Maximizing At-Scale AI Training Efficiency: The Power of Data Caching

Raphael Druon,
Sr. AI Solutions Engineer
rdruon@ddn.com
DDN AI400X2 – THE AI DATA PLATFORM PROVEN AT-SCALE

• **Turnkey appliance**, fully-optimized for maximum AI application performance, proven at the largest scale.

• **Predictable** performance, capacity, capability

• **Shared parallel architecture** maximizes infrastructure performance, streamlines workflows, eliminates data management overhead, scales limitlessly.

• **Feature-rich** data management and security: hot pools, hot nodes, encryption, multi-protocol data services.

• **Advanced capabilities** ideal for multi node and hyperscale AI infrastructure deployments with analytics.

---

**All-NVMe or Hybrid**

- 90 GB/s, 3M IOPS read
- 65 GB/s write
- HDR200 and 200GbE
- 2 RU, 2.2 KW, 7.5K BTU/hr
What is EMF? EXAScaler Management Framework

**Simplified Software Lifecycle Management**
Automate installation, upgrade and expansion of platforms, servers, utility nodes and clients

**Configuration Management**
Centrally store and enforce all EXAScaler related configurations (servers, utility nodes, clients)

**Simplified Command And Control**
Manage distributed components centrally using consistent commands (CLI, API, GUI)
NVIDIA Partners with DDN to Fast Track HPC and AI in Enterprise Data Centers Globally

DDN powers the largest AI supercomputers in the world

NVIDIA SELENE SuperPOD – 560 DGX A100s
40 DDN AI400X, 2 TB/s read throughput, 15 PB NVME capacity, 1EB on HDD
#5 on Top 500 and #2 on Green 500 Lists
Proven, Predictable, Best AI Data Performance
Simplicity of Scaling, Balanced Performance, Plug-And-Play Experience

900 GB/s
READ THROUGHPUT
OUT OF THE BOX

780 GB/s
BALANCED WRITE
PERFORMANCE

12 DDN AI400X2 with 20 NVIDIA DGX A100

© 2023
At-Scale NLP Workloads Common IO Features

- Large models require distributed training with hundreds or thousands of GPUs
- Data sets at least multiple petabytes and constantly growing
- Often small random read is the primary IO pattern
- MMAP POSIX function is very common for loading data
- Containerized applications require new data path optimizations
- Checkpoints are widely used to reduce impact of failures
- GPU systems are getting more powerful and AI models are getting larger which equates to a steady increase in data volumes and IO performance over time
Checkpoints take a snapshot of a model and store it in a non-volatile memory. They’re an important part of training long running AI models efficiently, especially at-scale.

**Register, save, pause and resume AI applications**
Resume at a particular step in the training process and recover from any failure, with all progress and energy used saved.

**Improve inference prediction accuracy**
In continuous learning, intermediate models are deployed for inference, while online training continues with new data sets and parameters.

**Relocate AI processes to different systems**
Easily migrate to another platform, ideal in case of infrastructure fault.

**Perform transfer learning**
Intermediate model states are used as seed to train for a different goal.

“10% of jobs run for at least 13.5 hours before they fail, and 1% of jobs fail after executing for not less than 53.9 hours. Many of these jobs require 128 GPUs spanning many nodes, that are very expensive to purchase, maintain and run.”

---

**Meta AI Research Team**
Multi-nodal distributed training

Limitation

• Nowadays, most of the datasets doesn’t fit in worker memory
• In this case, each workers need to read the dataset continuously from the shared storage
• Storage is kept busy while it could be used for other IO intensive task (such as checkpoint, ingest or other cluster activities)
Next-Gen **AI Data Caching** with DDN Hot Nodes *(Based on PCC)*

Leverage local flash to maximize benefits of unified, global shared namespace

- Achieve full AI application performance with data cached on local NVMe devices in client, without any manual and risky data management overhead.
- Automated data movement from shared space to local node with intelligent policy-based cache management makes the process entirely transparent for users.
- Delivers significant efficiencies and AI workload improvements with large number of nodes engaged simultaneously for training, especially for at-scale NLP.
Additional features

- Cache management tools
  - Fine control over cache behavior, policy-based cache management

- Statistics reporting
  - Cache utilization
  - Hit/Miss

- NVIDIA BC (Base Command) integration
  - Load data on preflight
  - Statistics visualization
12 DDN AI400X2 - SHARED AI STORAGE FOR NVIDIA DGX SUPERPOD

SHARED STORAGE READ THROUGHPUT

0.0 GB/s

SHARED STORAGE WRITE THROUGHPUT

0.0 GB/s

LET'S START A DISTRIBUTED AI TRAINING WORKFLOW.
At-Scale NLP with DDN: Megatron-LM on NVIDIA SELENE
Training Large Language Models Requires Fast Read and Write Performance and Large Capacity

• GPT3 model training using 128 DGX A100s and DDN shared storage
• 13B parameters in model (2020). Today, models are 40-50X larger.

• Read data set at beginning of every training job:
  • Up to 1 TB/s read from shared DDN storage during first iteration
  • Full data set is several TBs, each DGX assigned a different portion
  • Shared storage makes it easy access to entire data set without any copy

• Hot Nodes makes distributed training process more efficient:
  • On first read, data is delivered to application and copied to local DGX storage
  • Subsequent reads delivered from local storage, transparent to application
  • Shared storage available for checkpoints, ingest and other cluster activities
Product Features for AI
Highest Density All Flash with DDN’s new NVMeoF Enclosure

- **The Fastest Controller to Drive Your QLC**
  - Real-Time, Many-Core RAID Engine

- **High Density Flash**
  - Up to 7PB QLC in 10RU

- **Unprecedented Resilience**
  - SFAOS Data Protection, Intelligent data placement and SuperFast Rebuilds

- **EXAScaler Parallelism**
  - DDN Software extends outstanding performance across the network to your applications

- **No Single Point of Failure, Zero Midplane Design**

- **Simple Scale up and Scale out**
  - No External Switching with DDN Storage Fusion NVMe Fabric
DDN CSI Driver Security Features For Kubernetes

EMF enhancements simplify multi-tenant management and will introduce fine-grained authorization (RBAC)

Secure Containerized Applications with Full K8 Flexibility and Performance

- Per tenant authentication and access controls
- Central volume management (create/delete)
- Fine-grained quotas administration
- Dynamic publish & unpublish for volumes
- No root access required for tenants

Extra layer of security
Powered by EMF

EXA CSI Driver

kubernetes, Red Hat OpenShift
Further reading

- “Accelerating AI at-scale with Selene DGXA100 SuperPOD and Lustre Parallel Filesystem Storage” by Prethvi Kashinkunti and Julie Bernauer from NVIDIA, LUG 2021, (slides) (video)
- DDN Reference Architecture