Simulating Scale-Up: A critical step in AI infrastructure design

The importance of Large Language Models (LLMs) and Generative AI is pushing networking engineers to incorporate the role of **scale-up fabrics** in their networking simulations. Accurately simulating these tightly coupled, high-bandwidth, low-latency interconnects (such as UAlink or NVLink) is critical for understanding how the AI will perform in the real world network.

This white paper explores the significance of scale-up fabrics and explains how network engineers, infrastructure architects, and silicon designers can benefit from including scale-up in their network simulations to improve their design and deployment decisions. It also showcases that Scala Computing's simulation platform enables high-fidelity modeling across the full AI network stack at datacenter scale.

Why we need to Simulate Scale-Up

Modern AI workloads, particularly LLM training and inference, push the limits of compute, memory bandwidth, communications bandwidth and networking. The costs of training and inferencing make it well worth understanding how to optimize performance and network utilization.

AI workloads are too large for a single node. The compute and memory required is achieved by networking a number of nodes together. For best performance, those nodes must be connected together closely enough to operate as one unit. This tight grouping is called "Scale-Up" and becomes essential when AI workloads require tightly coupled processing or microsecond-scale inter-GPU communication.

Scale-up fabrics can also extend across chassis and racks. In these cases, network performance bottlenecks within a node or between a small group of nodes can have a disproportionate impact on application performance. Understanding these complex interactions, at scale, is difficult in a lab.

As Scala Computing is focused on simulating real world LLM workloads, it also needs to focus on simulating the Scale-Up Fabrics to match the behaviors of the real network. LLMs and Generative AI have dramatically increased the **complexity, need, and value** of high-fidelity Scale-up simulations.

- **Complexity is increasing**: In order to represent real-world networking, simulations must now include the scale-up fabrics, GPU communication patterns, sharding strategies, and multi-model orchestration that are operating in the real world.
- **Need is intensifying**: Entire data centers are being dedicated to AI training and inference with most endpoints being expensive GPU/accelerator heterogeneous systems. In addition to the increasing scale of the infrastructure, Generative AI models are growing and evolving (Transformers, Dense Models, MoEs, etc.) resulting in changes to how the algorithms are mapped to clusters and how they interact with RAG (Retrieval Augmented Generation) pipelines, and agentic environments.
- **Value is growing**: GPU infrastructure investment (CAPEX and OPEX) must now translate into tangible performance metrics like tokens per second, training time, and model quality, all of which are influenced by nuanced network behaviors.

Network Simulations to understand scale-up networking are increasingly critical for tightly coupled GPU workloads where latency and memory locality directly affect performance. With billions of dollars at stake, the need to optimize both data centers and LLMs is growing. The value of Simulation increases with the complexity of the problem that needs to be solved. With so many nuanced interactions, Simulation is the only practical method to explore these variables before deployment.

Modeling the Full Network Stack

Modern AI infrastructure involves three interrelated networks:

1. **Data Center Network;** storage, orchestration, control plane

2. **Scale-Out Network;** inter-node GPU collectives and communication
3. **Scale-Up Network or Fabric;** intra-node or intra-cluster GPU-to-GPU interconnects

Historically, simulations modeled only the scale-out network, assuming intra-node communication was fast enough to be ignored. The Scale Up fabric was rudimentarily modeled by assuming GPU chassis as the unit of deployment with very large bandwidth between GPUs (generally 8 per chassis). The goal was providing consistent low tail latencies for the ML training workload collectives, so that the network would not gate GPU performance and that performance is as close as possible to the optimal, based on the link speeds and switch topology. In these exercises the Data Center network was not modeled at all, except to confirm if it could be collapsed into the same Scale Out Network gear with no detrimental impact on metrics.

Today, accurate performance modeling of modern LLM workloads requires accurately simulating both scale-out and scale-up networks in combination, at data center scale, to capture:

- The **bandwidth ratio** between scale-up and scale-out paths
- The **latency sensitivity** of key collectives and inference steps or workloads
- The **topology options** for each network fabric, including oversubscription, re-order capable multi-pathing, rail optimized bandwidth, and flatter topologies at lower per MAC speeds

Critical Simulation Choices

Simulation enables exploration and evaluation of key architecture and deployment decisions:

Transport Layer Options

- Traditional **RoCE** with strict packet ordering
- **UET (Ultra Ethernet Transport)** with spray-based load balancing
- **Encapsulated Load/Store** protocols for scale-up fabrics

Topology Design

- Non-blocking vs. oversubscribed Clos designs
- One-hop Rail-optimized scale-up interconnects
- Use of highest possible MAC layer speed vs. lane-optimized throughput

Map Sharding Dimensions to Networks

- Use Scale-up for Latency-sensitive parallelization dimensions (e.g., tensor/model parallelism)
- Use Scale-Out for All-to-all or reduction collectives

Scale-Up Technology Selection

- **NVLink** for scales up to a rack or a few racks, but limited to a single vendor
- **UAlink** for flit-level scheduling with high predictability
- **UAlink mapped to Ethernet** as proposed by Broadcom SUE or AMD Infinity over Ethernet

Each choice affects latency, throughput, congestion response, and collective performance in different ways. Simulation is required to validate the impact at scale with LLM workloads.

Effective Scale-Up Simulation

To simulate scale-up effectively for AI workloads, platforms must support:

- **Low-latency transport modeling** (RoCE, UET, NVLink, UAlink)
    - In some cases, approximations or variations of these transports may be sufficient
- **Lossless Ethernet features**, such as PFC, ECN, buffer tuning, and headroom sizing
- **PCIe and GPU fabric topologies**, including NUMA effects and link contention
- **Packet-level collective communication patterns**, with fine-grained latency modeling

- **Scalable workload integration** using real-world traces (e.g., Chakra from MLCommons)

The Scala Compute Platform (SCP) is focused on providing these features, at data center scale, allowing researchers, operators, and vendors to simulate their real-world network behavior with high fidelity.

Why Scale-Up matters

Infrastructure Optimization

Operators, System Designers, and Silicon Designers can increase GPU utilization and reduce costs associated with purchasing GPUs and Networking equipment. They can design solutions that allow Generative AI infrastructure to operate efficiently for both Training and Inference purposes without the need for artificial infrastructure partitioning.

Protocol and Device Simplification

Simulations reveal which protocol and device features truly impact performance, and which can be eliminated, enabling simpler, more interoperable networks. Fewer required features translated into a wider and more flexible selection of vendors and reduced CAPEX and OpEX.

Clear Fabric Tradeoffs

Simulation results improve understanding of the tradeoffs and critical technology choices associated with emerging non-Ethernet protocols, and quantify the benefits. High-frequency interactions that require challenging speed and latency performance may be worth the cost of scale-up protocols and network components (e.g. GPUs), while independent parallel parts may be best handled by cost-effective and flexible scale-out network fabric topologies.

Workload Optimization

Simulation results can guide smarter workload optimization and improvements to collective libraries and workload distribution strategies, reducing incast, improving convergence and boosting utilization.

Simulating for the Future of AI

Being a leader in AI infrastructure hinges on the ability to simulate scale-up behavior. With billions of dollars being spent on developing the algorithms and hardware and assembling and operating the datacenters, the need to optimize is critical. Without modeling Scale-Up fabrics, like UAlink, on the Scala Compute Platform, engineers risk overspending for AI that will underperform.

Simulation empowers operators, system architects, and silicon designers to virtually explore performance outcomes with high fidelity before investing in expensive hardware. AI Engineers can improve their models to better utilize the networking hardware they will run on.

Scala Computing's simulation platform offers the fidelity, flexibility, and AI workload integration needed to guide your next generation infrastructure decisions.

About Scala Computing

Scala Computing provides advanced network simulation for AI-centric workloads at data center scale. Our Scala Compute Platform (SCP) enables accurate modeling of LLM training and inference workloads across transport protocols, topologies, device behavior, and scale-up fabrics, empowering infrastructure teams to design with confidence.