SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Demystifying Data Flows through typical LLM training

Bill Lynn
AMD Fellow

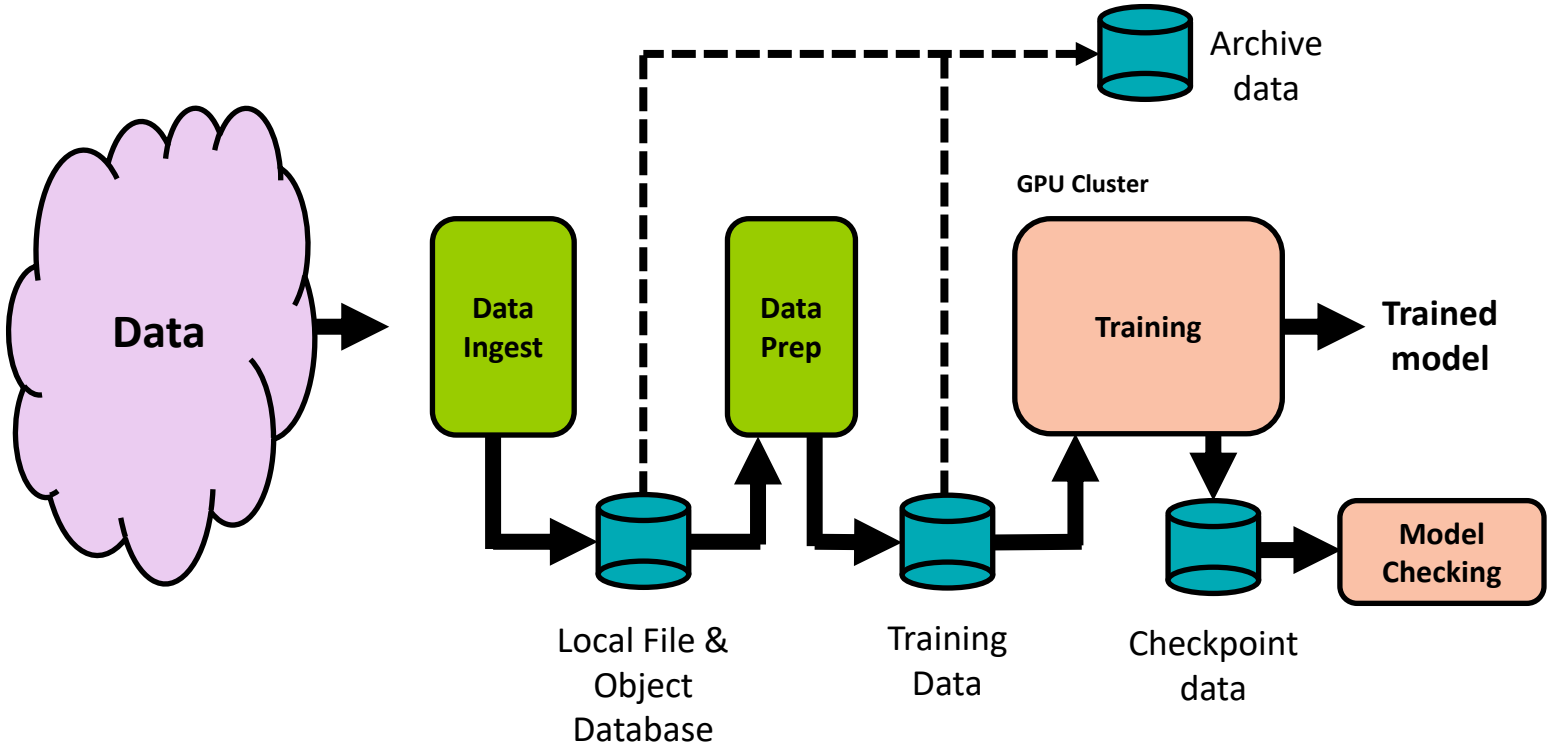
www.sniadeveloper.org

Introduction

Typical macro data flow through LLM training

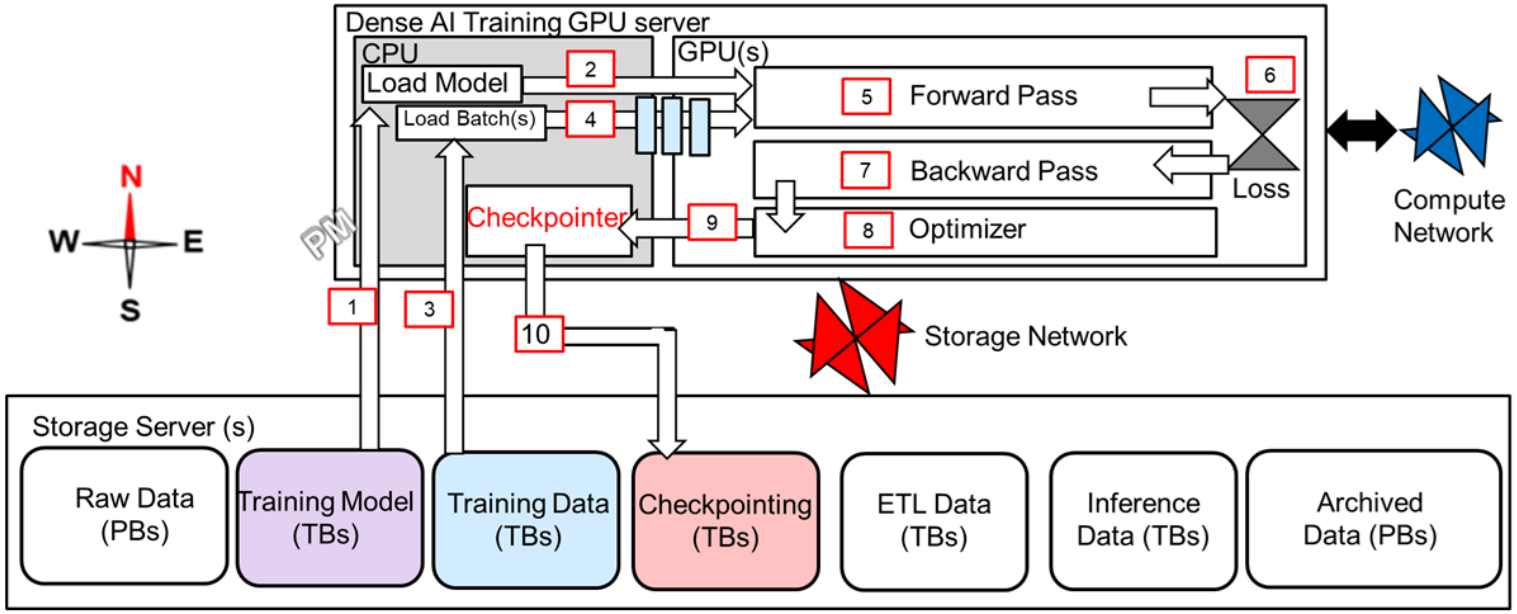
Training of complex AI models require massive amounts of data that can amount to multiple petabytes.

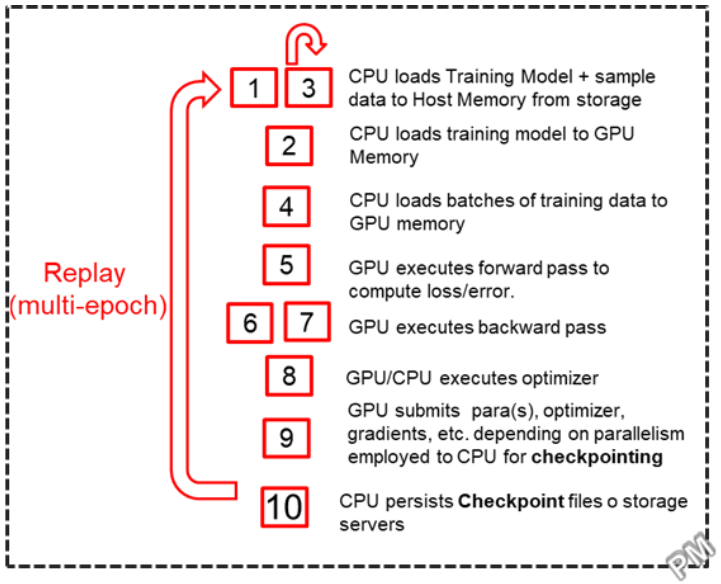
Data must be curated to deduplicate, ensure accuracy and remove any unwanted biases. Data must also be properly formatted (tokenized/tensorized) before it can be used in training



Today we will focus on how data moves through the GPU cluster

Introduction





For Meta's Llama 3 405B pre-training on 16K GPUs, the Model FLOPs Utilization is **~ 41%** [1].
Meta (BF16; 4D parallelism TP=8, CP=1, PP=16, DP =128)

- Extremely high memory req. (TBs) – (a) model states; (b) activations; (c) training data; (d) checkpoints.
- Highly bursty (intense) and periodic (line-rate of host-NICs 400Gbps), [2].
- Both E-W traffic (collectives) and N-S traffic (loads and checkpoints).

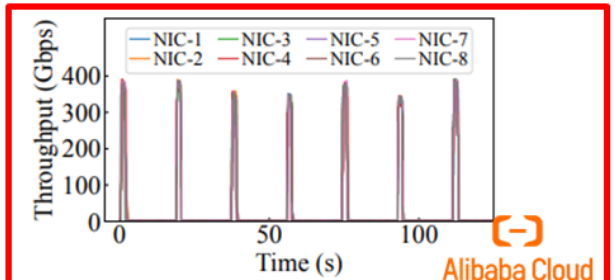


Figure 2: NIC egress traffic pattern during production model training.

[1] Dubey, Abhimanyu, et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783* (2024).
[2] Qian, Kun, et al. "Alibaba hpn: A data center network for large language model training." *Proceedings of the ACM SIGCOMM 2024 Conference*. 2024.

AI Model Fundamentals

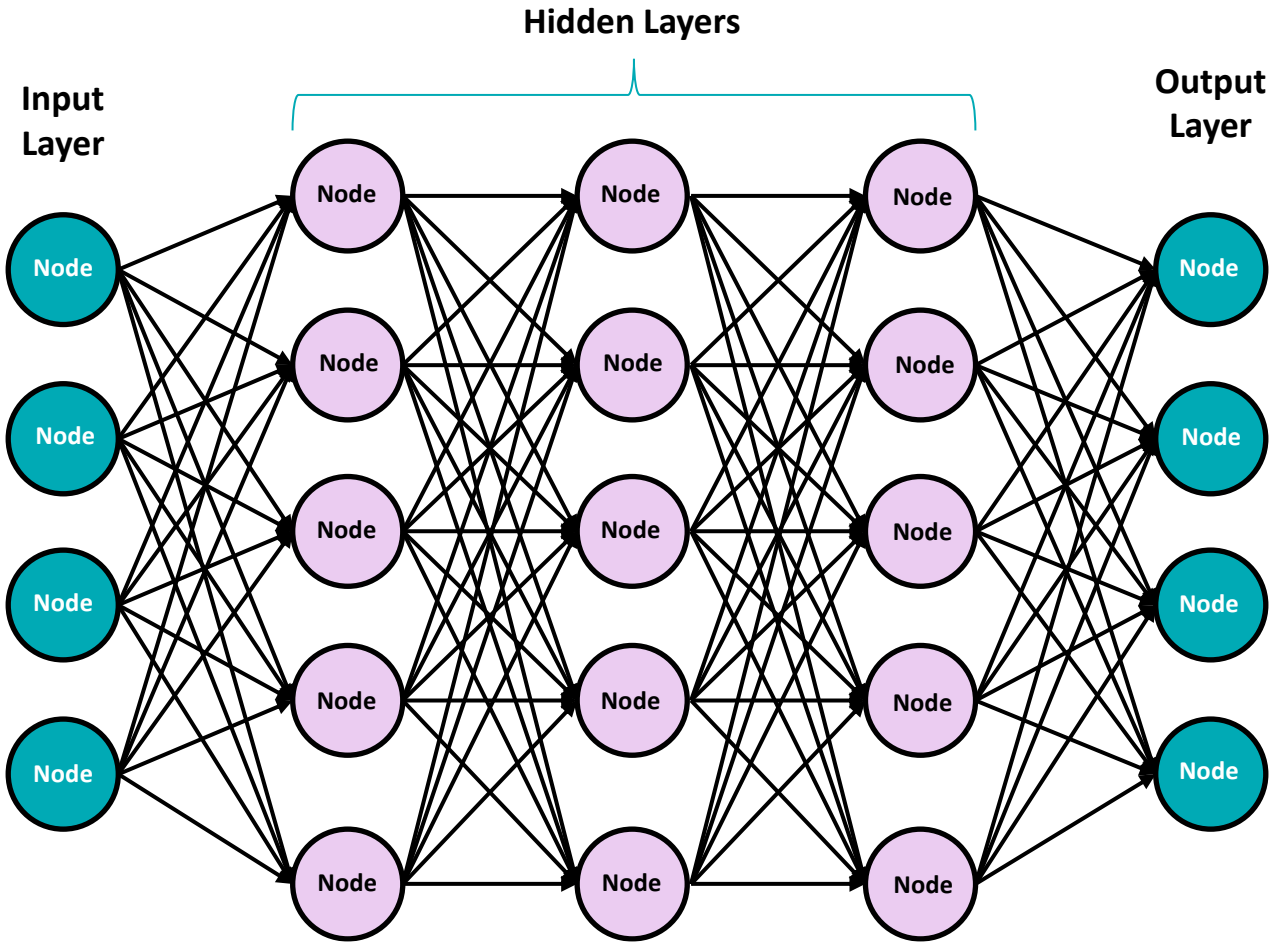
Simple AI models consist of neurons or nodes linked together to form a neural network. In more complex AI models, nodes consist of transformer blocks.

A neural network is made up of

- An input layer
- Some number of hidden layers
- An output layer

For example:

- Llama 3.1 8B - 32 layers
- Llama 3.1 70B - 80 layers
- Llama 3.1 405B - 126 layers
- Chat GPT 4 - 120 layers



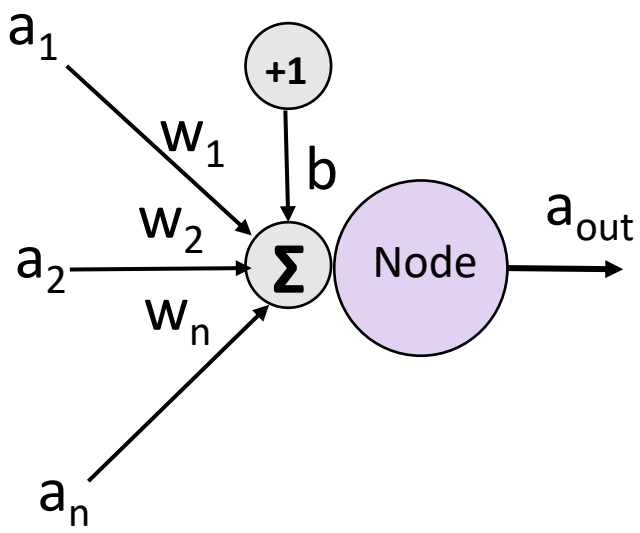
AI Model Fundamentals

Think of a node/neuron as a group of memory locations within the GPU.

The memory locations contain

- Weights for each of the inputs to the neuron
- A bias or offset value
- Values for an activation function

The GPU calculates the weighted sum of all the inputs plus the bias value. The GPU calculates the output of the neuron using the sum as input to an activation function g.



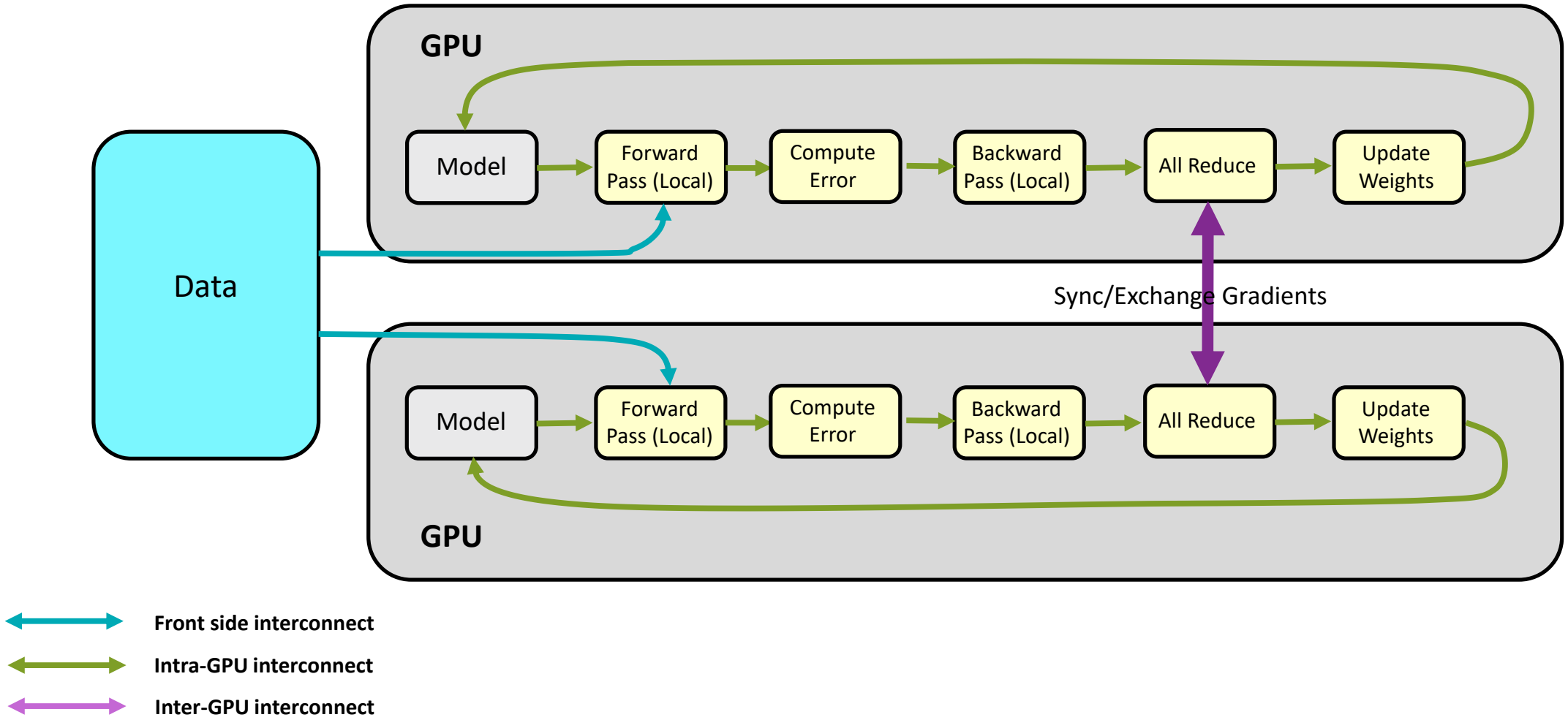
$$a_{out} = g\left(b + \sum_{i=1}^{n} a_i w_i\right)$$

Speeding up Training

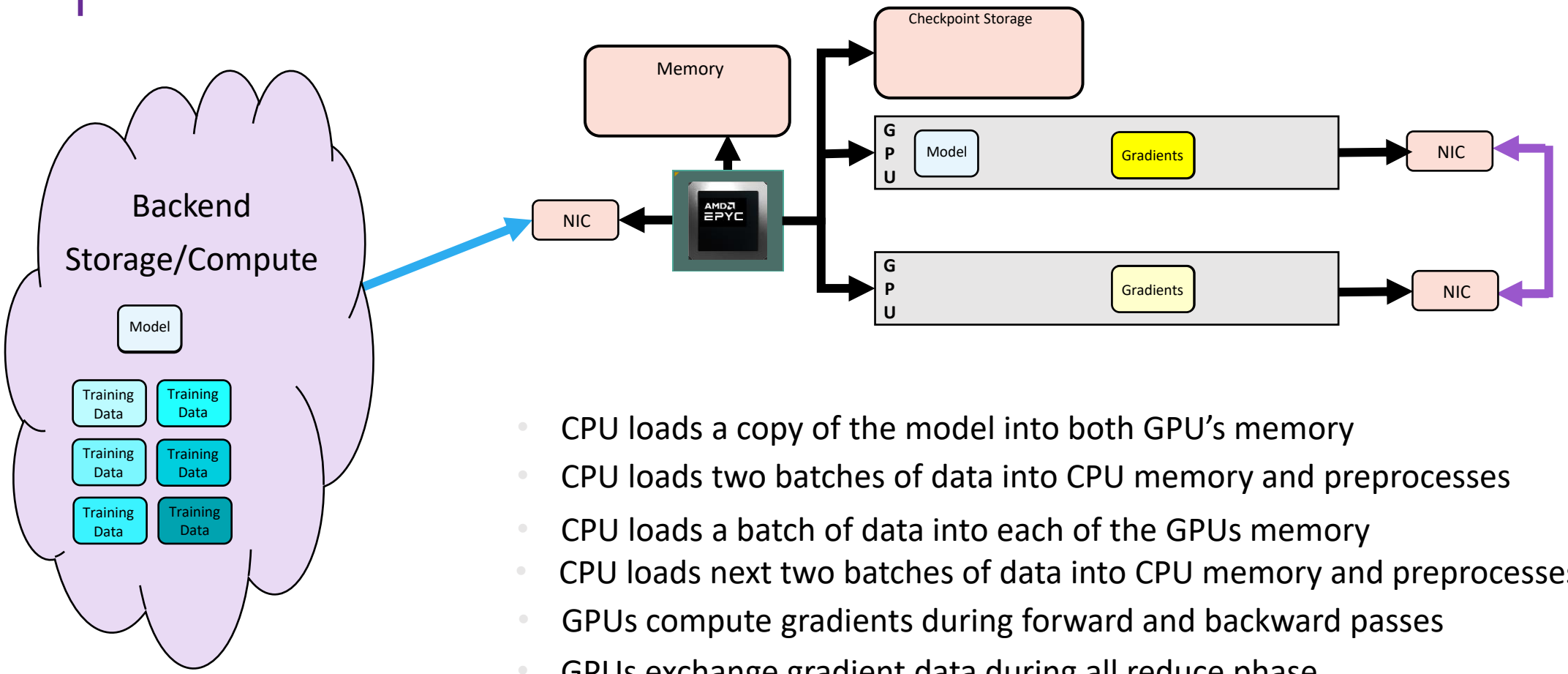
- LLMs require massive amounts of data to train (on the order of multiple PB)
  - Size of data depend on various factors (size of model, data precision, etc)

- The most straight-forward way to speed up training is to split the data between multiple GPUs and process in parallel.
  - LLMs are based primarily on linear algebra which lends itself to parallelization.
  - You do pay a penalty in having to synchronize gradients between passes
  - This assumes model will fit into GPU memory

Simple Data Flow



Simple Data Flow



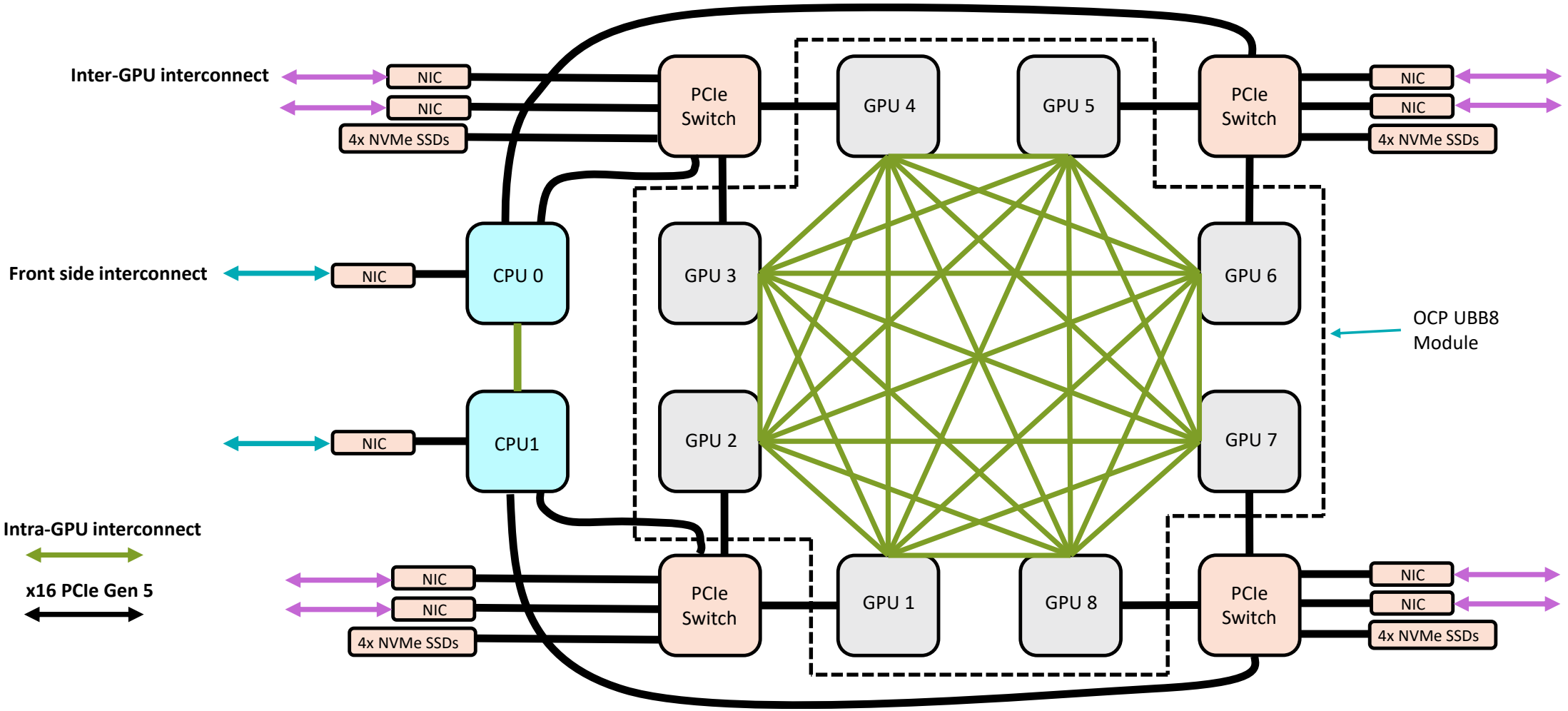
- CPU loads a copy of the model into both GPU's memory
- CPU loads two batches of data into CPU memory and preprocesses
- CPU loads a batch of data into each of the GPUs memory
- CPU loads next two batches of data into CPU memory and preprocesses
- GPUs compute gradients during forward and backward passes
- GPUs exchange gradient data during all reduce phase
- GPU incorporates updated model weights using gradient data
- CPU copies (checkpoints) the current model state into storage
- Repeat until all training data has been consumed

Scaling to Larger Model Sizes

- LLMs require large amounts of GPU memory
  - Amount of memory depend on various factors (size of model, data precision, etc)

- The most straight-forward way to accommodate larger model sizes is to increase GPU memory.
  - One approach is to increase on-chip HBM.
  - Another approach is to connect multiple GPUs together in a memory coherent mesh.

Scaling to Larger Model Sizes



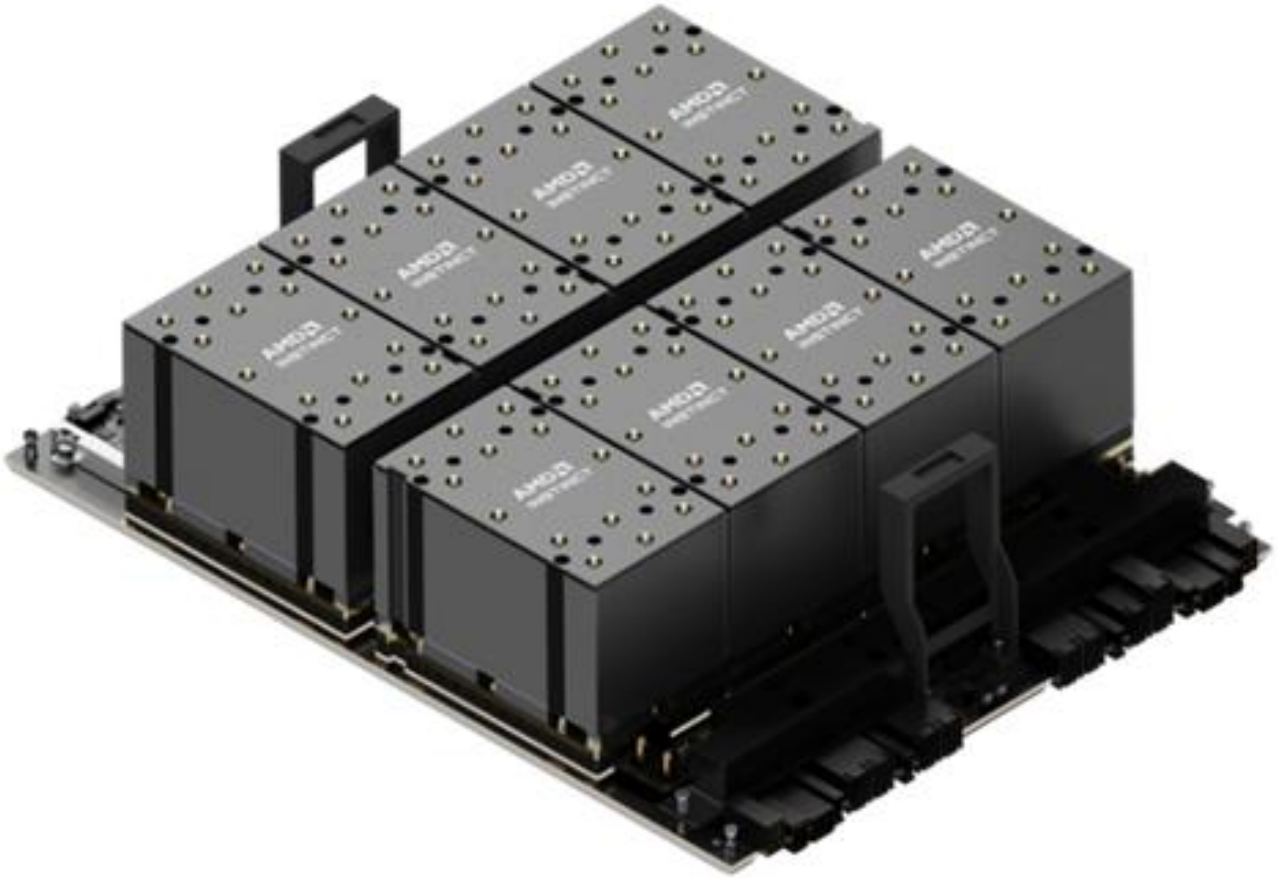
Scaling to Larger Model Sizes

Based on Open Compute Project (OCP) specifications

- Open Accelerator Module (OAM)
- Universal Baseboard (UBB)

UBB module with 8 MI300X OAM GPUs

- Seven x16 Infinity Fabric ports per GPUs (128 GB/s)
  - Scale up connectivity
- Single x16 PCIe Gen 5 port per GPU (128 GB/s)
  - Scale out connectivity
- 1.5 TB of HBM3
  - 192 GB of HBM3 on each OAM
- 5.3 TB/s max memory bandwidth (theoretical)
- 6 KW max power

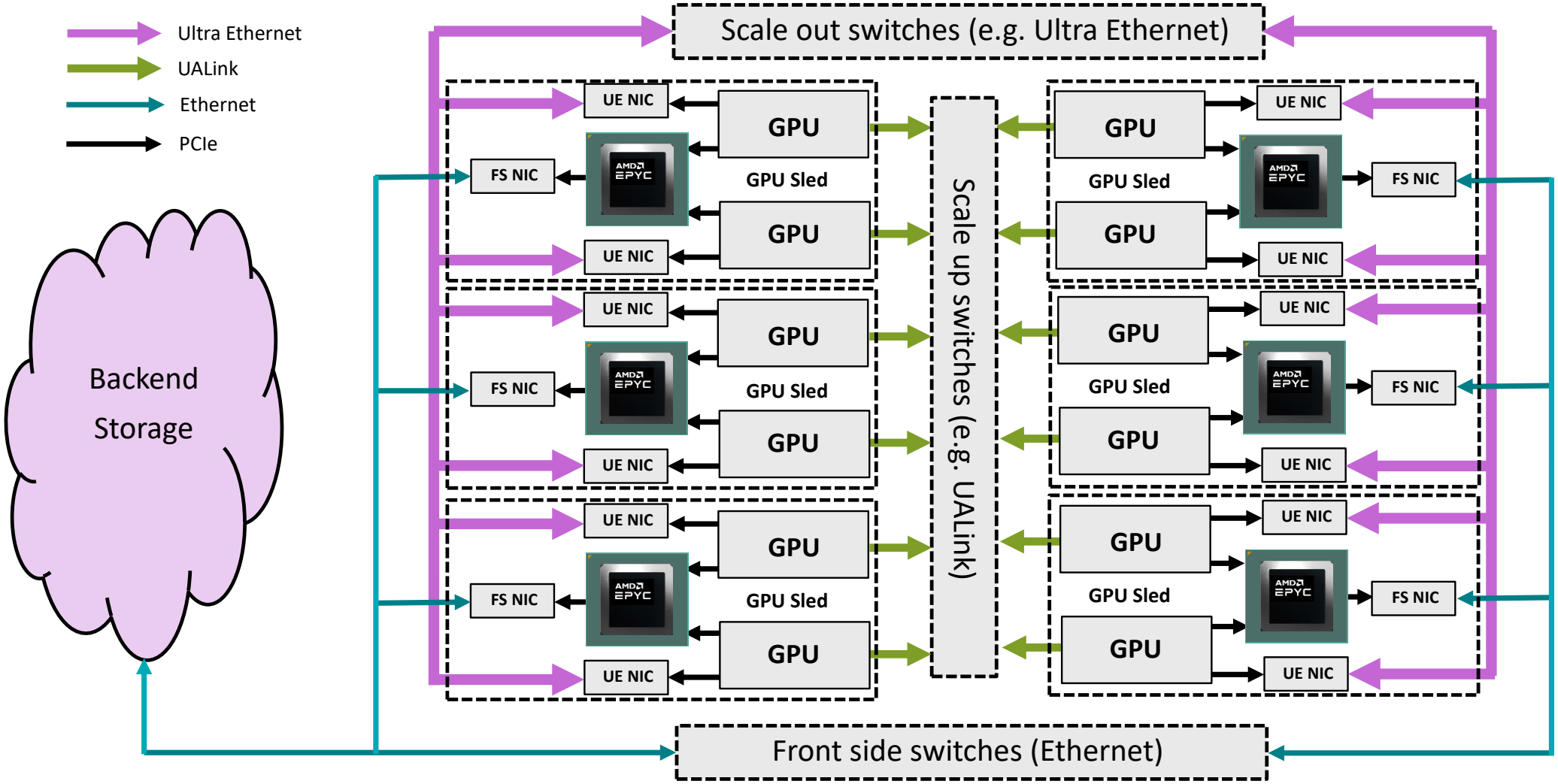


Scaling to Larger Model Sizes

MI300X Industry Standard OCP Server Designs for H100/200 and MI300

Dell PowerEdge XE9680	GIGABYTE G593-ZX1-AAZ1	HPE CRAY SC XD675	Lenovo ThinkSystem SR685a V3 Rack Server	Super Micro AS-8125GS-TNMR2 Server
UBB 2.0 6U	UBB 2.0 5U	UBB 2.0 8U	UBB 2.0 8U	UBB 2.0 8U
Dual CPUs with up to 56 cores per processor	Dual AMD EPYC™ 9004 Series Processors (with AMD 3D V-Cache™ Technology)	2x AMD EPYC up to 400W	2x 4th Gen AMD EPYC™ Processors	Dual AMD EPYC™ 9004 Series Processors
8x AMD Instinct MI300X Accelerators	8x AMD Instinct MI300X Accelerators	8x AMD Instinct MI300X Accelerators	8x AMD Instinct MI300X Accelerators	8x AMD Instinct MI300X Accelerators

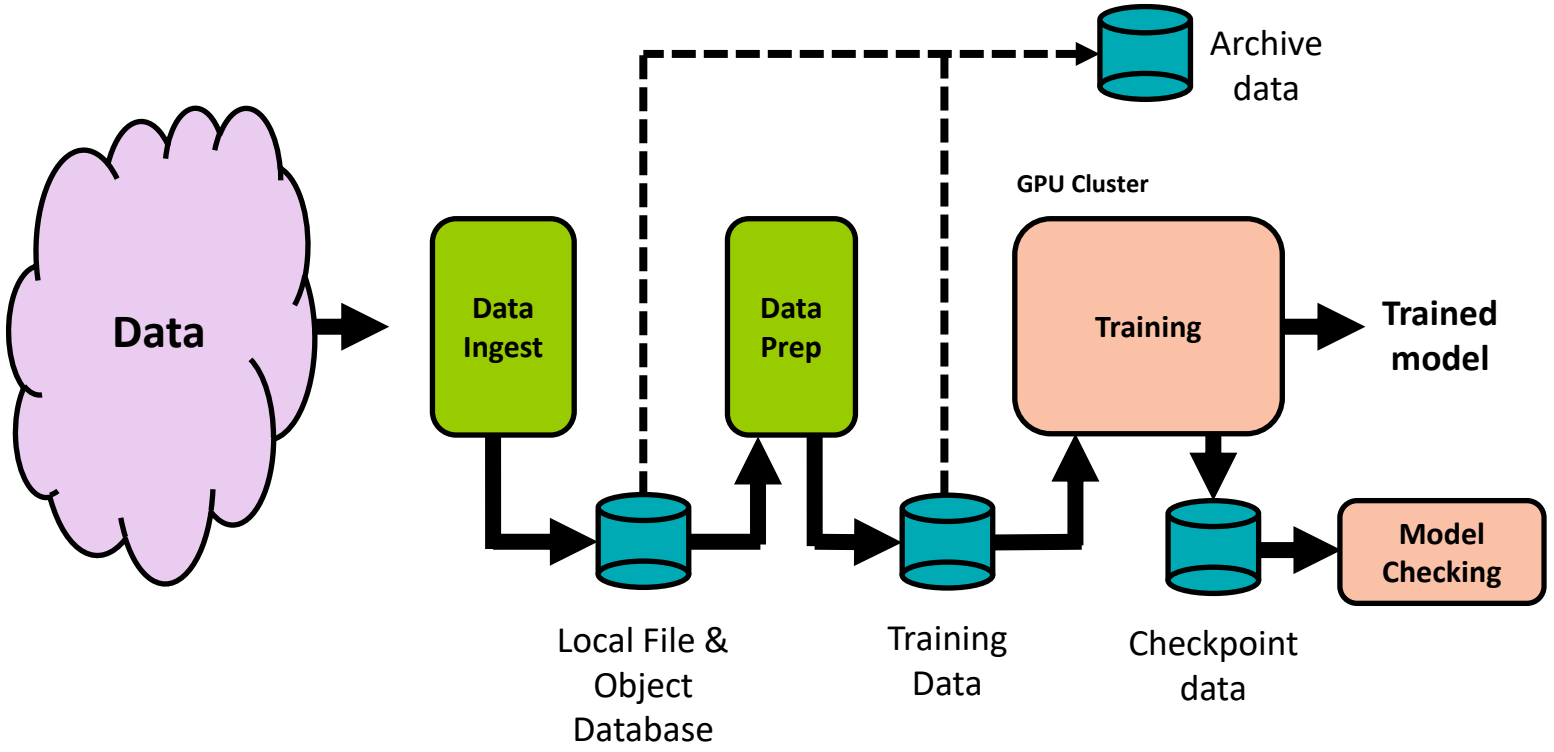
Scaling to the Future



Feeding the Beast

Training of complex AI models require massive amounts of data that can amount to multiple petabytes.

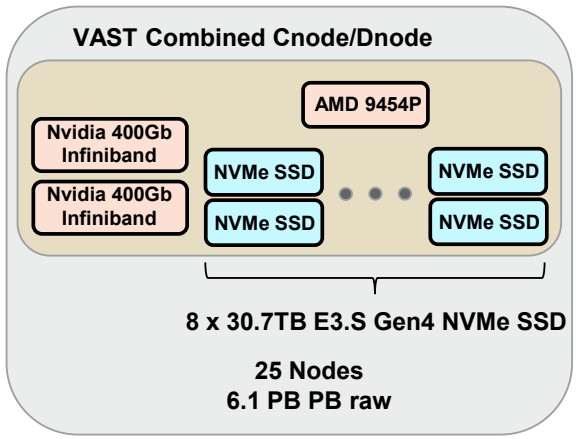
AI factories tend to use multiple storage tiers to supply data to the GPU clusters.



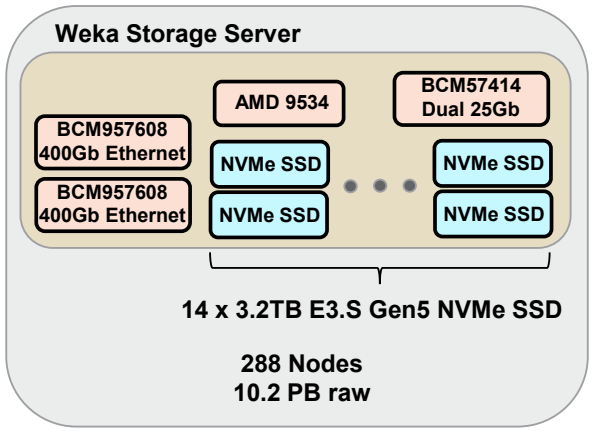
Feeding the Beast

Example 1: 9K GPU Cluster Storage Architecture

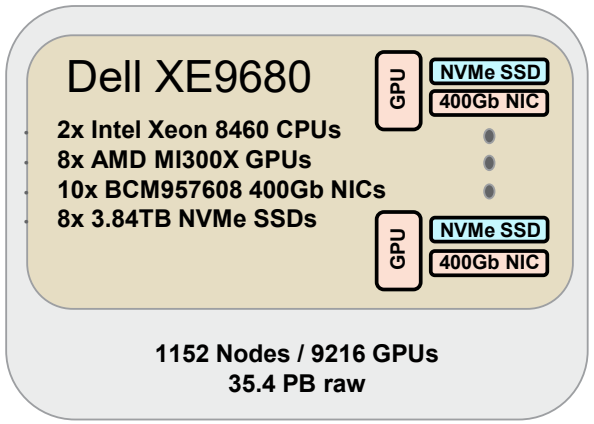
VAST



Weka



GPU Cluster



Tier 2 ingress/egress storage

- Mid-range storage
  - 6.1PB total capacity
- Located in a different building from the GPU cluster
- Used as staging area for customer/training data

Tier 1 scratch storage

- High performance scratch storage
  - 10.2PB total capacity
- Co-located in the same area as the GPU cluster
- Tightly coupled to the GPU cluster

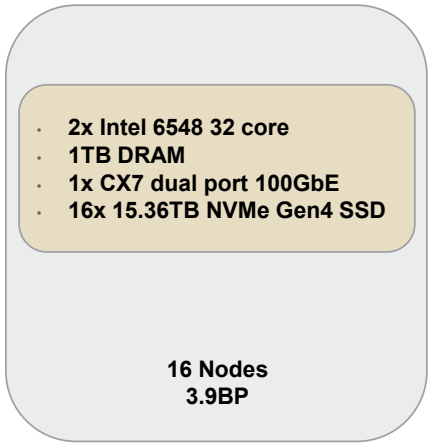
Tier 0 local storage

- Each node has 8 SSDs attached through PCIe switches.

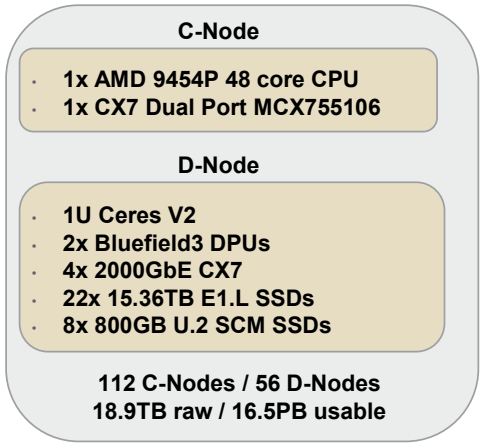
Feeding the Beast

Example 2: 13K GPU Cluster Storage Architecture

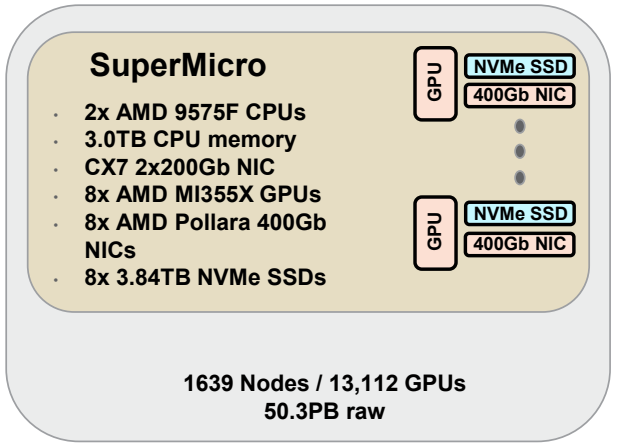
Lightbits



VAST



GPU Cluster



Persistent Disks

- Mid-range storage for OS deployment and customer use.
- Lightbits SDS software layered on top of generic servers.
- Presents block storage via NVMe-oF over TCP/IP.
- Block devices presented as virtio-blk devices to VMs on GPU servers.

Tier 1 Shared Storage

- High performance scratch storage
- Primary storage for customer data sets
- SCM SSDs are high endurance NVMe SSDs used for write buffering and meta-data updates.
- Co-located in the same area as the GPU cluster.
- Presents storage as either file or object to GPU servers.
- Filesystems presented to VMs as virtio-fs via hypervisor.

Tier 0 Ephemeral Disks

- Each node has 8 NVMe SSDs attached through PCIe switches to the EPYC CPU(s).

Conclusions

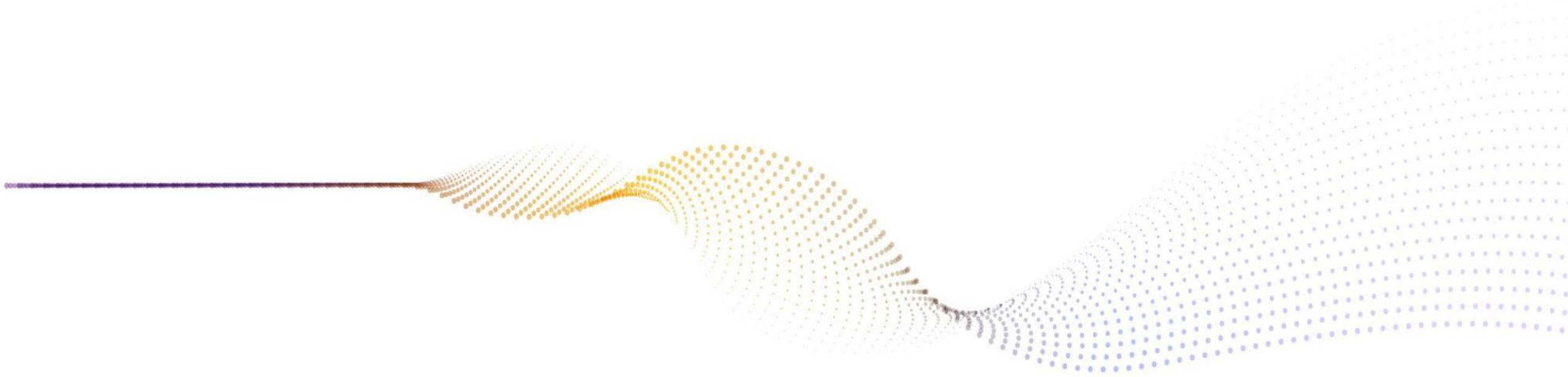
We are just at the beginning of the AI journey

- AI model architectures are evolving
  - LLM (Large Language Models) are getting bigger and better.
  - LBM (Large Behavioral Models) are starting to emerge for robotics.

- AI system architectures are evolving even faster
  - Rack scale architectures are more important than ever

- Multiple standards efforts are ongoing to address AI evolution
  - Ultra Accelerator Link (UALink)
  - Ultra Ethernet Consortium (UEC)
  - Storage Networking Industry Association (SNIA) Storage.AI
  - Open Compute Project (OCP)

Get Involved



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

