

The Strategic Importance of Chakra Workloads in Modern Network Simulation

Introduction

Artificial intelligence and machine learning workloads increasingly dominate data center traffic. The ability to accurately simulate and optimize network performance for these specialized workloads has become crucial for success in the industry.

Chakra workloads are directed acyclic graphs (DAGs) that represent the complex communication patterns of machine learning tasks across distributed computing environments. Unlike traditional simulation approaches, Chakra provides a detailed model of actual computation and communication patterns captured in real-world AI training environments. These Chakra traces have become the industries best method of representing real-world large language model (LLM) workloads for network simulations for high-performance computing applications.

Scala Computing is uniquely positioned to enable the advantages of Chakra for network simulations at data center scale. Scala is working with MLCommons, an open engineering consortium dedicated to improving machine learning (ML), to scale up Chakra for real world applications. By leveraging these traces in their simulations, network engineers can better understand how AI-driven applications interact with network infrastructures and optimize both the workload fit and the network resource allocation, congestion control, and data routing strategies.

Scala computing has already done the work of setting up a platform that can do the heavy lifting needed to simulate with Chakra workloads. This includes resource capacity for the memory requirements and CPU processing required for large scale-simulations with thousands of concurrent and interdependent traces. Scala also provides for the storage requirements for the chakra trace files and has

developed advanced cloud resource orchestration methods for simulating the heavy inter-node communication that grows exponentially as the Chakras are scaled up.

Motivation

Running benchmarks with full LLM workloads in the real world is expensive. There are only a handful of companies that have large scale physical labs with the thousands of high-end GPUs necessary for the task. Even within those companies, it is difficult for engineers to get large tests authorized. Companies like Open AI and Google invest millions in GPU time for optimizing architectures and training methods. Energy consumption alone is estimated to be more than the annual needs of 120 homes. The follow-up steps to fine-tune the workloads or optimize different parts of the stack require even more GPU hours.

Using simulation with Chakra to represent LLM workloads gives engineers a much more affordable method for optimizing the workload and network stack.

Key Benefits of Chakra Workload Simulation

Enhanced Realism

Traditional network simulations frequently rely on synthetic or overly simplistic workload models, which often fail to capture the nuanced complexity and dynamism of AI-driven applications, such as LLMs. As AI workloads increasingly dominate data center and cloud computing traffic, network research must account for their unique characteristics, including highly variable traffic patterns, burstiness, latency sensitivity, and high-bandwidth demands.

Chakra traces capture authentic machine learning communication patterns by recording actual GPU communications during training. Chakra workloads provide network architects with real-world traffic patterns rather than over-simplified approximations. Real AI workload characteristics, including compute cycles, memory usage, and communication patterns enable more accurate network simulations for network design and optimization.

Users can upload their own Chakra workloads, or they can use vendor captured or synthetic Chakras. The validated Chakra workloads are provided by industry leading hardware vendors and Hyperscalers so network architects can optimize their networks for the real world.

The network topology can be set up, in detail and at scale, with vendor validated device models that accurately represent the real networks devices, to match the real-world network being simulated. And then a mapping file can be used to map the chakra traces to their correct NPUs within that simulated network topology.

Granular Analysis

The Chakra traces capture the workload operations at every GPU in the network. Send/receive operations from every single communication node are explicitly included. A variety of Collective operations (All-to-All, All Reduce, Broadcast, etc.) are used, along with ring, tree, double tree, direct, sequential, interleaved, etc. These are all used in the simulation and then each Network Processing Unit (NPU) in the simulation can be analyzed, at that same granularity, for individual node performance, inter-node communication patterns, resource utilization and identifying Bottlenecks.

Various chakra files, network topologies or mapping strategies can be compared so researchers can deeply understand and optimize their networks.

Engineering Applications

Network architects using The Scala Computing platform can use Chakra workloads to validate their network topology designs for AI workloads. They can optimize buffer management strategies and Identify potential congestion points. They can scale up and scale out the simulation using elastic cloud hardware to test scalability at data center scale under realistic ML workloads.

> A key strength of data center networking research has been the reproducibility of work, enabling verification and further innovation within the community. Many hyper-scalers have open-sourced workload benchmarks, such as message size distributions, allowing researchers and vendors like Intel to

> evaluate their designs. However, these benchmarks no longer reflect the dominant workloads, particularly AI, where message sizes are significantly larger, more consistent, and require greater node coordination. A standardized, lightweight tracing method offers an ideal solution for representing communication patterns in these high-demand workloads. Chakra remains a strong candidate for this approach.
>
> > ~ Dr. Serhat Arslan, Intel Senior AI Network and Transport Layer Modeling Engineer

Scala Computing simulation, with Chakra, enables organizations to Fine-tune network parameters for specific AI/ML workflows and evaluate the impact of different network architectures, test failover strategies under realistic conditions and optimize collective communication patterns

Simulation allows users to understand how their workloads interact with their network, protocol and congestion algorithms in order to minimize execution times and identify blockages. By simulating their AI/LLM workloads over the network, engineers can identify sub-optimized workload coupling and then engineer the workload or the network or both to optimize efficiency and performance. Simulation allows engineers to simulate the effect of many parameters at protocol level, NICs and in network switches. ECN thresholds, properly adjusted by system/network engineers, can have a significant impact on the overall performance of the congestion control algorithms, and hence the completion time of the workload, across the network.

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.

Synthetic Chakra

Scala's ability to work with Chakras also opens up the use of Synthetic Chakras. Scala computing is working with MLCommons and other partners to provide clean and controlled synthetic Chakras that can be generated quickly and affordably, without limiting users to the real-world traffic and hardware configurations used during the real-world Chakra capture. Synthetic Chakras allow users to investigate what-if scenarios with a variety of workload designs without needing access to thousands of expensive GPUs. Workloads can be synthesized for any size network and the parameters can be adjusted to represent a broad range of AI workloads (GPT-2, LLaMA, Mistral 7B) or scale-up architectures (UALink, NVLink, CXL). These synthetic chakras can be validated against real-world chakras tested in the lab and then scaled up to represent a larger scale deployments that are too large to emulate. Engineers can adjust the workload characteristics, such as the communication patterns, to improve the coupling with the network.

For many users, an important advantage of Synthetic Chakra is that they allow investigating and comparing simulation results of different LLM workloads with other parties without exposing their own proprietary AI models.

Researchers at many Universities are involved in this project thru MLCommons and are focused on developing accurate synthetic traces that mimic real AI training/inference patterns.

The expert consensus is that the best approach is to run both types of simulations. Real captured traces can be used for studying real-world behavior in production environments or to validate that the synthetic traces match that real world behavior. There are also situations where the clear, concise and customizable communication patterns of the synthetic Chakras make it easier to isolate specific bottlenecks. Then network architects and engineers can leverage synthetic traces for scalability, flexibility and to simulate hypothetical future AI network behaviors.

Future Implications

Scala Computing's support for Chakra workloads as part of their network simulation solution represents a significant step forward in preparing network infrastructure for the demands of AI/ML applications. As these workloads become

more sophisticated, and the cost per training has grown into the tens of millions of dollars, the ability to accurately simulate their behavior will become increasingly critical for:

- Next-generation network architecture design
- AI infrastructure optimization
- Cost-effective capacity planning
- Performance prediction and optimization

Conclusion

Chakra workloads, as implemented by Scala Computing, provide an essential tool for understanding and optimizing network performance in the age of AI/ML. The combination of an accurate model of a customer's real-world workload (through Chakra) and an accurate representation of their network, using vendor validated device models, enables network engineers to identify areas of sub-optimized coupling between the workload and the network. This opens the possibility of engineering the workloads specifically to couple better with the network or configuring the network to optimize the performance of the workloads.

By enabling accurate simulation of complex distributed training patterns, Scala Computing is working to help our partners gain deeper insights into network performance and develop more efficient, scalable solutions for modern networking challenges. The cost of training AI models can be in the tens of millions of dollars, so even small performance improvements in ideal ratios can translate to significant recurring savings. This approach is essential for ensuring that next-generation network architectures can support the growing demands of AI and high-performance computing applications