

The Importance of Simulating Large Language Model Workloads with Network Simulations

The performance of LLMs has a strong dependency on network behavior. Simulation can guide network engineers in optimizing routing and transport protocols, bandwidth allocation and other critical network parameters for LLM traffic.

In the rapidly evolving landscape of artificial intelligence, large language models (LLMs) are at the forefront of transformative technologies. Their applications span from natural language processing to real-time decision-making systems, necessitating robust infrastructure to support their immense computational and networking demands. Simulating LLM workloads within network simulations has become a vital step in ensuring the efficiency, scalability, and reliability of these systems. Scala Computing has been working with Intel to develop a cutting-edge LLM solution, leveraging MPI-enhanced ns-3 for scalable network simulations, that is uniquely positioned to address this challenge

Scala's advanced orchestration efficiently provides the capacity to simulate the network behavior of the MoE models, with their large base sizes, high communication and memory demands. Customers can build their own LLMs, or import from vendors in the Consortium ecosystem. For their latest release, Scala is also including their own flexible ScalaLLM model in the product.

Simulating LLM workloads with network simulations brings transformative value to the networking industry by addressing the intricate interplay between computation and networking. LLMs, due to their vast size and computational intensity, depend heavily on robust, scalable network infrastructures. Network simulations provide a

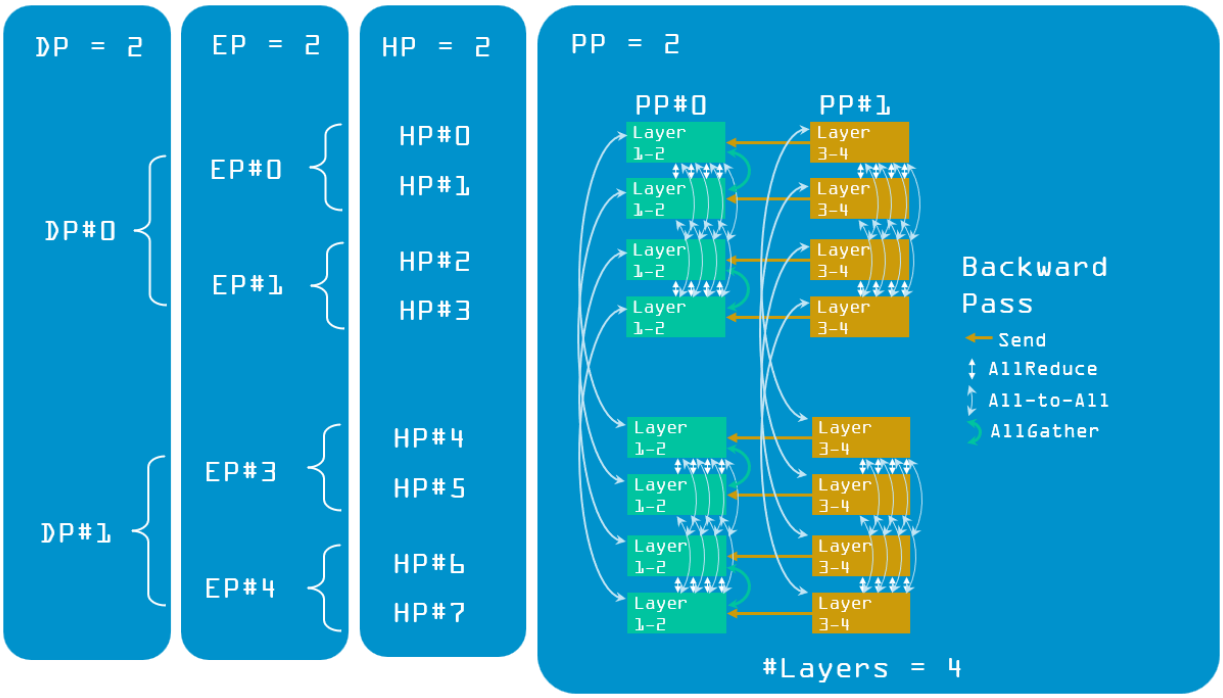
virtual environment to model, analyze, and optimize the networking aspects of LLM deployments. By simulating network behavior, researchers and engineers can focus on;

1. **Infrastructure Optimization:** Network simulations identify bottlenecks, improve data routing, and optimize resource allocation, leading to more efficient utilization of bandwidth and computing resources.
2. **Scalability Testing:** Scala Computing's networking simulations can scale to test scenarios with 64k+ nodes, providing insights that are prohibitively expensive or impractical to gather in live settings.
3. **Realism and Predictive Insights:** Incorporating coordinator models based on advanced LLM traffic algorithms, Scala enables the simulations to replicate realistic LLM traffic patterns, making the insights highly applicable to live deployments. *Our current development focus to support Chakra network traces, extracted from real GPUs, will only increase this realism.*
4. **Reliability and Innovation:** Simulations enable the testing of failover strategies, enhancing system reliability, while also shortening development cycles by addressing issues pre-deployment.
5. **Cost Efficiency:** By identifying inefficiencies early, organizations save costs on network and compute resources, while improving the speed of delivering AI solutions to market.

Scala Computing’s solution represents a paradigm shift in LLM and network simulation. Only Scala Computing is able to run the advanced models at this unparalleled scalability, supporting simulations of 64k+ nodes. This is critical for simulating modern AI workloads, which typically operate across vast networks.

The ScalaLLM workload

The ScalaLLM workload model, developed by Dr. Serhat Arslan at Intel, makes it easy to model the communication patterns generated by the mixture of forward and reverse passes used by advanced Hybrid Parallel LLM algorithms.



This hybrid parallelism LMM model allows the user to define a number of settings for the traffic coordinator. These include settings such as the Number of individual units of text that the model processes, the time to compute each state of the model layer and the 4 types of parallel activities. The parallel activities used in the algorithm are Data Parallelism (DP), Expert Parallelism (EP), Horizontal Parallelism (HP) and Pipeline parallelism (PP).

Running these simulations on various topologies can be used to validate their scalability or to demonstrate strong and weak scaling capabilities. They can prove optimizations delivering significant speedups as model and cluster sizes increase. They can be used to evaluate the network impact of algorithm memory and latency improvements. The ease of use and flexibility of the Scala Computing solution allows for research and deployment of increasingly sophisticated and effective systems. This all aligns with networking industry goals of reducing AI workload bottlenecks and optimizing traffic flows for massive networks, enhancing their ability to support next-generation AI systems.

The network charts, generated in real-time during the simulation, show the impact of the LLM algorithm traffic flowing between layers and make it easy to locate ECMP hotspots where uneven traffic distribution leads to congestion on specific network paths. This information can be used to test various topologies and

settings to locate and mitigate the Elephant Flows that are typical with AI/ML workloads. There are performance tables and charts for looking at the ideal times or for understanding the latency of each activity in the LLM. Scala Computing provides congestion charts to help you figure out exactly when the PFC storm begins so you can mitigate the issues. Buffer management solutions can also be simulated and compared to study their effectiveness.

Understanding the Complexity of LLM Workloads

LLMs are characterized by their substantial size, both in terms of parameters and data requirements. These models involve intricate computations that demand high-performance hardware, such as GPUs and TPUs, often distributed across large-scale data centers. Beyond computational resources, LLMs require robust networking to facilitate distributed training and inference, including seamless communication across nodes and efficient data transfer.

Traditional benchmarking approaches focus solely on computational metrics, often overlooking the critical role of networking in LLM performance. The interplay between compute and network resources is complex; even slight inefficiencies in data transmission or synchronization can significantly hinder overall performance. Given the high cost of GPU compute, failing to maximize efficiency translates directly into wasted resources and increased expenses. This underscores the need for holistic simulations that combine LLM workload characteristics with realistic network conditions.

Looking Ahead

As the demand for LLMs continues to grow, the importance of simulating their workloads in conjunction with network simulations will only increase. Scala Computing's innovative framework, powered by MPI enhanced ns-3, and utilizing flexible LLM workload models, positions it as a leader in this critical domain. We are currently expanding our LLM support to import real-world captured LLM workloads (Chakra traces) to be used in network simulations. We are also focused on scale-up and scale-out for reduced latency, high scalability and reduced cost. By providing the tools to simulate and optimize large-scale systems, Scala Computing is empowering organizations to harness the full potential of LLMs while ensuring efficient, reliable, and scalable infrastructure.

Combining LLM workload simulations with network simulations is no longer a luxury but a necessity. It bridges the gap between theoretical performance and practical deployment, enabling AI systems to operate seamlessly in an increasingly interconnected world. Scala Computing's solution exemplifies the synergy between cutting-edge simulation technology and real-world applicability, setting a benchmark for the future of AI infrastructure.

AI's growing demands are driving innovation and operational efficiency and Simulation with Scala Computing is there to help make it make sense.