Deep Science Publishing
https://doi.org/10.70593/978-93-49910-47-8_7

DeepScience
Open Access Books

# Chapter 7: The role of high-performance computing in scaling artificial intelligence-centric semiconductor architectures

## 7.1. Introduction

Fast-evolving artificial intelligence (AI) algorithms such as large language models have been driving the ever-increasing computing demands in today's data centers. Heterogeneous computing with domain-specific architectures (DSAs) brings many opportunities when scaling up and scaling out the computing system. In particular, heterogeneous chiplet architecture is favored to keep scaling up and scaling out the system as well as to reduce the design complexity and the cost stemming from the traditional monolithic chip design. However, how to interconnect computing resources and orchestrate heterogeneous chiplets is the key to success. This section will first discuss the diversity and evolving demands of different AI workloads. Then it will discuss how chiplet bring better cost efficiency and shorter time to market. It will further discuss the challenges in establishing chiplet interface standards, packaging, and security issues. Finally, it will discuss the software programming challenges in chiplet systems.

Computing workloads are evolving fast, and new demand is also emerging to make AI computable in a more efficient fashion. Next-generation AI algorithms and training settings are driven to explore better performance while pushing the limits of the hardware platform. Additionally, novel AI applications are also proposed to solve scientific computing challenges in a more efficient way. All these drivers of new workloads intend to maximize the performance on a certain task but also lead to diverse architectures ranging from specialized to general-purpose design. Such exploding diversities largely overwhelm the effectiveness and efficiency of the accelerator, making traditional architectural paradigms less applicable.

AI accelerators are already predominantly used to boost up the throughput of training and inference. It is expected that AI accelerators would continue evolving and become even more ubiquitous in broader areas including data analysis, simulations, etc. to achieve higher efficiency. However, with the architecture being dedicated to accelerating the AI workloads specifically, the workload capability of such accelerators must also evolve as the target AI algorithms. AI training requires huge amounts of data to be frequently moving across chips and chiplets. This would lead to escalating energy and consumption costs. An emerging workload is using memory side computing by rebaking the weight on DRAM chip 3D stack to alleviate the data exchange overhead. However, new power delivery and thermal issues arise that need to be researched and resolved (Ali et al., 2024; Poduval et al., 2024).

**Fig 7.1:** High Performance Computing Insight

### 7.1.1. Background and Significance

Recent years have witnessed a rise in the number and complexity of mechanical components in various hardware systems and applications, from mobile devices to cloud data centers. Such a trend results in outstanding performance improvements, signal integrity enhancements, and cost reductions, etc. However, it also leads to serious design challenges, including huge design complexity, extensive power and performance analysis with inefficiency, high cost in implementing a manufacturable design, debugging and verifying a robust and reliable design, etc.

To meet these design challenges, high-performance computing has been driving the advancement of very large scale integration in both hardware and software. Within the hardware framework, it gives birth to extraordinary massively parallel computing chips with several hundred million transistors, extremely fast junctions that can be fabricated very close to each other on a chip, extended scalable system architecture, etc. In the software system, evolving and maturing simulation tools for performance evaluation, design analysis, functional verification, etc., to support system modeling, simulation, and emulation at both a high level of abstraction and very high speeds, have become critical for the rapid convergence of the design cycle. On-chip accurate power and performance extraction tools, robustness analysis tools, and deep learning-based post-silicon validation tools further facilitate the evaluation and debugging of a design once it is implemented, resulting in a more robust, reliable, and manufacturable design.

Simulation techniques have helped tremendously and are expected to continue to dominate the increasingly more advanced hardware industry. However, the escalating complexity of hardware systems can quickly surpass the computing power of simulation techniques, and thus the limitations of simulation-based approaches have gradually been realized. For example, to avoid unnecessary high-level polynomial time complex simulation for performance evaluation, top-down statistical methods have been developed. Nevertheless, these sampling and statistical approaches always result in performance estimation rather than concrete performance value. This performance estimation impediment results in a large discrepancy between the performance estimation and simulation environment, which incapacitates various efficient design exploration and optimization techniques. Instead of work on performance evaluation, work has also been done directly on the design optimization side. For example, iterative fast gradient-based techniques have been developed to maximize the worst-case performance of a design, but still need a simulation to verify the improvement.

## 7.2. Overview of High-Performance Computing

Computing power is the basis of every computer program. Only a few years ago, the computational capacity of a typical laptop was considered extraordinarily high, now many AI applications require computing capacity that is thousands of times greater. This has led to increasing expenditures in computing power, teraflops, petaflops, exaflops, and even zettaflops clusters capable of executing quintillion and sextillion floating-point operations (FLOPs) per second. A rethinking of the architecture for computing systems is urgently needed to meet the demands of emerging AI applications. Recent history demonstrates that AI-centric computer architectures generally achieve more rapid advancement with lower cost-to-performance indicators than Von Neumann architectures. Nevertheless, AI-centric computer architectures often come with trade-

offs in programmability and expandability. Two basic considerations moving forward are chips designed with more special-purpose units than conventional chips and heterogeneous architecture designs integrating heterogeneous chiplets.

Chip scale. Progress in AI has inspired the looking back-and-forth exploration of computing architectures based on special purpose integration. Traditional Von Neumann computing counterparts are architecture-on-chip designs, including probabilistic chips; neuromorphic chips; and photonic chips. With phenomenal increase in the number of distributed chips, a heterogeneous chiplet architecture would take the advantage of seamlessly combining many rapidly evolving application specific architecture design paradigms to cope with the fast-evolving nature of AI algorithms. A heterogeneous chiplet architecture would tremendously simplify the collision of the expansive design space with the growing vulnerability of Moore's Law, as it would provide water for the forward engineering iteration process of many individual and diverse designs.

Chiplet system challenges. To harness tremendous computing power on a chiplet platform, several key technologies need to be conceived and further developed. Organizing computing units into a high-performance and efficient interdie communication network is the critical first step for chiplet success, because mismatched communication rings would cause load imbalance and degradation of system throughput. A 3D chiplet needs dedicated circuit designs to match high-density inter-chiplet wiring fabric. Die displacement on the wafer is a drastic challenge for wafer-level chiplet packaging technology that should be managed beforehand to enhance yield ratio. On-chip communication protocols are also crucial for cost-efficient and fast communication. A heterogeneous architecture adopts many accelerators on chip, which are also heterogeneous computing architectures on each chip. The interoperability, scheduling and load balancing of heterogeneous architectures would benefit from the elucidation of fine-grained AI models and workloads to introduce chiplet exploitability frameworks and frameworks to aid architecture evaluation just as Windows and Android expand the performance accessibility of heterogeneous mobile processors.

### 7.2.1. Definition and Importance

High-performance computing (HPC) has become a broad-encompassing field that encompasses high-performance hardware architectures (processor, memory, storage, media, etc.), high-performance interconnects and networks, job scheduling and resource allocation middleware, high-performance operating systems, high-productivity programming languages, compilers and libraries, high-quality numerical and software engineering, fault tolerance, and so on. It is critical to clearly establish the concept and boundaries of HPC before discussing its major role in scaling AI-centric semiconductor architectures, as the term might be interpreted to include faster computing by better

software, applications, or data. With the rapid rise of AI and deep learning (DL), the term "AI" has become a default prefix of many terms or technologies in HPC as in most other fields, which, as a consequence, contributes to the cloudy definition of HPC.

HPC infrastructures and technologies are critically needed to keep pace with the fast growth and huge computing/scaling demands of AI algorithms. In particular, new AI-centric semiconductor architectures need to be designed and fabricated to satisfy the heterogeneous computing demands from divergent AI algorithms, data formats, and methodologies. Such an effort involves the co-design of the hardware architecture, the architecture-middleware-software stack, the corresponding tool chain, and the AI algorithms. Designing semiconductor chip architectures is usually a multi-year and even decade-long effort. Even for a simpler design, it usually takes one to two years to fabricate a chip. It is thus a big challenge to scale the AI-centric semiconductor architectures to satisfy the rapidly growing computing demands of AI.

### 7.2.2. Historical Development

Research efforts aimed at "Heterogeneous Chiplets" for semiconductor architectures began as early as 1984, when various types of programmable logic structures and problems to chip interconnection were discussed. In the next decade, an adaptive mesh-like chiplet design was proposed that optimally distributes connections among chiplets and locales to a larger chip. A similar design called "circuit on demand" was later developed to reconfigure chiplets dynamically via programmable circuits while each chiplet is statically designed. More elaborate heterogeneous chiplet systems were investigated in the following decades, suggesting connectivity activation before integration, a hybrid on-chip interconnection design, dynamic reconfiguration of chiplets based on online workload characteristics, and a modular architecture allowing chiplets either to remain static with predefined connections or dynamically migrate to on-chip locations. Practical examples of homogeneous chiplets like 3D DRAM or CPU to GPU interconnects are readily found today. Fully heterogeneous chiplets are comparatively fewer, with designs taking square FPGA chips and sequentials for operating GPUs for IC logic and cache design. A 28nm 256Gbp 64×32 die stack combined with FET memory chips was suggested. The complexity in placement was discussed, proposing placements of chiplets in accounts of the fine-grain heat control and small thermal design costs thanks to the homogeneous chiplets. The channel on transmitting more than independent chiplets over interfacing capacitors was discussed, showing external and per-diem ground noise. Recently, the design of a brain-inspired and heterogeneous HPC architecture over CPU chips, sensors, racetrack and optical memory chips VLIW architectures was detailed. The design has been used for analysis by several other researchers. The critical power gain over networks' throughput on hybrid comparing

networks on chips with recursive placements among identical devices through meshes and examined distance's effect on throughputs.

### 7.2.3. Current Trends

As AI algorithms quickly transition from prototype and academic stages to enterprise-ready systems in cybersecurity, finance, healthcare, and self-driving cars, the underlying infrastructure to satisfy the ever-increasing demand for AI computing resources remains a challenging and multifaceted problem. In data centers, there have been multiple co-design avenues to scale up and scale out the AI computing capabilities including domain-specific hardware accelerators, memory architecture co-designs, heterogeneous systems, and even new computation paradigms. An interdisciplinary discussion is presented about the major trends in hardware architecture and compact interconnects for enabling large-scale and AI-centric computing systems.

As analyzing the unique characteristics of AI workloads considering both emerging model complexity and increasing data volume, it delves into the computing, memory, and communication resource requirements for AI workloads. The challenges of monolithic chip scaling and architecture diversity in IPs and architecture types are explored. It is focused on the opportunities and challenges brought by heterogeneous chiplet architecture and ecosystem. The M-Scale AI-Driven High-Performance Lifecycle Computing is proposed, which calls for close collaboration among chip, board, system, co-design stacks, co-simulation, operating system, and compiler infrastructure. It is anticipated that interdisciplinary consensus building is urgently needed to address the increasing complexity of hierarchical computing systems.

## 7.3. AI-Centric Semiconductor Architectures

While semiconductor chip-based deep learning has achieved impressive results over the last decade, massive amounts of data and complex models have fueled the demandability for fast, efficient, and high-performance chips in terms of compatibility for recent large-scale AI models such as ChatGPT. To accelerate these inferences, the typical design style for chips is highly scalable, leading to a large number of processing units. As a consequence, sophisticated interconnects are often required to hold together the programming. Combined with the ever-increasing data movement, it leads to large power and bandwidth overheads. To mitigate these problems, emerging Distributed Architecture design is discussed, which aims to disintegrate chips into multiple smaller chiplets composed of heterogeneous domain-specific architectures, with different chiplets serving workload-specific purposes. This design strategy not only significantly reduces the connectivity and packaging costs but also improves adaptability, reusability,

and flexibility. However, it necessitates an efficient interconnections framework across chiplets, which is nontrivial, given the high-performance and high-bandwidth demands.

In addition, a well-designed chiplet architecture is key to enabling energy-efficient AI-capable computing solutions. High-level synthesis brings merit by allowing designers to work at a level of abstraction closer to algorithms/protocols instead of circuit topologies. Such abstraction can make it easier to explore alternative architectural designs. However, it is more challenging to guarantee bugs in hardware implementations as the abstraction level rises. Moreover, high-performance computing demands of many AI workloads can form significant computational bottlenecks in chip-grade AI acceleration. Thus, new hardware designs and performance models are needed to facilitate the delivery of chip-grade AI accelerators, which is an enabling step crucial to make effective use of the rapidly advancing fabrication technologies.

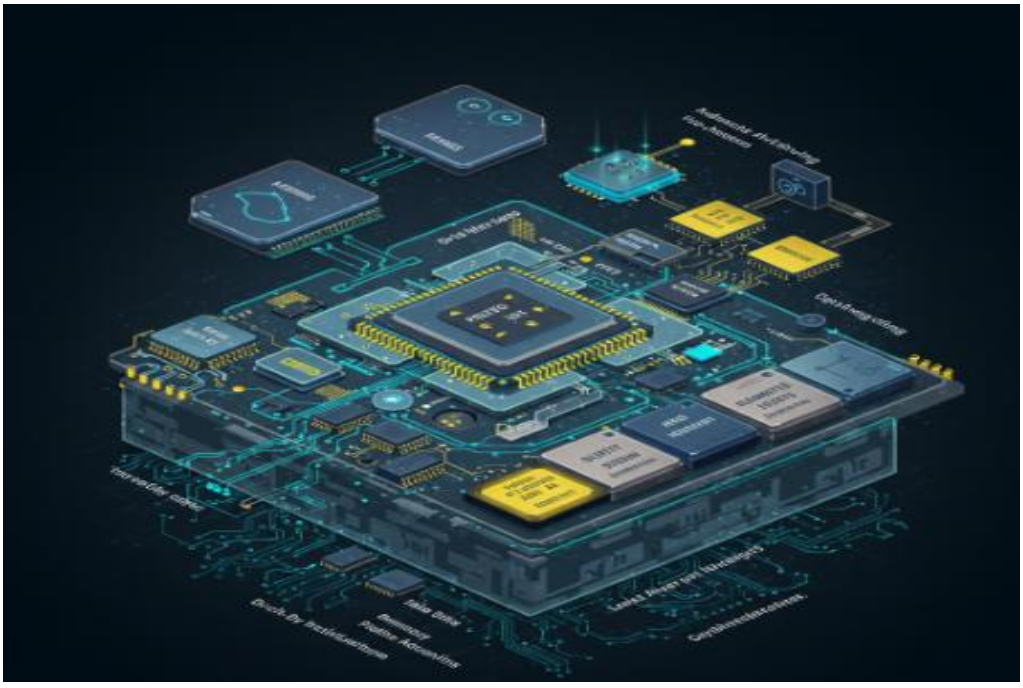

**Fig 7.2:** AI-Centric Semiconductor Architectures

### 7.3.1. Definition and Characteristics

AI (Artificial Intelligence) models, particularly generative large language models (LLMs), are driving an increased focus on high-performance computing (HPC) in both commercial and research environments. There is a need to frame AI workloads broadly in order to characterize computer, memory, storage, and hierarchical interconnects,

spanning from the edge to the data center. In terms of computer architectures, there is a need to profile custom ASICs for accelerating AI inference, exploring heterogeneous architectures including CPUs, GPUs, and TPUs for scaling up training, and studying High-Performance Computing and fast-evolving AI architecture collaboration scenarios on exascale supercomputers. As concerns about processor power and capacity density increase, there is a need to design high-density architectures for efficient data movement and computation during training and inference. In terms of communication and network architectures, there is a need to characterize data distribution and model partitioning features for co-designs of communication and network architectures based on AI model characteristics. The performance bottlenecks of existing interconnects should also be analyzed and explored, along with emerging and next-generation application-aware switched architecture networks. Dynamic reconfiguration capabilities for efficiently composing the most suitable internode communication topology should be investigated, as well as task- and data-centric communication abstractions using mixed programming models and programming paradigms coupled with heterogeneous architectures.

AI algorithms demand intricate model architectures that can no longer fit into the memory of a single training processor. For good performance, data parallelism and pipeline parallelism must be well-tuned together to minimize the inter-node communication costs. Also, feed-forward layers within the model must be well-paced in the middle stage between communication and computation. Many existing deep learning training systems struggle to optimize performance in such models. To address this gap, an easy-to-use and high-performance system based on the backpressure mechanism is proposed. The newly proposed conversations, making it easier to represent a model training process and construction of software frameworks with new AI-HPC workflows, along with transition designs to leverage existing training systems are also discussed. This software is also successfully integrated with existing big computing clusters and widely used by the AI community. AI is transforming every sector of society, from medicine, finance, and transportation to entertainment. Heightened interest has been paired with rapidly-evolving generative AI model types, architectures, and associated software tools and frameworks. AI algorithms and models differ greatly in terms of their architecture, layer and data tensor characteristics, and system requirements. These differences, alongside wildly different model training frameworks, present a substantial obstacle to the integration of AI and HPC workflows.

### 7.3.2. Key Players in the Market

AI-centric semiconductor architecture is in the center of high-performance computing or supercomputing as computing, memory, and communication technologies are scaling to exascale and beyond. AI-centric semiconductor architecture can use a similar

architecture, algorithm, and methodology to massively parallel and distributed architectures for traditional supercomputers with algorithm and architecture customization. The custom AI-centric semiconductors then feed the value-added AI computing and its applications in simulation, cloud computing, and autonomous everything.

Global companies including semiconductor and system vendors, cloud and datacenter services providers, AI chip and software developers comprise the AI computing supercomputing ecosystem. These players and their products, technologies, platforms, and opportunities are summarized. AMD, Intel, and Nvidia compete for the datacenter CPU market. AMD's EPYC CPUs and AI chip MI250 and MI300 use chiplet technology for scalability and efficiency. Intel's Sapphire Rapids and Granite Rapids CPUs and Gaudi AI chips with on-package high-bandwidth memory target the AI datacenter together with Habana's Gaudi2 chips with similar technology. Nvidia's CPUs and chips including Hopper GPUs drive the AI datacenter market, supported by its association acquisition and omniAI initiative. Custom semiconductor chip and IP vendors include Arm, Graphcore, Renesas, and Tenstorrent. Software stack and ecosystem providers include software stack vendors ETA and Dwanze, chip and software co-design IDEs Semihalo and Kumo, hypervisors Domospace, scheduling middleware scheduler and Coigone, x86 programming language extensions Projex and Exsolv. 9886 startups from AI software to AI chips leverage these vendor products and services. Enterprises deploy these technologies.

The companies from semiconductor to cloud computing hardware and programming languages dominate the first two layers. Consumer innovation, manufacturing optimization, and combination are opportunities in the third layer. Low-cost and energy-efficient optimal smart hardware devices, new unique smart hardware products including AIoT wearables, sensor, and actuators with algorithm customization, and flexible and recyclable quantum chips and devices with physically unclone-able function and their architecture, computing, and programming language remain challenges to avoid sleepless nights.

### 7.3.3. Recent Innovations

Recently, the performances of both software and hardware for training and inference of neural networks have seen revolutionary improvements. They range from new ResNet architectures and new optimizers with better convergence properties to innovative chip architectures and training algorithms that facilitate utilization of large supercomputers. These advances enable exceptionally faster training, such as the first training of a 1 trillion parameter model on supercomputers in only a few days. Several extremely successful AI companies taking advantage of these breakthroughs have emerged, and

the silicon valley AI community has faced an enormous paradigm shift. Unlike the traditional software industry with various software-based products, the most modern and successful AI companies also need to release prevalent new chips. Traditionally chip design and fabrication duration used to be on the decade scale, these white knights should form fierce rivalries with long-standing semiconductor giants. There still remains a huge knowledge base and unresolved challenges for chip architecting and designing. This greatly motivates researchers from both hardware and software communities to provide new insights to advance this growing new frontier. Although many newer devices have emerged, these revolutionary improvements on massively parallel chip architectures and novel computational paradigms for disparate neural network training all help to turbocharge this chip-centric golden age.

Deep learning (DL)-based AI technologies are rapidly revolutionizing every aspect of modern society, significantly outperforming their traditional counterparts in diverse fields such as computer vision (CV) and natural language processing (NLP). To keep up with the ever-accelerating performance demand, massive parallel platform strategies from single GPU model parallelism to distributed data parallelism across thousands of GPUs, chips, or nodes are extensively investigated. However, the software and hardware limitations are making it more and more challenging to scale to trillions of parameters and millions of GPUs. Most emerging large-scale training supercomputers are composed of more than 400,000 A100 or H100 GPUs with 8–80 GPUs per node all using PCIe interconnect. AI-on-CPU-centric architectures are investigated to meet vast computational demand and address dramatic interconnect scalability challenges.

### 7.4. The Intersection of HPC and AI

Machine learning (ML) from the perspective of high-performance computing (HPC) is one of the hottest topics in today's field of computing. It is being used in cutting-edge applications from seismic analysis of subterranean oil reservoir data to reconstruction of the cosmic web from cosmic microwave background sky maps to imaging the first black hole in history through the Event Horizon Telescope experiment to modeling polycyclic aromatic hydrocarbons in catalytic chemical processes and many more. On the one hand, the complexity of the scientific endeavors from these fields pushes the ML algorithms to the very edge of computing systems, which leads to the need for a larger scale of ML with more compute resources and ingenious utilization of these resources, in other words, a scaleup of ML. On the other hand, this wave of adoption of ML brings new challenges on how to efficiently scale ML algorithms on large supercomputers and utilize cluster management methods usually developed for traditional computing workloads in production environments for both large-scale cluster deployment and utilization. Consequently, a tremendous number of efforts have been invested in these

areas including massively parallel ML libraries and systems such as Lbann, Dask-ML, and Horovod, MPI-based sparse deep learning and hyperparameter optimization libraries, attempts to enable the efficient scaling of third-party ML libraries such as BigDL and LAMBDA on supercomputers either by creating distributed training backs or designing novel parallel training methods. However, the education and communication gap between the two communities is still prevalent. For instance, HPC clusters are equipped with HPC-oriented middleware and libraries to facilitate the needy support of scientific computing workloads. What if the ML workloads run on multiple nodes using the same backend MPI? Native applications usually call upon HPC libraries such as ScaLAPACK for high-dimensional large-scale tensor decomposition of necessary input data? Although this endeavor can, in general, alleviate many implementation details, bringing in these competencies into the ML world may shorten the time to science enormously. Vice versa, ML techniques have been extensively employed in numerous aspects of HPC. The combination of these two worlds is actually yet to be enriched.

### 7.4.1. Synergies and Benefits

The advancement of AI technologies has fueled the rapid growth of their computing demands. Understanding how to effectively deploy a high-performance AI-centric semiconductor architecture to address these soaring demands and expand its deployment and application in other domains lies at the intersection of heterogeneous processing technologies and high-performance computing (HPC) workloads. On the processing technology side, traditional approaches mostly rely on optimized monolithic architectures, which successfully improve efficiency and performance but cannot cope with the exponentially growing demands of new-age applications, such as AI inference. Thus, as conventional scaling approaches become infeasible, scalable and chiplet-based heterogeneous chip architectures have begun to take the stage, leading to a shift from a “design once for all” homogeneous architecture paradigm to a more flexible heterogeneous architecture. As an interesting twist, this evolution towards heterogeneity creates opportunities for new approaches in high-performance computing by enabling the co-design and co-optimization of both architectures and workloads, particularly for AI-HPC workflows .

Traditionally separated HPC and AI have recently converged, and a new paradigm with the interplay of traditional structure-based numerical simulations and modern data-based machine learning has begun to gain attention. This trend is primarily driven by a strong demand for scientific analytics in a more effective and efficient way for better decision-making. On the one hand, service providers of exascale systems are actively exploring the deployment of AI models to facilitate the monitoring of workloads, helping improve

system performance, utilization/effectiveness, and reliability. As targeted workloads shift from pure simulations to AI-coupled workflows, it is of great interest to study the needs from both sides and develop better orchestration strategies for improved performance and efficiency. On the other hand, AI workloads require low-latency and high-bandwidth inter-chip communication, meaning that abundant PHYSICAL chip-hops should be designed and implemented in an area-efficient way.

### 7.4.2. Challenges in Integration

With the rapid increase in the Inference workloads generated by Generative AI, semiconductor technology nodes, computational structures, AI algorithms as well as heterogeneous architectures are all moving towards different directions. New memory architectures that are compatible with these nodes, structures, algorithms and architectures are critical for unleashing their full potential. In this talk, the challenges facing emerging AI-centric semiconductor architectures beyond computing and memory will be briefly discussed. AI workloads in Generative AI are characterized with large models, diverse data types and real-time model compositions. The capabilities of various hardware accelerators beyond computing are summarized against the characteristics of Generative AI workloads. In particular, co-optimizing the many-core fabrics, on/off-chip memory and high-performance interconnect will be discussed in detail. Emerging chiplet technology enables heterogeneous chip design across different IP vendors, and hence increases the availability and variety of customized DNN accelerator architectures. A well-trained neural network model can have millions of parameters and hence impractically large storage footprint. Complicated dataflow and extensive data movements inefficiently utilize on-chip memory bandwidth under high precision weights. A DNN model can be designed, trained and retrained with an increased depth and width to achieve a target accuracy. However, scaling up a model increases compression difficulties in Post-training Quantization (PTQ), minimizes the trade-off between precision and performance in Federated Learning (FL), and limits the representation ability in model pruning. Logic-in-Memory (LiM) devices, with non-volatile semiconductor materials to implement memristors, and novel architectures show good capacity in enhancing the endurance and speed but struggle to utilize less power and cost. They can evolve as a product but rich co-design strategies are still needed to unlock the technology potential. Under the rapidly changing technologies, a chosen architecture may not be the most efficient choice in exascale DNN training implementation. Existing chiplet design flow needs to be modified to enable self-tuning for a more efficient chip design.

### 7.5. Scalability Issues in AI-Centric Architectures

High-performance computing (HPC) enables the exploration of various architectures for scaling AI-centric workloads. The goal of HPC is to realize the inherent parallelism in workloads, whether it is deep learning or traditional scientific computing, and take advantage of that parallelism across compute cores. The parallelism is explored at different levels and corresponding systems are built that feature high customizability and performance on the workloads of a given class. Within the TPU family, systolic arrays are used for matrix multiplications and model training, while a bridging chip based on custom architectures is used for embedding updates and user assignments .

**Fig :** High-Performance Computing in Scaling AI-Centric Semiconductor Architectures

There are architectural components of a system that can be off-loaded to hardware to support customizability or performance. Systems off-load performance-critical tasks of the workloads like training or inference to compute clusters consisting of multiple nodes where multiple chips collaboratively work on compute-intensive tasks. The computer nodes take advantage of scalable interconnects, snapshots, switches, protocol-offloading chips, and data-agnostic bandwidth. Future work will focus on addressing the loopholes in the architecture and improving the overall scalability and performance. These systems usher in deep learning computing infrastructure with the ability to scale in accordance with compute demand. Heterogeneous computing with domain-specific architectures (DSAs) opens up many opportunities while scaling up and scaling out the computing system. A heterogeneous chiplet architecture is favored by cloud service providers to

keep scaling up and scaling out the system as well as to reduce the design complexity and the cost stemming from the traditional monolithic chip design.

### 7.5.1. Performance Bottlenecks

With the ever-increasing compute demand of ever-more complex AI models, performance bottlenecks are expected to arise. The architectural key resources for performance scaling are bandwidth and computation (i.e., the capability to execute FLOP). Traditionally, the scaling of bandwidth and computation follow Moore's law as feature sizes shrink and, together with the foundry/nodes technology, scaling further advances at an ever-increasing pace. However, as feature sizes are reduced to a novel node, particular design-related technical challenges need to be addressed, which collectively result in RFIs and large performance penalties, especially on both bandwidth and computation. Hence, heterogeneous designs, such as a high-performance chiplet for the chiplet architecture need to be created for performance scaling.

Moreover, in the post-Moore’s law regime architectures scale computation localities by preserving a strict decentralized bisectional bandwidth balance which enables scaling the partitioning of algorithmic and hardware aspects on single and multi-node massively parallel architectures. For in-memory/fabric HPC-AI and GPU-free architectures, algorithms and novelty heterogeneous architectures execute computations on unconventional memory-centric elements within the integrated memory and accelerate AI-completions and re-completions cycles on-chip on unconventional memory-centric elements within high-speed fabric chiplets.

### 7.5.2. Resource Management

Effective management of distributed resources is crucial in executing applications spanning multiple computing resources. Many scientific applications for AI model training and inferencing require immense computational resources and often rely on multiple machines or computer architectures, such as CPU-GPU, heterogeneous or hybrid cloud-geographical computing resources, etc. To address the resource management challenge, a multi-resource resource management service for orchestration and scheduling of applications was built, with a focus on scheduling of loosely-coupled applications across heterogeneous resources. In addition to scheduling procedures, the design of application, resource and task models was discussed, serving as the foundation of more sophisticated resource management schemes addressing resource contention and scheduling compliance.

The increasing processing demands of emerging multi-model workloads and high-performance computing requirements pose significant challenges for current chip monolithic multi-chip module (MCM)-based AI accelerators. To address these issues, the requirements for scheduling workloads on heterogeneous AI MCM architectures were identified, and a general scheduling model was proposed. It has been shown that optimal scheduling is NP-hard. A set of fast heuristics to minimize schedule length and energy-delay product were developed, which codify advanced techniques such as pipelined execution and work-sharing strategies. The proposed scheduling method demonstrated effective and efficient scheduling for MC-MCMs, far exceeding previous homogeneous MCM scheduling approaches .

## 7.6. Conclusion

There is an urgent need to maximize the compute efficiency, performance, and capability of AI-centric semiconductor architectures, deployed in the cloud to support large-scale AI workloads will remain ever-present. The constant demands of high-performance training and inference of real-time recommender systems, large foundation models, and highly efficient AI accelerators, etc., will drive the innovation for faster and more power-efficient chips.

AI-centric semiconductor architectures in data centers can have multiple common features, including scalable OMI-based distributed memory, on-chip buffering that reduces memory access energy and latency, and on-chip triple-TLBs, a key enhancement to improve the efficiency and robustness of high-bandwidth memory architecture. A novel and comprehensive on-chip bank partitioning of adder trees across various AI chips will be demonstrated to reduce HBM memory bandwidth and, hence, energy. The performance and energy benefit of global context switching for weight-stationary multi-dimensional tensor core-based DNN inference are also discussed.

Lastly, a few paradigms, such as multi-fidelity deep learning, forecasting AI for future multiphysics models, and compact modeling for fast transistor-level simulation of very-large SoCs, can be explored to broaden the current applications of field (and speed) programmable gate arrays (FPGAs) in enhancing the design and outputs of other computational resources. Lastly, promising future directions such as the advancement of fabrication and packaging technologies to mitigate the impact of parasitics on performance and robustness, addressing the temperature gradient and non-uniform efficiency of AI OMI with a joint consideration of chip architecture, fab technology, and architecture into circuit design to break the limitations of mats connectivity between chips, etc., are outlined.

### 7.6.1. Emerging Technologies

The emergence of generative AI has brought new challenges for hardware design. The scalability and efficiency of generative AI greatly rely on high-performance computing. Beside accelerators, the communications architecture, especially on-chip heterogeneous interconnects designs have become critical. Major cloud service providers are proposed to build dedicated GPU clusters over 1000X larger than the currently deployed chip with thousands of GPU and Infiniband nodes. However, the scalability and costs of HPC is becoming a bottleneck for high-end generative AI workloads. Recently, there is also a trend of exploring advanced Memory Contributions, from Memory Centric to Memory Near. HPC can be in the (near) memory with 2D/3D area-efficient dense processors or 3D chiplets with HBM interfaces.

Generative AI workloads scale dramatically to meet the increasing model/parameter size. The costs go massively higher and the energy consumption also scales exponentially per unit AI generation. Training typically takes weeks on hundreds of GPUs and there are growing concerns on the effectiveness of carbon-neutrality. With chiplet technology, such GPU replaced predecessors requires hundreds of on-chip high-speed pairs for interconnects compared to many paired channels on chip with networks-in-memory, showing 3D a faster performance scale-up.

## References


Poduval, P. P., Ni, Y., Zou, Z., Ni, K., & Imani, M. (2024). NetHD: Neurally Inspired Integration of Communication and Learning in Hyperspace. Advanced Intelligent Systems, 6(7), 2300841.

Ali, H. M. S., Jalal, S. K., Saab, M. W., Sulaiman, S. K., Ghno, G. S. N., Mustafa, S. I., & Azimov, B. (2024, April). Deciphering the Implications of Swarm Intelligence Algorithms in Efficiently Managing Drone Swarms. In 2024 35th Conference of Open Innovations Association (FRUCT) (pp. 112-123). IEEE.