SCALA COMPUTING

Accelerating Distributed Machine Learning Training

Industry Challenges

- With the exploding growth in data, ML Training must be more frequent and needs hyperscale datacenters to reduce time-to-results and maintain accuracy.
- Deploying distributed ML Training on thousands of servers and networks at scale is complex, iterative, expensive, and time-consuming.

Scala Computing Benefits

- Users can cost-effectively emulate and simulate distributed ML Training cluster network performance before deploying elaborate infrastructure.
- Networking providers can reduce time-to-market and deliver the best operating cluster network performance for complex ML workflows.

Deploying distributed Machine Learning (ML) is challenging

ML, a major part of Artificial Intelligence (AI), is rapidly evolving and significantly improving growth, profits, and operational efficiencies in almost every industry. However, with the rising volume, velocity, and variety of data, ML models are getting larger and more complex and must be trained more frequently. So, to reduce time-to-results, distributed ML training across thousands of nodes is typical in cloud computing and social media companies with massive, hyperscale datacenters.

However, deploying these datacenters is expensive, arduous, and fraught with trial and error, and unforeseen changes can have disastrous ripple effects on the business. Moreover, it is tough to forecast ML production cluster network performance at scale accurately, and current approaches to building small physical labs or running sub-scale simulations just don't cut it.

What's needed are accurate predictive simulation solutions that enable organizations to understand the impact of different strategies, topologies, and hardware on ML cluster network performance. Scala Compute Platform (SCP) delivers unique, heterogenous, cloud-based solutions at scale to accelerate distributed ML training with emulation and simulation.

Distributed ML training is a complex, iterative, and time-consuming process that must be frequently repeated

ML trains computers to do what is natural for humans: learn from experience. These algorithms learn directly from data to build the Trained Model (say a Neural Network), whose accuracy improves as the number of data samples available for Training increases. This Trained Model can be used to make Inferences on new data sets.

Training a model with a billion parameters can take days/weeks unless properly optimized and scaled. To complete Training in a few hours, one typically needs a high-performance infrastructure and scalable data and model parallel algorithms that distribute the ML computational kernels over thousands of processors.

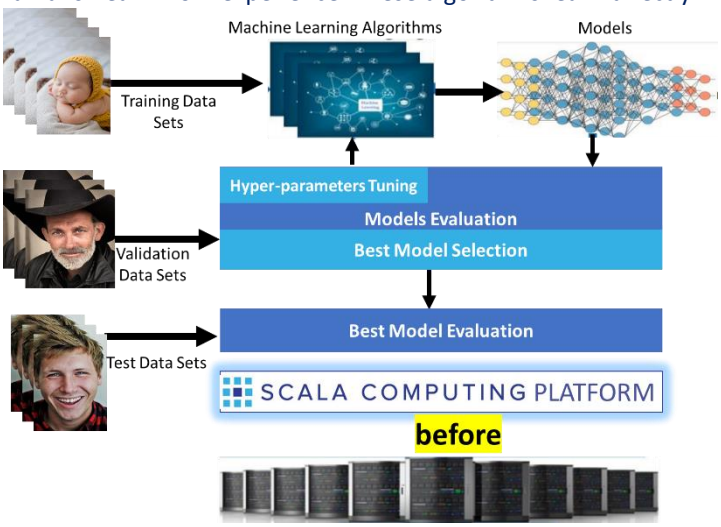

Figure 1: High-level Training process for a facial recognition workflow

Organizations must repeat the Training process (Fig.1) to experiment with different topologies, networks, algorithms, and hyper-parameters to reach the desired accuracy level and to reduce network communication overheads. SCP's network simulator enables users to experiment virtually at scale **before** physically deploying expensive high-performance infrastructure.

Unmatched heterogenous cloud-based solutions to accelerate distributed ML at scale

SCP is the only high-performance solution to run highly accurate discrete event-driven network simulations at scale for datacenter operators and networking equipment manufacturers. Its emulation and simulation tools can model large-scale networks that are not practical to build until deployment and help simulate next-generation devices for features like links speed and capacities not yet available. It includes a highly collaborative ecosystem and provides simplified, on-demand access to scalable, secure compute clusters on the cloud.

Data scientists can quickly launch network simulations and run emulated distributed ML Training workloads for 1000s of endpoints and improve performance 10-fold. Users get various insights and analytics to quickly evaluate and optimize several network communications factors that impact ML cluster network performance.

Evaluating network communication factors that impact distributed ML Training workloads

During Training, the vital computational kernels are numerous matrix operations (Fig. 2 shows this for a matrix of size three – a tiny number) throughout the recurring forward and backward propagation steps in the Neural Network. The amount of computation (and network communication) depends on the algorithm and rapidly grows with the size of the input data, the number of layers in the Neural Network, the number of outputs, and compute cluster size. For large-scale models this can be huge.

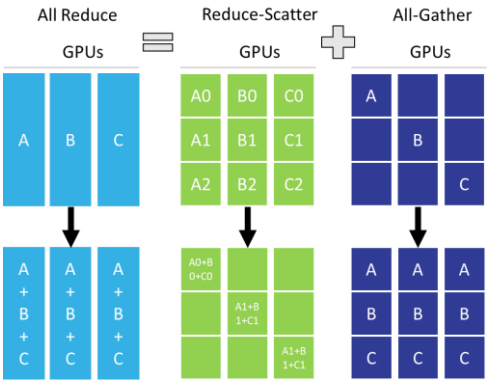

Figure 2: Collective communication pattern for distributed ML

In addition, many factors can significantly impact training time and the datacenter. These include:

1. *Distribution strategy:* ML communications patterns (Fig. 2) with Message Passing Interface (MPI) collectives such as Scatter/Gather, AllGather, AllReduce, AlltoAll, etc. These collectives could use graphics processing units (GPUs) with NVIDIA Collective Communications Libraries (NCCL).
2. *Topology:* number of stages, multipath routing scheme
3. *Placement of workers:* could be bare metal or virtual machines. Optimal placement can reduce congestion points and accelerate ML training.
4. *Network transport:* Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE), Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and so on.
5. *Networking hardware:* Packet switches, cell fabrics, Network Interface Card (NIC), I/O busses, and so on.
6. *Concurrency:* Optimal use of multiple parallel connections or flows to exploit network path diversity, multi-core processing, and latency hiding through pipelining. NCCL Fast Sockets exploit and enable some of this.

How clients use Scala Computing to accelerate distributed ML in their unique environments:

1. Debug/test different distribution strategies in parallel to quickly understand optimal strategies
2. Tweak strategies discovered in Step 1 to take advantage of the network topology or test Step 1 strategies against various topologies. Identify congestion and optimize placement of "workers" to reduce bottlenecks.
3. Understand which protocols/stacks/hardware reduce training times i.e., quickly compare RoCE vs. TCP using the strategies discovered in Step 2.

Client 1: A disruptive AI Chip accelerator venture: Training is done over a dedicated or a converged network with TCP, and switches/routers use IP over Ethernet with shared memory output queued architectures. This system matches merchant market systems, or voice output queuing (VOQ) cell fabric deep buffer systems that fit specialized original equipment manufacturer (OEM) chassis.

Client 2: A leading social media giant: Training is done over a dedicated network, the transport layer is a RoCE RC service, and switches/routers use UDP over IP over Ethernet with shared memory output queued architectures that match merchant market systems, or VOQ cell fabric deep buffer systems that match specialized OEM chassis.

About Scala Computing

Scala Computing is an industry-leading cloud computing software firm that helps organizations deploy, manage, visualize, and optimize complex workloads. Our highly skilled team has received many prestigious awards for working with clients to solve extreme challenges in the world's most scale-intensive and complex environments. The Scala Compute Platform is the industry's first datacenter scale end-to-end, secure, High-Performance Computing network simulation solution that enables enterprises to radically reduce the costs, risks, and time to deploy applications on specialized infrastructure.

