GTC 2024 - Getting Storage Right for AI Application

S62476 3/20 2:00

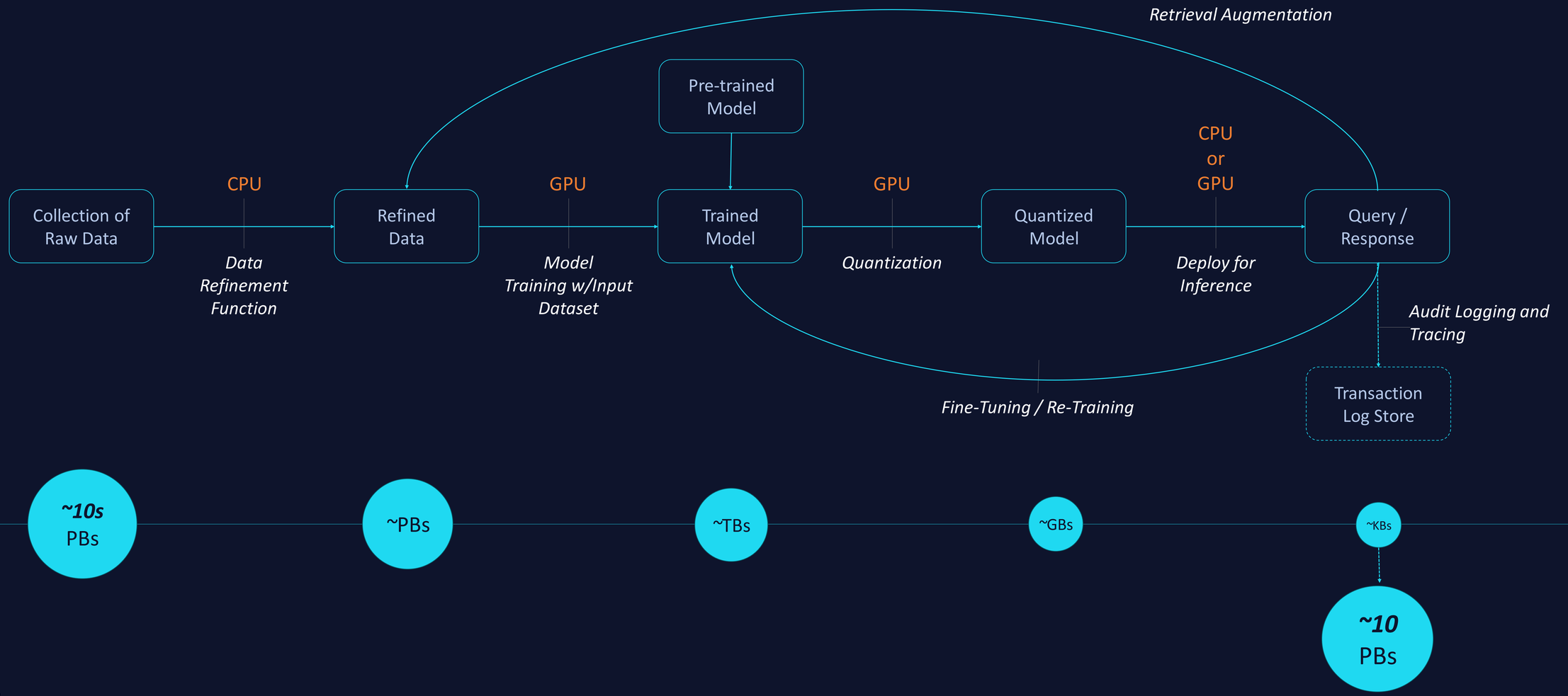
AGENDA

- Introductions
  - Colleen Tartow Ph.D. - Field CTO, VAST Data
  - Pranoop Erasani - Vice President of Engineering, NetApp
  - Sven Oehme - Chief Technology Officer, DDN
  - CJ Newburn - Distinguished Engineer, IO and Security Architect, NVIDIA
  - Jason Duquette - Distinguished Engineer and Chief Platform Architect PowerScale (Isilon) product line, Dell Technologies

- Q/A
  - Panel and audience

Data Requirements for AI Pipelines End-to-End

Storage is key, but only one piece of the story.
How do you account for its place in the AI journey?



AI Storage Challenges

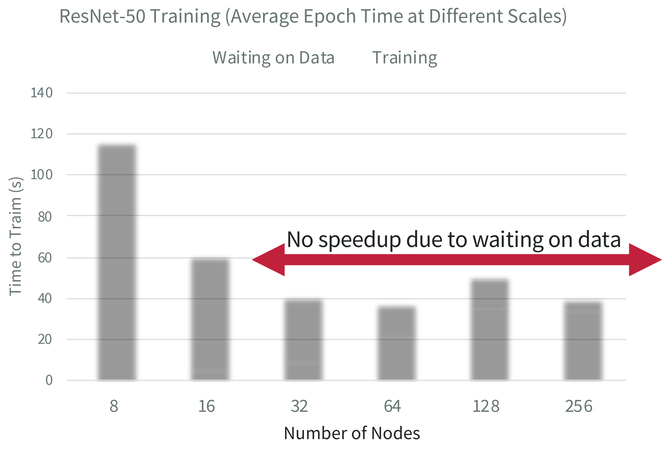
NetApp's perspective

- Diverse performance needs for various phases
  - Data Preparation (Data Lake)
  - Model development/tuning (Training)
  - Model deployment (Inference)
  - Model generation (AI Data Lifecycle)
- Data Lake vs Training Optimized
  - Single system vs Multiple systems
- Need for continuous data movement
  - Data Transformation (Labeling, Compression)
  - Data Copies (Model versioning, Test and Dev)
  - Data Lake (Edge to Core to Cloud)

Training At-Scale – Small Factors Make a Big Difference

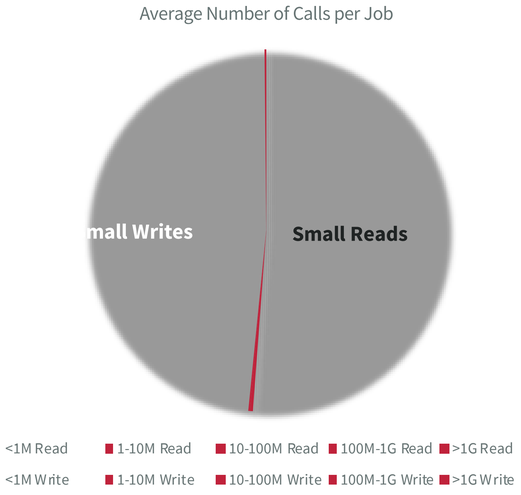
Waiting on Data Can Become a Critical Factor

As models scale across nodes data movement for multi-epoch training can become the dominant factor in training time.



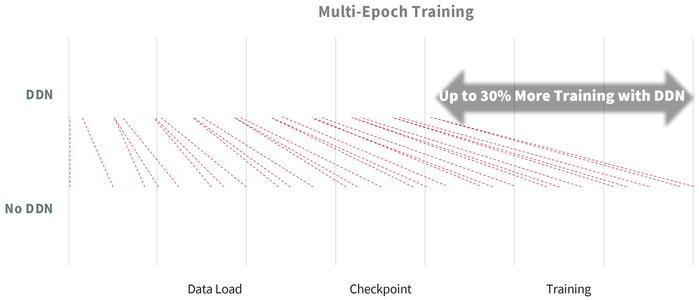
Machine Learning Needs More Than Just Fast Read Access

Research concludes ML uses equal number of reads and writes. NVIDIA sees similar balance on their EOS system.



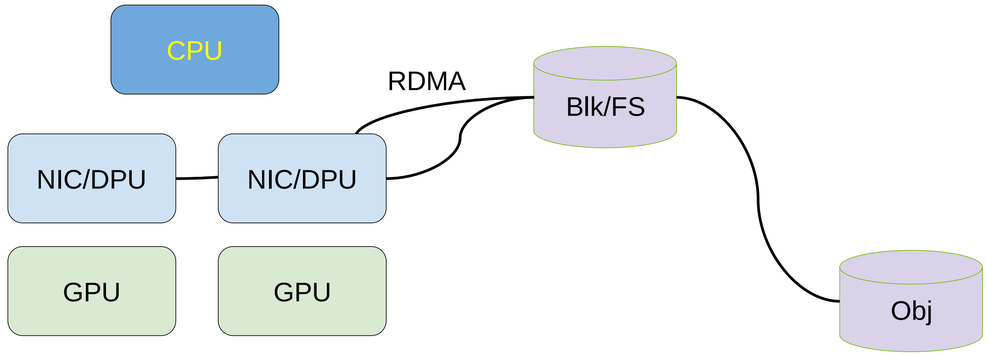
Data Loading and Checkpoints Can Have a Dramatic Impact

The right storage can significantly reduce training time and increase the productivity of GPU clusters.



Accelerated, Secure Storage

New: GPU-initiated IO; DPUs facilitate management and security



- Accelerated
  - Feed GPUs: bandwidth for DL, IOPs for GNNs
  - Direct into GPU with GPUDirect Storage
  - GPU-initiated accesses to memory/storage
- SW/HW Architecture
  - In-node: PCIe switch avoids CPUs' peer-peer and 2x PCI tree bottlenecks
  - Disaggregated: GB/s, manageability, security, utilization
  - Tiering: accelerated/near with RDMA as files or keys; bulk/far as objects
  - Serverless: avoid overspecifying name, location, format; incl data services, orchestration
- Security
  - Shift storage clients out of untrusted compute node into DPU proxy
  - Potential for zero copy with managed key service in DPU proxy
  - Potential for inline acceleration of encryption for data in transit and at rest

Standards, Standards, Standards

Protocol Changes:

- Metadata Access
- KV Store Methods
- RDMA/GPUDirect/etc

Client Side Access:

- Drivers
- DPU Abstractions (API)

Benchmarks:

- Full Workflow Suite
- Input from Industry (Storage)
- Ensure “fair” results/reporting

Audience Q&A