Scale-within, Scale-up and Scale-out Fabrics Enable Efficient AI Acceleration

Karl Freund, Founder and Principal Analyst
Cambrian-AI Research
June 23, 2025

Introduction

While GPU performance has been the focus in data centers over the last few years, the performance of fabrics has become a key enabler or bottleneck in achieving the throughput and latency required to create and deliver artificial intelligence at scale. Nvidia's prescient acquisition of Mellanox has been a critical component of its success over the last few years, enabling scalable performance. However, it's not just scale-up (in-rack) performance and scale-out (rack-to-rack) connectivity; latencies in those, as well as scale-within network-on-chip (NoC), have become essential for achieving high AI throughput and improved response times.

Nvidia first provided the networking hardware and software necessary for large-scale AI infrastructure with the proprietary NVLink switched fabric. Despite the recent announcement of NVLink Fusion to expand the ecosystem, the industry remains eager to see a more open alternative that enables a broad ecosystem of AI chip vendors, including AMD, Intel, Qualcomm, and numerous startups. Additionally, the explosion of chiplets requires a network on the chip to connect memory, and CPU cores require fabric technology intellectual property (IP) to link pre-built chiplets to new logic blocks designed for custom application support.

This paper examines the markets of fabric IP and related software that facilitate the scale-within, scale-up, and scale-out networking needed by the HPC and AI infrastructure landscape.

The Importance of Advanced Fabric Solutions

The computing landscape has undergone significant changes with the advent of Artificial Intelligence, evolving from a loosely coupled network of independent computers to a highly integrated fabric of collaborating, accelerated computing nodes. Three levels of scale require such interconnects: chip-to-chip, rack-to-rack, and data center-to-data center. Each compute element must share data with its neighbors and beyond over a low-latency, high-bandwidth communication channel to maximize performance and minimize latency.

As the specialization of computing also increases due to the need for performance, efficiency, and scalability, the complexity of design grows accordingly. Control-flow processing with the CPU and data flow processing (initially visual or graphical) with the GPU, as well as more advanced neural processing with the NPU, require different computing elements to work together in synchrony. However, the more critical aspect of this is how data movement is done between them.

Fabrics on system-on-chips (SoCs) tend to be isolated; they are designed to connect a specific domain on the chip, which works well until data needs to be moved to another domain, creating latency-inducing "hops" due to bridging and other limitations. Common examples of NoCs are:

- Coherent CPU-to-CPU fabric
- A non-coherent fabric that connects memory to specific functions
- Protocol fabrics such as NVLink, UE, and UALink for AI
- An I/O coherent fabric
- A multi-socket coherent fabric for connecting various levels of cache

A unified NoC provides a single transport mechanism for the various protocols for each fabric, as in the figure below. Transport is separate from protocol layers, minimizing wires and logic in building a unified fabric that supports coherent, non-coherent, and custom protocols for maximum efficiency, lowest cost, and reduced power consumption.

Figure 1: A Unified Fabric across a SOC links the various NoCs with a single transport layer.

Scale-within Fabrics

On-chip fabrics connect processor cores, accelerators, and cache memory within a single or multi-chip module. As SoCs become more complex, integrating tens or even hundreds of cores or IP blocks, a single NoC often cannot provide the required bandwidth and scalability. Multiple NoCs, or subnetworks, are used to manage traffic within each System-on-Chip (SoC) or chiplet, each potentially optimized for specific data types or communication patterns. Additionally, NoC needs to be aware of the design to ensure efficient communication between chiplets. For example, one NoC might handle high-bandwidth data transfers between compute chiplets that understand software coherency across these domains, while another manages control signals or memory access. As chiplet-based designs gain wider adoption, these NoCs become the bottleneck of chiplet-to-chiplet communication and data sharing, potentially leading to unnecessary cost and energy inefficiency due to suboptimal on-die and cross-die memory systems.

However, current NoCs were not designed to be integrated into a single transport layer, thus presenting potential barriers to performance and latency on the SoC. A unified fabric approach could enable "Scale-within" designs that can more easily extend to Scale-Up and Scale-Out implementations. These chiplet-ready fabrics also optimize the extra data movement or minimize the caching or resources associated with cache coherency in compute subsystems.

Distinct chiplets often use different process nodes and architectures (e.g., CPU, GPU, memory, I/O). A unified fabric can carry standardized protocols (e.g., UCIe, BoW) for interoperability and resolve protocol mismatches (e.g., AMBA CHI vs. AXI). For example, the Baya Systems' unified NoC fabric is used in the Tenstorrent accelerator to reduce hop counts compared to traditional mesh topologies and cut latency.

A unified fabric significantly enhances latency and bandwidth in chiplet-based systems by streamlining communication across fragmented networks-on-chip (NoCs) and optimizing physical interconnects. Here's how it achieves these improvements:

Latency Reduction

1. **Hop Minimization**
   Traditional multi-chiplet designs require data to traverse multiple NoC hops across separate chiplets.
2. Intelligent Routing
   - Baya Systems' WeaverPro software dynamically selects shortest paths using predictive congestion avoidance algorithms, bypassing traditional XY routing limitations.
   - Protocol Overlay: Supports AMBA CHI, AXI, and custom protocols on shared wires, minimizing protocol translation delays.

Bandwidth Improvements

1. Unified Namespace
    - Eliminates redundant data copies by treating distributed chiplet memory as a single address space.
    - Enables 512-core coherent scaling (vs. 64-core limits in fragmented designs).
2. Parallel Fabric Planes
    - Deploys multiple independent fabric layers (e.g., 4x200G lanes) for concurrent data streams.
    - Achieves 6.4 Tbps aggregate bandwidth in Tenstorrent's RISC-V chiplets.
3. Congestion Management
    - Configurable QoS: Prioritizes critical traffic (e.g., GPU tensor data) over background tasks.
    - Credit-Based Flow Control: Prevents buffer overflows, maintaining >95% link utilization.

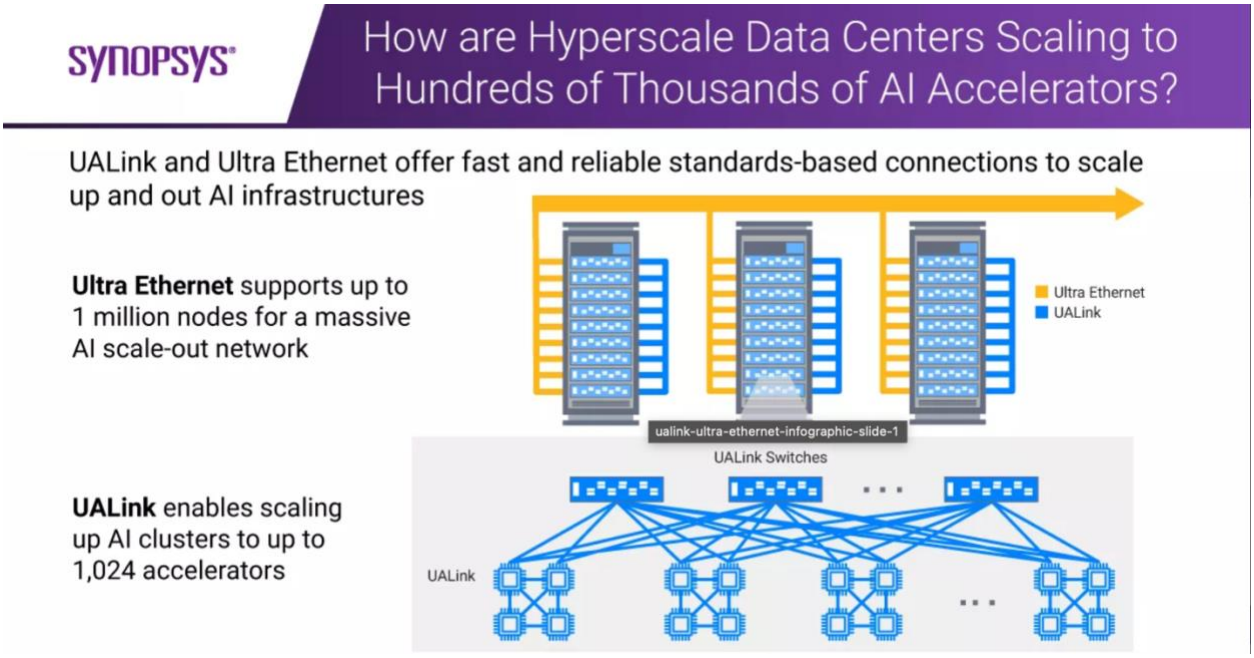


Figure 2: Positioning UALink and UltraEthernet

Scale-Up Fabrics

Scale-up fabrics connect accelerators (GPUs, AI processors) within a single rack or AI pod, prioritizing ultra-low latency and high bandwidth communication. Scaling up with NVLink has been the go-to standard, but the industry needs an open alternative, such as UALink, to interconnect accelerators from other vendors. UALink is a memory-semantic interconnect standard led by the UALink Consortium, enabling accelerators to

share memory directly. Its four-layer protocol stack supports single-stage switching, reducing latency and congestion. UALink will deliver up to 200 Gbps per lane and memory-sharing capabilities to scale (up) accelerator connectivity. The Consortium recently approved the V1.0 Specification of UALink in April 2025, and the first silicon is expected to be available later this year, with volume production scheduled for 2026.

Scale-Out Fabrics

Scale-out fabrics interconnect multiple racks or pods, enabling the distribution of workloads across large clusters. Nvidia offers both Ethernet and InfiniBand networking to connect racks for east-west traffic. As for scale-up alternatives, the industry is standardizing a high-bandwidth open networking protocol called Ultra Ethernet tailored for AI workloads across as many as one million heterogeneous nodes.

Ultra Ethernet IP solution will enable 1.6 Tbps of bandwidth for scaling (out) massive AI networks. UALink will deliver up to 200 Gbps per lane and memory-sharing capabilities to scale (up) accelerator connectivity.

Baya Systems' Fabrics

Baya has architected a semiconductor IP and software portfolio to enable designers of SoCs, systems, and data center scale infrastructure to build high-performance AI technology quickly and efficiently. Baya Systems fabrics are designed to address both on-chip, scale-up, and cross-system (scale-out) networking challenges. Its flexibility and modularity position it for broader applications, potentially integrating various processing units and accelerating communication in diverse, high-performance environments. The Baya fabric supports multiple protocols, including AMBA, UCIe, UALink, and UltraEthernet. WeaveIP also enables large clusters of coherent processors that can support up to 4TB/s of coherent bandwidth on a single die while reducing the silicon footprint compared to standard mesh designs.

Baya Systems' primary focus is on enabling designers to explore the optimal system configuration, from memory hierarchy development to chiplet partitioning, and the rapid development and deployment of System-on-Chip (SoC) and System-of-Chiplets. This approach overcomes the real challenges of complexity, performance readiness, risk reduction, Time to Market (TTM), and future-proofing.

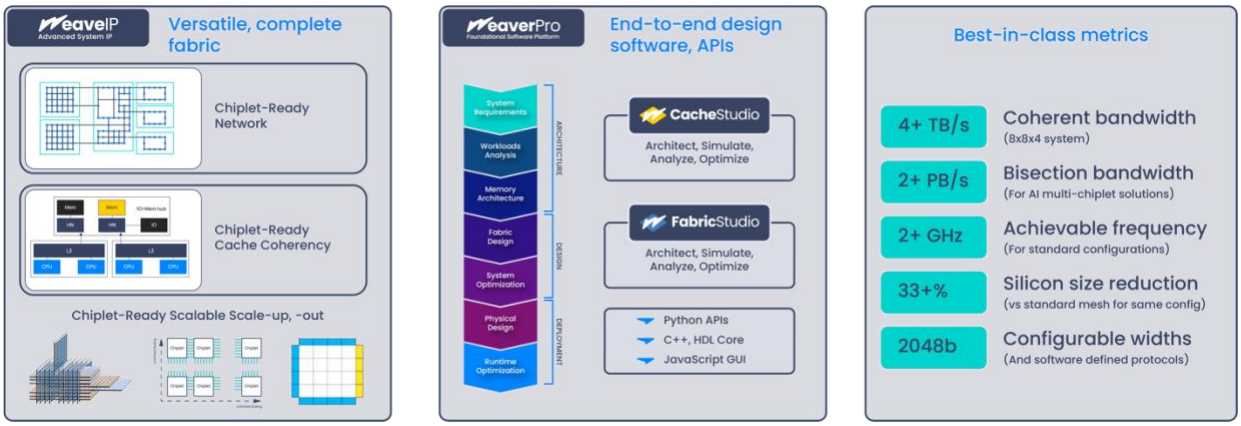


Figure 3: Baya Systems has created a comprehensive fabric that supports popular protocols for scale-within, scale-out, and scale-up.

Baya Systems' technology is built around high-performance fabric architectures designed to interconnect components:

- **Scale-within Networking:** Within a chip or a single package, Baya's technology enables ultra-fast, low-latency connections, improving on-chip communication between cores, accelerators, or memory controllers. This is crucial for achieving higher performance in compact, integrated designs. Scale-within supports AMBA and UCIe protocols.
- **Scale-up and -out Networking:** Baya extends these principles beyond a single system, enabling high-speed interconnects across multiple chips or systems. This is particularly relevant in data centers and distributed computing environments where efficient, high-bandwidth communication is essential. UALink switches enable Scale-up, and Ultra-Ethernet switches enable Scale-out. Baya enables both of these switches to achieve higher performance (higher radix) by facilitating chiplet layouts that will allow "lego-blocking." This approach can unblock data center bottlenecks to achieve future Scale-up and Scale-up goals.
- **Advanced Interconnect Protocols:** Leveraging state-of-the-art signaling techniques and interconnect standards, the solutions are designed to optimize data flow, reduce bottlenecks, and support demanding bandwidth requirements.
- **Integration and Flexibility:** The fabric is software-defined, enabling "correct by construction" generated as RTL, which can be integrated with existing semiconductor processes and architectures, allowing manufacturers to adopt the technology without a complete system overhaul.

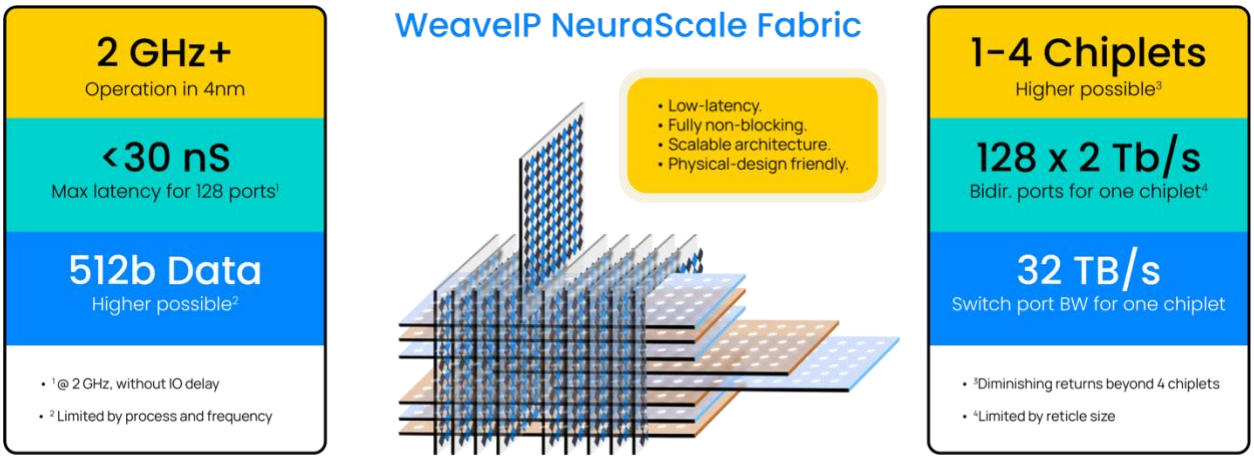
The foundation of the Baya System fabric technology is the WeaveIP family of network IP. By offering complete chiplet-ready IP and tools, Baya has created a fabric framework that SoC designers can rely on for reliable, high-performance communication, allowing them to focus on designing their specific value-added features.

Baya's also addressing the bigger picture with WeaverPro. The software platform enables a much faster, fit-for-purpose, future-proofed, and correct-by-construction outcome, thereby derisking some of the most complex systems being designed. They do this by supporting:

- Static and Dynamic data-driven design: Built-in simulation supports running actual trace workloads and traffic patterns to identify the exemplary architecture and microarchitecture, from memory and chiplet partitioning to final subsystem tuning.
- Correct-by-construction fabric: The software-driven fabric generation addresses design correctness, and the virtual channel (VC) approach also protects against deadlocking.
- Physically aware, tiled design: The platform enables simplified tiling to create a larger fabric, and the tiles can be oriented as needed for the best fit in floor plans while providing full awareness to designers of the impacts.
- Future-proofing and Quality of Service: Advanced monitoring and telemetry enable silicon systems in the field to perform route planning, optimization, and fairness.
- Advanced Reliability and Functional Safety: For data centers, infrastructure, automotive, and industrial systems, reliability and functional safety are mandatory. Baya Systems supports both of these aspects.

NeuraScale Scalable Switch Fabric

NeuraScale is a scalable fabric solution based on Baya's WeaveIP technologies, providing a non-blocking switching function between UALink or UltraEthernet for emerging scale-up and scale-out systems. Extreme port density is achieved while maintaining the near-theoretical lowest latency, a tight latency-bandwidth curve, and vastly reduced physical design (PD) complexity. The architecture enables hardening a single small switch tile and then building the switch fabric by utilizing an array of hardened tiles. Supporting all three major industry fabric protocols should help Baya reach the maximum set of potential customers. Each chiplet using the Baya NeuraScale fabric can deliver 32 TB/s of interconnect performance. According to the company, this is not a limitation of the architecture, and more capable designs are possible with the fabric.



Conclusions

The modern data center is evolving rapidly, both in its compute elements (chiplets, chips, CPUs, and GPUs) and in its networking, to enable these systems to scale to hundreds of thousands of nodes and support AI applications. To accommodate this scale, as well as the performance of the chiplet complex, various vendors and consortia are developing switching networks to compete with Nvidia NVLink and InfiniBand for scale-up and scale-out applications. Chiplet interfacing UCIe, scale-up networking standard UALink, and scale-out networking extensions to standard Ethernet (UltraEthernet) depend on semiconductor technologies developed by companies such as Baya Systems to reach the largest market.

While Nvidia's new NVLink Fusion will enable non-Nvidia CPUs and GPUs to participate in the company's Nvidia rack-scale architecture, fueling its growth, hardware vendors and hyperscalers will continue to seek an open fabric alternative to an ecosystem controlled by a single firm. Consequently, we envision a significant increase in these heterogeneous fabric technologies as AMD, Intel, and hyperscalers adopt them to build out their own AI Factories, both with and without Nvidia hardware.

IMPORTANT INFORMATION ABOUT THIS PAPER

Author: Karl Freund, Founder, and Principal Analyst at Cambrian-AI Research

Inquiries: Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

Disclosures: This paper was commissioned by Baya Systems, Inc. Cambrian-AI Research provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Disclaimer: The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.