, Xsight will offer several X2 variants (SKUs) with bandwidths down to 6.4Tbps. The 9.6Tbps and 8.0Tbps versions handle 32x200G and 48x100G downlinks, respectively, while also providing 4x800G uplink ports. A 6.4Tbps version provides 128x50G PAM4 serdes, providing compatibility with older NICs and 400G optical modules. The X2 family allows OEMs to replace aging switch designs based on 16nm silicon with lower power chips that also offer programmability.

To support both OEM and hyperscale customers, Xsight developed an SDK that offers three different API levels as well as off-the-shelf SONiC support. The SONiC network operating system (NOS) leverages the OCP Switch Abstraction Interface (SAI), which Xsight's XSAI layer implements. As Figure 3 shows, the next level down is the XSW API, which provides functional control for standard features. Customers wishing to implement custom features can access device primitives using the hardware-abstraction layer (XHAL) or directly access the switch silicon using the low-level driver (XDRV). The X-PCI driver provides a packet interface over PCI for control-plane software. Xsight also provides an instruction-accurate behavioral model (XBM) that allows customers to develop and test software in advance of silicon integration. The company offers an X2 evaluation system that integrates 16xOSFP800 ports and a COMe CPU module in a 1U chassis.

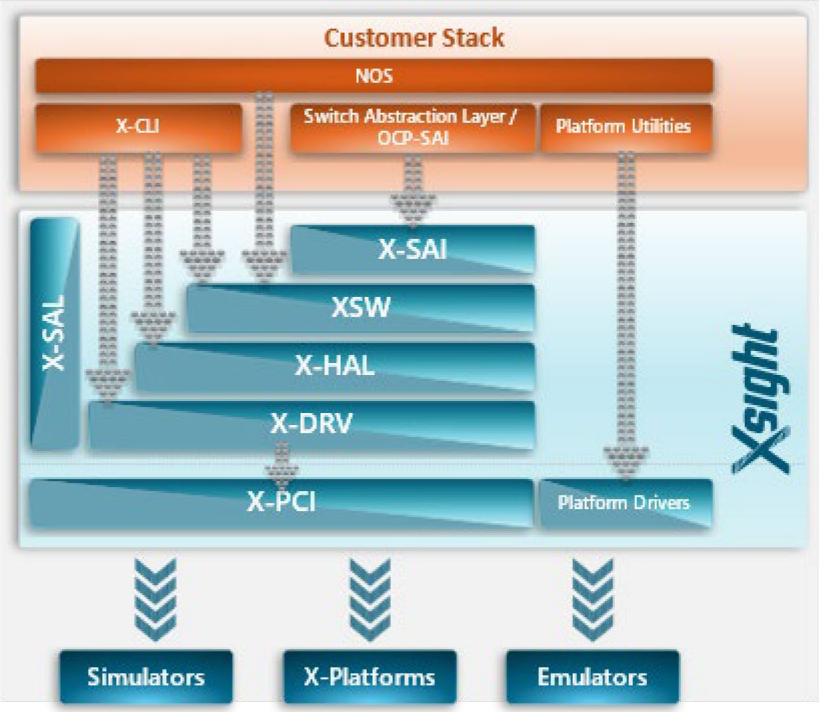

FIGURE 3.XSIGHT SDK ARCHITECTURE
(Source: Xsight Labs)

For datacenter operators, X2-based systems promise lower operating expenses through power savings as compared with first-generation 12.8T switches. Starting with the silicon, the X2 should save 149W over Broadcom's 16nm Tomahawk3, which dissipates 365W (maximum) in the low-latency mode these customers prefer. Using 4x800G instead of 8x400G optical modules saves around 32W for the fabric uplinks. As discussed above, connecting new server NICs with 100G/lane interfaces to existing 12.8T switches (with 50G/lane serdes) requires AECs with gearboxing, adding about 240W for 24 server links. In total, an X2-based deployment should save about 424W per switch, excluding reduced fan power. Although power costs vary widely across customer types and geographic regions, power savings translate directly to savings in operating expenses. In addition, many racks are now power limited rather than space limited. Any network-power savings free up power for revenue-producing compute resources.

Xsight's competitors could develop new switch chips optimized for cloud-ToR designs, but they have instead focused elsewhere. Broadcom, Cisco, Marvell, and Nvidia have all developed 51.2T switch chips in 5nm process technology, but they haven't derived new ToR-optimized silicon from these designs. Broadcom comes closest with its new Trident5-X12, a 16Tbps switch chip with 100G/lane serdes and built in 5nm technology. The Trident line, however, targets enterprise networks that require massive table scale and advanced security features. As a result, Trident5-X12 dissipates an estimated 350W (maximum), essentially the same power as the 16nm Tomahawk3. Vendors withhold die size, but power alone suggests Broadcom's new chip is much larger than Xsight's X2.

Rightsizing the Cloud ToR

Despite the rapid evolution of cloud data centers, 12.8T ToRs will serve cloud compute for another generation. Multiple factors drive the need for ToRs with lower bandwidth than leaf/spine switches. First, most customers favor the ToR configuration because it enables the use of low-cost and low-power

DAC connections between servers and the switch. Using larger switches in a middle- or end-of-row configuration instead requires the use of optical links, raising cost and power. Additionally, the ToR configuration limits the switch-failure "blast radius" to a single rack, meaning only servers in that rack will be affected. Finally, whereas AI back-end networks are pushing NIC-speed limits, compute racks are only now moving to 200G speeds with 400G on the horizon.

As the incumbent Ethernet switch-chip vendors pursue relentless density increases, they have left the 12.8T cloud-ToR behind. Broadcom's 16nm Tomahawk3 reached production nearly six years ago, yet it remains at the heart of most 12.8T switch systems designed for cloud datacenters. These systems also use older 50G/lane module form factors, whereas leaf/spine switches are moving to 100G/lane form factors that support the latest 800G optics. This situation leaves an underserved opportunity for a modern ToR that is in step with the datacenter fabric. Furthermore, existing cloud-ToR systems may lack features needed to optimally support new protocols such as UET.

With its X2, Xsight is delivering a forward-looking cloud-optimized design positioned to serve the ToR role in modern network fabrics. The chip aims for the level of programmability and table scale required in this application, reducing power and cost as compared with designs that also serve enterprise networks. In the competitive Ethernet switch-chip market, Xsight's strategy is unconventional and differentiated. Rather than face the large incumbents head on, the startup identified a neglected high-volume segment for its new switch chip. Xsight has reemerged as a pure-play chip vendor with a unique offering, endeavoring to establish a beachhead in the cloud datacenter.

Bob Wheeler is an independent industry analyst covering semiconductors and networking for more than two decades. He is currently principal analyst at Wheeler's Network, established in 2022. Previously, Wheeler was a principal analyst at The Linley Group and a senior editor for Microprocessor Report. Joining the company in 2001, he authored articles, reports, and white papers covering a range of chips including Ethernet switches, DPUs, server processors, and embedded processors, as well as emerging technologies. Wheeler's Network offers white papers, strategic consulting, roadmap reviews, and custom reports. Our free blog is available at www.wheelersnetwork.com.