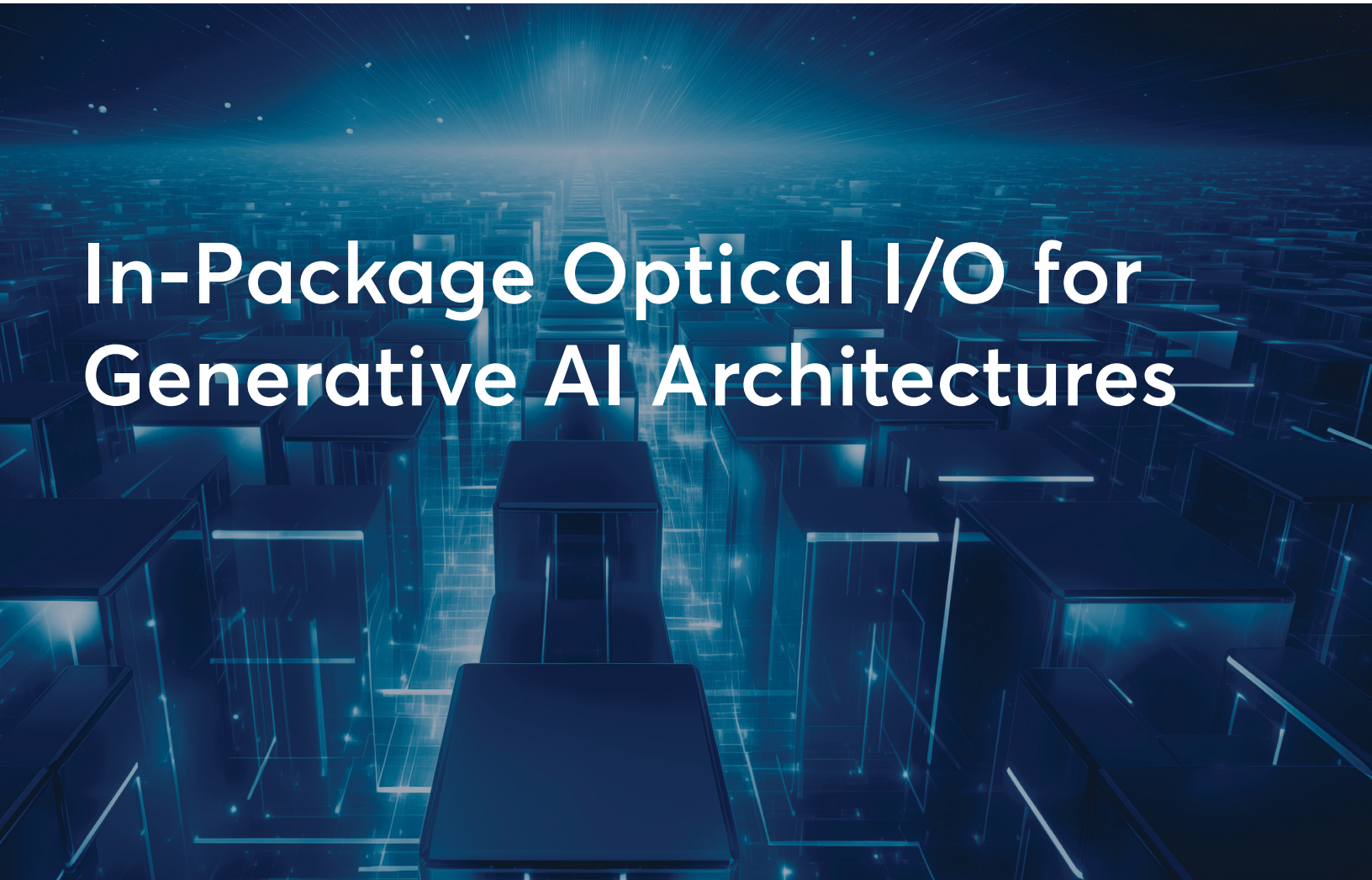
AyarLabs

Solution Brief



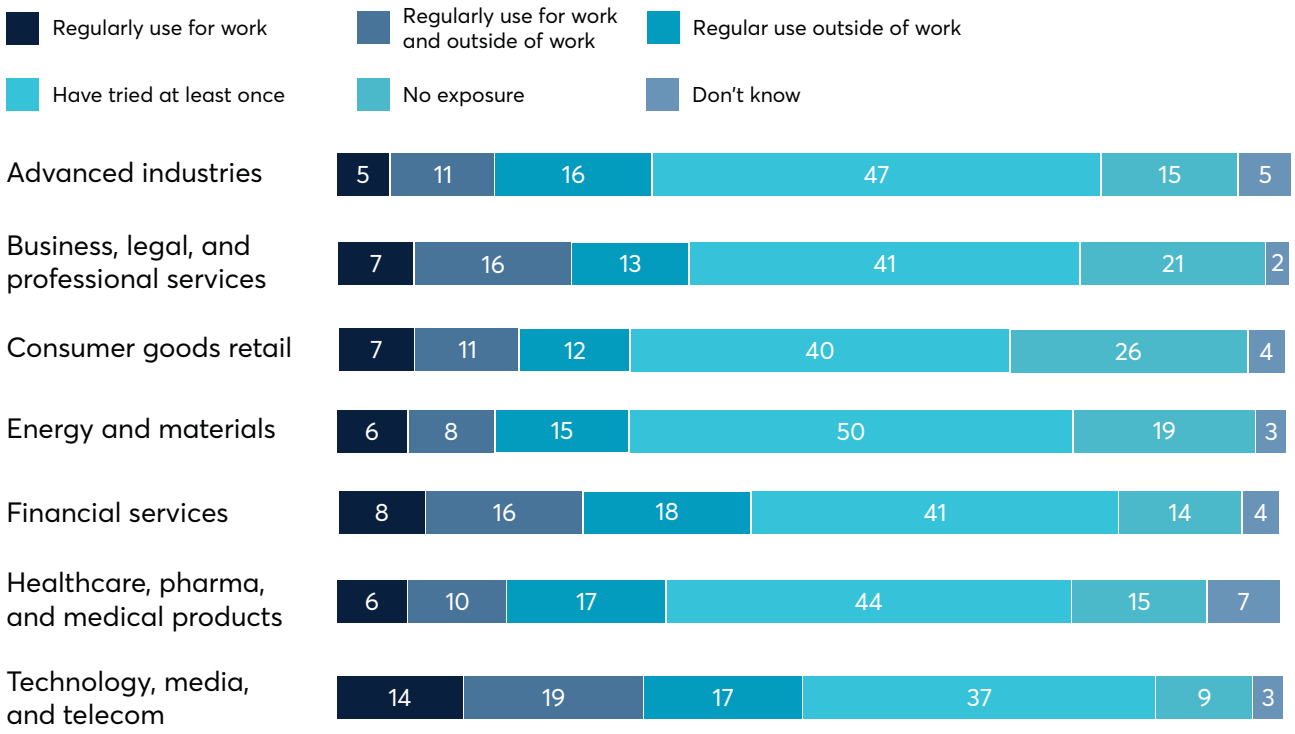
In-Package Optical I/O for Generative AI Architectures

Introduction

The versatility of generative artificial intelligence is staggering. From financial institutions using it for analysis and fraud detection, to pharmaceutical companies relying on the technology to speed up drug discovery, generative AI is being explored in multiple industries.

A VentureBeat survey from July 2023 found that 54.6 percent of institutions are experimenting with generative AI on a small scale to see if it can fit their needs. What's more, the McKinsey Global Survey from August 2023 found that 79 percent of respondents say they have at least some exposure to generative AI — either in their everyday life or at work — and 22 percent said they regularly use it at work.

Reported Exposure of Generative AI Tools by Industry, Percent of Respondents



Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11-25, 2023

Figure 1: Generative AI tools in use by industry.

Despite the growing demand, adoption at scale is exclusive to massive cloud providers and hyperscalers with the resources and hardware to support the complex demands of this technology. Generative AI architectures require high-performance hardware — such as GPUs, TPUs, or ASICs — and large amounts of memory and storage to handle complex tasks and large datasets.

Efficient data transmission and resource allocation are crucial for the performance of any AI system. Networking infrastructure must be able to handle increased demand for data transfer and communication between different components of AI systems.

For generative AI specifically, due to the large models used, many GPUs must essentially function as one giant GPU. This puts a higher strain on interconnections between these GPUs that simply cannot be handled by a traditional network. Thus, a more efficient interconnect is required.

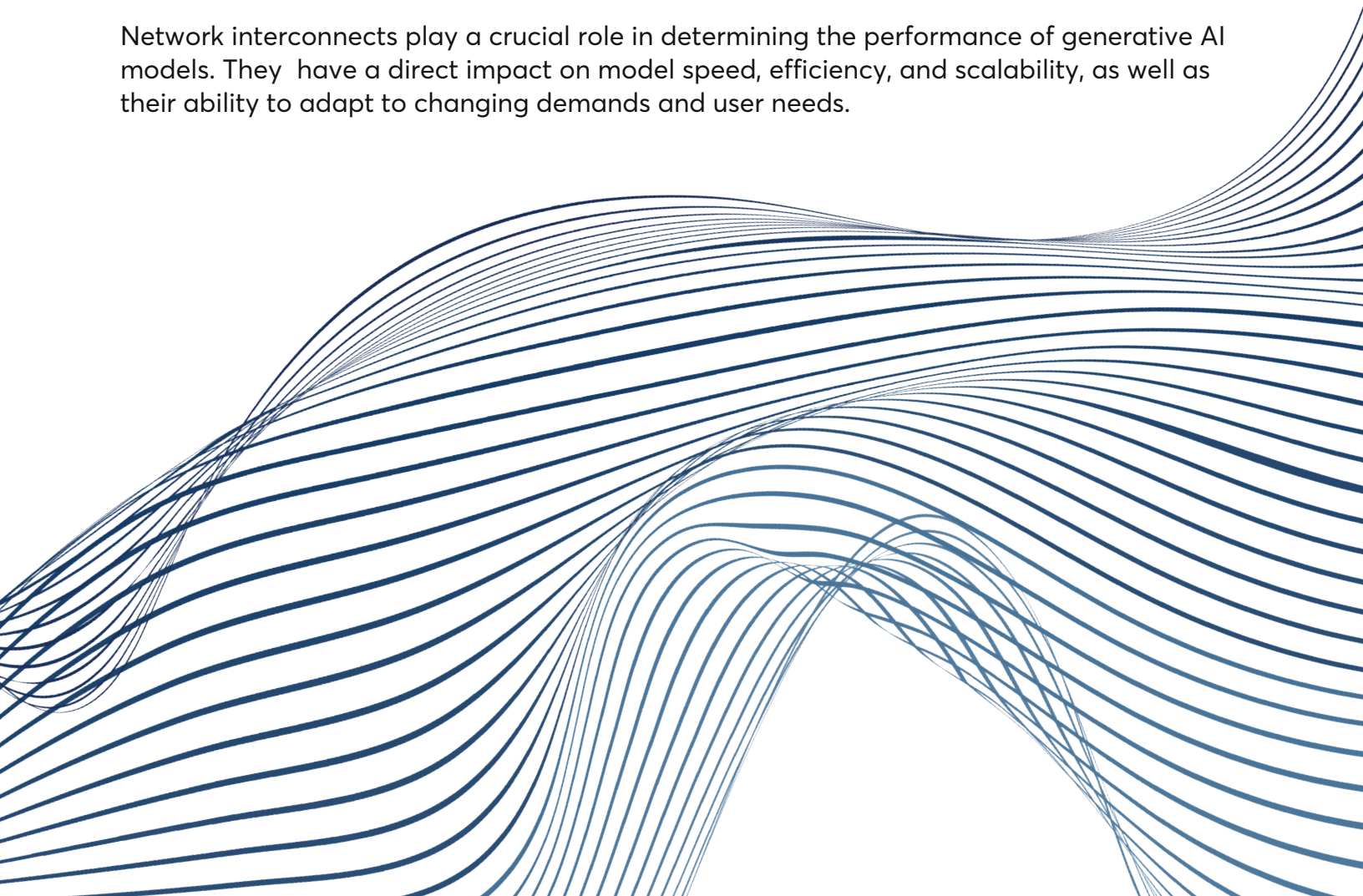
To address these challenges, high-bandwidth, low-latency, and power-efficient interconnects are needed. These advanced interconnects enable efficient data transmission, seamless scalability, and efficient collaboration between devices, ensuring optimal performance and efficiency for generative AI applications.

Unfortunately, electrical I/O combined with pluggable optical interconnects are vastly insufficient, and these traditional interconnects cannot provide what generative AI work requires. To truly unlock the power of generative AI, organizations must look to in-package optical interconnects, which provide distance-insensitive system-wide package-to-package connectivity.

This brief explores:

- The specific networking challenges presented by generative AI architectures
- Why traditional interconnects — electrical interconnects with pluggable optics — are a suboptimal solution that will keep generative AI from reaching its full potential
- How in-package optical I/O can overcome these challenges

Network interconnects play a crucial role in determining the performance of generative AI models. They have a direct impact on model speed, efficiency, and scalability, as well as their ability to adapt to changing demands and user needs.



Scaling Challenges Presented by AI

Large language models (LLMs) are some of the largest within generative AI — and they are growing. LLMs themselves represent foundational or generative AI models simply because they capture more than just language syntax and can be used to derive other applications on top (for example, search, vision, etc.).

AI and Memory Wall

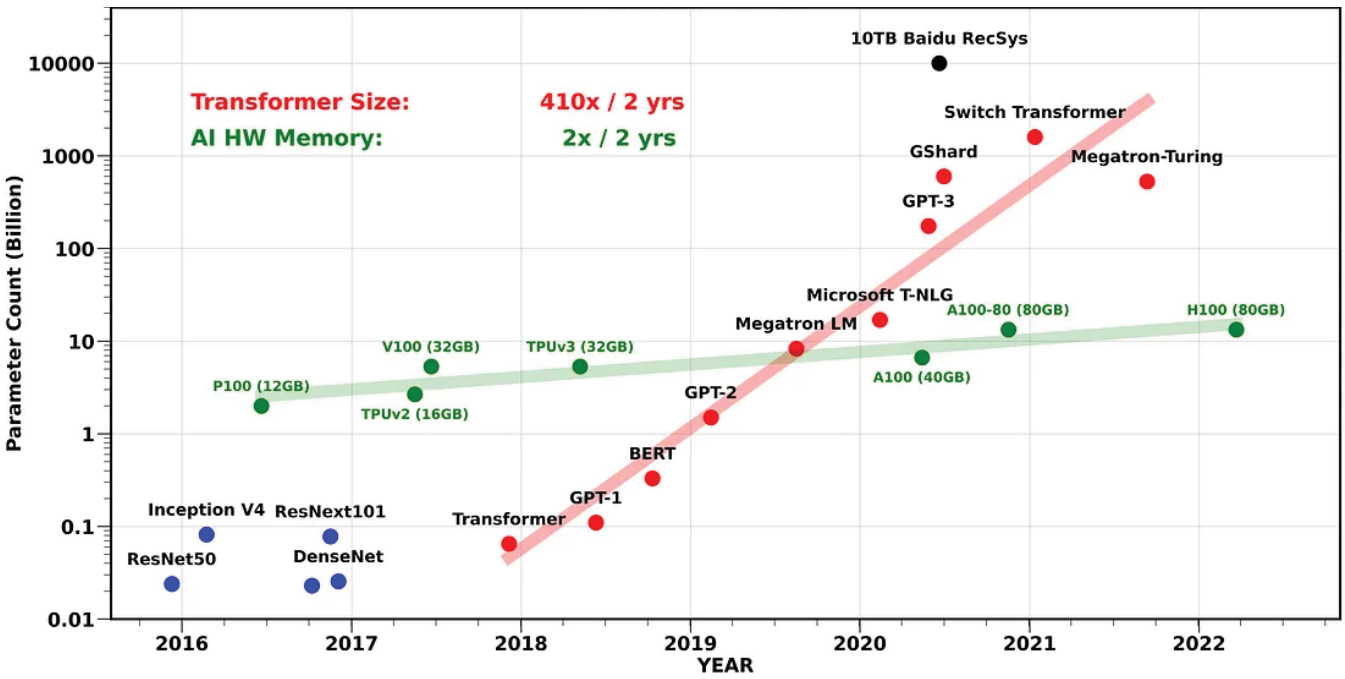


Figure 2: The evolution of the number of parameters of SOTA models over the years, along with the AI accelerator memory capacity (green dots). The number of parameters in large transformer models has been exponentially increasing with a factor of 410x every two years, while the single GPU memory has only been scaled at a rate of 2x every two years. Source.

GPT-4 Has More than a Trillion Parameters, a report from THE DECODER, stated that OpenAI's new GPT-4 language model is said to be based on eight models with 220 billion parameters each, which adds up to 1.76 trillion parameters. For reference, GPT-3 had 175 billion parameters in 2020 and GPT-2 had 1.54 billion in 2019.

As these parameter sizes increase, so too does the need for high throughput. As a general rule, throughput can only be increased in two ways — adding nodes or increasing the speeds of each node.

The process of training, fine-tuning, and inferencing is another concern. In general, inferencing already requires tens of GPUs, fine-tuning hundreds of GPUs, and training thousands of GPUs. As the models continue to grow rapidly in size and complexity these numbers are likely to increase by at least an order of magnitude in each category in the next generation. Since a typical chassis holds at most eight GPUs, and a rack holds 16-32 GPUs, this means that even inferencing is a rack-scale operation with a high demand for efficiency.

A data center is size- and power-limited — for example, think of a 100 MW data center. With typically ~1 kW per GPU (with associated memory, storage, and interconnect), this means you can have up to ~100 K GPUs per data center. A data center operator needs to decide what to use them for. Given that inference is most frequent and brings in money as a service, you can fit up to 10,000 x 10 or 1,000 x 100 GPU inference systems. Fine-tuning requires 100-1,000 GPUs per system, so again you can only fit so many of them within the 100 K GPU footprint.

These enormous compute requirements mean that generative AI architecture is a prime candidate for disaggregation. This would give systems the flexibility to allocate GPUs to different tasks dynamically. But, again this requires efficient interconnect fabric.

Generative AI systems need global communication across many GPUs, and therefore demand low latency and high bandwidth. GPUs need to be able to talk to each other beyond a single chassis or rack in a data center. Properly performing training, fine-tuning, and inference demands a lot of resources — one rack for inference, tens of racks for fine-tuning, and hundreds of racks for training. All of these GPUs need to interconnect to complete the generative AI task. In a sense, the entire distributed compute operation needs to look like a single virtual GPU.

This requires low latency of data transport, as the GPUs must be able to communicate efficiently without waiting on each other's workloads. Adequate bandwidth ensures that the large volumes of data associated with generative AI tasks can be transmitted swiftly between GPUs.

This need for global communication across GPUs, along with the ever-increasing parameter sizes of these generative AI models, demands high scalability and uniform latency. While low latency itself is important, uniform latency is crucial as organizations scale up the number of GPUs in their systems. Without uniform latency, the system's scalability is limited by some GPUs experiencing significant communication bottlenecks and even sitting idle, while others perform efficiently. Uniform latency is key to having a more efficient software/programming model, which in the end limits system scalability.

Under all of these performance challenges lies the ever-important consideration of energy efficiency. Recent research into GPT-3 found training the model took 1.287 Gigawatt hours (Gwh) of energy to train, generating 552 metric tons of CO_2 and other greenhouse gasses, for an energy cost of about $193,000.

NVIDIA founder and CEO Jensen Huang made some interesting points about energy efficiency during his keynote speech at COMPUTEX 2023. Huang discusses the accelerated computing architectures required for LLMs. He points out that $10 million can buy 960 CPU servers, which need 11 Gwh to train one LLM. However, that same $10 million can buy 48 GPU servers, which need 3.2 GWh to train 44 LLMs. Even more interesting, $400 k can buy two GPU servers, which need 0.13 GWh to train one LLM. So, for the same training performance, GPUs are 25x ($10 M/$400 k) more cost-effective and 85x more energy efficient (11 GWh/0.13 GWh).

While his discussion was about efficiency, Huang brings his point home with a quote that can easily be applied to interconnects: "The GPU server is no longer the computer. The computer is the data center."

Generative AI tools require a large amount of energy and compute resources to function. Staying relevant within this space will demand a heavy focus on scalability and a consideration for the data center as a whole.

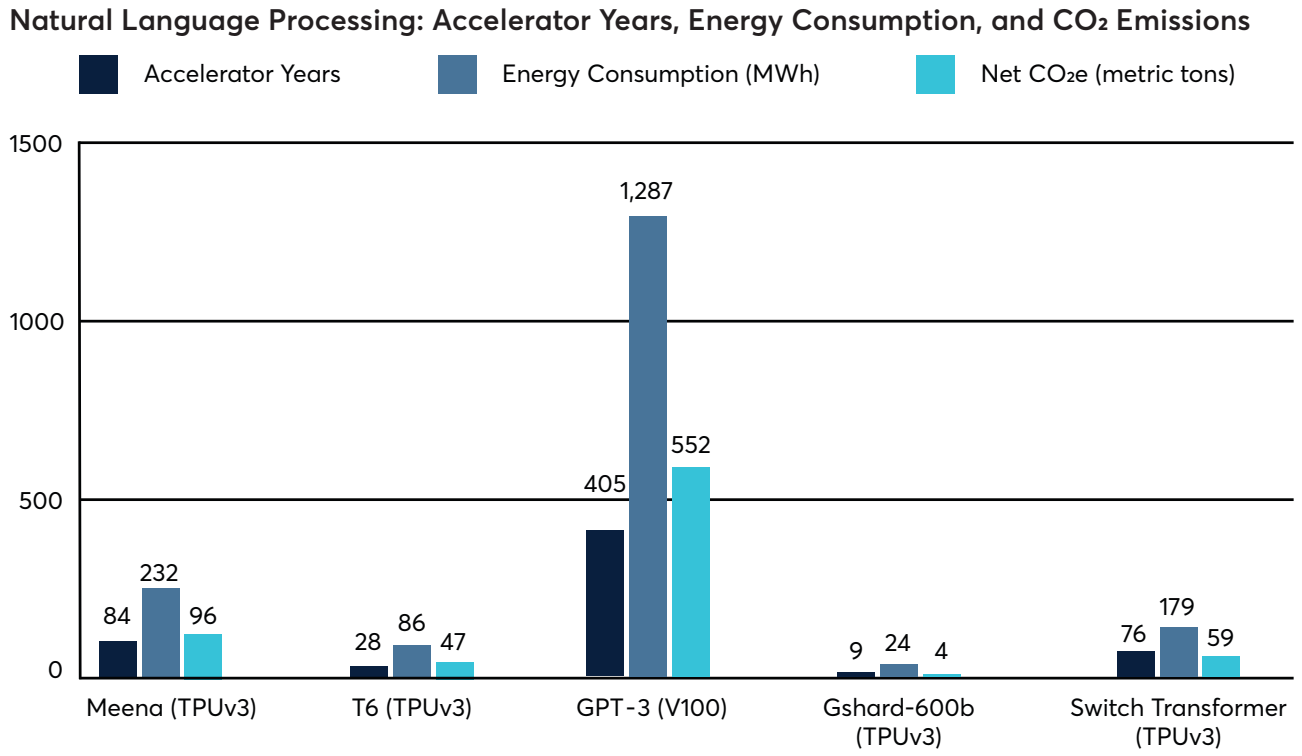


Figure 3: Accelerator years of computation, energy consumption, and CO_2e for five large natural language processing (NLP) deep neural networks (DNNs). Source.

Bandwidth Density and Scalability: Key Challenges for Interconnects in Generative AI Architectures

Many organizations are currently running generative AI tasks on systems that rely on traditional interconnects — electrical I/O with pluggable optics. This solution is provably insufficient at handling the data transfer speeds and throughput that will enable practical, efficient generative AI.

Electrical interconnects suffer signal degradation over longer distances, and are confined to a single chassis at current 100 G and upcoming 200 G links. Pluggable optic connections between chassis are necessary here as connectivity requirements expand beyond the chassis to cross-rack and multi-rack scale. When attempting to provide global communication from every GPU to every other device in the system, traditional interconnects comprise electrical interconnects combined with pluggable optics.

As a system scales up, more of these connections with higher bandwidth need to be made from any GPU to any other GPU in the system through the system fabric. This is hard to do with traditional interconnects because of the large energy and physical footprint of pluggable optics. They simply do not scale efficiently enough for the needs of generative AI.

Pluggables also fail in terms of power consumption. The pluggables-based GPU-to-GPU link consumes a total of 30 picoJoules per bit (pJ/b) in the way described in the following graphic. For reference, an in-package optical I/O solution that directly connects two packages uses less than 5pJ/b.

With copper cables, any data rate increase shortens the reliable signal transmission distance. Signal integrity can also suffer from dense electronic installations that require shielding against electromagnetic interference (EMI), which adds cable size and weight to network systems. In comparison, optical cable is virtually immune to electrical interference, transfers more data over longer distances, and is significantly lighter weight and more compact. This is particularly important for base station installations in which power constraints, the physical volume available for deployment, and weight are often vital factors for system architects.

In-package optical I/O provides the interconnect that enables high-bandwidth and low-power connectivity between antenna/sensing elements and the digital signal processing infrastructure within the base station in a compact monolithic package (see Figure 2). Optical I/O also supports the versatility of beamforming to enable massive MIMO (multiple input, multiple output) implementations.

Pluggables are bulky modules. Their edge bandwidth density is 10x lower than in-package optical I/O, and their area density is 100x lower. This limits how much bandwidth from the GPU can be taken to the rest of the system. As future parameter sizes increase, this limited bandwidth density will create bottlenecks that can make generative AI tasks unviable.

All of this combines to prove that pluggable optics are difficult to scale. The space and power constraints of pluggable optics make it hard to accommodate the growing needs of generative AI models. The monetary costs of pluggables also scale poorly, hovering around $1-$2/Gbps. This needs to be roughly 10x lower for cost-effective generative AI.

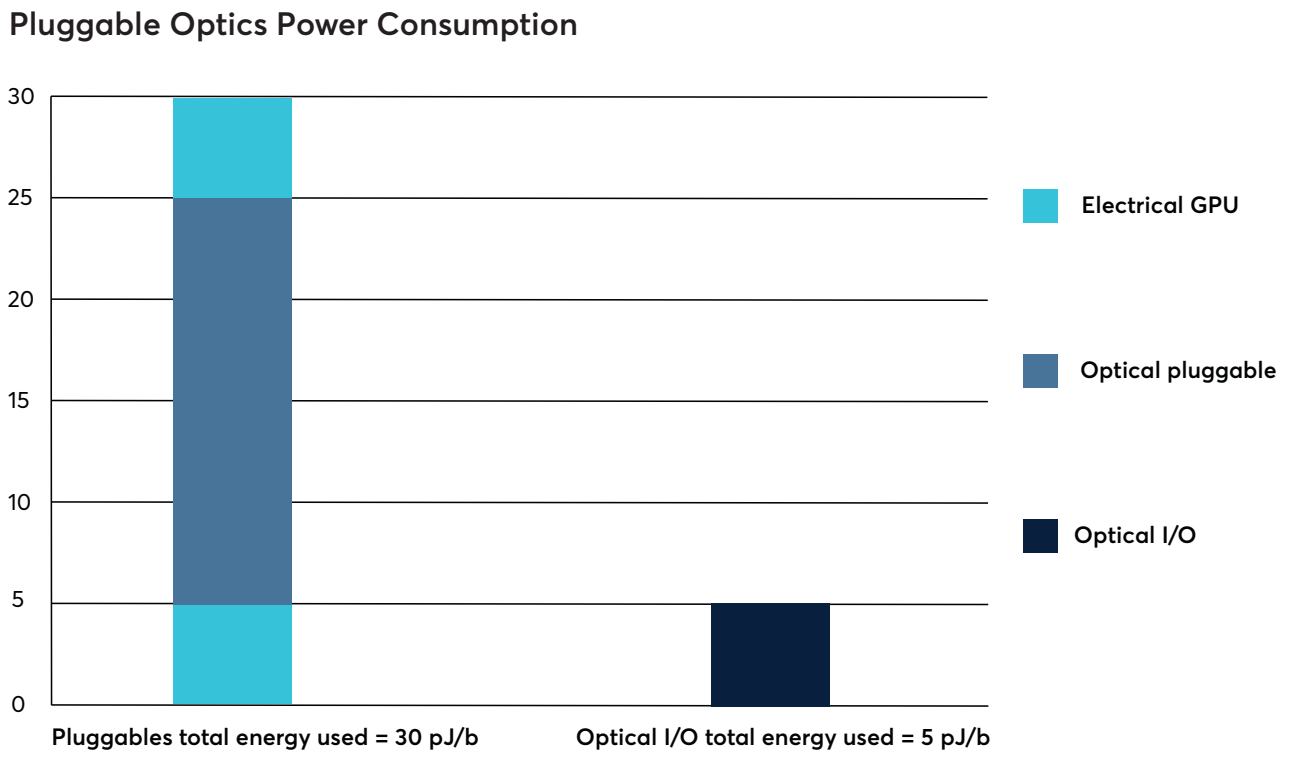


Figure 4: Pluggable optics power consumption vs. optical I/O.

Optical I/O Presents an Exciting Solution

Generative AI requires processing large datasets and parameter sets that demand high bandwidth interconnects between GPUs, and optical I/O provides the bandwidth necessary. Ayar Labs' TeraPHY™ in-package optical I/O chiplet combined with the SuperNova™ light source meets this need by providing 4 Tbps of bidirectional bandwidth.

Optical I/O solutions also affect the package-level metrics and enable scaling of compute-socket bandwidth to beyond 100 Tbps. As Table 1 indicates, the bandwidth scaling roadmaps based on UCIe, CW-WDM MSA, and microring-based optical I/O chiplets provide paths toward 100+ Tbps of off-package I/O bandwidth, while enabling a connectivity radix of up to 128 ports per package.

Ayar Labs TeraPHY Optical I/O Chiplet Roadmap

Gen	Electrical Interface (Advanced Package)				Optical Interface (CW-WDM)			Optical Chiplet Bandwidth (Tx+Rx)	Off-Package I/O Bandwidth (4-8 Chiplets per Package)	Off-Package Radix (Number of Ports)
	I/F	Mod	Tx / Rx I/Os	Data Rate: Gbps per I/O	Ports	λs / Port	Data Rate: Gbps/λ			
1	AIB	24	20/20	2	8	8	16	2 Tbps	8-16 Tbps	32-64
2*	AIB	16	80/80	2	8	8	32	4 Tbps	16-32 Tbps	32-64
3	UCIe	16	32/32	8	8	16	32	8 Tbps	32-65 Tbps	32-64
4	UCIe	16	64/64	8	16	16	32	16 Tbps	65-131 Tbps	64-128
5	UCIe	16	64/64	16	16	16	64	32 Tbps	131-262 Tbps	64-128

Gen = Generation of Ayar Labs TeraPHY optical I/O chiplet. *Indicates current generation.

I/F = Interface

Mod = Modules

Table 1: Generation roadmap for optical I/O chiplets.

This combination of high radix and high bandwidth per port, coupled with low optical I/O link latency (<10ns + time of flight [ToF]), provides an unprecedented level of flexibility in architecting all-optical fabric connectivity solutions for the large-scale distributed compute systems needed to support generative AI. It allows the design of large-scale system fabrics that have low, uniform latency and high throughput, keeping the compute nodes fully utilized.

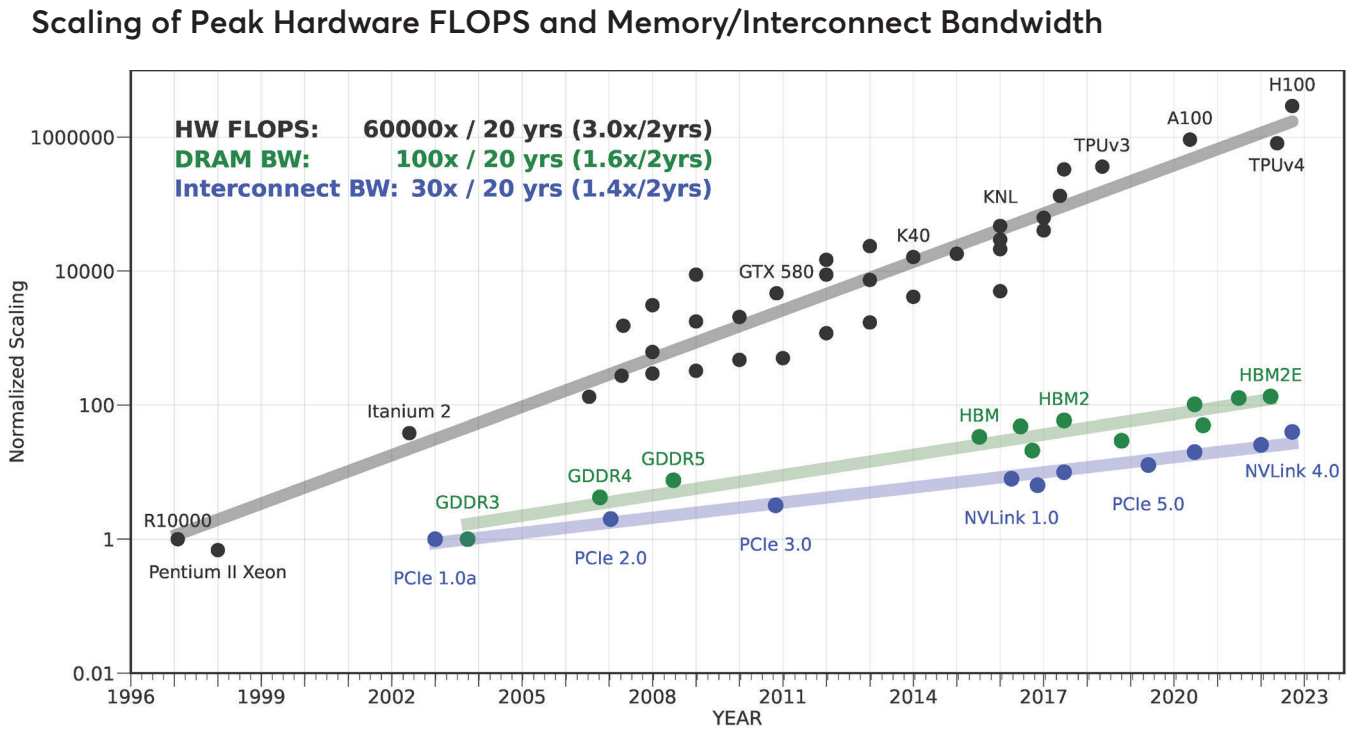


Figure 5: The scaling of the bandwidth of different generations of interconnections and memory, as well as the peak FLOPS. As can be seen, the bandwidth is increasing very slowly. Source.

The high bandwidth provided by optical I/O is especially relevant in terms of the limiting effect interconnect bandwidth can have when discussing distributed systems. As shown in Figure 5, peak FLOPS has increased by 60,000x over the past two decades while interconnect bandwidth has only increased 30x. For the scalable distributed systems required to support generative AI, fixing the interconnect bandwidth problem is of the utmost importance.

Enabling Interoperability Through Thoughtful Standardization

As optical I/O matures, industry standards will be critical for ecosystem development and widespread adoption. Universal Chiplet Interconnect Express (UCIe) aims to standardize the physical layer by defining a die-to-die interconnect for chiplets using optical I/O. Continuous-Wave Wavelength Division Multiplexing Multi-Source Agreement (CW-WDM MSA) looks to standardize optical wavelengths for dense bandwidth scaling. Meanwhile, Compute Express Link (CXL) standardizes the software layer by enabling pooled memory and cache coherence between GPUs, CPUs, and accelerators. This allows them to fully utilize optical I/O's high bandwidth and low latency.

Thoughtful standardization across both the physical layer through UCIe and CW-WDM MSA, and the software layer through CXL, will enable seamless integration between diverse hardware and frameworks. With momentum building around these key standards, optical I/O is poised to become a vital enabler of next-generation generative AI computing architectures by providing the scalable, efficient interconnects needed for advanced workloads.

Conclusion

Generative AI is one of the most disruptive technological developments in recent years, and for good reason. The sheer usability and versatility of it rivals the introduction of the internet, and institutions in a wide variety of fields who wish to remain relevant will need to find a means of integrating generative AI in a way that works for them.

Traditional interconnects are not sufficient for generative AI architectures. The requirements for high bandwidth, low latency, high throughput, and energy efficiency all point to a need for in-package optical I/O.

Additional Resources

Artificial Intelligence | Optical I/O for AI Overview

Ayar Labs Optical I/O: Shattering the Barriers to AI at Scale | Video

Ayar Labs' Optical I/O Enables Disaggregated Architectures for Cloud, AI, and HPC | Video

In-Package Optical I/O: Unleashing Innovation | Video

Scalable and Sustainable AI: Rethinking Hardware and System Architecture | Webinar Video

Optical I/O Chiplets Eliminate Bottlenecks to Unleash Innovation | Technical Brief

Speed Meets Sustainability: Dr. Satoshi Matsuoka on the Future of AI and Supercomputing | EE Times

Intel & Ayar Labs 4 Tbps Optical FPGA Demo | Video

In-Package Optical I/O Solutions | Solution Overview

Sources

M. Wade, C. Sun, M. Sysak, V. Stojanović, P. Tadayon, R. Mahajan, B. Sabi, "Driving Compute Scale-out Performance with Optical I/O Chiplets in Advanced System-in-Package Platforms," HOT CHIPS 2023. https://ieeexplore.ieee.org/document/10254699

C. Franzen, VentureBeat 2023 AI survey, 2023. https://venturebeat.com/ai/more-than-70-of-companies-are-experimenting-with-generative-ai-but-few-are-willing-to-commit-more-spending/

McKinsey Global Survey, 2023. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year#/

M. Bastian, "GPT-4 has more than a trillion parameters — Report," THE DECODER, 2023. https://the-decoder.com/gpt-4-has-a-trillion-parameters/

D. Patterson, J. Gonzalez, Q. Le, C. Liang, L. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon Emissions and Large Neural Network Training," arXiv, 2021. https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf

J. Huang, "NVIDIA Keynote at COMPUTEX 2023", YouTube, 2023. https://www.youtube.com/watch?v=i-wpzS9ZsCs&t=1190s

A. Gholami, Z. Yao, S. Kim, MW Mahoney, K. Keutzer K, "AI and Memory Wall," RiseLab Medium Blog Post, University of California Berkeley, 2021, March 29 (updated). https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

AyarLabs

Headquarters
695 River Oaks Parkway
San Jose, CA 95134

Emeryville Office
5909 Christie Avenue
Emeryville, CA 94608