

PHOTONICS spectra®

Connectivity for the AI Data Center Revolution

Features

26

Integrated Optics: Breaking the Bandwidth Bottleneck for Hyperscale AI

by Sunil Priyadarshi and Vijay Vusirikala, Arista Networks

Innovative and readily deployable optical strategies are helping to overcome the limits of copper interconnects. Large-scale AI systems are flourishing.

34

Synchronization: The Key to and Enabler of Quantum Network Fidelity

by Jonas Heidrich, Andrea Pertoldi, and Florian Emaury, Menhir Photonics

From fiber links to satellites, quantum networks demand timing precision far beyond classical systems. This article explores what it takes to synchronize photons, protocols, and nodes in the emerging quantum internet.

40

Advanced Photonics Enable the Next Generation of AI Data Centers

by Christian Urricariet, Lumentum

A set of advanced photonics technology platforms is forming a converging road map toward more efficient, flexible, and sustainable data centers.

48

Lasers Are Shaping the Automation Revolution

by Andreas Thoss, Contributing Editor

Lasers enabled modernized, automated production of cars, smartphones, and microprocessors. Now, in the age of AI, their importance for higher productivity continues to increase.



54

Photonics Illuminates the Path to Next-Generation Quantum Positioning, Navigation, and Timing Devices

by Matt Himsworth, Aquark Technologies Ltd., and Prajwal Doddaballapura, Eblana Photonics Ltd.

Incorporating the precision of quantum measurements into positioning, navigation, and timing (PNT) data offers unmatched precision across application spaces.



Subscribe or Renew Today!



www.photonics.com/subscribe

PHOTONICS SPECTRA ISSN-0731-1230, (USPS 448870) IS PUBLISHED MONTHLY BY Laurin Publishing Co. Inc., 100 West Street, PO Box 4949, Pittsfield, MA 01202, +1 413-499-0514; fax: +1 413-442-3180; email: photonics@photonics.com. TITLE reg. in U.S. Library of Congress. Copyright © 2026 by Laurin Publishing Co. Inc. All rights reserved. Copies of *Photonics Spectra* on microfilm are available from University Microfilm, 300 North Zeeb Road, Ann Arbor, MI 48103. *Photonics Spectra* articles are indexed in the Engineering Index. **POSTMASTER:** Send form 3579 to *Photonics Spectra*, 100 West Street, PO Box 4949, Pittsfield, MA 01202. Periodicals postage paid at Pittsfield, MA, and at additional mailing offices. **CIRCULATION POLICY:** *Photonics Spectra* is distributed without charge to qualified scientists, engineers, technicians, and management personnel. Eligibility requests must be returned with your business card or organization's letterhead. Rates for others as follows: \$122 per year, prepaid. Overseas postage: \$28 surface mail, \$108 airmail per year. Inquire for multiyear subscription rates. Publisher reserves the right to refuse nonqualified subscriptions. **ARTICLES FOR PUBLICATION:** Scientists, engineers, educators, technical executives, and technical writers are invited to contribute articles on optical, laser, fiber optic, electro-optical, imaging, optoelectronics, and related fields. Communications regarding the editorial content of *Photonics Spectra* should be addressed to the senior editor. Contributed statements and opinions expressed in *Photonics Spectra* are those of the contributors — the publisher assumes no responsibility for them.



An evolving context for 'co-packaged'

In the March 2024 issue of *Photonics Spectra*, Ayar Labs cofounder and CTO Vladimir Stojanovic authored an article that defined, and distinguished between, in-package optical I/O and optical co-packaging. From Stojanovic's perspective, differentiating between the two necessitates more than a technical comparison. Whereas co-packaged optics is a strategy that aims to replace pluggables by overcoming their limitations, in-package optical I/O is an interconnect solution.

The point is not simply semantic: Stojanovic ultimately does provide a comparison between co-packaged and in-package (which ought to be required reading for those who are unfamiliar). Rather, the intentionality of his phrasing supports his argument that co-packaged optics is a technological evolution in silicon integrated photonics. As a solution, in-package optical I/O is driving toward a revolution in interconnectivity.

Once again, the viewpoint goes beyond vocabulary. Co-packaged optics (in very broad terms) is a product of society's now-urgent need to optimize how it connects servers inside data centers to meet bandwidth requirements. As with many technologies of the moment, how it is designed and deployed is best understood in relation to the approach that it is replacing.

Since March 2024, co-packaged optics has advanced from a technology worth knowing to a technology that is unmissable.

NVIDIA's 2025 unveiling of its co-packaged architectures marked a turning point. Meanwhile, ongoing innovations at Broadcom, Intel, Cisco, and others make a strong case that there will be more than just one.

With these and other companies driving us toward a future powered by co-packaged optics, the present context for the technology is also shifting. The advent of co-packaged optics has placed pluggables in the past. But there are other approaches on the rise, which Arista Networks' Sunil Priyadarshi and Vijay Vusirikala discuss on page 26. Linear drive pluggable optics, near-packaged optics, and co-packaged copper are three alternatives with operational benefits of their own.

Whether and how these approaches influence the technology arc of co-packaged optics remain open questions. But as Priyadarshi and Vusirikala make clear, "breaking the bandwidth bottleneck" leaves room for multiple design innovations. In addition to looking to the past and the future, let's endeavor not to lose sight of the present.

jake.saltzman@photonics.com

Editorial Advisory Board

Robert R. Alfano, Ph.D.
City College of New York

Joel Bagwell
Elbit Systems
of America LLC

Robert Bourdelais
MKS/Newport Corp.

Walter Burgess
Power Technology Inc.

Federico Capasso, Ph.D.
Harvard University

Richard A. Crocombe, Ph.D.
Crocombe Spectroscopic
Consulting

Stephen D. Fantone, Ph.D.
Optikos Corp.

Dr. rer. nat. Bruno Gross
Thorlabs Inc.

Earl Hergert
Hamamatsu Corp.

Eliezer Manor
Shirat Enterprises Ltd., Israel

Ellen V. Miseo, Ph.D.
Miseo Consulting

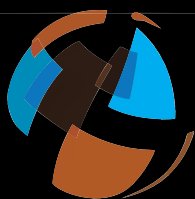
William Plummer, Ph.D.
WTP Optics

Ryszard S. Romaniuk, Ph.D.
Warsaw University of
Technology, Poland

Katie Schwertz
Edmund Optics

Joseph Spilman
Optimax Systems Inc.

**OPTIMIZED
SIGNAL-TO-NOISE
PERFORMANCE
ANOTHER REASON
TO CHOOSE
ALLUXA**



Alluxa

YOUR OPTICAL COATING PARTNER

alluxa.com

PHOTONICS spectra®

Contributors



Prajwal Doddaballapura

Prajwal Doddaballapura is an applications engineering manager at Eblana Photonics. He focuses on laser technologies for communications, quantum, and other applications. Page 54.



Sunil Priyadarshi

Sunil Priyadarshi is senior director of engineering at Arista Networks. He has held senior technical and leadership positions at Intel, Broadcom, Coherent, Source Photonics, and other leading firms. Page 26.



Florian Emaury

Florian Emaury is CEO and cofounder of Menhir Photonics, a developer of low-noise gigahertz ultrafast lasers bridging photonics and electronics, founded in 2018. He holds a Ph.D. from ETH Zürich. Page 34.



Barry Silverstein

Barry Silverstein directs the Center for Extended Reality (CXR) at the University of Rochester. He has held leadership roles at META's Reality Labs and now chairs the AR Alliance's LDAR working group. Page 67.



Jonas Heidrich

Jonas Heidrich is product line manager at Menhir Photonics, developing ultralow-noise femtosecond laser systems for scientific and quantum technologies. He holds a Ph.D. from ETH Zürich. Page 34.



Andreas Thoss

Andreas Thoss, Ph.D., is a laser physicist, founder of THOSS Media, and a *Photonics Spectra* contributing editor. He has been writing and editing technical texts, with a focus on photonics, for two decades. Page 48.



Matt Himsworth

Matt Himsworth is the chief scientific officer at Aquark Technologies. He previously spent five years at the U.K.'s Defence Science and Technology Laboratory's quantum laboratory. Page 54.



Christian Urricariet

Christian Urricariet is senior director of product and strategic marketing at Lumentum, with more than 25 years of industry experience and a degree in electrical engineering (optoelectronics specialization). Page 40.



Ivan Nikitskiy

Ivan Nikitskiy, Ph.D., is program manager at EPIC, where he brings both academic and industrial experience in quantum technology, optoelectronics, and semiconductor components. Page 63.



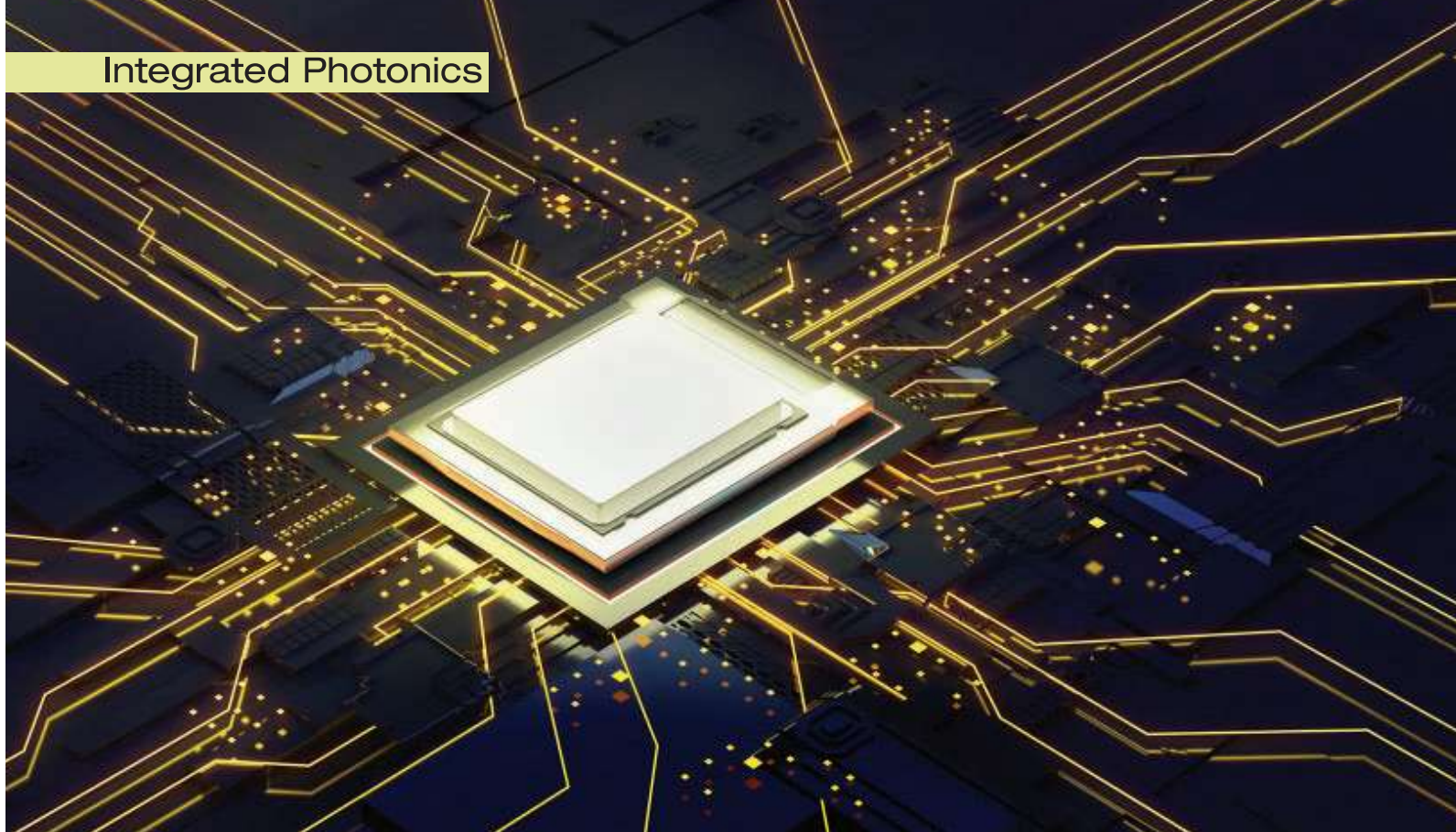
Vijay Vusirikala

Vijay Vusirikala leads AI Systems and Networks at Arista Networks, focusing on networking for hyperscalers and large AI clusters. An Optica Fellow, he serves on the OFC Long Range Planning Committee. Page 26.



Andrea Pertoldi

Andrea Pertoldi, Ph.D., is a laser application engineer at Menhir Photonics, supporting partners in frequency-comb stabilization and low-noise ultrafast laser system development. Page 34.



Integrated Optics:

Breaking the Bandwidth Bottleneck for Hyperscale AI

Innovative and readily deployable optical strategies are helping to overcome the limits of copper interconnects. Large-scale AI systems are flourishing.

BY SUNIL PRIYADARSHI AND
VIJAY VUSIRIKALA
ARISTA NETWORKS

Modern AI is reshaping the future of data centers, driving a shift toward architectures capable of handling unprecedented demands for compute, memory, storage, and interconnect bandwidth. Training trillion-parameter models, such as ChatGPT and Gemini, while delivering real-time inference to billions of devices, requires massive infrastructure optimized simultaneously for scale, efficiency, and performance. Large-scale AI systems are central to this evolution, processing vast data sets across domains such as natural language processing, computer vision, and genomics while ensuring high data throughput for both training and inference.

Advanced techniques, such as hyperparameter tuning and ensemble learning, are emerging to enhance model accuracy, accelerating breakthroughs in research

and cross-disciplinary innovation. At the same time, these systems support latency-sensitive applications, including autonomous driving, interactive AI, and robotics. They have become indispensable across industries such as manufacturing, logistics, and finance, enabling AI-driven automation for more efficient and dynamic operations.

To sustain this momentum, data centers must evolve beyond traditional designs and adopt new approaches to hardware, networking, cooling, and energy efficiency to meet AI's insatiable demands.

AI models: Doubling every six months

During the rise of the social networking era, content service providers built massive hyperscale data centers to support general-purpose workloads such as web hosting, cloud applications, video streaming, and enterprise software. Conventional central processing units (CPUs) — similar to those found in laptops — powered these facilities. While effective, these CPUs were optimized for versatility rather than specialization.

The advent of AI has shifted the focus toward highly specialized processing units designed specifically for AI workloads. As a result, hyperscale data centers are undergoing a transformation: CPUs now primarily handle orchestra-

tion and networking, while GPUs carry the immense computational burden of AI training and inference.

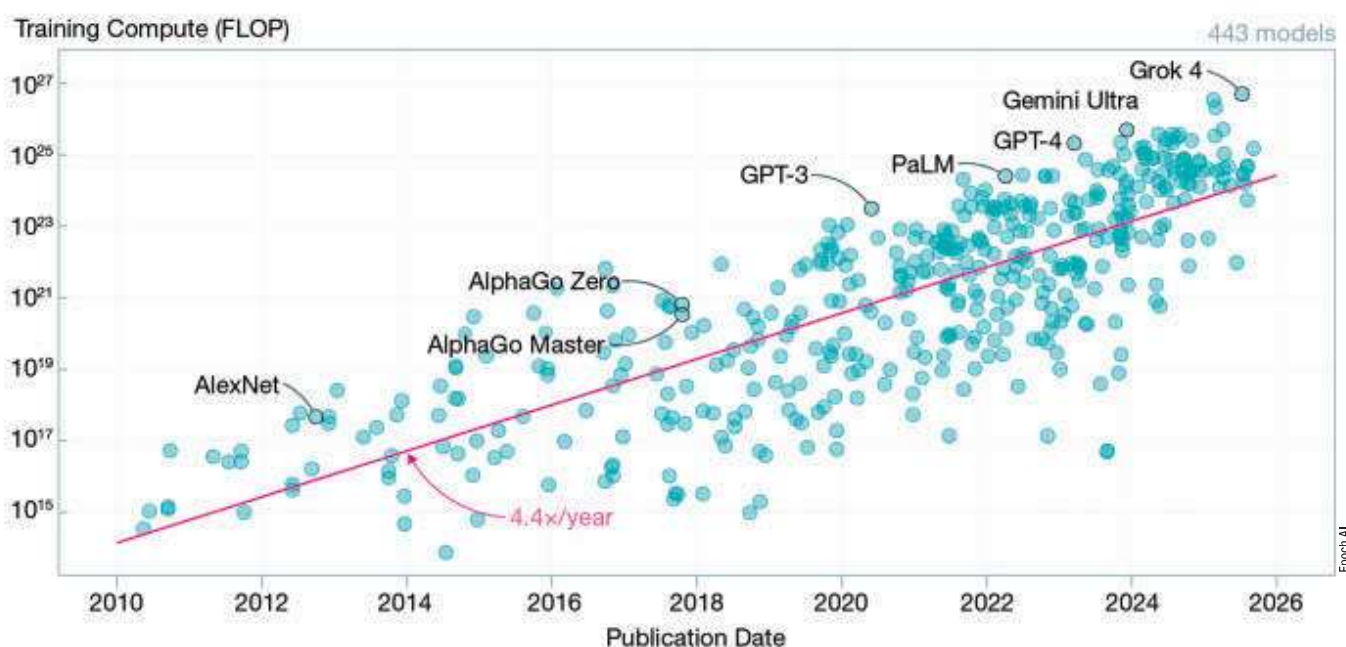
Purpose-built for parallel processing, GPUs enable the training of trillion-parameter models and deliver real-time inference at global scale. A single large AI training run can consume >100 MW of power — rivaling the energy usage of a small city and highlighting the scale of resources required. Inference has become a ubiquitous background process, integrated into nearly every digital interaction, from personalized recommendations to autonomous systems.

To support this shift, data centers are evolving into specialized environments designed for compute density, massive parallelism, and high-speed data movement. Increasingly, these AI factories are powered by custom accelerators, advanced interconnects, and modular architectures that consume unprecedented amounts of power per rack.

Since 2010, the compute required to train state-of-the-art AI models has grown at an average rate of $4.4\times$ per year (Figure 1)¹. In contrast, hardware bandwidth doubles every two to three years. Figure 2 illustrates how Ethernet/switch port bandwidth has increased markedly.

The result is a widening gap between model demands and system capabilities

Figure 1. The graph shows AI training compute demand growth through the progression of trillion-parameter models, measured in floating-point operations (FLOPs). Adapted with permission from Reference 1.



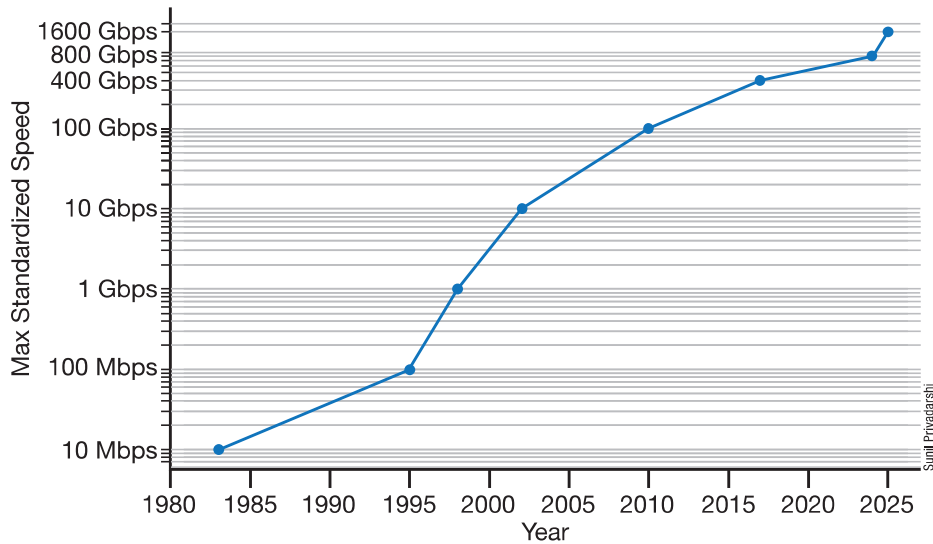
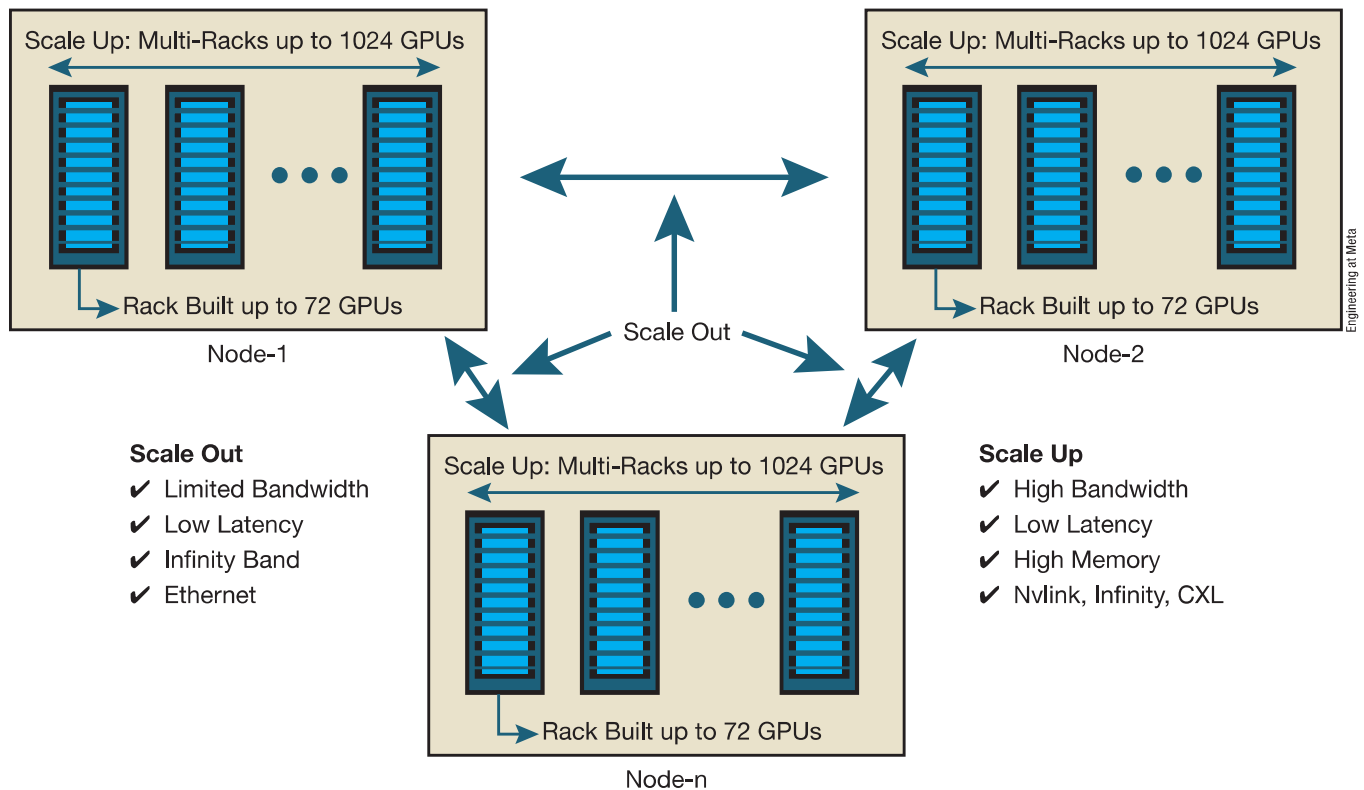


Figure 2. Maximum standardized Ethernet/switch port speed has doubled about every two to three years (**left**). The graph shows this progression over a period of >40 years.

Figure 3. Large-scale AI systems, via scale-up and scale-out strategies (**below**). Adapted with permission from Reference 3.



— a critical bottleneck for large-scale AI training and inference.

Scale-up and scale-out strategies

Large-scale AI systems use scale-up and scale-out strategies to meet the immense computational demands of AI workloads. Scale-up, also known as vertical scaling, compute fabric, and/or memory-semantic fabric, refers to the network architecture

within an AI compute node. It focuses on maximizing intranode communication bandwidth and minimizing latency among GPUs within a node.

Scale-out, also called horizontal scaling and/or backend networks, expands capacity by adding multiple nodes horizontally, which is ideal for distributed computing environments. Horizontal scalability enables the seamless integra-

tion of large numbers of GPUs to meet demand. Scale-out architectures enhance load balancing, improve fault tolerance, prevent bottlenecks, and provide elasticity — allowing rapid adaptation to changing workloads. This makes scale-out well suited for dynamic applications and cost-effective deployment. Figure 3 illustrates how large-scale data centers are built using scale-up and scale-out strategies².

To scale up, GPUs are connected using high-speed interconnects. System architects have traditionally used onboard printed circuit board (PCB) traces, backplanes, or passive or active direct-attached copper cables because of their low cost and power efficiency. Copper interconnects remain the dominant choice for scale-up networks, while optics are primarily used for scale-out networks.

For example, in NVIDIA's NVL72 system, 72 GPUs are linked through 18 NVLink switches in a 2.2-m-high rack, using >5000 unique copper cables at 100 Gbps per link in each direction (200 Gbps bidirectional)³. However, as GPU port speeds and counts continue to increase — even within a single rack — these copper-based solutions are reaching their limits in both bandwidth and reach.

Drivers for the copper-to-optical transition

Comparing copper trace performance of 224 Gbps versus 448 Gbps highlights several critical challenges of operating at higher data rates. At 224 Gbps, copper traces can manage short distances using advanced design techniques. However, as data rates increase to 448 Gbps, copper traces face bandwidth limitations due to frequency-dependent attenuation in copper transmission lines, which struggle to support such high frequencies. Additional bottlenecks arise from signal loss and reduced signal-to-noise ratios.

Current scale-out networks widely use pluggable optics at 400 Gbps, 800 Gbps, and, soon, 1.6 Tbps. Over the years, pluggable optics bandwidth has increased due to advancements in modulation techniques, such as pulse amplitude modulation (PAM4), and the use of retimers and digital signal processors (DSPs) to improve error detection and correction.

However, these improvements have also significantly increased power consumption. For example, a fully retimed 1.6-Tbps pluggable optical transceiver in the octal small form-factor pluggable (OSFP) format can consume up to 25 W, of which 15 W could be just for the DSP.

Scale-up networks require connectivity between every GPU and therefore demand much higher interconnect bandwidth per GPU than scale-out networks. Today, scale-up networks are generally limited to a single rack due to the cost-effectiveness and power efficiency of copper-

based interconnects. Expanding to a multi-rack scale-up network is desirable from a performance perspective, but the limited reach of copper prevents this. Consequently, the need for denser, lower-power, cost-effective, and mass-manufacturable optical interconnects is critical. For example, an AI cluster with 128,000 GPUs — each requiring 1.6-T scale-out bandwidth and 12.8-T scale-up bandwidth — would require more than a million transceivers for the scale-up network alone.

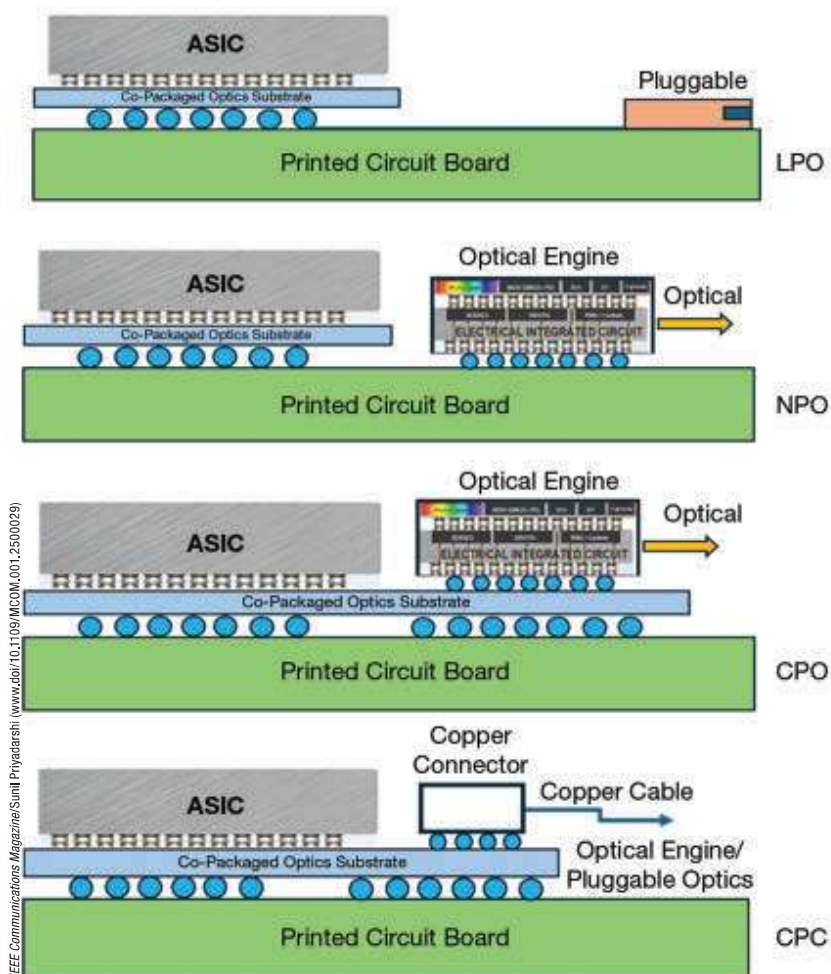
Lower latency is another key requirement, enabling enhanced memory semantics in high-performance computing, advanced memory architectures, and improved data processing capabilities. These optical interconnects must also be future-proof, supporting advanced network topologies while ensuring the performance and scalability of next-generation AI systems.

The industry is actively developing emerging solutions based on linear interfaces, including linear pluggable optics (LPO), co-packaged optics (CPO), near-packaged optics (NPO), and co-packaged copper (CPC). Each approach aims to deliver optics with reduced power consumption, lower latency, and increased bandwidth density.

Pluggable and co-packaged optics

LPO is gaining traction in AI systems due to its similarity to traditional pluggable modules, with the key difference being the omission of the internal DSP. This design choice has significant implications

Figure 4. Different optical interconnect architecture types are compared and contrasted. The architectures are linear pluggable optics (LPO), near-packaged optics (NPO), co-packaged optics (CPO), and co-packaged copper (CPC). ASIC: application-specific integrated circuit.



for system performance, as LPO is more sensitive to host signal integrity and the overall channel loss budget, which can limit reach and reduce interoperability across systems. Architecturally, LPO maintains compatibility with existing network infrastructure, enabling straight-forward integration and serviceability. The removal of the DSP directly results in lower latency, reduced power consumption, and lower cost. LPO also benefit from the established multivendor pluggable optics manufacturing ecosystem, making them a highly cost-effective option. LPO offers the lowest power consumption among transceivers, using ~10 W per 1.6-Tbps module for direct-reach optics.

NPO positions optics close to the ASIC on the same PCB, improving signal integrity and reducing overall channel loss, which enables better power efficiency by eliminating DSPs. This approach is especially suited for high-density applications where pluggable modules, such as LPO, are not ideal. Compared with CPO, NPO requires only moderate system changes rather than a full redesign, offering a practical balance of flexibility, scalability, and power efficiency.

In CPO, optics are co-packaged with the ASIC on the same substrate, further improving signal integrity and reducing channel loss to deliver even better power efficiency. Although CPO solutions require more complex designs and have greater thermal management and manufacturing challenges, they are ideally suited for high-performance environments where maximum efficiency and density are crucial — such as in AI data centers.

CPC is a newer approach in which a high-speed copper connector, rather than optics, is co-packaged with the ASIC on the same substrate. Optics or other ASICs are then connected via copper flyover cables, which have much lower loss and avoid ball grid array (BGA) transitions from the ASIC substrate to the PCB, as well as the higher loss of PCB traces. This reduces channel loss and improves signal integrity, similar to NPO, while allowing optics to be mounted farther from the ASICs.

As an emerging technology, CPC is still under active development but shows strong potential for supporting higher



Figure 5. Meta's Catalina ORv3-HPR rack shown from the front (left) and rear. It delivers up to 140 kW/rack for next-generation AI workloads. Adapted with permission from Reference 5.

data rates — such as 200 G/lane LPO and, eventually, 400 G/lane — by simplifying channel design. CPC provides a more cost-effective and easier-to-implement alternative to full CPO.

The power efficiency of LPO, NPO, CPO, and CPC is similar, as none of these approaches use retimers or DSPs on the optics side (Figure 4).

New and improving protocols

To further support optics for scale-up using these integrated optics technologies, the networking industry is advancing new protocols and enhancing existing ones.

Ethernet for Scale-Up Networking (ESUN) is an open industry collaboration focused on optimizing standard Ethernet technology for high-performance AI scale-up infrastructure. ESUN leverages and enhances the existing Ethernet ecosystem to provide high bandwidth, ultralow latency, and lossless transport for AI workloads within a single cluster. ESUN leverages these characteristics to optimize the network and transport layer, resulting in simpler protocols, smaller headers, and unified, low-overhead memory access among XPU.

This modular framework is built on three key building blocks: common Ethernet headers for maximum interoperability; the Open Ethernet Data Link Layer, which provides a high-performance foundation for AI collectives using

standards-based mechanisms; and the ubiquitous Ethernet PHY Layer, which assures vendor interoperability across various optical and copper interconnect options.

Meanwhile, Ultra Ethernet, spearheaded by the Ultra Ethernet Consortium, is creating a next-generation Ethernet architecture designed precisely for AI and high-performance computing workloads. Ultra Ethernet aims to offer higher performance, improved congestion management, lower latency, and enhanced security compared with traditional Ethernet.

These efforts, combined with advancements in linear-drive optics and in advanced packaging and interconnection techniques, such as CPO and CPC, are laying the foundation for next-generation AI infrastructure. Together, these innovations will provide the performance, scalability, and efficiency required for AI systems to handle massive data throughput and real-time processing demands. This, in turn, will ensure that networks pace with the unprecedented scale of modern AI workloads.

Breaking traditional limits

Power delivery and thermal management have become central to modern data center design. Traditional racks once consumed only 2 to 5 kW, but AI-focused racks now routinely demand 80 to 150 kW, with future deployments projected to reach 300 kW to >1 MW/rack. To support these massive requirements, data centers are increasingly adopting high-voltage distributed current power distribution (400 to 800 V), reducing both conversion and transmission losses.

Current hyperscale facilities often distribute power through multiple low-voltage conversion stages that rely on bulky and lossy components, such as inductors. Solid-state transformers (SSTs) provide a breakthrough by enabling direct, high-frequency conversion from 800 V to lower voltages such as 48 V or 12 V, reducing conversion steps, minimizing losses, and improving power density. SSTs are smaller, lighter, and more efficient than conventional transformers. They also provide advanced monitoring, fault isolation, and bidirectional control, supporting renewable energy protocols and backup systems.

Advanced cooling technologies are

also emerging to manage extreme thermal loads from GPUs, switches, and dense servers. Solutions include direct-to-chip cold plates, rear-door heat exchangers, and full-system immersion cooling. Energy reuse systems capture and repurpose waste heat, improving sustainability. The convergence of high-power AI hardware, innovative cooling, and next-generation power architectures enables data centers to scale efficiently for AI workloads.

A leading example is Open Rack V3 (ORv3) from the Open Compute Project, which modernizes infrastructure with efficiency, modularity, and serviceability⁴. ORv3 introduces a 48-V distributed current/distributed power architecture, reducing copper losses compared with traditional 12-V systems. It supports 30 to 60 kW/rack and accommodates air and liquid cooling, which is ideal for cloud and enterprise compute environments.

The ORv3-HPR (high power rack) extends ORv3 for GPU-dense AI and high-performance compute workloads.

Designed for 80 to 150 kW/rack, it reinforces the busbar, integrates liquid cooling with cold plates and coolant units, and strengthens mechanics for high-density deployments. Meta's Catalina Rack, supporting NVIDIA GB200 Grace Blackwell GPUs, exemplifies this evolution (Figure 5)⁵. It delivers up to 140 kW/rack for next-generation AI workloads.

As rack power climbs toward a megawatt, next-generation data centers are adopting 400- to 800-V distributed current/distributed power to reduce copper losses and boost efficiency. A 300-kW rack at 48 V would draw ~6250 A — far beyond what copper busbars can efficiently manage. By distributing power at higher voltages and using SSTs for localized conversion, data centers can significantly reduce current, resistive losses, and infrastructure complexity.

At the same time, traditional cold plates are reaching their thermal limits because heat must still travel through multiple material layers. Microfluidic cooling, which channels coolant within microns

of the silicon, delivers much higher heat-transfer efficiency and significantly mitigates hotspots. Combined with high-voltage DC power distribution, these advanced cooling technologies enhance reliability, thermal performance, scalability, and operational cost — paving the way for next-generation AI-optimized data centers.

Reliability and environmental awareness

The unprecedented scale and complexity of AI data centers place extreme demands on reliability. Any downtime or performance degradation can negatively affect AI training time, inference latency, and operational costs. To ensure continuous operation, data centers employ advanced monitoring, predictive maintenance, and automated failover mechanisms. The co-design of hardware and software further enables dynamic resource allocation and fault isolation to mitigate risks.

Environmental sustainability is also a key priority. Data centers are among the largest consumers of electricity globally, and AI workloads are driving increased

WHEN ORDINARY DIODE LASERS WON'T DO...

**QPC LASERS IS
ALKALI VAPOR
PUMPING**

QPC Lasers Inc. • info@qpclasers.com
Tel: +1 818 986 0000 • www.qpclasers.com

QPC LASERS 

power demand. Ongoing and planned industry initiatives focus on renewable energy integration, water-efficient cooling, carbon-neutral operation, and circular economy practices, such as component recycling and reuse. Advancements in power delivery, cooling, and optics technologies contribute directly to these goals by improving efficiency and reducing waste.

Conclusion

The shift from CPU-centric to GPU-accelerated AI factories demands innovative networking solutions, particularly in optical interconnects, to overcome the limitations of copper at scale. Emerging technologies such as LPO, LRO, NPO, CPO, and CPC, along with optical protocols such as ESUN and Ultra Ethernet, are critical to meeting these challenges.

Simultaneously, power and cooling systems are evolving with high-voltage power distribution, SSTs, and sophisticated microfluidic cooling techniques. These advancements, combined with a

focus on reliability and sustainability, are shaping the future of data centers that can support the next generation of AI workloads.

As AI models continue to increase in size and complexity, data centers must remain agile and innovative. Designers must embrace new architectures, protocols, and technologies to sustain performance while minimizing environmental impact. Ultimately, the future of AI depends not only on algorithms and models but also on the foundational infrastructure that enables them.

Meet the authors

Sunil Priyadarshi is senior director of engineering at Arista Networks, leading optical networking architecture for data centers and large AI systems. He has held senior technical and leadership roles at Intel, Broadcom, Coherent, Source Photonics, Lucent Technologies, and NEC; email: spriyadarshi@arista.com.

Vijay Vusirikala leads AI systems and networks at Arista Networks, focusing on networking for hyperscalers and large AI

clusters. He is an Optica Fellow and serves on the OFC Long Range Planning Committee; email: vijayv@arista.com.

Acknowledgment

The first author wishes to thank Andy Bechtolsheim for his invaluable suggestions, insights, support, and encouragement.

References

1. Epoch.AI. Data on AI models/Epoch.AI AI model database, www.epoch.ai/data/ai-models.
2. S. Priyadarshi. (2025). Advancing AI scalability and performance with optical interconnects. *IEEE Communications Magazine*, Vol. 63, No. 11, pp. 22-27.
3. Nvidia.com (cloud and data center). Unlocking real-time trillion-parameter models, www.nvidia.com/en-us/data-center/gb200-nvl72.
4. Opencompute.org. www.opencompute.org.
5. D. Rabinovitsj and O. Baldonado. Facebook (Meta) Engineering. Meta's open AI hardware vision, www.facebook.com/2024/10/15/data-infrastructure/metasp-open-ai-hardware-vision.

CUSTOM COATINGS

Precision • Durability • Performance

DPT
MRF
PATTERNS

REYNARD CORPORATION
Protecting Your Light™

ISO 9001:2015 • ITAR • CMMC
reynardcorp.com

MCL
MAD CITY LABS INC.

Think Nano®
Positioning | Microscopy | Solutions

Closed Loop Nanopositioning
Micropositioning
AFM & Quantum Scanning Microscopy
Single Molecule Microscopes
Custom Solutions

Visit APS Global Summit #1600
sales@madcitylabs.com • www.madcitylabs.com