Xsight

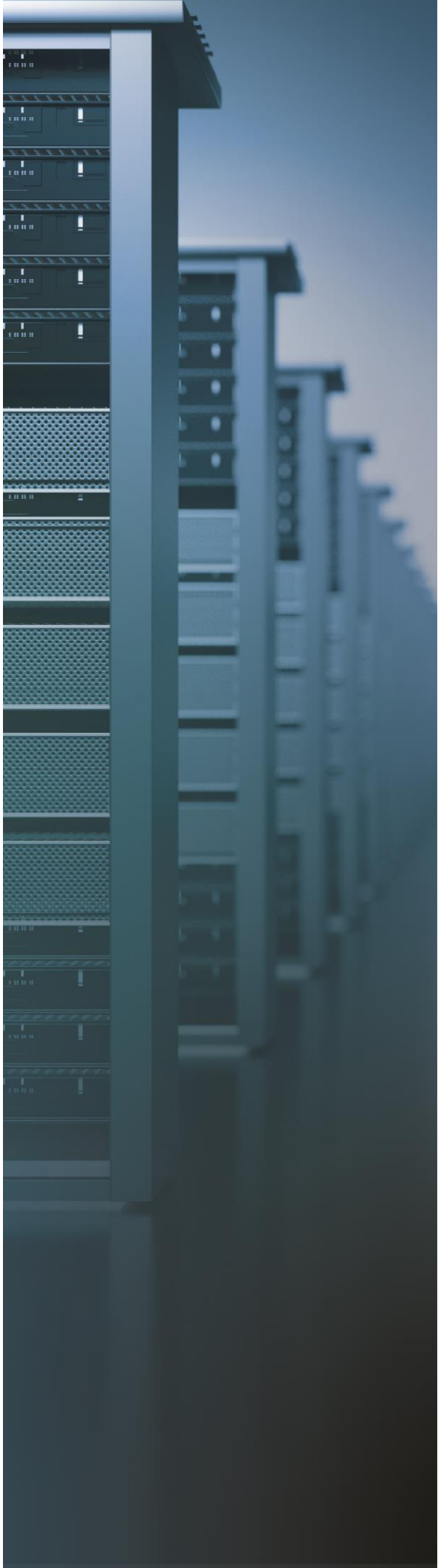
WHITE PAPER

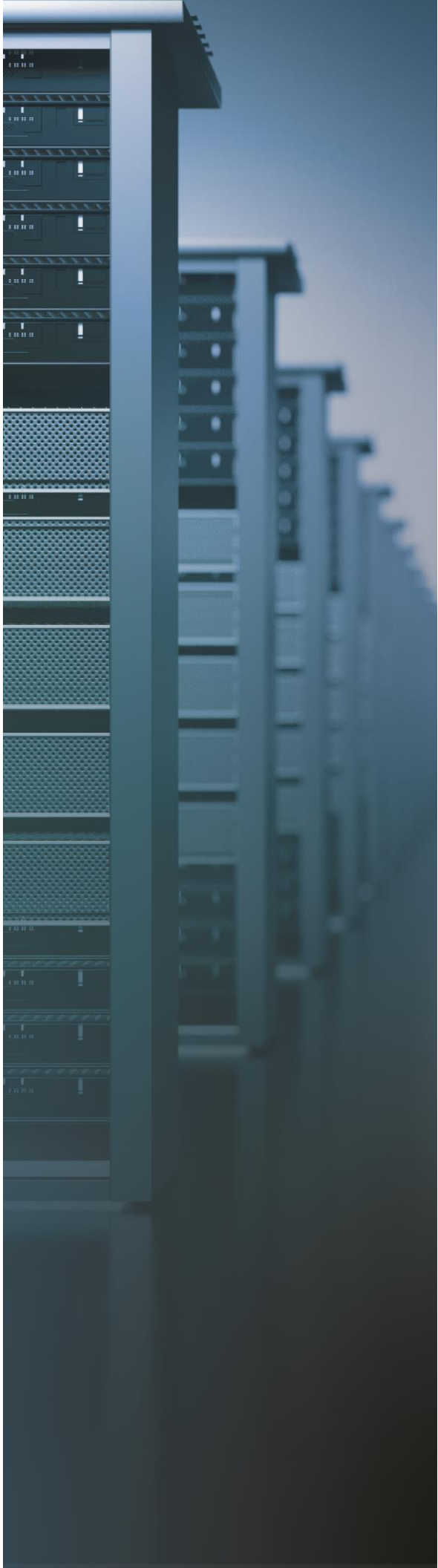
Accelerating Data Center Growth

X1 for the Hyperscale Data Center

DECEMBER, 2020

Table of Contents





Abstract

Modern data centers, driven by current industry trends, are currently growing faster than ever. This paper focuses on the network infrastructure needed to support these massive scale data centers and offers an overview of switch requirements needed to address these challenges. Data centers designed to evolve will enable seamless and painless growth spurts.

Xsight Labs introduces the X1 family of fully programmable switching ASIC devices, optimized for data center deployments and workloads. This paper focuses on the different methods the X1 family uses to mitigate these network infrastructure challenges.

Data Center Boom

Data center traffic is growing exponentially due to a number of different reasons such as: enterprises moving their infrastructure to the cloud, ubiquitous AI services that analyze petabytes of data, Video/VR streaming services of unprecedented quality, autonomous vehicles, smart cities, an endless list of IoT devices to name just a few.

Ultra-low latency networking becomes very important as storage latency is reduced with the adoption of FLASH and NVMe over fabric technology.

To support these sophisticated services, the data center network must provide throughput and loss guarantees and unprecedented availability. The huge scale of the data center network requires highly automated management. The network nodes must therefore be capable of providing verbose visibility into traffic characteristics and network faults.

An Ocean of Data

People, distributed applications, and machines are creating and consuming data faster than ever. All these together are creating an unprecedented data explosion. Rapidly evolving technologies — Blockchain, IoT, 5G and AI — add to this exponential data growth [1]. There is tangible cost associated with data transfer at hyperscale data centers. Massive data growth is creating a "data gravity" phenomenon where compute resources, applications and business logic are drawn closer to physical data location and thus are creating mega data centers[2]. Consolidation of compute and data resources are creating an urgent need for fast, flexible and more reliable transparent networks.

The Rise of Artificial Intelligence

Oceans of data allow industries, science, and businesses to fully adopt data driven decision making and business models. Data is extremely valuable for making decisions as well as for advancing business and technology. However, data sets still need to be processed and prepared for use. Moreover, the exploding amount of data makes its processing and analysis by human or traditionally focused programs an impossible task. Artificial Intelligence (AI) solves these problems by applying sophisticated data processing and decision making. Data hungry AI applications are placing enormous workloads on the data center and AI cluster network infrastructure.

Edge Compute

IoT, smart cities, machine-to-machine (M2M) in general, and autonomous vehicles, in particular, all drive a need for extremely fast data processing and response. Edge data centers are emerging to address fast data processing in close proximity to the data source and are driving the requirement for high bandwidth and low latency networks.

Global Events

Recent events, such as the current global pandemic, accelerate the growth of data centers and create a new landscape of distributed workplaces, online education and automation. This trend, may not be as short-lived as initially thought since businesses and

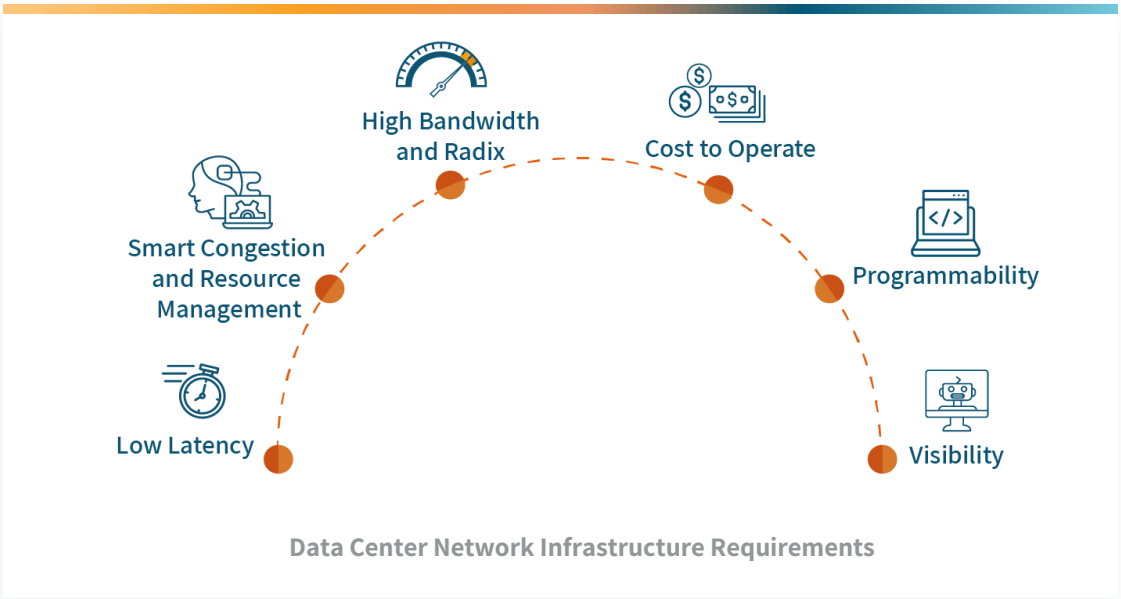
enterprises, driven by long term economic advantages, are considering keeping their workforce fully or partially remote. In summary, hyperscale data centers, immense and unpredictable workloads, data transfer cost and latency, processing latency, as well as AI, require network infrastructure to evolve and enable continuous data center growth.

Data Center Network Requirements

The following are the main requirements for a network infrastructure to enable and accelerate data center scale:

High Bandwidth and High Radix

Data explosion, distributed applications, distributed storage, new data-heavy technologies, and business logic as well as other sectors moving into cloud-based storage create a need for ever growing network bandwidth. Compute and networking nodes must be constantly evolving and provide the highest possible throughput. While high throughput is essential for a high-performance distributed system, it is not the only parameter needed to enhance the overall data center performance. Adding compute nodes requires more network connections. A smaller switch radix will require one or more network tier additions, which will increase the network diameter and therefore also increase the total network latency, introduce more queueing delays, and increase flow completion time (FCT). Increasing the switch high radix will flatten the network and therefore boost overall system performance.



Data Center Network Infrastructure Requirements

Cost to Operate

A hyperscale data center may consist of tens of thousands of hosts and hundreds of switches such that the cost to operate is always a concern. Cost can be broken down into two main categories Operation Expenses (OPEX) and Capital Expenditures (CAPEX). Switch vendors must target the minimization of both.

POWER CONSUMPTION

The switch power consumption directly drives network infrastructure OPEX. Increased switch throughput and the number of SerDes on the device significantly increase power consumption. Another aspect of switch power consumption in data centers is the overall power budget. Massive scale data centers are closing in on physical building and rack-power limits. Increased power consumption of network nodes may lead to an inability to add compute nodes due to power budget limits. The hyperscale data center switch architecture must be power-optimized in order to minimize the power consumption of the network node as a whole. Long reach SerDes eliminates the need for retimer devices and enables in-rack passive attach. True programmability delivers power savings per deployment. Monolithic die design for high throughput devices decreases power consumption.

FUTURE PROOFED ARCHITECTURE

Data center infrastructure is scaling rapidly in an attempt to absorb the newly created mounds of data oceans. Ethernet evolution is accelerating, having leapt from 25 GbE to 800 GbE in only 6 years. Switch ASICs advanced from 12.8 Tbps to 25.6 Tbps and next generation 51.2 Tbps devices are almost here. With relatively frequent infrastructure upgrade cycles, switch architecture must be ready to support the current waves of 100 GbE to 400 GbE connectivity and be ready for the next 800 GbE Ethernet standard without upgrading network infrastructure. In addition, switching ASIC architecture that is able to scale to the next generation (51.2 Tbps) with only minor changes, enables smooth and fast upgrades. 800 GbE readiness and scalable architecture allow data center operators to make the most out of their infrastructure investment and thereby minimize CAPEX.

Smart Congestion Management and Elastic Buffer

The incessant increase in the number of applications and their distribution (driven by servers' virtualization and evolution), micro-services architecture, distributed storage, and the concurrent nature of AI and high performance compute (HPC) services lead to traffic bursts and high traffic loads in the data center.

Bursts fill network node queues and lead to an increase in queueing time and queue overrun. The network is required to handle bursts, since bursts and micro-bursts may create a cascading effect on data center operation and affect business operations.

In lossy networks, packet drops caused by queue overrun lead to TCP retransmission and therefore to a significant increase in FCT. Retransmission is extremely costly for overall network latency and may cause significant performance degradation for

distributed applications and micro-services. Distributed storage cannot tolerate an FCT increase caused by TCP retransmission and therefore requires a lossless network. Lossless networks use UDP-based Remote Direct Memory Access over Converged Ethernet (RoCEv2). While UDP is scalable and doesn't carry TCP management overhead, it doesn't provide built-in reliability and therefore requires a lossless network. Lossless switching for RoCEv2 is implemented by Priority Flow Control (PFC) and Explicit Congestion Notification (ECN). However, increased queue time caused by PFC, leads to Head of Line (HoL) blocking and potential network deadlocks [3]. As a result, an increase in FCT degrades distributed services performance. One way to minimize FCT is to improve the switch's burst absorption. The data center switches' internal packet memory (buffer) is an extremely valuable resource since it cannot grow indefinitely. However, elastic switch internal memories directly lead to improved burst absorption capability. Elasticity of this resource enables its complete utilization and does not waste any unused memories (in a given deployment and/or place in network), features, and/or control tables.

Improving burst absorption by increasing the buffer size or by implementing elastic memories, is only part of the solution. In a multi-tier network, PFC has a built-in HoL blocking problem, that in some cases may lead to network deadlock[3]. PFC supports only eight priorities, and in many switches deployed today, only up to four are practically supported due to buffer management, size, and architecture limitations. The granularity of eight priorities is far from ideal. Many traffic flows are mapped to a single PFC priority because there are a small number of priorities compared to the large number of flows. Therefore, a large number of queues along with fine-grained flow control is required to minimize FCT.

Programmability

Data center network nodes are critical and enormously valuable. One of the benefits of network node programmability is its ability to adapt to future network protocols and processing. However, programmability can bring tangible value even today. Every available memory and packet processing power must be fully utilized (without "islands" of idle unused logic and memories) to serve its place in network and deployment requirements. Full and real programmability of networking devices deliver uncompromised network utilization. In addition, full and uncompromised programmability, as opposed to configurability, drives both power and latency reduction for a given deployment.

Low Latency

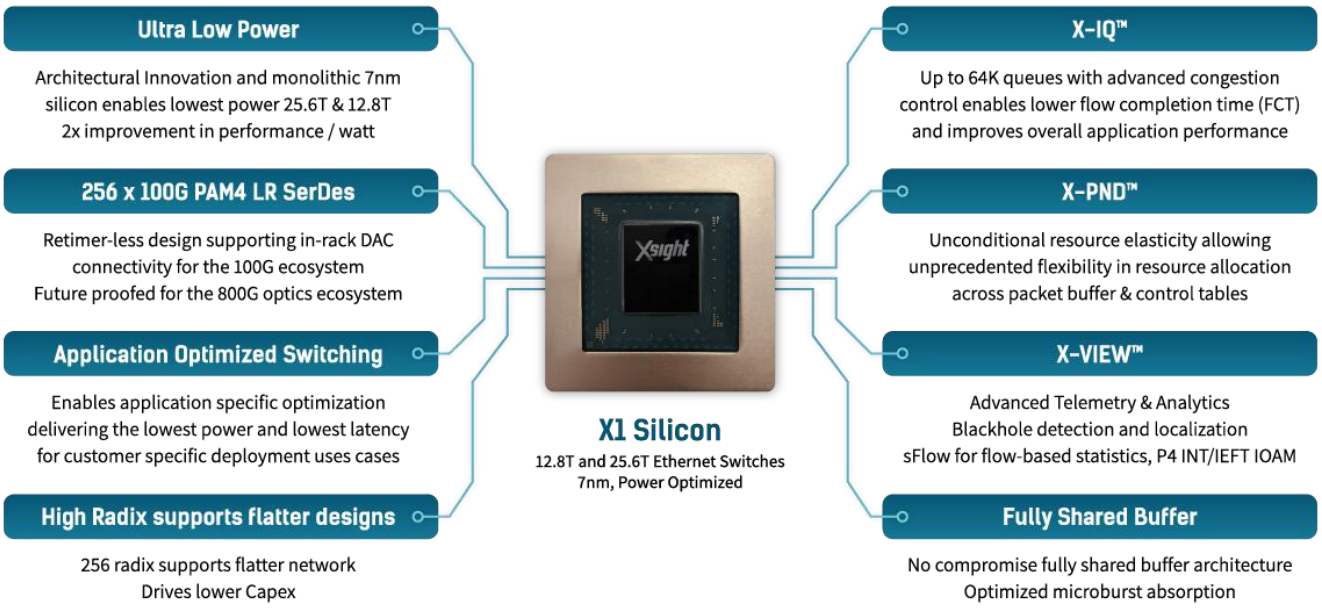
Distributed applications and storage, edge computing, AI, mission critical applications drive a need for lower latency of data center network infrastructure to accelerate performance.

Network Monitoring and Troubleshooting

Modern data centers are extremely complicated and scaled. Local malfunctions, unpredicted traffic drops, incorrect configurations, etc., may significantly affect data center operation. Finding the problem source and obtaining a clear picture of network flows is critical when attempting to mitigate issues of network automation, and traffic management applications provided by the network infrastructure platform in the shape of comprehensive telemetry features to address both network troubleshooting and monitoring.

X1 Devices Enable and Accelerate Data Center Growth

The X1 family of switching devices delivers best-in-class 25.6 Tbps full-duplex throughput, (with a robust 12.8 Tbps variation), ultra-low power, low latency, revolutionary traffic management, and uncompromised programmability. Its scalable groundbreaking architecture is highly optimized for the modern data center and delivers both OPEX (ultra-low power) and CAPEX (tangibly future-proofed for current infrastructure and next



X1 Architecture Highlights

generations). Xsight's revolutionary Intelligent Queueing (X-IQ™) along with its uncompromised shared and elastic buffer (XPND™) minimizes FCT and provides unmatched granularity of congestion management, and boosts the overall performance of data center services and applications. Application Optimized Switching delivers tangible benefits such as ultra-low power and low latency along with a tailored feature set per deployment.

Ultra-Low Power

The X1 monolithic die 25.6 Tbps design delivers a comprehensive feature set and large memories. The combination of revolutionary architecture, monolithic die, and Application Optimized Switching delivers ultra-low power for typical parallel computing and data center use cases.

High Radix

High radix switching is an important factor for today's massive scale data centers. High radix allows connecting a number of server racks to a single switch, thus enabling the system to scale. High radix also reduces the number of network nodes required to interconnect distributed compute and storage, offering a flatter and reduced diameter network. A flatter network drives overall cost down, reduces system complexity, and significantly minimizes FCT (lower network diameter), effectively boosting overall system performance.

X1's best-in-class radix of 256 ports allows creating massively scaled data centers that contain tens of thousands of compute nodes. For example, X1's high radix allows connecting up to 32,768 hosts in 2 network layers, and a massive 4,194,304 hosts in 3 layers.

100G LR SerDes

The X1 family of devices incorporates industry leading 100G LR PAM4 SerDes, enabling the design of in-rack DAC connectivity for the 100G ecosystem without the need for retimers while future-proofed for 800G optics. It enables in-rack passive copper attach and minimizes optical connectivity.

The X1 100G PAM4 and 50G NRZ LR SerDes enables interoperability and seamless integration into existing infrastructure with support for 400G and 800G modules. The X1 supports flexible port configurations using 100, 200, and 400 GbE speeds for port densities of up to 256 x 100 GbE, 128 x 200 GbE or 64 x 100GbE.

The devices deliver future-proofed systems by enabling the design of high scale systems with existing infrastructure that are ready to transition to 800G connectivity without upgrading network infrastructure.

X-IQ™

PFC's native lack of granularity, HoL blocking, and possible deadlock significantly increase FCT in bursty data center networks, all leading to an increase in tail latency and degraded system performance.

X1 X-IQ™ introduces 64K on-chip queues and fine-grained channelized flow control protocol with XFC™. This unprecedented granularity of congestion control along with its comprehensive set of traffic

management mechanisms significantly reduces queueing time and minimizes HoL blocking.

Application Optimized Switching

X1 Application Optimized Switching enables ultra-low power (for typical data center use cases) along with low latency and fully optimized packet processing.

X-PND™: ELASTIC RESOURCES

Switches must be able to absorb and handle traffic loads and bursts gracefully. Packet memory and traffic management is critical, since burst management directly affects performance of data center services and applications.

There are two main memory components in any switch architecture: packet memory and control table memory. Some architectures dedicate separate memories for packet storage and for control tables. The latter is further partitioned into a set of control tables, most of them with their own dedicated memory. Such inflexible memory architecture poses two problems. First, the inability to increase packet memory size at the expense of unused control tables, since not all tables are used or/and utilized in full in the different deployments, reducing the network's ability to sustain larger loads and bursts. Second, the lack of memory flexibility within control table blocks (inability to grow deployment critical tables at the expense of unused or under-utilized ones) decreases the network's ability to fully address deployment requirements and creates islands of unused critical memory resources. Certain switch architectures allow flexibility within the control table blocks. This approach addresses only the second problem, yet does not address the main problem.

X1's XPND™ introduces complete and uncompromised resource elasticity. X1's fully shared and elastic memory can be partitioned without limitations. This approach enables a single architecture to tailor resource allocation per data center use-case, deployment, and place-in-network in order to avoid under-utilization.

Allocating resources to a packet buffer alone is not sufficient to mitigate performance degradation caused by bursts and FCT increase. Thus, X-PND™ alongside smart buffer management, X-IQ™, and XFC™ deliver an optimized solution that enables the data center to scale, while running distributed services and applications.

FULL PROGRAMMABILITY

Traditional, "hardcoded" pipeline packet processing switch architecture was created to address a features-heavy enterprise environment at relatively small scales. Legacy architectures carry the same architecture, "a little bit of everything", approach into a cloud world. The hyperscale data center needs "lean and mean" networks that are fully utilized without legacy architectures overhead of unused memories and logic that carryover power, latency and cost penalties.

Some architectures took steps forward by introducing configurable pipeline stages and flexible memories. This approach mitigates an overhead of unused control memories and some logic, however,

it still carries the penalty of power, latency, and packet memory inflexibility.

X1 architecture delivers tangible network programmability and introduces uncompromised programmability across its processing logic, memories, and queue management. Full programmability of the X1 device delivers Application Optimized Packet Processing without overheads that lead to tangible power and latency advantages.

LOW LATENCY

Distributed applications and storage, AI, and edge computing require a low latency network in the data center. X1's revolutionary architecture components such as a monolithic die, Application Optimized Switching without overheads, and X-IQ™ deliver low and deterministic latency in modern data center deployments.

X-VIEW™: Traffic Analytics and Telemetry

The network is a critical component of parallel compute clusters. It requires application optimized network nodes and application optimized traffic management sub-systems configurations. In order to troubleshoot and optimize cluster performance, network analytics data and telemetry support is vital. X1's XVIEW™ delivers a comprehensive analytics and telemetry suite. It incorporates in-band telemetry, any-cause verbose mirroring, black hole detection and localization, real-time statistics histograms and microburst detection.

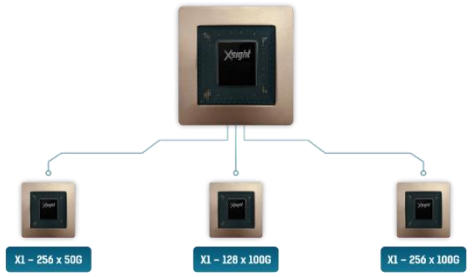
X-MON™: On-Chip Health Analytics

Data explosion drives massive growth of data centers, which in turn drives demand for large quantities of switching silicon in order to address infrastructure needs. Networking devices with 25.6 Tbps throughput, comprehensive feature sets, and large memories (such as X1) are large and complicated. Quantities along with the enormous scale of such devices makes quality and reliability more important than ever. Network reliability is critical since it directly affects parallel computing system operation. Chip vendors invest in comprehensive silicon level test coverage, screening processes, and quality and reliability monitoring processes, yet, test escapes and latent defects are a reality. Such issues often manifest themselves as in-field failures. Defective Parts Per Million (DPPM) is never 0 for chips at this scale. Prediction ability and root cause analysis of such in-field failures is practically non-existent today.

X1's X-MON™, powered by proteanTecs, is a novel approach to this problem. It delivers an in-field reliability assurance by providing readable data that enables predictive maintenance, alerts before failure, thereby extending system lifetime. In addition, X1's UCT™ dramatically reduces latent defects in deployment, significantly improves DPPM, increases in-field reliability, and reduces network down time.

The X1 System

The X1 product family is comprised of 3 different variations, all of which share the same software and feature set and are full interoperable.



X1 Family of Devices

Device Part Number	Maximum Throughput	Network Facing SerDes (Gbps)	Port Configuration Examples (GbE)
XLX1A256A	25.6 Tbps	256 x 100	• 256 x 25/50/100 • 128 x 200 • 64 x 400
XLX1A128A	12.8 Tbps	128 x 100	• 128 x 25/50 • 64 x 200 • 32 x 400
XLX1A256B	12.8 Tbps	256 x 50	• 256 x 25/50 • 128 x 100 • 64 x 200 • 32 x 400

X1 Hardware

The X1 family of devices is optimized for interconnecting data center networks. It enables building compact 1RU switches with large port densities of up to thirty-two 800G QSFP-DD/OSFP. The X1 based 1 RU system is built and backed by a leading ODM. A 1RU production-ready, cost effective system is available in multiple configurations: 12.8 Tbps (32 x 400G), 12.8 Tbps (16 x 800G), and 25.6 Tbps (32 x 800G) dual-face plate configurations — 32xOSFP or 32xQSFP-DD — enabling smooth infrastructure integration. The 1RU system's retimer–less design delivers improved power efficiency.

X1 1-RU Hardware

X1 Software

The Xsight Software Development Kit (X-SDK) delivers a comprehensive feature set. The multi-layered SDK design enables multiple integration models with different NOS types. The same X-SDK and feature set are consistent across the entire X1 family of devices. X1 software embraces open networking with SAI and SONiC integration.

Summary

The X1 family of devices are best-in-class ultra-low power switches providing throughputs of up to 25.6 Tbps. The devices are powered by flexible 100G LR SerDes, unconditionally shared buffers, data center Application Optimized Switching, X-IQ™, comprehensive traffic management subsystem, X-PND™, X-VIEW™, and X-MON™ technologies. They deliver an extremely low power and low latency, highly optimized solution for the hyperscale data center network, by addressing and minimizing current network infrastructure problems.

References

[1] Rich Miller. *Data Gravity is Shifting the Data Center Network. But in Which Direction?* https://datacenterfrontier.com/data-gravity-is-shifting-the-data-center-network-but-in-which-direction/
Accessed on 10/19/2020

[2] Rich Miller. *Study: Data Gravity Will Guide the Global Digital Economy*. https://datacenterfrontier.com/study-data-gravity-will-guide-the-global-digital-economy/
Accessed on 10/21/2020

[3] Shuihai Hu, Yibo Zhu, Peng Cheng, Chuanxiong Guo, Kun Tan, Jitendra Padhye, Kai Chen. *Deadlocks in Datacenter Networks: Why Do They Form, and How to Avoid Them*. Microsoft, Hong Kong University of Science and Technology.
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/rdmahotnets16.pdf
Accessed on 11/01/2020

About Xsight Labs

Xsight Labs is a fabless semiconductor company headquartered in Kiryat Gat, Israel with additional offices in Tel-Aviv and Binyamina. In the United States, Xsight Labs has offices in Boston, MA, Raleigh, NC, and San Jose, CA.

Founded in 2017, Xsight Labs has assembled a world-class engineering team to re-architect the foundation of cloud infrastructure by delivering a broad portfolio of products that enable end-to-end connectivity. Xsight Labs' technology delivers exponential bandwidth growth while reducing power and total cost of ownership.

Building on over 20 years of experience in developing and productizing multiple generations of cloud infrastructure products, the Xsight Labs executive team is focused on tackling modern data center challenges.

Xsight Labs
Leshem 1, Kiryat Gat, Israel
Contact a representative at sales@xsightlabs.com