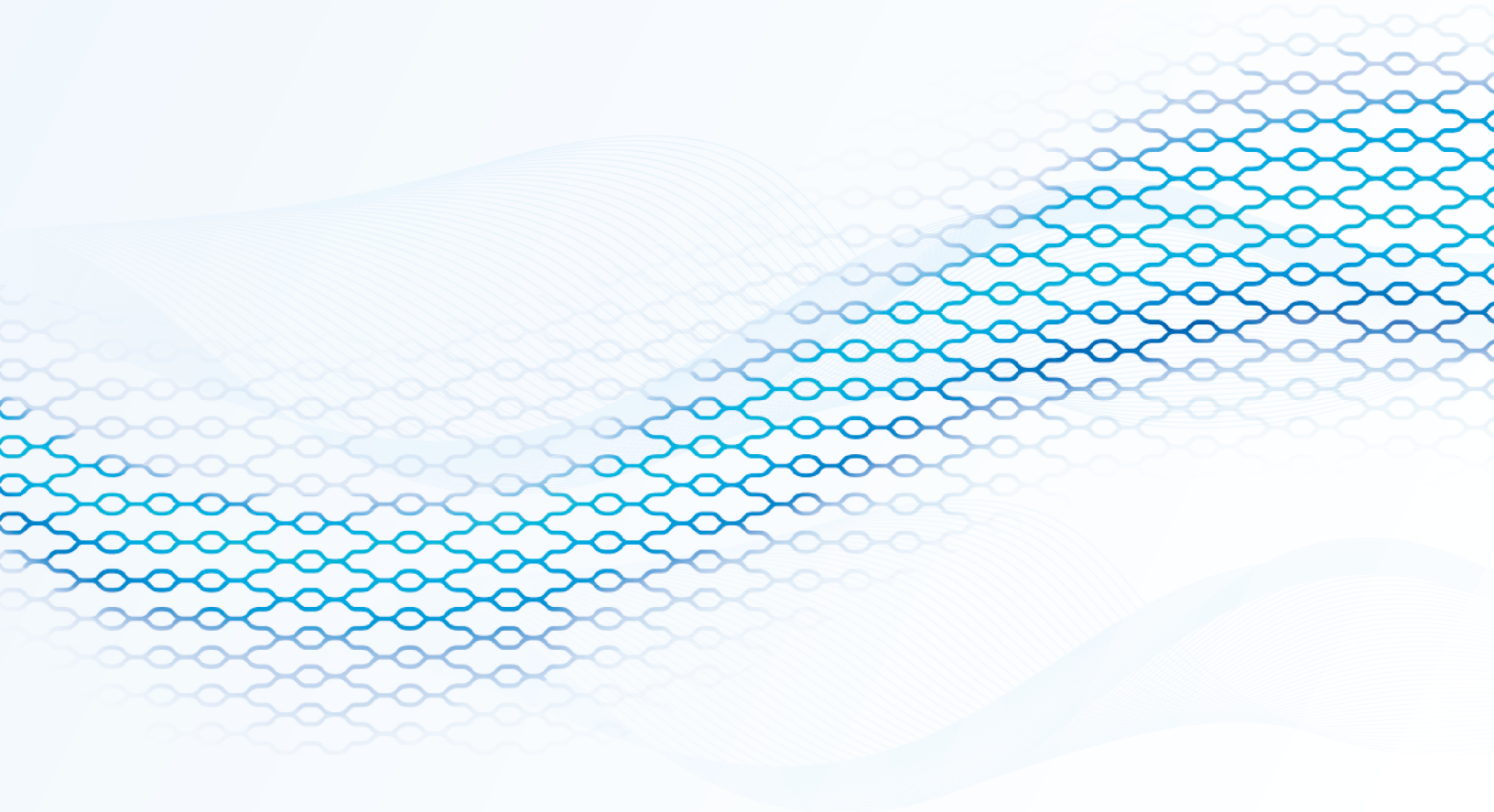
LIGHTELLIGENCE

Large-scale Optoelectronic Integration Enabling Intelligent Computing Power Networks



Content

1 Computing Power Demand - Trends and Issues

With the explosive growth of artificial intelligence (AI) applications such as intelligent transportation, industrial brain, autonomous driving, and the internet of things (IoT), human society generates massive amounts of data every day. To analyze and extract valuable information from this data, powerful data storage, transmission and processing capabilities are required, posing unprecedented challenges to the computing capabilities of existing data centers and edge devices. Meanwhile, AI models are rapidly evolving and expanding in attempts to leverage this data. Further, the explosive growth of model parameters means greater computing power demand for each unit of input data to the model. According to the data published by OpenAI, as shown in Figure 1, in recent years, the average size of AI models increases tenfold every year. The demand for AI training follows a similar trend[1].

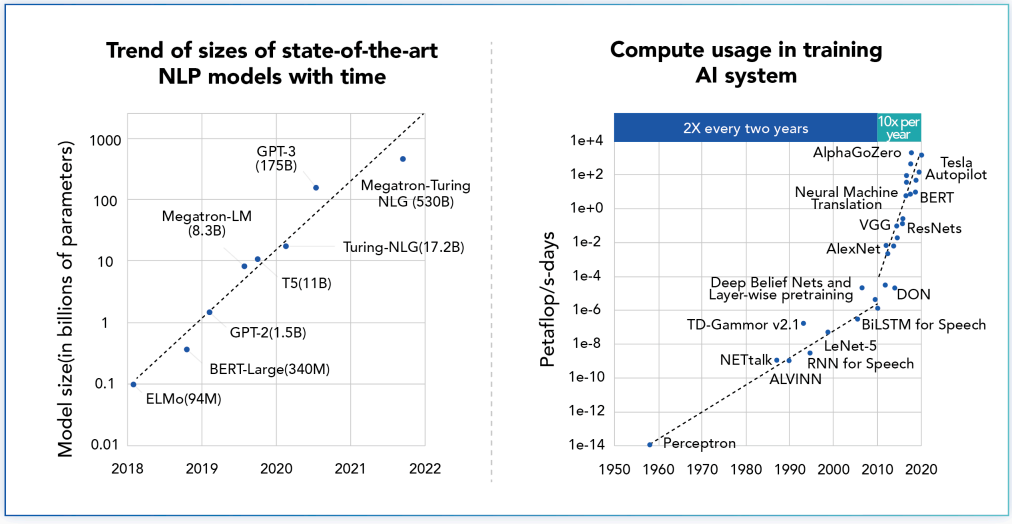


Figure 1 AI Model size and compute usage in training AI systems[1]

However, performance growth of traditional computing chips has encountered bottlenecks. Since the start of the semiconductor industry 60 years ago, computing power improvement can be described by Moore's Law and Dennard Scaling. Moore's Law predicts that transistor density doubles every 18 months, while Dennard Scaling[2] states that power density remains unchanged as transistor density improves. The combination of Moore's Law and Dennard Scaling allows CMOS chips to continuously improve computing power while maintaining constant energy and area consumption. However, as the chip manufacturing process moves to 5nm and 3nm, transistor density is close to its physical limit. Moore's Law is slowing down and is expected to end in the 2020s[3]. Dennard Scaling ended as early as 2004[4], resulting in power supply and heat dissipation challenges as transistor density improves. This is also known as the "Power Wall". In addition, the tape-out and design costs of advanced processes are getting higher and higher, which creates a "Cost Wall". The traditional single-chip computing power improvement path is now unsustainable.

Even if Moore’s Law and Dennard Scaling continue, the exponentially growing demand for computing power cannot be met by discrete computing systems. The consumption of resources to meet computing demand is increasing every day. It is necessary to deploy large-scale distributed computing systems in data centers.

This whitepaper will first introduce the semiconductor industry's existing efforts to increase and better utilize computing power, as well as the challenges encountered to do so. Afterwards, a new data center paradigm based on large-scale optoelectronic integration technology will be proposed. This new paradigm aims to provide a new idea for the evolution of next-generation data centers.

2 Development of Computing Power - Status and Challenges

Hyperscalers and data centers have made substantial efforts to improve computing power and efficiency. Among them, the Computing Power Network (CPN) has become one of the most widely recognized concepts in the global community[5]. CPN is a new type of information infrastructure that flexibly allocates and schedules computing, storage, and network resources on demand according to business needs. Its ultimate goal is to abstract hardware resources into computing power so that users can purchase computing power from data centers according to actual computing needs without buying or leasing hardware equipment. Thus, using computing power is equally convenient as using other utility services (such as water, electricity, and gas).

To realize this vision, CPN may be composed of massive high-performance computing nodes with efficient data interconnection between nodes. While each computing node is scaled up to construct a solid computation foundation of CPN, the computing power can also be scaled out through efficient data interconnection to form a huge computing capacity. CPN is expected not only to solve the problems of low computing power utilization and scalability, but also address the difficulties of computing power migration and usability. Through the flexible scheduling of hardware resources, CPN can realize more fine-grained sharing within the network. This chapter will briefly introduce the current challenges faced by the industry in terms of scaling-up and scaling-out of computing power.

2.1 Scale-up in a single computing node

In data centers, a computing node usually refers to a single server with one or two CPU sockets that may contain many CPU cores. As the increase of transistor density slows down, the performance improvement associated with semiconductor process advances is becoming more and more limited. Thus, various ideas have been proposed to improve computing power in a single node. The first is to prioritize higher computing efficiency over the versatility of computation found in CPUs through chip-level architecture innovations. Based upon these innovations, some researchers even seek to go beyond traditional von Neumann architecture and CMOS technology, and search for disruptive new computing paradigms to boost computing power far beyond Moore's Law. Another strategy is to leverage advanced packaging technology and put multiple chiplets into a single package which breaks the performance barrier imposed by the reticle size and achieves higher computing power.

2.1.1 Heterogeneous computing architecture innovations

Early architectural innovations were mainly driven by increasing the instruction-level parallelism (ILP) in general-purpose computing architectures to take advantage of on-chip transistor resources brought about by Moore's Law. For example, with abundant transistor resources, the superscalar CPU architecture can enable multiple-instruction issuing and out-of-order execution. Long and deep pipelining allows more computing operations on a single unit input of data. At the same time, deepening of the pipeline reduces compute operations in each pipeline stage, thereby achieving a substantial increase of CPU frequency.

As semiconductor technology evolves, the increase of chip area also enables more logic functions to be integrated on the chip. These logic functions, including more hierarchical data caches, larger data buffering, and data prefetching blocks, improve performance which has been long troubled by the "Memory Wall" problem due to the unbalanced development of computing and memory speeds. Recently, large-scale efficient data transfer technology has become a new driving force for performance advancements. Innovative technologies, such as High Bandwidth Memory (HBM), have brought a new wave of performance improvements to computing architectures. This trend is well reflected in Google's paper[6] which summarizes the evolution of several generations of Tensor Processing Unit (TPU) architectures.

Finally, the transistor boom also makes it possible to enable multi-threading, multi-core, and context-based kilo-threading architectures. In combination with single instruction multiple data (SIMD) and other vectorization technologies, thread-level parallelism (TLP) has resulted in a multiple magnitude performance leap in modern computing architectures.

Built on top of these general-purpose architecture improvements, domain specific architecture (DSA) has also achieved an accelerated pace of development. As technological innovations such as artificial intelligence, 5G, autonomous driving, and VR/AR continue to advance, the requirements for chip computing power, functions, power consumption, cost, and security in different applications are increasingly diversified. Under the trend of wide-ranging computing power requirements, DSA emerges timely as the customized architecture targeted for certain specific application domains. It can include special computing units with unique parallel mechanisms, data types, and domain-specific languages, etc. The DSA architecture achieves performance

speed-ups by sacrificing versatility of the architecture to effectively utilize native computing capabilities of the hardware, thus gaining better energy efficiency than general-purpose computing architectures.

For example, Nvidia's latest Hopper graphics processing unit (GPU)[7], based on a typical TLP architecture, uses more powerful Tensor Cores to conduct matrix multiplications. Moreover, Tensor Core has added more domain-specific technologies to scale up computing power, including fine-grained sparse computing and dynamic programming algorithm optimization. Compared with its previous generation A100 GPU, the Hopper H100 GPU has a performance improvement of 2 to 4 times on a set of AI training tasks. Another DSA example is Google's TPU, in which the systolic array is designed to optimize matrix multiplication. Systolic array significantly improves the compute density by increasing multiple operations on a unit input of data while alleviating the Memory Wall effect.

However, due to the nature of customization, domain-specific architecture usually lacks computational completeness. Therefore, the heterogeneous computing architecture is used to combine CPUs and multiple types of DSAs. By allowing each DSA to maximize performance in its own domain, heterogeneous computing achieves the highest performance and best energy-efficient computing.

While domain-specific architecture has achieved a substantial increase in computing power, it is still limited by the underlying computing components and von Neumann architecture. The underlying components in traditional computer chips are based on CMOS transistors. The core working mechanism of CMOS is to control the current in transistors by voltage signals.

However, with improvement of the CMOS manufacturing process, transistor size is shrinking and the quantum tunneling effect reduces the efficiency of controlling current. Thus, new underlying computing mechanisms are required to break through this bottleneck.

At the same time, traditional computer design is typically based on von Neumann architecture, where computing and data are separated into different functional blocks. After data is moved to the computing unit through sequential control logic, calculations are performed. The main issue of this architecture is that the calculation is delayed due to data fetch and data movement, which also increases power consumption. These effects gradually reveal the Memory Wall problem. Even though modern architecture innovations continue to advance through vectorization, hyper-threading, pipelined parallelism, and multi-core architectures, the performance potential in von Neumann architecture is getting smaller and smaller.

Along with general architecture innovations, non-von Neumann architecture has begun to flourish. Given that non-von Neumann architectures are no longer based on any sequential control flow execution (e.g., biological computing and quantum computing), or often are derived to overcome the core bottleneck of von Neumann architectures (e.g., memristor-based in-memory computing), these novel architectures prove disruptive and create a huge space for performance and energy efficiency improvement due to their new underlying computing mechanisms. For example, the near-memory computing engine[8], based on 3D hybrid packaging, can dramatically improve the performance of AI accelerators on recommendation models by directly connecting multi-storage memory to computing logic units.

2.1.2 Chiplet architecture

Because the increase of transistor density is gradually slowing, boosting computing power within a single package can only be achieved by expanding the total chip area once architecture improvements are exhausted. However, due to limitations of the CMOS process, maximum area of a single chip is capped by the size of the reticle, which is generally around 800mm^2. To further expand the total area of the chip, it is necessary to find new ways to break through the upper limit of the area of a single chip, thus emerges the new idea of chiplet systems.

Thanks to the progress of advanced packaging technologies, it is possible to package multiple chiplets with domain-specific functions on the same substrate. In the past three years, the world's leading chip companies have increasingly been deploying chiplet architectures to high-end computing systems. For instance, Intel's Ponte Vecchio GPU[9] consists of more than 40 dies with a total area of over 3,000 mm^2. Similarly, Cerebras' Wafer Scale Engine (WSE)[10] has an area of over 40,000 mm^2, which is 50 times larger than the area of the current reticle size.

In addition to increasing the total chip area, modularity of chiplets also allow computing nodes of different processes or different fabs to be contained within a single package. This not only enables flexibility of disaggregated architecture design, but also greatly improves yield and reduces manufacturing costs. Furthermore, chiplet architecture helps realize domain-specific heterogeneous computing at the chip level, which is more accommodating to various computing tasks, as opposed to the card level.

The current chiplet products are primarily based on proprietary architectures. However, there are many industrial organizations (such as UCIe and BOW) actively promoting

chiplet interconnection standards for on-chip heterogeneous computing. For example, in 2021, Google released the specification of Open Chiplet architecture[11], to further promote development of the chiplet ecosystem.

As total area of the chip becomes larger, the distance of data transmission by electricity also increases, as does latency and energy consumption of data transfer. In pure digital circuits, to reduce the cost of data transfer, each computing unit generally only transmits data to its nearest neighboring computing units. Therefore, data transfer across multiple computing units often requires a few hops. Since a large computing task usually needs to be mapped to multiple computing units, the process requires very complex algorithms that optimize mapping to avoid long-distance data transportation.

2.2 Scale-out with massive computing nodes

2.2.1 Challenges of current large-scale distributed computing

Architectural innovation and computing power improvement within a single computing node alone are insufficient to meet the growing demand for large-scale compute, therefore, tens of thousands of computing devices are often deployed in data centers. However, simply stacking up of many computing nodes doesn't necessarily lead to efficient use of computing resources due to network congestion. Especially with data-intensive applications, multiple parallel tasks conflict with each other in the communication network, which causes additional delay and performance loss, resulting in low resource utilization of the overall system. Moreover, data centers often plan the hardware architecture for close to peak computing power demand, which may be several or even dozens of times higher than usual. This leads to hardware planning and

deployment carried out according to peak computing power requirements, resulting in underutilized computing equipment.

In addition, the configuration of computing resources such as CPUs, GPUs, and memory inside data center servers are relatively fixed, but different computing tasks have different resource requirements. Once the hardware configuration is pre-determined, certain types of resources are frequently underutilized because computing power cannot be flexibly scheduled. As observed in Alibaba and ByteDance data centers[12,13], the utilization of GPUs is about 40% when measured on the GPU card as a whole and only 10% when measured on the Streaming Multiprocessors (SMs) within the GPUs.

A better approach is to integrate computing resources and then allocate them to computing tasks in a flexible and efficient manner. This idea of integration is often referred to as resource pooling. The traditional resource pooling mechanism, including computing and memory pooling, mainly focuses on resource sharing within a single computing node. This sharing mode partitions CPUs, memories, and other resources to different virtual machines through hypervisor technology and then implements multi-tenant resource isolation and sharing of virtual machines on the physical host. Therefore, the traditional pooling model does not break through the physical boundary of a single node and cannot take full advantage of resource sharing among large-scale computing nodes. In general, the challenges of existing distributed server architectures can be summarized as follows:

- Inflexible computing resource configuration often results in unbalanced use of resources in the system.
- Different applications have different usage patterns of computing resources, so it is hard to accurately match the granularity of scheduled computing resources (such as

GPU cards or SMs) with the computing power requirements of tasks, resulting in underutilization and waste of computing resources.

- Efficient task mapping and optimization with distributed computing resources requires sufficient understanding of difficult-to-ascertain proprietary architectures. This challenge often leads to inefficient optimization of computing resources.
- The isolation and fault tolerance of computing resources are relatively poor with current architectures, so it is difficult to achieve fine-grained resource sharing.

2.2.2 Disaggregated composable resource pooling

In response to the above problems, a new computing paradigm has emerged in the industry: disaggregated composable computing resource pooling. The main goal of computing resource pooling is to achieve scalability and flexibility of computing power which in turn improves resource utilization by disaggregating computing resources and dynamically sharing them.

The underlying technology of resource pooling is to disaggregate multiple resources such as computing units, memory, and storage within the traditional computing framework, and then form an independent resource pool for each type of resource. Flexible allocation of computing resources and elastic expansion of computing power are implemented through high-bandwidth, low-latency interconnection technologies (such as the latest CXL standard[14]) within and between resource pools.

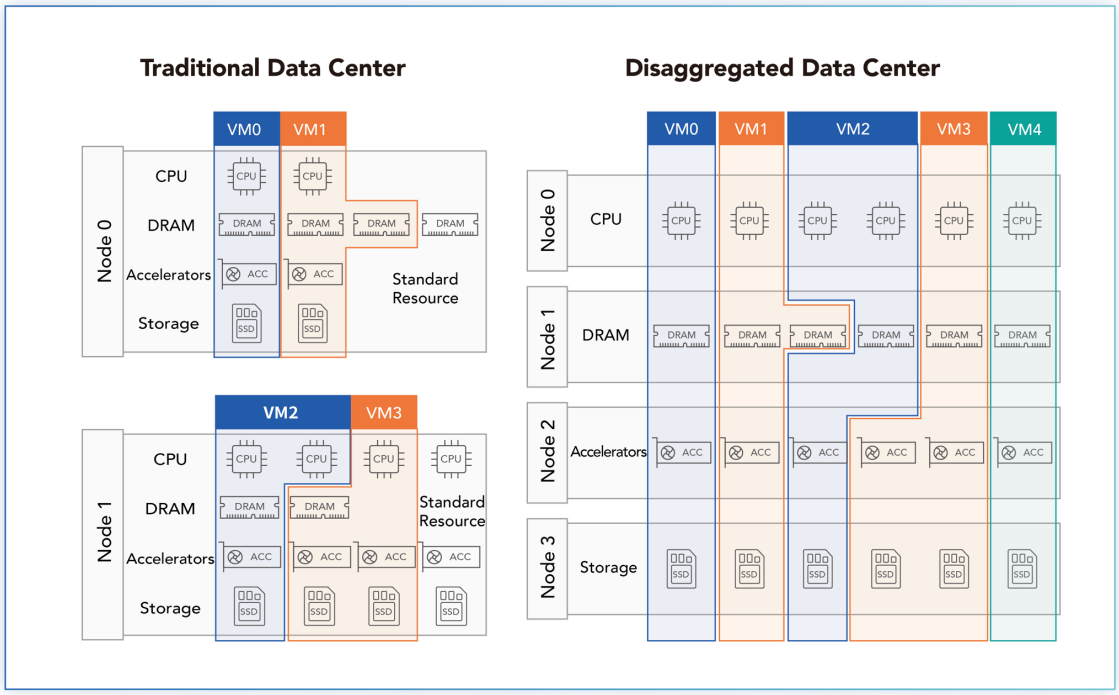


Figure 2 Dynamic resource allocations in traditional vs disaggregated data centers

As shown in Figure 2, resource pooling in traditional data centers is typically limited within a single physical server, easily leading to underutilization of local resources. With the disaggregated composable architecture, each type of resource is clustered into its own pool, such as a CPU pool (physical node 0), a memory pool (physical node 1), a heterogeneous accelerator pool (physical node 2), and a storage pool (physical node 3). The scheduling system allocates appropriate computing resources according to the actual needs of computing tasks and recomposes computing instances (such as virtual machine 0, virtual machine 1, etc.) in a dynamic composition mode. With the new disaggregation paradigm, the stranded resources, unused otherwise in the traditional data center, can be recomposed to create a new virtual instance (VM4 in the right side of Figure 2), resulting in 25% more VMs with the same provision of resources.

However, the infrastructure of the current data centers makes it difficult to support efficient computing resource pooling. This is because the physical distance between large-scale distributed computing devices is relatively long, so Ethernet-based data communication is usually deployed between racks in current data centers. Due to the bandwidth limitation and high communication delay of Ethernet-based data transportation, it is extremely challenging for data centers to achieve the near-linear expansion of large-scale computing power. For example, as the HARP experiment[12] shows, when there is a deterioration of network latency and bandwidth, there is a direct and sizable deterioration in training performance.

2.3 Key challenges in computing power network

Computing Power Network not only needs continuous advancement of domain specific architectures to scale up the computing power within a single node, but also needs to scale out computing power through high performance interconnection technology for expanding the compute capacity of resource pools. In addition, the hardware-software co-design and co-optimizations of computing architectures are also a necessary means to boost computing performance and energy efficiency.

Overall, the future technological breakthroughs needed for more powerful and efficient computing power networks can be summarized as follows.

- New computing paradigm beyond the traditional CMOS technology, and a heterogeneous computing architecture that better matches the compute demands in industry-scale use scenarios.

- High performance and scalable chiplet system, including physical layer innovations beyond traditional electrical interconnection and easy-to-use software stack with high adaptability.
- High-bandwidth, low-latency cross-rack interconnect technology, including hardware innovation and advancement of interconnect protocols.

3 New Paradigm of Computing Power Network

Breaking through bottlenecks and overcoming challenges at the data center requires innovation in underlying technology. Large-scale integrated silicon photonics technology has the potential to surpass traditional technologies both vertically by increasing single node computing power and horizontally by improving the efficiency of large-scale distributed computing. Lightelligence, founded in 2017, focuses on optoelectronic hybrid computing solutions. Lightelligence is a pioneer in this field and has developed a number of key technologies enabling efficient computing power networks. This chapter will briefly introduce the new paradigm of data centers based on large-scale integrated silicon photonics technology. More specific information will be discussed in thematic white papers later in this series.

3.1 Scale-up in a single computing node

3.1.1 New principle of disruptive computation: Optical Multiply Accumulate (oMAC)

In order to maintain the continuous improvement of chip computing power, we need to look at the underlying physical principles with a new lens. Digital chips are now limited by the physical limits of the underlying component – the CMOS transistor. However, optical signals and devices follow different physical principles. The interactions of optical signals with scattering mediums are typically linear and therefore can be mapped as linear calculations.

There are many phenomena of optical linear calculations in everyday life, for example, the lens of an optical camera. An optical signal passes through the lens by completing two two-dimensional spatial light Fourier transforms which are then imaged on a photosensitive element. The camera lens can be regarded as a non-programmable optical linear computing unit given these inputs and outputs. But for computing units with practical value, the system must be programmable.

Given that matrix multiplication occupies a core position throughout current mainstream data center computing tasks, such as artificial intelligence, numerical simulation, etc., efficient matrix multipliers that go beyond Moore's Law will have a wide range of commercial prospects. Matrix multiplication, which can be regarded as parallel multiply accumulate operations, is a typical linear operation that can be accelerated using photonic computing units. Therefore, as Moore’s Law decelerates, programmable optical Multiply Accumulate (oMAC)[15] technology is expected to support the continuous improvement of computing power, providing a new pathway for hardware infrastructure in the digital economy era.

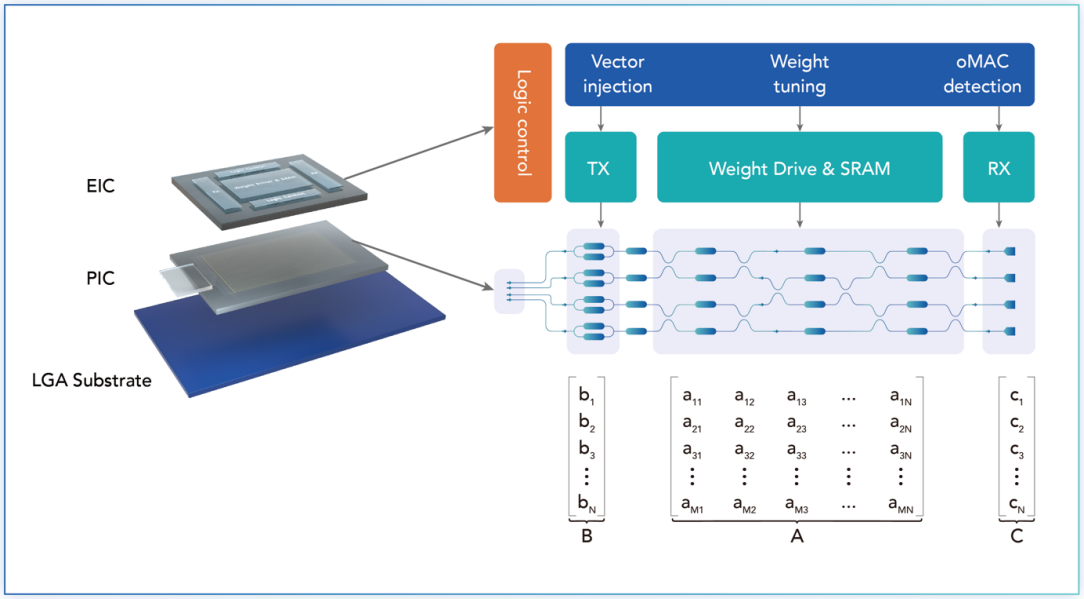


Figure 3 Diagram of programmable oMAC system

Figure 3 shows one implementation of a programmable oMAC hardware. At the physical level, the system includes optical and electronic chips packaged together by a 3D flip-chip method. At the functional level, it includes three parts: signal input, signal processing, and signal output. After the optical signal enters the optical chip, the input vector $\vec{b}$ is converted into optical signals by the modulators, and these optical signals will pass through the programmable optical matrix A. The output optical signal $\vec{c}$ is the result of the matrix operation. All optical devices are integrated on an optical chip, and the logic control circuits and memory of the optical chip are deployed on the electronic chip.

The most significant advantage of photonic computing over traditional CMOS digital circuits is low latency. As shown in Figure 4, for a digital $N \times N$ matrix unit, the latency is proportional to $O(N)$. Some digital architectures that specifically optimize latency, can achieve latency approaching $O(\log N)$, particularly when the matrix size is small. For

oMAC on the other hand, since the process of calculation is the process of transmitting the optical signal array in the chip, the time of the calculation itself is the time of light transmission across the chip, generally below 1 ns. The time consumed by oMAC operations mainly comes from optoelectronic conversion and digital-to-analog conversion, which is about several clocks and is independent from the size of the matrix, which is equivalent to $O(1)$. Therefore, in the case of large N, the latency advantage of photonic computing is obvious. Another dimension of oMAC's latency advantage is that oMAC chips can achieve a global main frequency significantly greater than that of digital chips at a given process node.

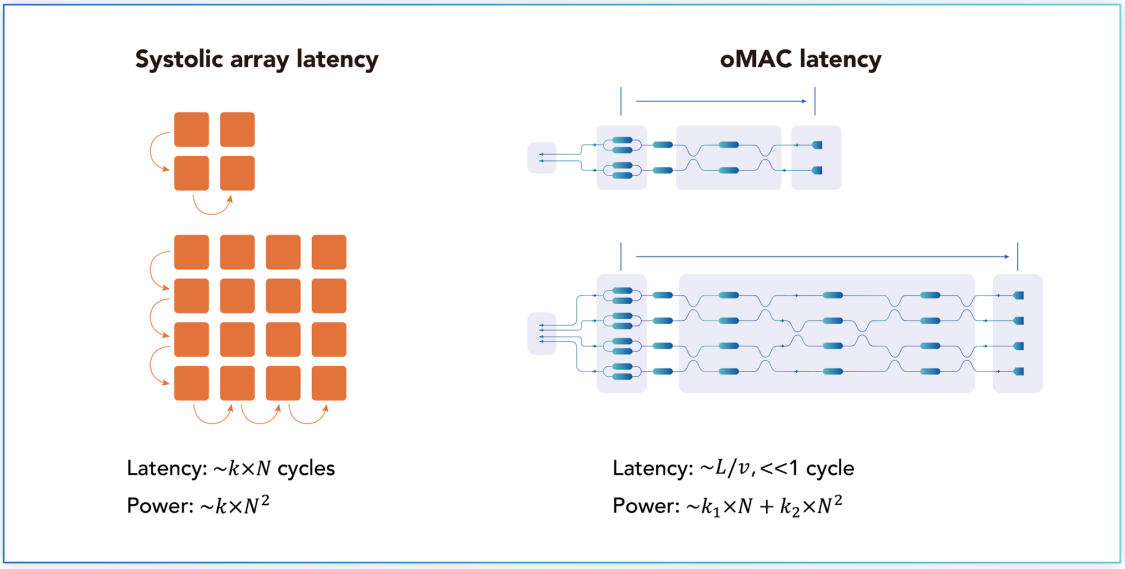


Figure 4 Schematic diagram comparing systolic array and oMAC latencies

In addition to the latency advantage, photonic computing also has the characteristics of low energy consumption. For an $N \times N$ digital matrix operation unit, its energy consumption is $k \times N^2$, where k is related to the power consumption of a single multiplication and accumulation. The overall power consumption is proportional to

$O(N^2)$. For the optical matrix multiplier, its power consumption can be $k_1 \times N + k_2 \times N^2$, k_1 is related to the power consumption of the vector input and receiving end, and k_2 is related to the power consumption of the matrix weight part. In the case that the update speed of the matrix itself is much lower than the vector input, its energy consumption mainly comes from the first half, so it is proportional to $O(N)$. Under the premise that optical devices and their control circuits are well optimized, the energy efficiency of optical computing based on relatively traditional process nodes is comparable to or even surpasses that of digital chips with advanced process nodes.

Optical computing has some caveats compared to digital computing. For example, optical computing, as an analog calculation, cannot support floating-point numbers. For fixed-point numbers, when accuracy exceeds 8 bits, the advantage of energy efficiency diminishes. Therefore, algorithms based on floating-point numbers or fixed-point numbers above 8 bits need to be quantitatively adjusted for photonic computing hardware to show energy efficiency advantages. Also, the light source required for a photonic computing system takes up a certain volume. The current development of light source miniaturization can reduce the volume of the light source used in each server to the same size as a few coins.

Fortunately, most mainstream artificial intelligence inference algorithms are developed with fixed-point numbers below 8 bits and can take advantage of optical computing. For scientific computing where high-precision floating-point numbers are required, software optimizations will need to be developed to take advantage of optical computing.

3.1.2 Enable high efficiency chiplet systems: Optical Network On Chip (oNOC)

In addition to optical computing technology, large-scale optoelectronic integration can also enable large-scale chiplet systems. Chiplet systems improve the computing power and efficiency of a single node through larger on-chip area and more heterogeneous units. However, this scale-up of chiplet systems has bottlenecks in data communication.

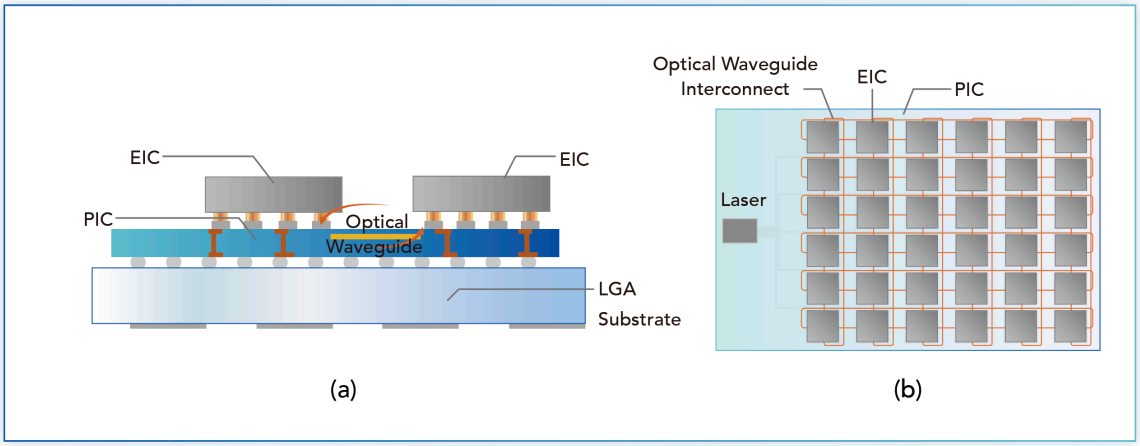


Figure 5 Cross-sectional view (a) and top view (b) of oNOC system where electronic chips are interconnected by optical waveguide based links

One way to solve the data communication bottleneck is to use optical interconnect to replace the electrical interconnect between electronic chips. As shown in Figure 5(a), two electronic chips are stacked on the same photonic chip, and data communication between electronic chips is performed by waveguide based optical links on the photonic chip. Since optical interconnect is not sensitive to distance, the on-chip optical network can enable many long-distance channels. As shown in Figure 5(b), the photonic chip can be extended to the entire wafer, thereby realizing a wafer-level oNOC system, which can support tens of electronic chips, thereby achieving 2D torus or other types of isotropic interconnect network topology (as shown by the orange line in Figure 5(b)). This could

simplify mapping computing tasks to different chips and achieve higher computing resource utilization. Moreover, oNOC could also provide high bandwidth and low latency on-chip interconnect infrastructure with polymorphic computing architecture[16] for future AI accelerators.

3.2 Compute scale-out: Optical inter-chip Networking (oNET)

Currently, distributed computing based on Ethernet is limited by interconnect latency and bandwidth. There is room for improvement in overall efficiency with optoelectronic integration. As shown in Figure 6, in traditional data center architecture, the external optical interconnect of computing chips needs to pass through a network interface card (NIC). One way to optimize the latency and bandwidth of data interconnect is to remove the NIC and directly connect the computing ASIC to an optical module through electrical-to-optical/optical-to-electrical (EO/OE) conversion. This type of optical interconnect concept optimized for computing has not yet formed an industry standard. There are several different names for this type of optical interconnect including "Optical I/O", "Optical Compute Interconnect" and so on[17,18,19]. In the following text, Lightelligence calls this type of optical interconnect between computing chips as "optical inter-chip Networking", in short as "oNET", distinguishing from the aforementioned "oNOC" technology. Realizing low-latency, high-bandwidth, and low-power inter-chip optical networks requires innovations in both the physical layer and interconnect protocols.

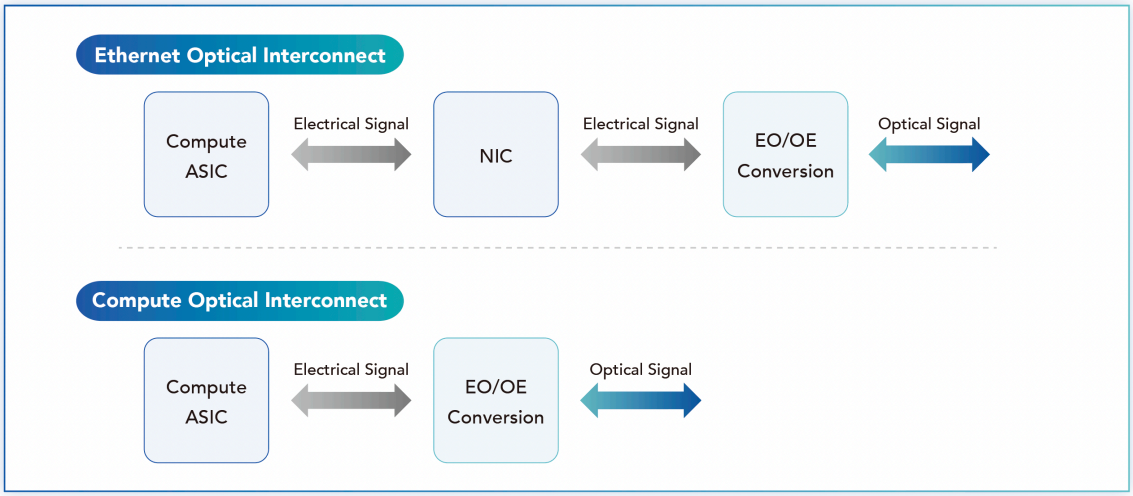


Figure 6 Ethernet optical interconnect and compute optical interconnect

3.2.1 Physical layer innovation

In the interconnected system, the signal will be affected by loss and crosstalk during propagation which will degrade the signal quality. A variety of techniques can improve the signal quality and meet the requirements of bit error rate (BER). The most intuitive way is to increase the strength of the signal itself, but this usually means higher power consumption. Another approach is to use some error correction algorithms to reduce the BER, such as forward error correction (FEC), but this usually means higher latency. For example, the FEC algorithm currently used in Ethernet will take an additional latency of 100ns-200ns[20]. Therefore, to meet the requirements of low latency and low power consumption at the same time, the best technique is to reduce the signal degradation during propagation.

Computing chips usually output electrical signals, and the transmission loss of electrical signals is sensitive to distance. Therefore, shortening the distance between computing

chips and optical modules can help reduce system power consumption and latency. As shown in Figure 7(a), in a traditional server, the external optical communication of the computing chip usually uses a pluggable optical transceiver module. A better approach is to place the optical module on the main PCB as shown in Figure 7(b), as close as possible to the computing chip, thereby forming an on-board optics (OBO) module. The ultimate solution is to package the optical transceiver module and the computing chip on the same substrate. This approach is called Co-Packaged Optics (CPO).

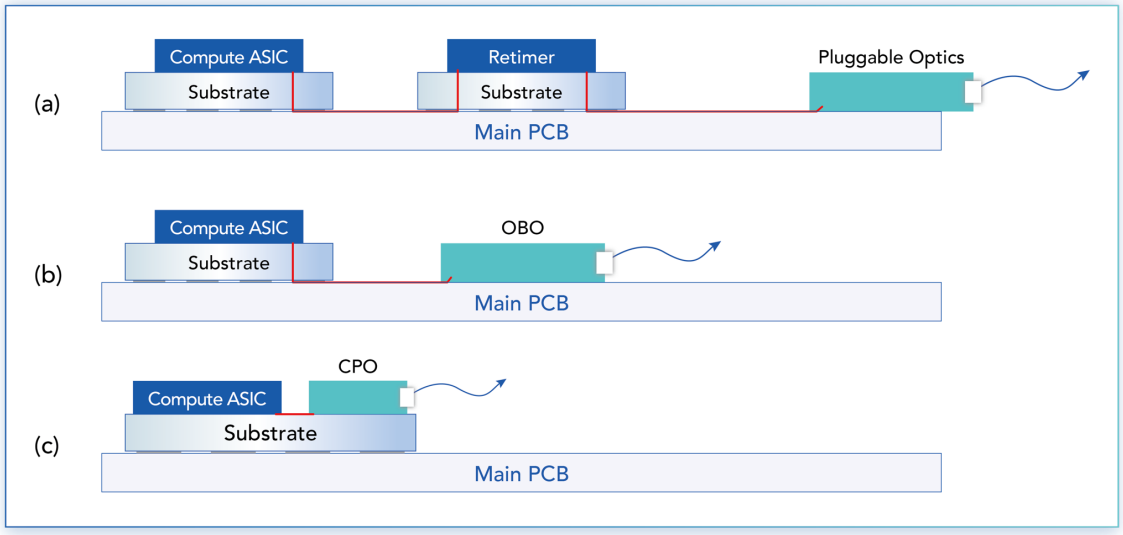


Figure 7 Evolution of interconnect between computing chips and optical transceiver modules

For CPUs, communication is implemented through the PCIe protocol. However, most of the optical interconnect solutions in the current data center are designed for Ethernet. Optical interconnect solutions based on PCIe are not widely available. As shown in Table 1, compared to Ethernet, PCIe applications have more channels, lower single-channel bandwidth, different modulation methods, and much less tolerance to latency. Therefore, current Ethernet-based optical interconnect solutions cannot be directly applied to PCIe

applications. A new optical interconnect standard optimized for PCIe needs to be developed.

Table 1 Comparison of Ethernet and PCIe/CXL optical interconnect solutions

	Ethernet(400G)	PCIe/CXL(Gen 5.0)
Data Rate	4×100G	16×32G
Modulation Format	PAM4	NRZ
Latency Tolerance	Not Sensitive	<100ns

Given large numbers of PCIe channels and lower latency tolerance requirement, silicon photonics-based optical interconnect provides a better solution compared to others. Figure 8 shows one type of system architecture. Optical interconnect is realized by a set of 3D stacked electronic and photonic chips. The combined structure can be packaged around the computing chip, realizing a co-packaged optics module. This type of optical module can also be mounted on a PCB to form an OBO module.

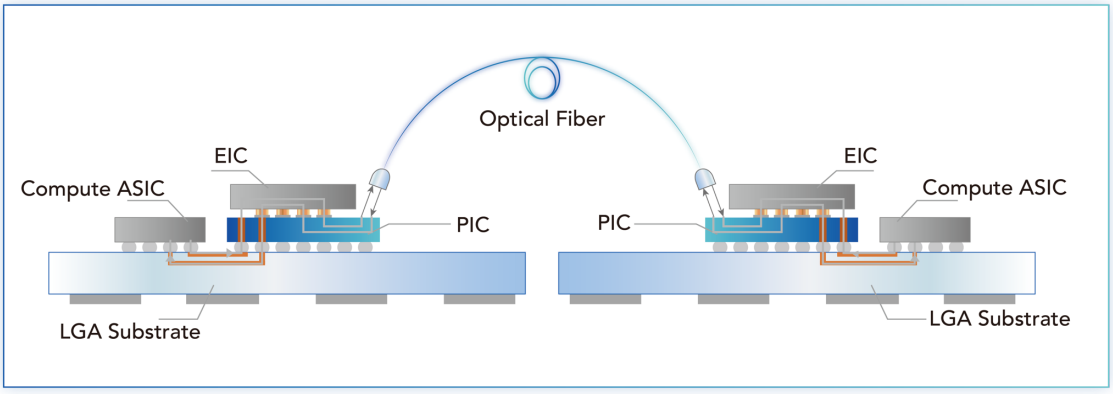


Figure 8 System architecture of silicon photonics-based CPO solution

3.2.2 Protocol layer innovation

The current mainstream distributed computing system mainly uses Ethernet-based software and hardware ecosystems. There is a lot of headroom for improvement. Realizing a distributed computing power network requires efficient data parallelism and synchronization mechanisms. Current Ethernet-based solutions require the use of memory barriers or software defined critical conditions, resulting in performance overhead, long delays, and even deadlocks under complex control processes.

One way to solve these potential issues of Ethernet protocol is CXL protocol (Compute Express Link). The protocol is based on the PCIe physical layer and emphasizes high bandwidth and low latency. CXL has gained widespread support since its first release in 2019. CXL board members include almost all major internet and semiconductor companies. CXL provides efficient data synchronization, which can greatly simplify software management and reduce the overhead of CPU network processing functions. Point-to-point transmission latency can be reduced from the order of 10us with Ethernet to the order of 100ns with CXL.

3.3 New paradigm of computing power network

As shown in Figure 9, photonic computing provides a computational power enhancement path beyond Moore's Law; Wafer-level on-chip optical networks enable the new paradigm of computing chips to cooperate effectively with traditional electrical and memory chips to improve computing power in a single node. In addition, cross-cabinet optical networks based on CXL protocol support efficient resource pooling,

making large-scale distributed computing systems more efficient, flexible, and energy efficient than ever before.

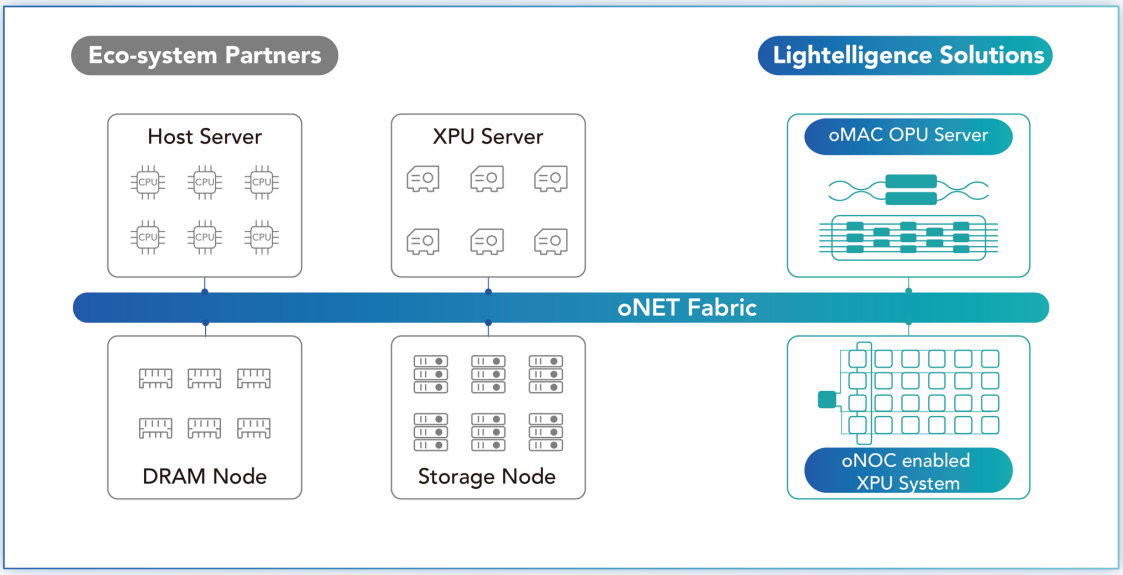


Figure 9 Schematics of a new data center architecture with integrated silicon photonics technology

By combining optical computing, on-chip and inter-chip optical networks, and other technologies, a new paradigm of data center architecture will become possible.

4 Opportunities and Prospects

Humans' thirst for computing power is never-ending.

As productivity of society continues to grow, more people and things are included in the digital space. People generate more data and rely on more complex models to analyze and use data to further improve productivity. However, the traditional way of improving computing power is limited by physical principles, and new ways of improving computing power will inevitably become the focus of the information technology industry.

Opportunities always come with challenges, and new technological revolutions are brewing. Compared with traditional digital circuits, large-scale optoelectronic integration based on silicon photonics introduces novel information processing and interconnection capabilities, thus providing a new computing paradigm.

Like the development of all new technologies in history, this new computing paradigm will go through a transitional stage in supply chain, ecosystem, and business models. Innovations are needed from low-level components all the way to top-level application software development. However, despite challenges, the future looks bright with the promise to revolutionize the way people compute.

Glossary

Abbreviation	Full name
AI	Artificial Intelligence
ASIC	Application Specific Integrated Circuit
BER	Bit Error Rate
BOW	Bunch of Wires
CMOS	Complementary Metal-Oxide-Semiconductor
CPN	Computing Power Network
CPO	Co-Packaged Optics
CPU	Central Processing Unit
CXL	Compute Express Link
DSA	Domain Specific Architecture
EIC	Electronic Integrated Circuit
FEC	Forward Error Correction
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
ILP	Instruction-Level Parallelism
IO	Input Output
LGA	Land Grid Array

NIC	Network Interface Card
NLP	Natural Language Processing
OBO	On-Board Optics
oMAC	Optical Multiply Accumulate
oNET	Optical inter-chip Networking
oNOC	Optical Network On Chip
OPU	Optical Processing Unit
PCIe	Peripheral Component Interconnect Express
PIC	Photonic Integrated Circuit
RX	Receiver
SIMD	Single Instruction Multiple Data
SM	Streaming Multiprocessors
SRAM	Static Random-Access Memory
TLP	Thread-Level Parallelism
TPU	Tensor Processing Unit
TX	Transmitter
UCIe	Universal Chiplet Interconnect Express
VM	Virtual Machine
WSE	Wafer Scale Engine
XPU	CPU, GPU, ...

Reference

[1] Dario Amodei, Danny Hernandez, et al. "AI and compute," 2019 [Online]. Available: https://openai.com/blog/ai-and-compute/

[2] Dennard, Robert H., et al. "Design of ion-implanted MOSFET's with very small physical dimensions," IEEE Journal of solid-state circuits 9.5 (1974): 256-268.

[3] David Rotman, "We're not prepared for the end of Moore's Law," MIT Tech Review, 2020 [Online]. Available: https://www.technologyreview.com/2020/02/24/905789/-were-not-prepared-for-the-end-of-moores-law/

[4] Johnsson, L., and Gilbert Netzer. "The impact of Moore's Law and loss of Dennard scaling: Are DSP SoCs an energy efficient alternative to x86 SoCs?," Journal of Physics: Conference Series. Vol. 762. No. 1. IOP Publishing, 2016

[5] Yukun Sun, et al, "Computing Power Network: A Survey," 2022 [Online]. Available: https://arxiv.org/pdf/2210.06080.pdf

[6] Norman P. Jouppi et al., "Ten Lessons From Three Generations Shaped Google's TPUv4i : Industrial Product," ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), pp. 1-14,2021

[7] Micheal Andersch, et al, "Nvidia Hopper Architecture In-Depth", 2022 [Online]. Available: https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/

[8] Dimin Niu, et al, "184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near- Memory Engine for Recommendation System," ISSCC, 2022

[9] Wilfred Gomes et al., "Ponte Vecchio: A Multi-Tile 3D Stacked Processor for Exascale Computing," ISSCC, 2022

[10] Sean Lie, "Wafer-Scale Deep Learning," Hot Chips, 2019

[11] Ricmib, "OCP Google OpenChiplet spec," 2021 [Online]. Available: https://github.com/google/open-chiplet/blob/main/docs/open-chiplet.md

[12] Pengfei Fan, et al., "HARP: An efficient and elastic GPU-sharing system," O'Reilly Conference TensorFlow World, 2019

[13] Yibo Zhu, "Maximizing GPU utilization in Large Scale Machine Learning Infrastructure," Nvidia GTC, 2022

[14] CXL Consortium,"Compute Express Link: The breakthrough CPU-to-Device Interconnect," [Online]. Available: https://www.computeexpresslink.org/

[15] Yichen Shen, et al., "Deep learning with coherent nanophotonic circuits," Nature Photon 11, 441–446, 2017

[16] Weifeng Zhang, "Polymorphic Architecture for Future AI/ML Applications," OCP Future Technology Symposium, San Jose, 2022

[17] Mark Wade, "TeraPHY: a chiplet technology for low-power, high-bandwidth in-package optical I/O," Hotchips, 2019

[18] Joris Van Campenhout, "Silicon photonics technology for terabit-scale optical I/O, ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP), 2021

[19] Eduard Roytman, "HPC/AI system opportunity with integrated photonics chiplets," HiPChips Chiplet Workshop @ ISCA Conference, 2022

[20] Ilya Lyubomirsky et al., "FEC Latency and Power/Area Tradeoffs for 100G KR/CR," IEEE P802.3ck Meeting, Indianapolis, 2019