DDN Importance of Checkpoints with LLM

The AI Data Company

DDN Creates and Delivers Enabling Data Storage Solutions for the Digital Transformation of Organizations, Government and Academia

11,000 + Customers
Strategic Enterprise, Government, Academia

Broad Markets
Research, AI, Healthcare, Cloud, Finance, Gov

Global Presence
US, EMEA, APAC
Customers, Labs, People

World Class Team
1,000 Team Members
600 Engineers

Stability and Longevity
20+ Years in Business
Run by Engineers

Leadership at Scale
Proven Execution in Innovation and Delivery

DGX H100 Workload To Storage Sizing Guidance

NVIDIA uses “Light IO Workload” as default requirement for BasePOD and SuperPOD RAs

	Storage Sizing	Common Applications	Data Sets
Light IO Workload **NVIDIA DEFAULT RECOMMENDATION**	1 GB/s read 750 MB/s write per GPU	• Multi-node training of natural language models (up to 128 nodes). • CoE / small research environments. • Matches the EOS configuration.	• Datasets generally fit in local cache.
Medium IO Workload	2 GB/s read 1.5 GB/s write per GPU	• Distributed training of large natural language models (more than 128). • Image processing with compressed video/images file formats. • Long-running jobs (over 12 hours) with checkpoint.	• Some datasets fit in local cache, many don’t. • Big first epoch read, and checkpoint write requirements.
High IO Workload	4 GB/s read 3 GB/s write per GPU	• Distributed training of large natural language models (more than 256). • Image processing with uncompressed video/images file formats. • Long-running jobs (over 24 hours) with checkpoint.	• Datasets are too large to fit in local cache. • Massive first epoch read, and checkpoint write requirements.

DDN AI400X2 – THE AI DATA PLATFORM PROVEN AT-SCALE

All-NVMe

90 GB/s read

65 GB/s write

HDR200 and 200GbE

2 RU, 2.2 KW, 7.5K BTU/hr

- **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.

Checkpointing Improves AI Model and Workflow Performance

Checkpoints take a snapshot of a model and store it in a non-volatile memory.

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.

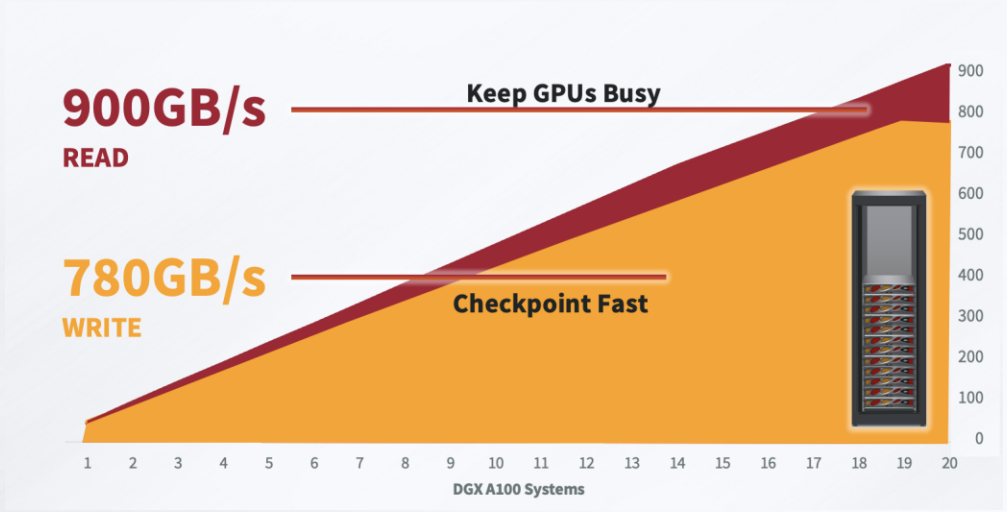
Perform transfer learning

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

AI Training Needs Checkpoints and so AI Storage much be Optimized for Fast Writes

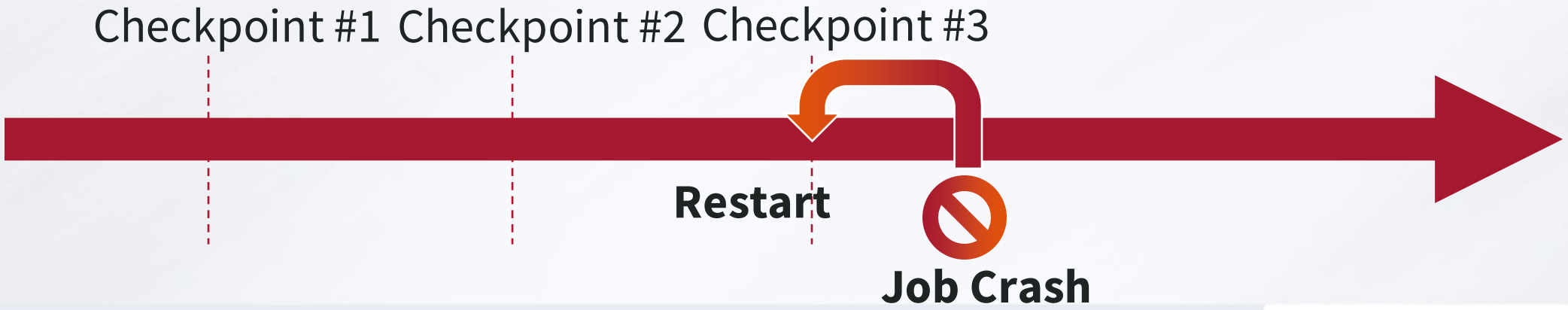


"What we are also seeing and what we put in the effort around DGX SuperPOD stack is around checkpointing. So a lot of our frameworks continue to do checkpointing and that can be very heavy on the storage side."

Premal Savla - Senior Director of NVIDIA DGX

What are Checkpoints?

- An intermediate dump or a snapshot of a model's entire internal state, including
  - Weights
  - learning rate
  - number of epochs executed, etc.,
- Checkpoints are a jumping-off point so that the framework can pick up on its training from here whenever needed.



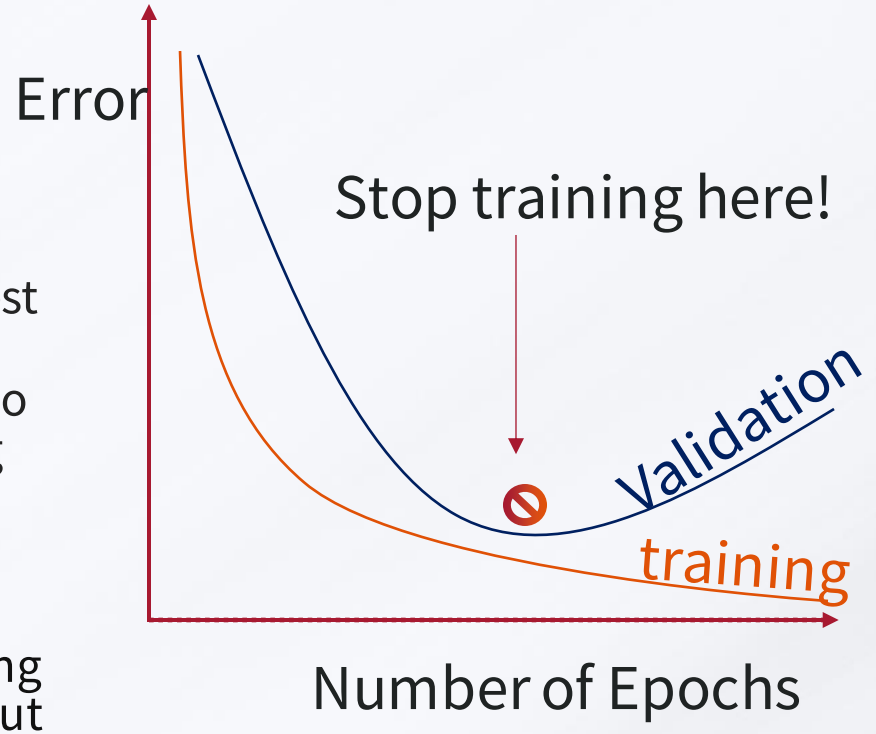
Uses of Checkpoints

Checkpoints are generally used for **interval training** because you can stop, pause, and resume training from specific states in your training job. Aside from this (and other small applications within this process), checkpoints can also be used for the following:

- **Prediction Accuracy** Improvement
  - Checkpointing can be used to improve inference prediction accuracy–this happens when the learning rate is lowered to increase the model's accuracy. That means that even while the model is still being trained, you can use it to make predictions.
- **Multi-System Training**
  - When you pick up from a checkpoint, you can either continue training the model with the existing dataset or start with training across different nodes or clusters. This is helpful when you need to do a training job that requires input from multiple systems.
- **Transfer Learning**
  - At some point during a long training job, you might find that your goals have changed. When this happens, you can use checkpoints to perform transfer learning.

Uses of Checkpoints

- **Early Stopping**
  - The longer you train, the lower the loss on the training dataset. But at some point, the error stops decreasing
  - For large models, without sufficient regularization, the error on the evaluation dataset might even start to increase.
  - → need to go back and export the model that had the best validation error.
  - This is also called *early stopping*. The only way you can do early stopping is if you have been periodically evaluating and checkpointing the model.
- **Better Fine Tuning**
  - If you need to retrain your model on fresh data, you'll want to emphasise the new information, instead of pulling from the past datasets. With checkpoints, you can pick out certain states from which you can easily start your new training or experiment with new paths.



Uses of Checkpoints

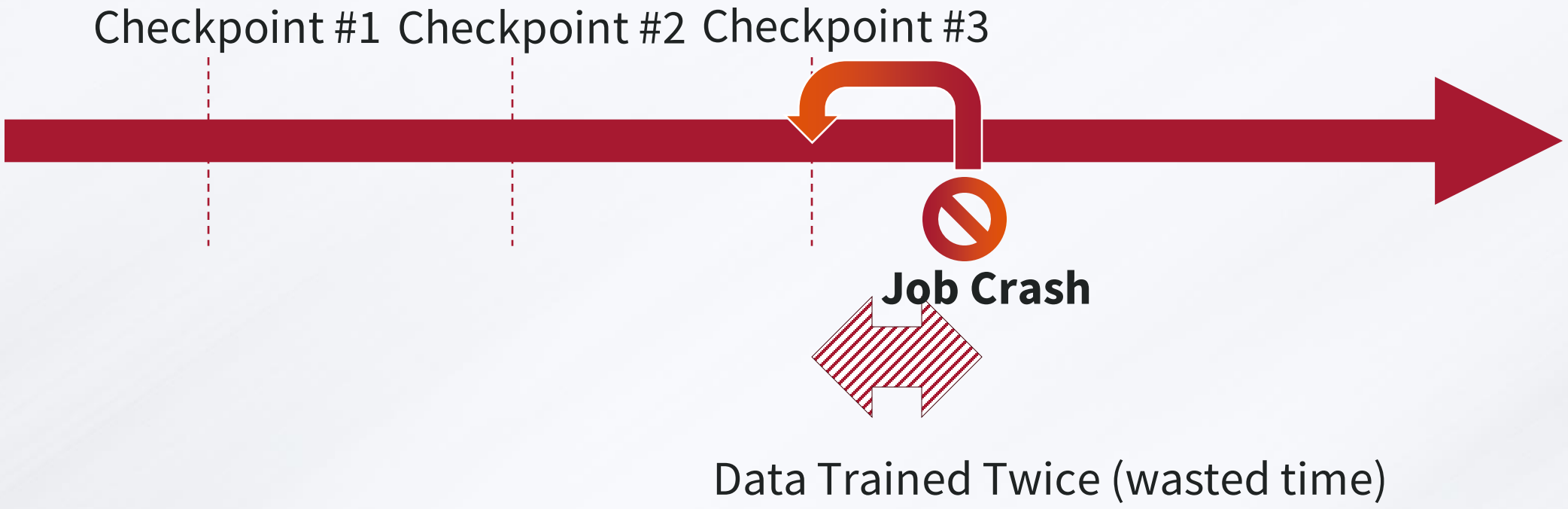
Restart following failure

- Training Jobs can take weeks across thousands of GPUs to complete.
- Long training jobs are subject to a lot of risks, with the likelihood of machine failure increasing the longer it goes on. With checkpoints, you can resume from the last saved state instead of having to work from the very beginning of the project.

"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."

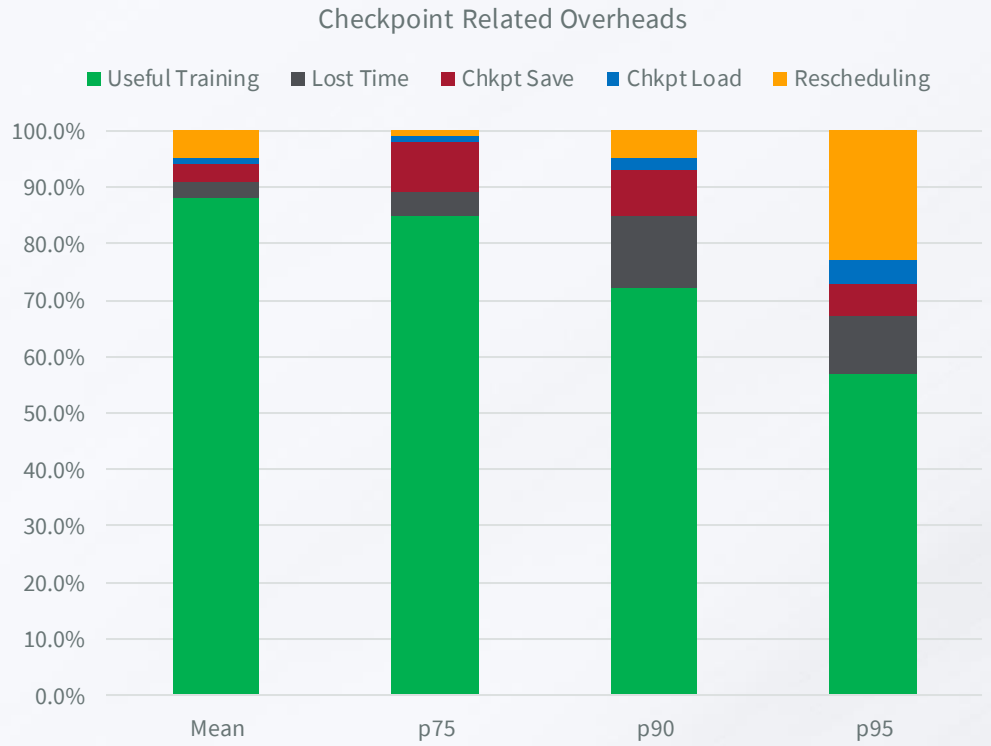
Meta AI Research Team

Regular Checkpoints minimize Wasted Training Time

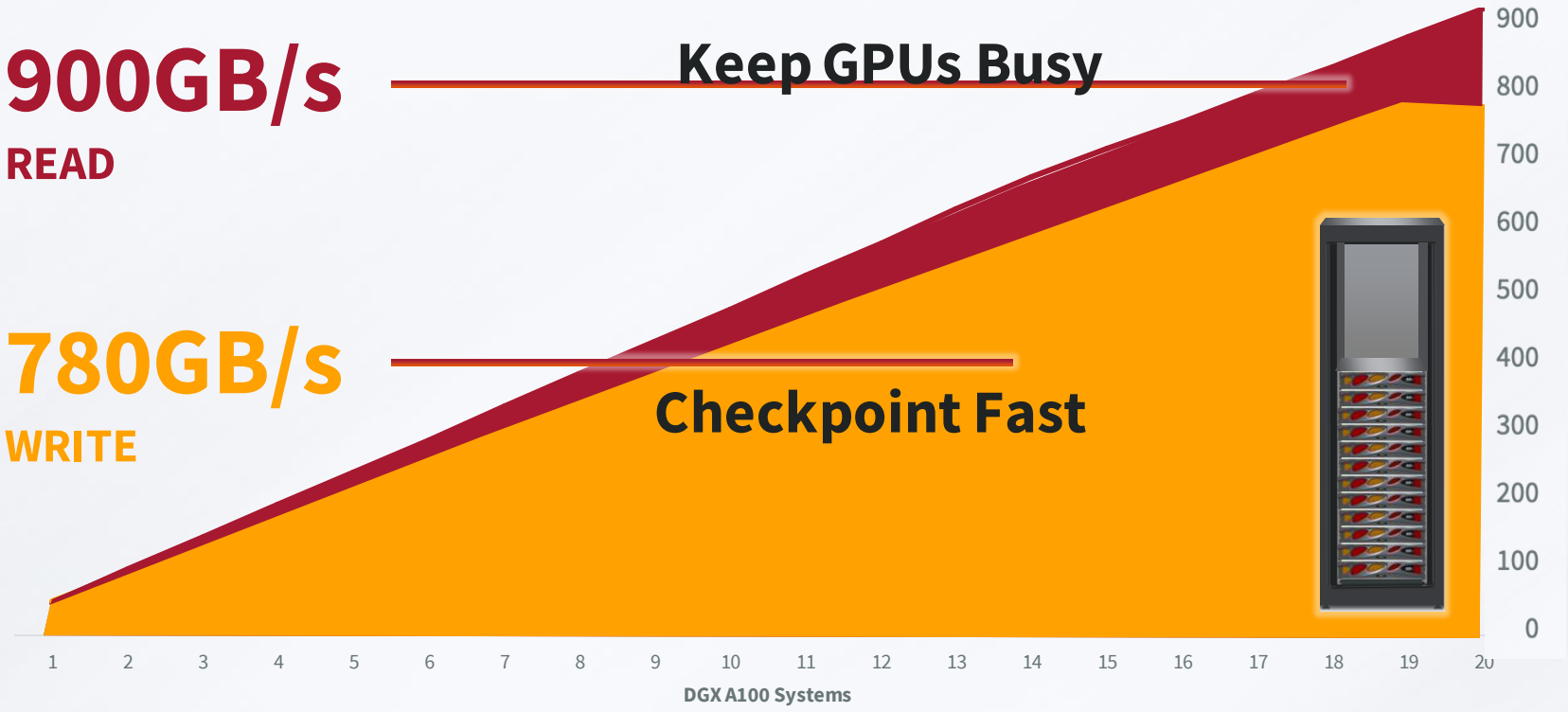


Why Choose a Filesystem with FAST Writes?

- On average, checkpoint-related overheads can account for 12% of total training time and can rise to as much as 43% (Maeng et al., 2021).



Strong Writes, Strong Reads, Linear Scale



Real measurements provided by


ddn