The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

XSIGHT SAMPLES LEADING-EDGE SWITCH

25.6Tbps X1 Designed to Handle Next-Generation Optics

By Bob Wheeler (January 18, 2021)

Xsight Labs emerged from stealth last month to challenge Broadcom at the Ethernet switch market's high end. Led by a greatly successful team, the Israeli startup is already sampling its first chip, called the X1. It claims the monolithic 7nm design reduces power dissipation relative to competing chips. The X1 also offers programmable packet processing combined with a unique architecture that shares memory between packet buffers and lookup tables.

The leading-edge 25.6Tbps Ethernet switch chip uses 100Gbps PAM4 serdes, which connect to next-generation 800Gbps optical modules. Logically, the 32x800Gbps optical modules can break out to 64x400G or 256x100G Ethernet ports. The X1 targets multiple data-center applications, including leaf/spine fabrics, top-of-rack (ToR) switches, and high-performance clusters in AI and HPC. As a result, Xsight offers two 12.8Tbps X1 variants in addition to the 25.6Tbps model.

Differentiation is important, as the start-up enters a crowded field. As Figure 1 shows, seven competitors are shipping 400G Ethernet switch chips, and Broadcom is already in production with the first 25.6Tbps chip (see *MPR 1/11/21,* "Broadcom SmartToR Ups Flow Scale"). Competitors packing 100Gbps serdes include Broadcom's Tomahawk4-100G (TH4-100G) and Innovium's Teralynx 8 (see *MPR 5/25/20,* "Innovium Debuts 25.6Tbps Switch").

Although Xsight touts the X1 for multiple applications, we see it initially serving only a handful of the largest data-center operators. With Broadcom and Innovium already supplying these customers, the newcomer must now demonstrate the superiority of its design. Otherwise, these demanding operators will take the lower-risk path of using proven suppliers, as market leader Broadcom continues to deliver new products at an impressive cadence.

Willenz on a Winning Streak

Xsight chairman Avigdor Willenz originally founded switch-silicon pioneer Galileo Technology, which Marvell acquired in 2001. The serial entrepreneur later cofounded Annapurna Labs (sold to Amazon in 2015), was the lead investor in Leaba Semiconductor (acquired by Cisco in 2016), and co-founded Habana Labs (which Intel purchased in 2019).

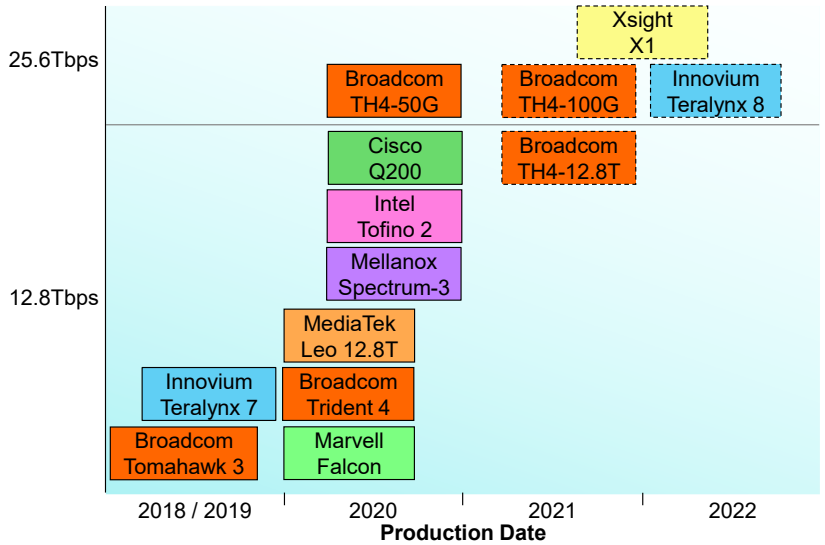

Figure 1. Roadmap for 400GbE switch chips. By sampling the X1, Xsight joins Broadcom and Innovium at the leading edge of 400GbE switch chips. Dotted outlines indicate models that have yet to reach production. (Source: The Linley Group)

Price and Availability

Xsight is sampling three X1 versions: the 25.6Tbps XLX1A256A, the XLX1A128A with 128x100Gbps serdes, and the XLX1A128B with 256x50Gbps serdes. The company withheld pricing. For more information, access *www.xsightlabs.com*.

Willenz founded Xsight in early 2017 along with several EZchip veterans after Mellanox acquired that network-processor pioneer and later terminated its NPU development (see *MPR 10/19/15,* "EZchip Gives Mellanox Brains"). The startup's CEO, Guy Koren, was EZchip's long-time CTO.

The Xsight founders' vision of a single switch architecture serving multiple applications sounds akin to that of Leaba, but the earlier startup targeted routing in addition to switching. Xsight instead aims to connect heterogeneous components in the data center, including servers, storage, and accelerators. As a result, it designed the X1 to deliver the lowest power as well as the lowest tail latency.

Since its founding, Xsight has raised $116 million from venture and strategic investors, including Intel, Microsoft, and Xilinx. Intel led the second round of $80 million, closed in May 2020. Given it's shipping the 12.8Tbps Tofino 2 switch acquired with Barefoot, Intel appears to be hedging its bets (see *MPR 1/7/19,* "Barefoot Joins 400GbE-Switch Club"). Xsight has grown to more than 120 employees, with most located in Israel and the US.

X1 Is Memory-Game Champ

Like many of its competitors, the startup disclosed few X1 architecture details, instead branding various features. As Figure 2 shows, the chip includes a programmable packet processor, enabling parsing and editing of a packet's first 256 bytes. Xsight doesn't support the P4 programming language, and it withheld how customers can program the device. The chip has a three-level hierarchical scheduler that manages 256 queues per port (64K queues in total) with active queue management (AQM).

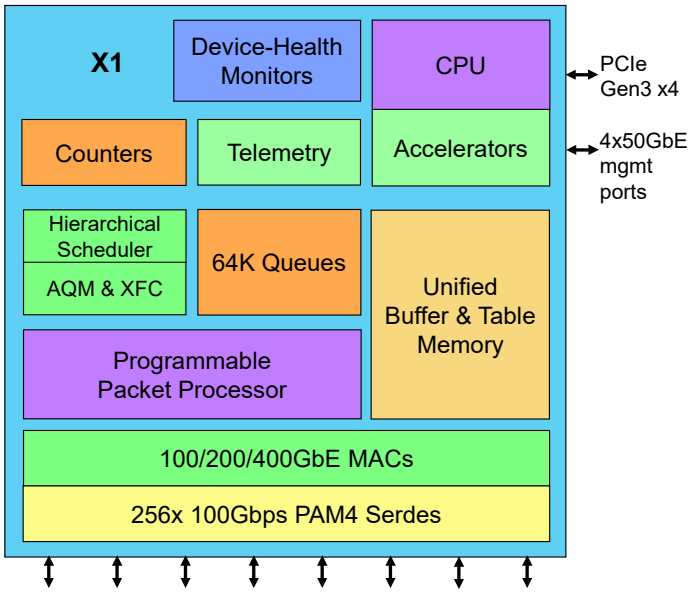

Figure 2. Xsight X1 block diagram. The unified buffer/table memory stands out as a unique element, enabling customers to optimize their installations.

To address latency-sensitive applications such as machine-learning clusters, the startup focused on reducing flow-completion time (FCT). Long tails that increase FCT are generally a result of packet loss owing to incast congestion, which occurs when many senders target one receiver. The X1 optionally implements a proprietary end-to-end protocol for congestion management, dubbed XFC, that the company claims adds little handshake overhead. The chip also handles standard congestion-notification schemes including Explicit Congestion Notification (ECN).

The X1's unique memory architecture helps it handle various applications. Branded X-PND, the unified packet-buffer and table memory enables application-specific memory partitioning, trading off buffer depth versus table entries. Xsight can create memory profiles for leaf/spine, top-of-rack (ToR), and cluster installations. For example, HPC clusters typically employ small tables, so most memory can be allocated to packet buffers to maximize performance. On the other hand, a ToR switch might require large tables but less performance.

The startup claims its packet buffer is fully shared, meaning any port can consume the entire buffer, which aids in handling incast congestion. This subject is controversial, however, as most competitors also claim to implement fully shared packet buffers. Because vendors withhold memory-architecture details, effective buffer depths and fairness problems usually surface only when vendors benchmark competing products.

Network telemetry is important to large data-center operators in diagnosing performance problems and failures. Like other data-center switch chips, the X1 includes comprehensive statistics counters, queue monitoring, packet mirroring, and event triggers. For in-band telemetry, it handles the P4-INT and IETF IOAM standards. To protect against denial-of-service attacks, up to 256 queues are available for scheduling and shaping management traffic. An on-chip CPU of undisclosed type supports telemetry and other real-time processing, augmenting an external host processor attached using PCI Express. The X1 also has 4x50GbE management ports, providing an unusually large amount of management-plane bandwidth.

We seldom associate device-aging concerns with data centers that have equipment-replacement cycles of less than 10 years. Xsight says its customers care about in-field failures, however, for two reasons. First, unlike servers, switches can have a large "blast radius," meaning one switch failure can bring down many connections. Second, as power levels continue to rise, reliability decreases. For

these reasons, the X1 includes intellectual property from another Israeli startup, ProteanTecs, to monitor active circuits for degradation. It claims this monitoring can predict failures before they occur, thereby avoiding network downtime.

Forging a Tomahawk Alternative

Externally, the X1 is nearly identical to Broadcom's Tomahawk 4-100G, as Table 1 shows. Its 256 long-reach serdes handle direct-attach cabling (DAC) and 800Gbps optical modules without external retimers. In addition to 100Gbps PAM4 operation, they support 50Gbps PAM4 as well as 25Gbps and 10Gbps NRZ for backward compatibility. The chip's MACs also provide up to 256 ports of 100GbE as well as lower Ethernet rates. In addition to the 25.6Tbps chip (XLX1A256A), Xsight offers a version with half as many 100Gbps serdes (XLX1A128A), suitable for 100GbE ToR switches, as well as a version with 256x50Gbps serdes (XLX1A128B) for systems using 400Gbps optical modules.

Xsight offers a 25.6Tbps customer-evaluation system in a 2U form factor, and it's developing a 1U system fit for manufacturing with 32xOSFP or 32xQSFP-DD800 modules. Its software stack is unremarkable, providing a low-level API (XHAL) and a high-level API (XSW) for network-operating-system integration. The XSW layer implements the Switch Abstraction Interface (SAI) for compatibility with open source Sonic. The Xsight software-development kit also supports script-based programming of the packet processor, although the company expects cloud customers will use its off-the-shelf profiles.

Because Xsight withheld the scale of its buffer/table profiles, directly comparing these metrics with the TH4-100G's is impossible. Broadcom's design includes a massive 114MB buffer, and its algorithmic longest-prefix-match (LPM) tables handle at least 850K IPv4 or 360K IPv6 routes. We speculate the X1 offers similar maximum buffer and table sizes but reduces total SRAM area by trading off the two. Fortunately, the X1 provides superior scheduling, with far more queues and a deeper hierarchy than the TH4-100G. In leaf/spine fabrics, for example, customers can use fine-grain scheduling combined with congestion-avoidance protocols and load balancing to manage queue depths.

Although Xsight emphasizes tail latency and FCT, it claims to deliver cut-through latencies near those of competitors; we estimate the TH4-100G's minimum latency at around 600ns with forward error correction (FEC) enabled. Until the startup reveals more about its packet processor, we remain skeptical of any minimum-latency claims.

Power-dissipation claims are also problematic, as no standard exists for normal test conditions, and vendors are unable to disclose their customers' conditions. Variables include total-bandwidth load, packet-size mix, and other traffic characteristics, as well as standard IC variables such as temperature and voltage. Understandably, vendors are reluctant to estimate maximum power, which is scary high in this class of chips. Broadcom rates its shipping Tomahawk4-50G, which sports 512x50Gbps serdes, at about 500W maximum and 300W typical. Surprisingly, the 100G version dissipates more power (350W typical) despite having half as many serdes. Xsight claims its 25.6Tbps X1 undercuts the TH4-100G's power by about 14%, but it withheld any specific innovations that enable this feat.

Another competitor, Innovium, preannounced its 25.6Tbps Teralynx 8, but it has yet to deliver samples (see *MPR 5/25/20,* "Innovium Debuts 25.6Tbps Switch"). Compared with Xsight, the company has the advantages of proven software, existing customers, and deep pockets after raising $170 million last year. Regardless, we expect Broadcom will qualify the TH4-100G for production before either startup achieves that milestone. None of the other 12.8Tbps-switch vendors in Figure 1 has announced a 25.6Tbps chip or even a schedule for one.

Cloud Shows Xsight-ing Prospects

With advanced scheduling and congestion-control features, the X1 stands out for AI and HPC clusters. Those applications employ carefully designed networks, and customers are often willing to use proprietary protocols such as XFC. That willingness also represents a downside, however, as customers may instead choose Ethernet alternatives. Nvidia leads in merchant cluster interconnects through its Mellanox acquisition, supplying end-to-end InfiniBand networks including system-level switches and host adapters. Some deep-learning accelerators also have cache-coherent interconnects for creating small clusters, obviating the need for high-performance Ethernet. Notably, Google uses this approach with its TPU clusters (pods), which scale beyond 1,000 chips. Thus, we see a limited market for Ethernet switches in HPC and AI.

For mainstream applications, large data-center operators can employ the X1's underlying features for their own congestion-management schemes. Like those with AI/HPC

	Xsight X1	Broadcom Tomahawk4-100G
Bandwidth	25.6Tbps	25.6Tbps
Serdes (PAM4)	256x100Gbps	256x100Gbps
Network-Port Configurations	64x400GbE, 128x200GbE, 256x100GbE	64x400GbE, 128x200GbE, 256x100GbE
Host Interface	PCIe Gen3 x4	PCIe Gen3 x4
Buffer Memory	Undisclosed	114MB
IPv4 Addresses	Undisclosed	>850K routes
Queues	64K	3K
Latency (L3 w/FEC)	Undisclosed	600ns*
IC Process	TSMC N7	TSMC N7
Power (typ)	300W	350W
Samples	4Q20	3Q20

Table 1. Comparison of 25.6Tbps switch chips. FEC=forward error correction. Externally, the X1 is nearly identical to the TH4-100G, but it introduces less visible architectural innovations. (Source: vendors, except *The Linley Group estimate)

clusters, operators using NVMe over Fabrics to disaggregate storage are concerned about tail latencies, and the X1's hierarchical scheduler could help prevent packet drops. Owing to the large scale that cloud storage requires, operators favor standard Ethernet hardware over more-expensive InfiniBand or proprietary interconnects. Power dissipation is important both as a limiting factor for rack density and as a contributor to operating costs. Thus, Xsight's claimed power advantage is a strong motivation, but customers will need to validate it under their own test conditions.

The X1 is an impressive first effort, sampling close behind Broadcom's TH4-100G. Assuming the startup can move it to production by early 2022, the chip will be well timed for the transition to 800Gbps optical modules. Microsoft implicitly endorsed the product by investing in Xsight, so it's an obvious first customer among the hyperscale cloud operators. Although customers desire alternatives to Broadcom, they ultimately adopt only those that are superior in some way. With the X1 now in their labs, they can fully assess its unique attributes for their workloads. ♦