NVIDIA

Accelerating AI at-scale with Selene DGXA100 SuperPOD and Lustre Parallel Filesystem Storage

Julie Bernauer and Prethvi Kashinkunti - Datacenter Systems Engineering, NVIDIA

Lustre User Group Meeting, May 19th, 2021

AGENDA

Challenges of AI at scale

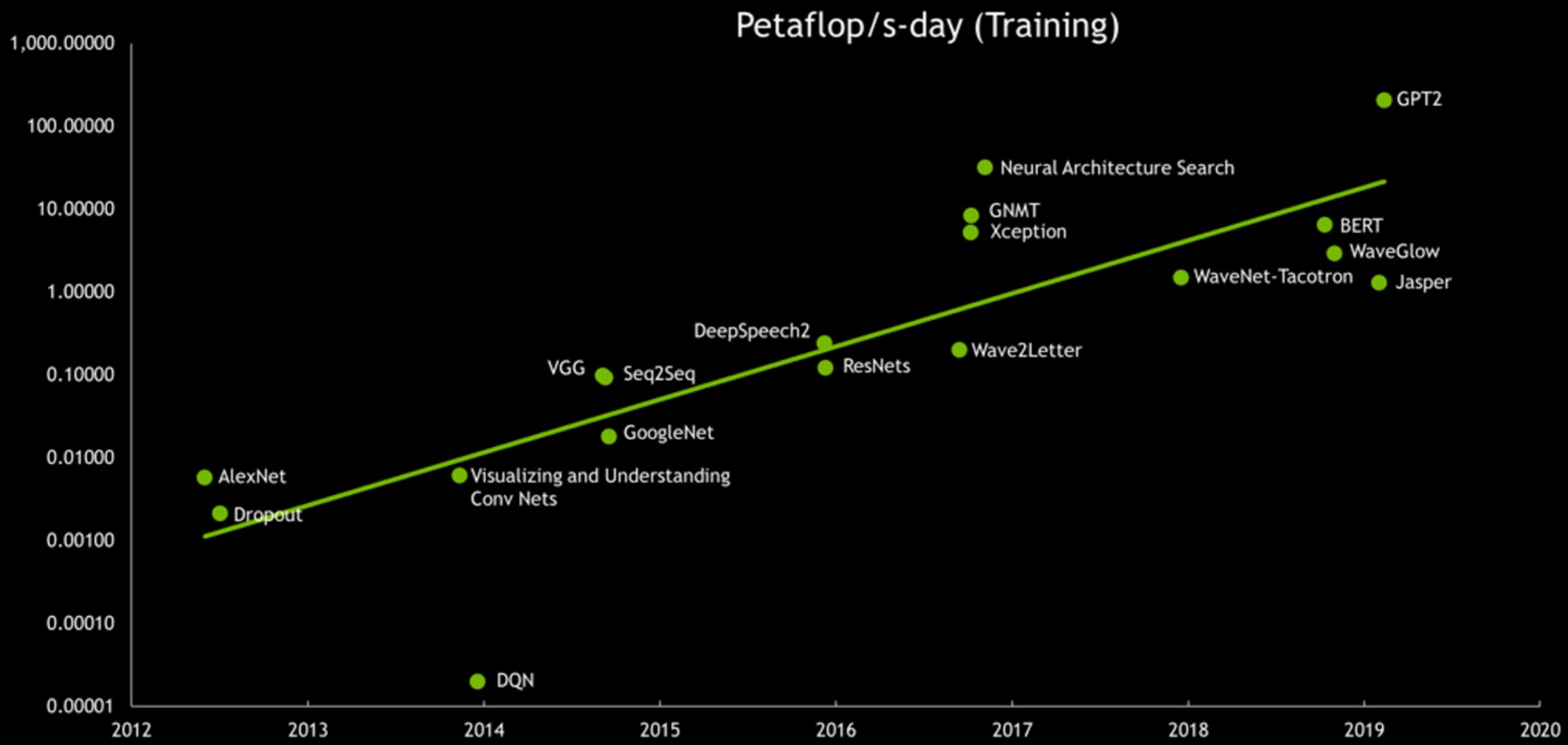
DGXA100 and Selene

Discussion on Selene Storage architecture

Synthetic and Real Application Performance

Client caching: a new feature for workload perf?

MODELS GETTING MORE COMPLEX



Source: OpenAI and NVIDIA

DATASETS GETTING LARGER

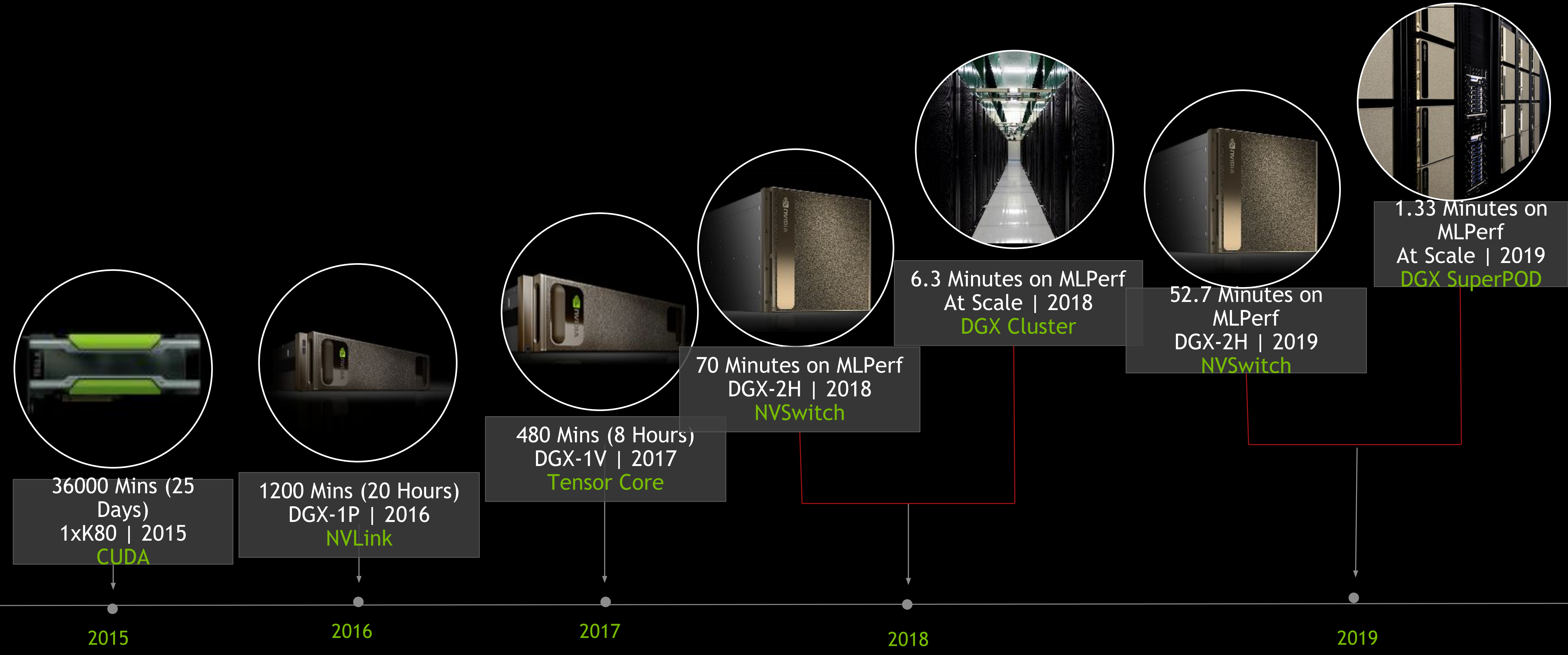
Unlabeled data:

- Language model: BooksCorpus (800M words), English Wikipedia (2.5B words), WebText (8M documents, 40 GB), C4 (Common Crawl, 745 GB)
- GAN: unlabeled images and videos
- Reinforcement learning: unsupervised self-play generates unlimited data

Labeled data:

- ImageNet (2012) - 1.3M images, 1000 categories
- Open Images (2019) - 9M images, 6000 categories
- Semi-autonomous vehicles: 0.5-1.1TB of data for every 8h driving

DL TRAINING: FROM SINGLE GPU TO MULTI-NODE



ResNet50 v1.5 training

DGX SuperPOD with DGX-2

SELENE

DGX SuperPOD Deployment

#1 on MLPerf for commercially available systems

#5 on TOP500 (63.46 PetaFLOPS HPL)

#5 on Green500 (23.98 GF/watt) - **#1** on Green500 (26.2 GF/W) - single scalable unit

#4 on HPCG (1.6 PetaFLOPS)

#3 on HPL-AI (250 PetaFLOPS)

Fastest Industrial System in U.S. — 1+ ExaFLOPS AI

Built with NVIDIA DGX SuperPOD Architecture

- NVIDIA DGX A100 and NVIDIA Mellanox IB
- NVIDIA’s decade of AI experience

Configuration:

- 4480 NVIDIA A100 Tensor Core GPUs
- 560 NVIDIA DGX A100 systems
- 850 Mellanox 200G HDR IB switches
- 14 PB of all-flash storage

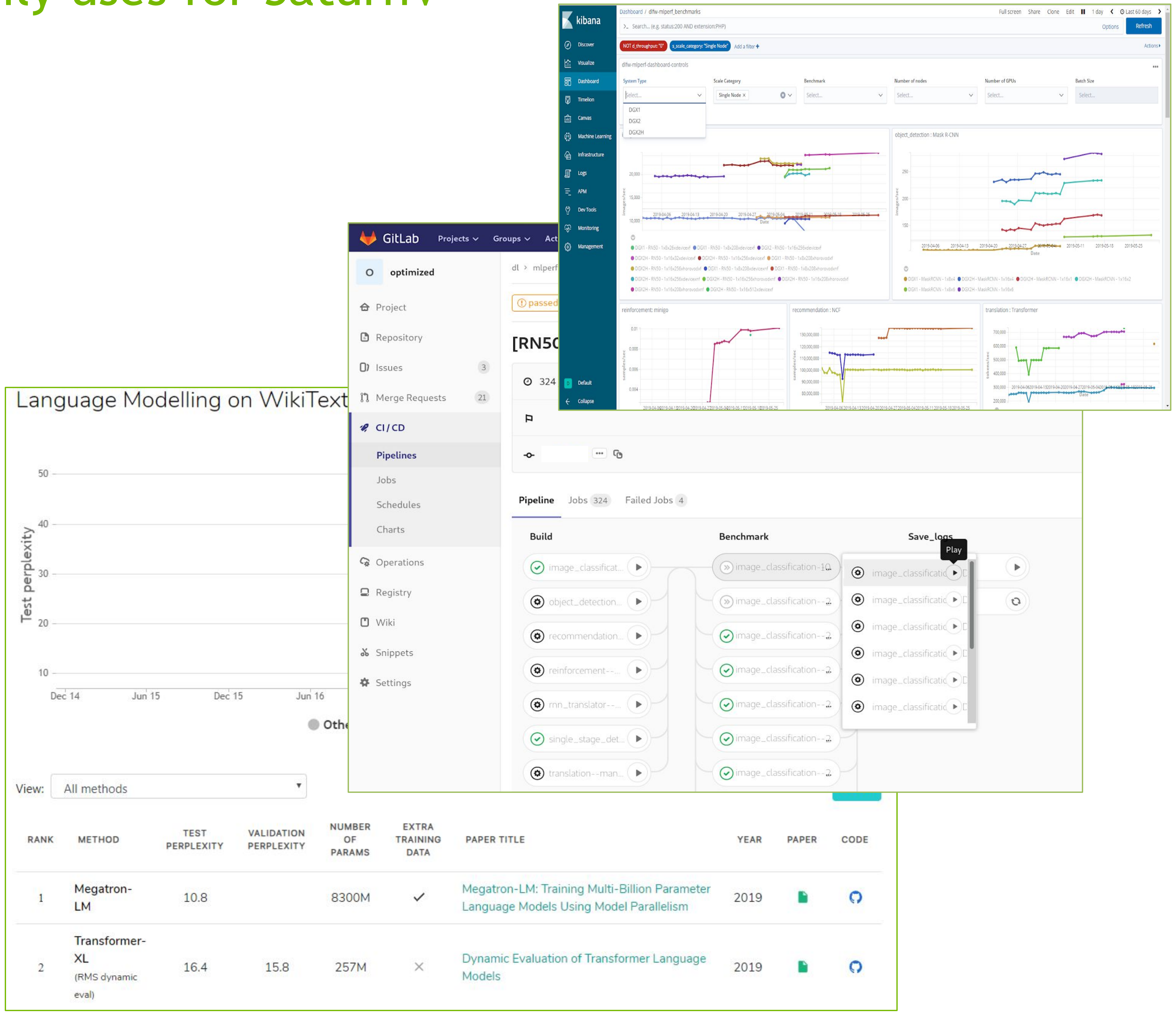
CLUSTERS AT NVIDIA

A wide variety of daily uses for SaturnV

Supporting a wide community of users

- supercomputer-scale continuous integration for software
- research
- “big iron AI” work (e.g. Megatron, ASR)
- automotive
- QA

Need for performance at scale and flexibility



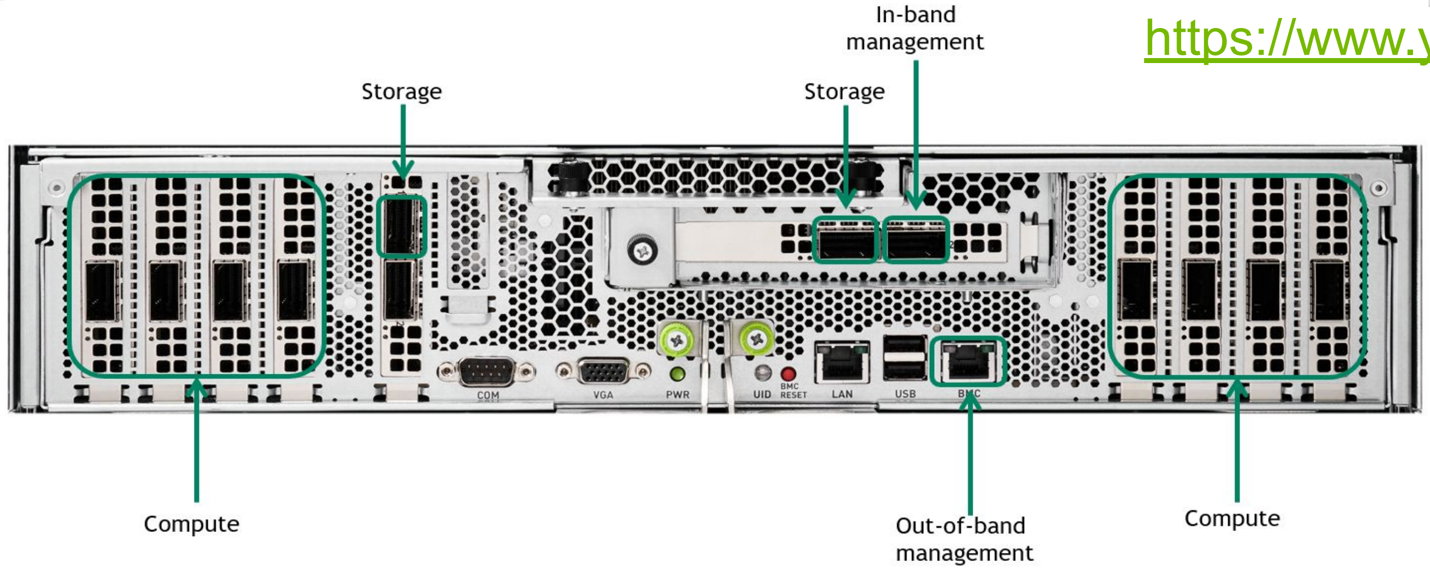
DGXA100 and Selene

A NEW GENERATION OF MACHINES

NVIDIA DGXA100

GPUs	8x NVIDIA A100 80GB
GPU Memory	640 GB total
Peak performance	5 petaFLOPS AI | 10 petaOPS INT8
NVSwitches	6
System Power Usage	6.5kW max
CPU	Dual AMD Rome 7742 128 cores total, 2.25 GHz(base), 3.4GHz (max boost)
System Memory	2TB
Networking	8x Single-Port Mellanox ConnectX-6 200Gb/s HDR Infiniband (Compute Network) 2x Dual-Port Mellanox ConnectX-6 200Gb/s HDR Infiniband (Storage Network also used for Eth*)
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 30TB (8x 3.84TB) U.2 NVME drives
Software	Ubuntu Linux OS (5.4+ kernel)
System Weight	271 lbs (123 kgs)
Packaged System Weight	359 lbs (163 kgs)
Height	6U
Operating temp range	5˚C to 30˚C (41˚F to 86˚F)

https://www.youtube.com/watch?v=TJcKYUTaBtg



Utilize Multi-rail as a way to get to perf

The DGXA100 SuperPOD

An extensible model

1K GPU POD Cluster

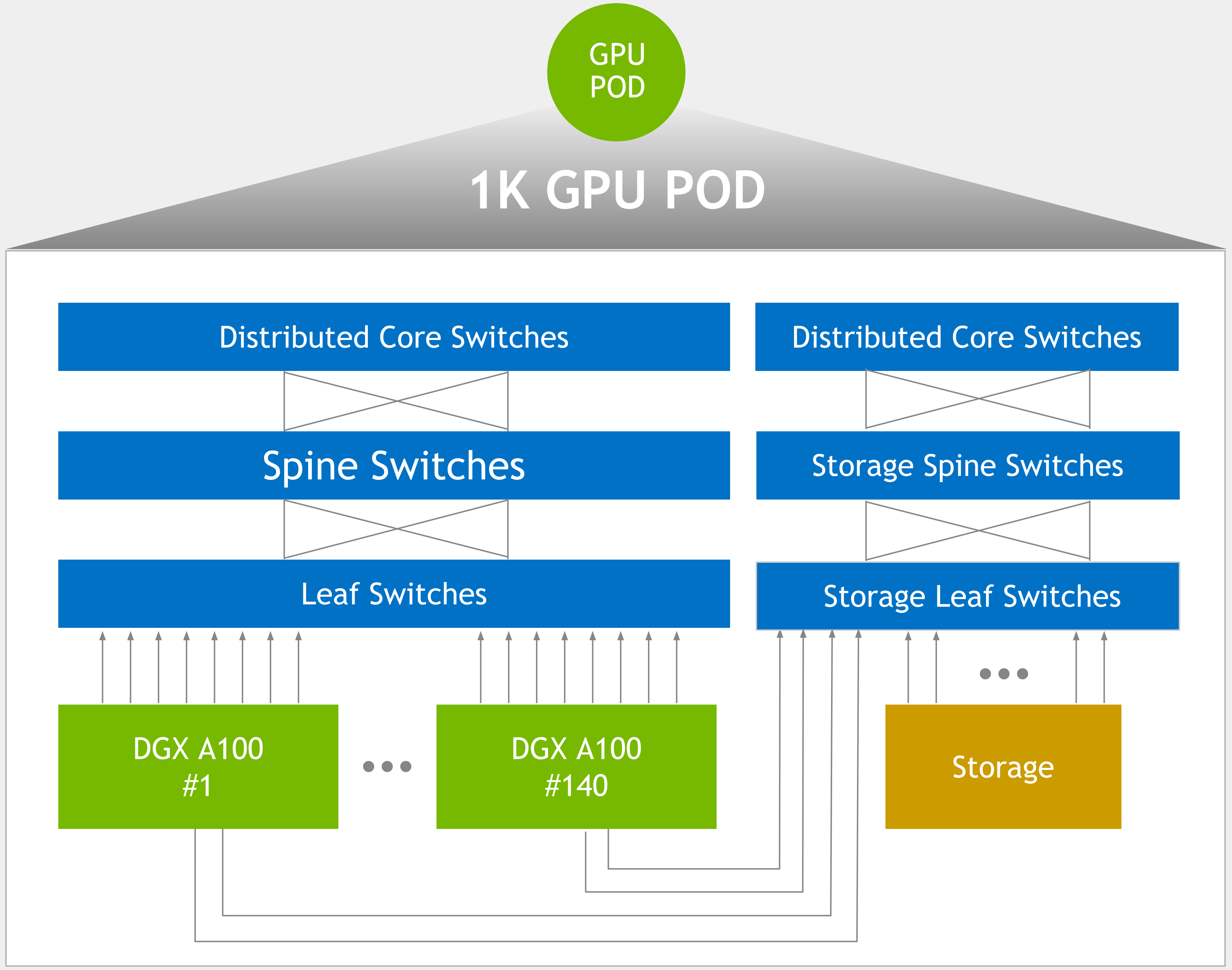
- 140 DGXA100 nodes (1120 GPUs) in a GPU POD
- *1st tier fast storage - DDN AI400X with EXAScaler*
- Mellanox HDR 200Gb/s InfiniBand - Full Fat-tree
- Network optimized for AI and HPC

DGXA100 Nodes

- 2x AMD 7742 EPYC CPUs + 8x A100 GPUs
- NVLINK 3.0 Fully Connected Switch
- 8 Compute + 2 Storage HDR IB Ports

A fast interconnect

- Modular IB Fat-tree
- *Separate network for Compute and Storage*
  - *Needed to achieve 1TB/s Storage BW requirement*
- Adaptive routing and SHARP support for offload

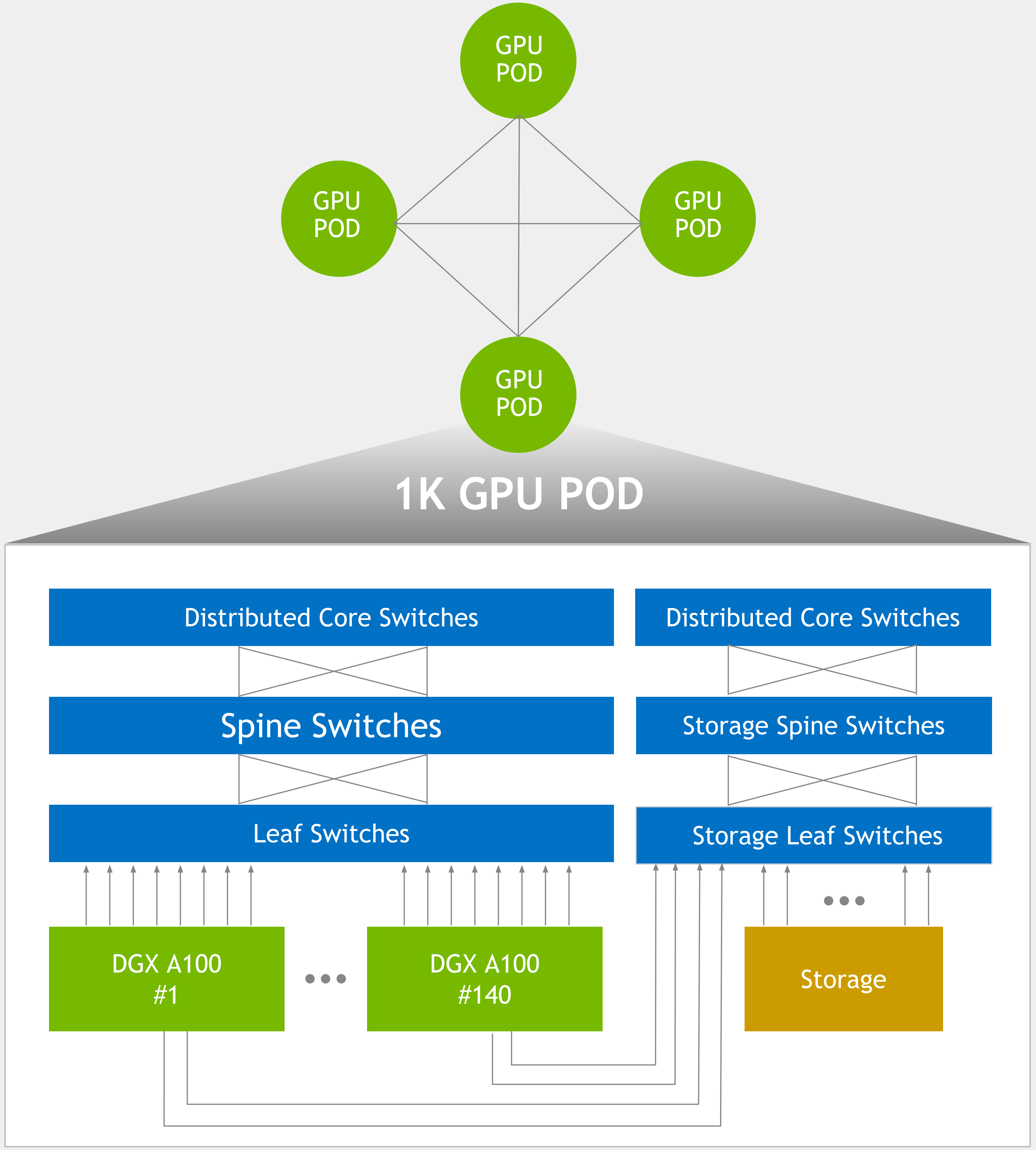


The DGXA100 SuperPOD

An extensible model

POD to POD

- Modular IB Fat-tree
    - Core IB Switches Distributed Between PODs
    - Direct connect POD to POD
- *Separate network for Compute and Storage*
- Adaptive routing and SHARP support for offload

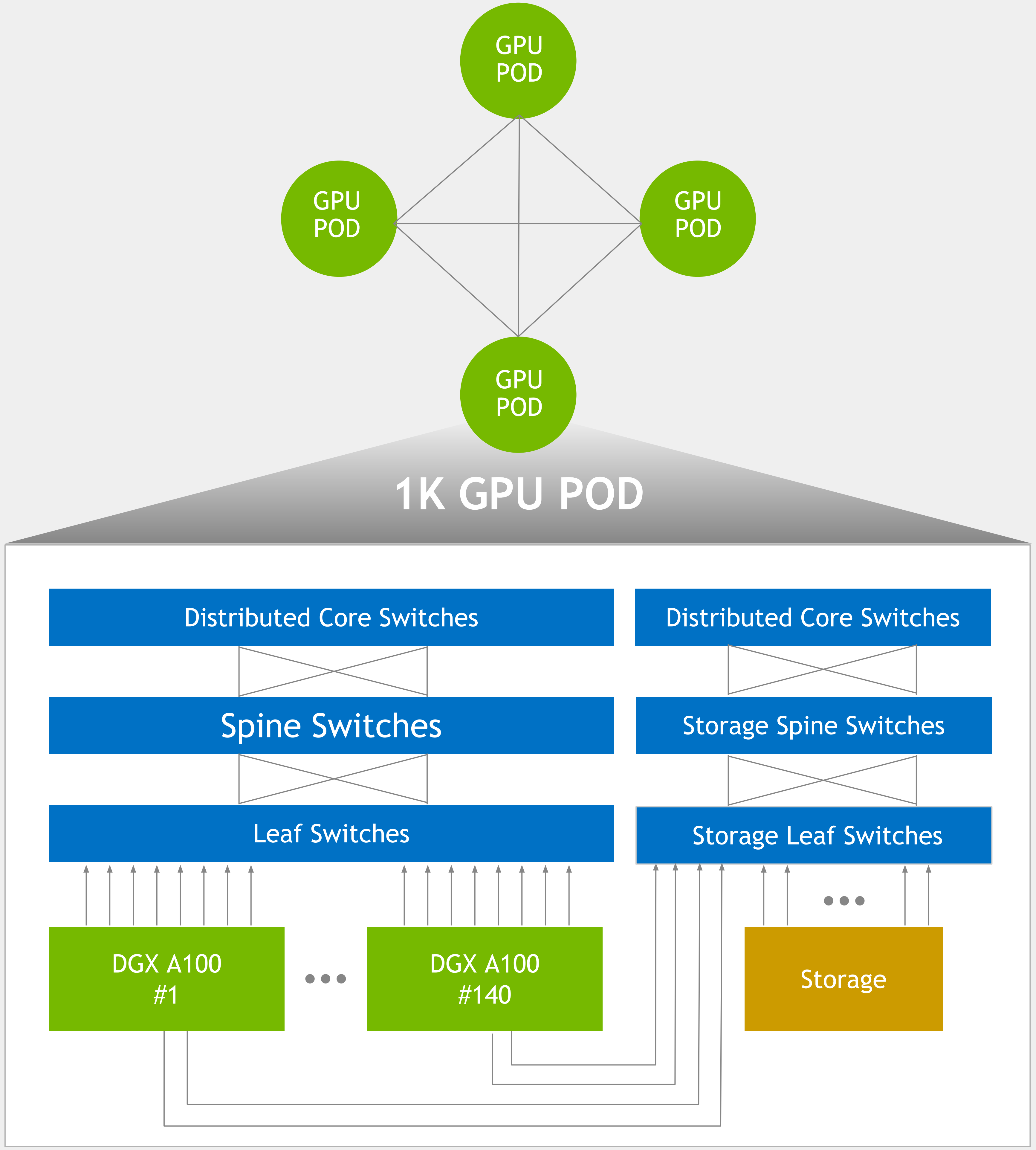


The DGXA100 SuperPOD

An extensible model

POD to POD

- Modular IB Fat-tree
  - Core IB Switches Distributed Between PODs
  - Direct connect POD to POD
- *Separate network for Compute and Storage*
- Adaptive routing and SHARP support for offload



Selene Storage Architecture

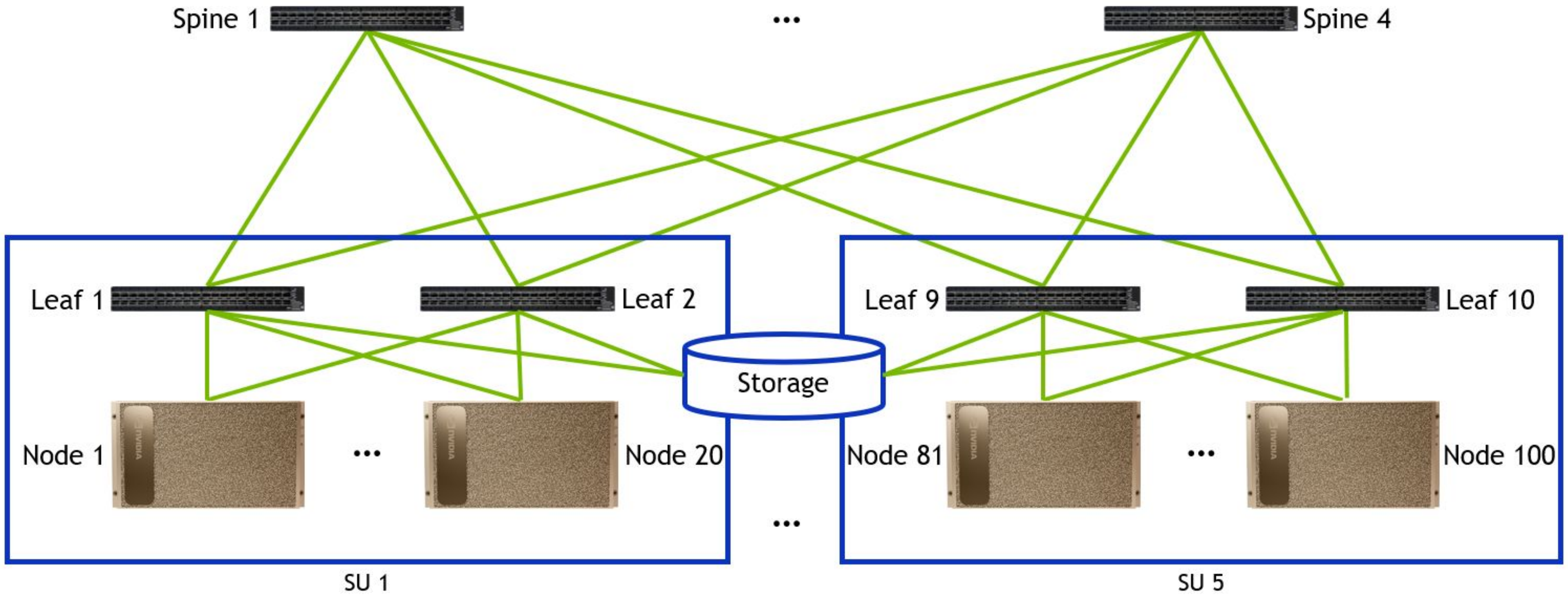
A POD at any scale

Growing with Scalable Units (SU)

Nodes	SUs	Storage Ports	QM8790 Switches		Cables			Subscription Ratio
			Leaf	Spine	Leaf	Spine	Storage	
10	1/2	4	2	1	20	20	4	1:1
20	1	8	2	1	40	32	8	3:2
40	2	16	4	2	80	64	16	3:2
80	4	32	8	4	160	128	32	3:2
100	5	40	10	4	200	160	40	3:2
140	7	56	14	8	280	224	56	5:4

Storage fabric with different ratios

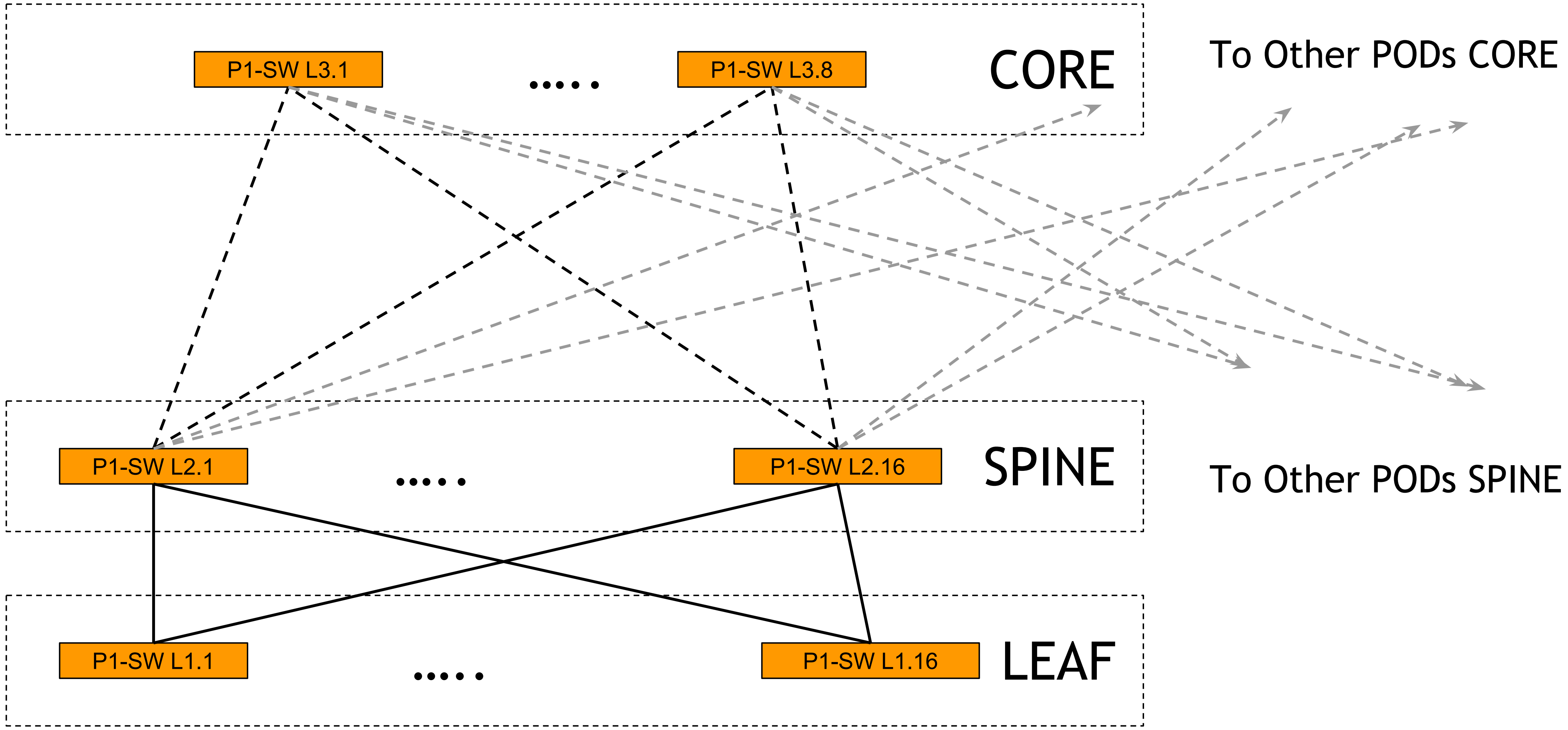
100 node example



NVIDIA

Selene SuperPOD
Close up on the Storage Fabric

a.k.a "how did we cable it?"



- 2 HDR200 Per Node
- 140 Nodes Per POD
- 3-to-2 Full Fat Tree Per POD
  1-to-1 Uplink ratio between SPINE/CORE
- 8 100G connections for each for AI400x
- InterPOD BW*: 3200 GB/s (128*HDR)
- Resilient to **switch failures** at spine/core level: **max 7% perf hit to peak BW** when down one switch

➔ 16 LEAF Switches / 16 SPINE Switches / 8 CORE Switches
➔ 24 links down from LEAF
➔ One link between each LEAF <->SPINE (16 ports between SPINE to LEAF)
➔ One link between each SPINE<->CORE (16 ports from SPINE to CORE), 8 ports unused on SPINE, 8 ports unused on CORE
➔ Interleaved connected, odd-to-odd and even-to-even

*InterPOD BW is the uni-directional bandwidth from a single 1K POD to another

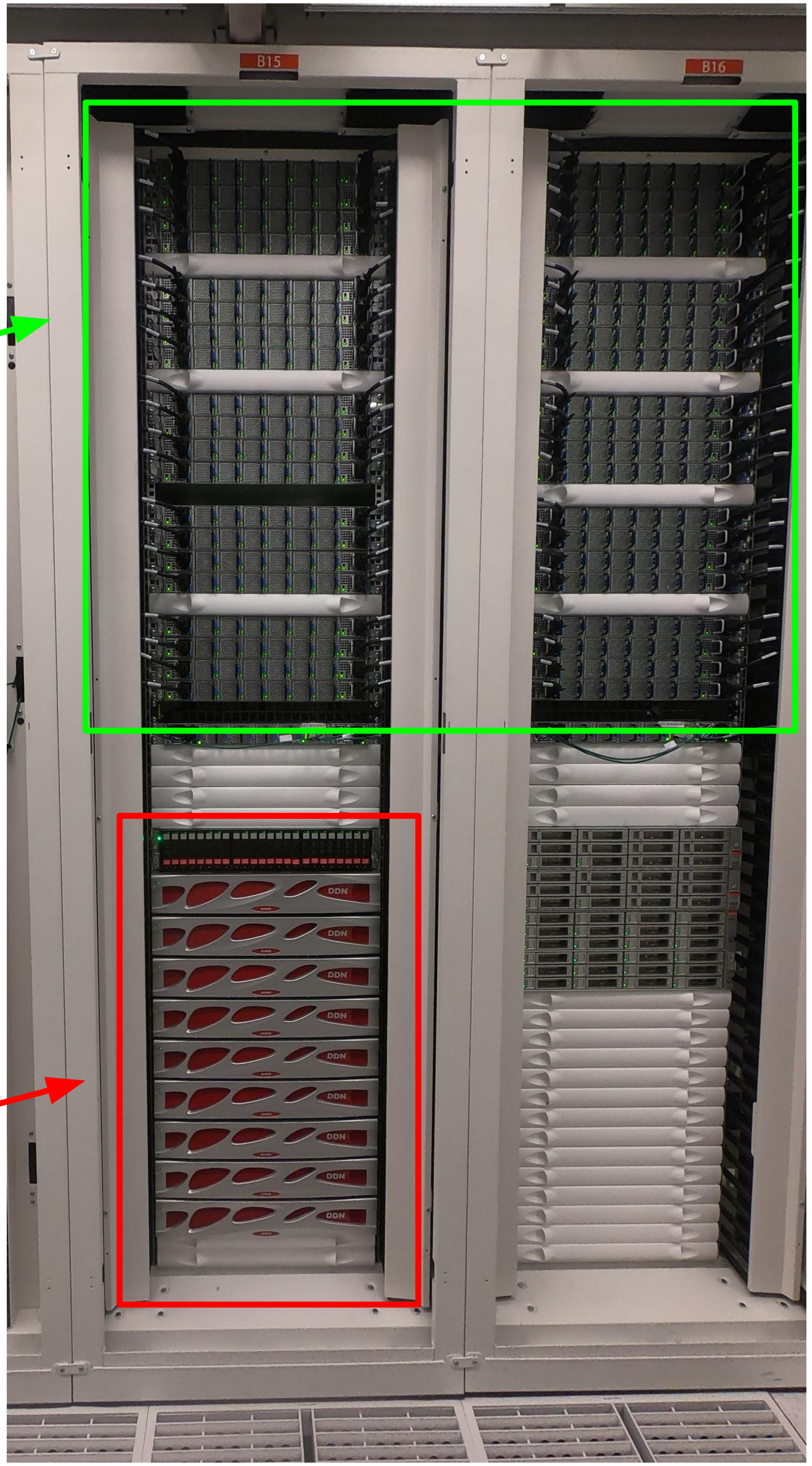
Performance Storage

DDN AI400X Appliances

- 2 racks provisioned for storage per GPU POD
  - Selene is configured with four GPU PODS
- High-performance storage system (per POD)
  - 10 DDN AI400X appliances
  - All-NVME drives, unified namespace
  - 2.4 PB useable capacity
  - Peak performance read/write: 500/350 GB/s
  - 80 HDR100 interfaces
  - 20 RU, 16.6 KW, 57K BTU/hr
- High-performance storage system (Selene)
  - 40 DDN AI400X appliances
  - 10 PB useable capacity
  - Peak performance: 2 TB/s read, 1.4 TB/s write

Storage IB Switches

AI400X Appliances



Performance Storage

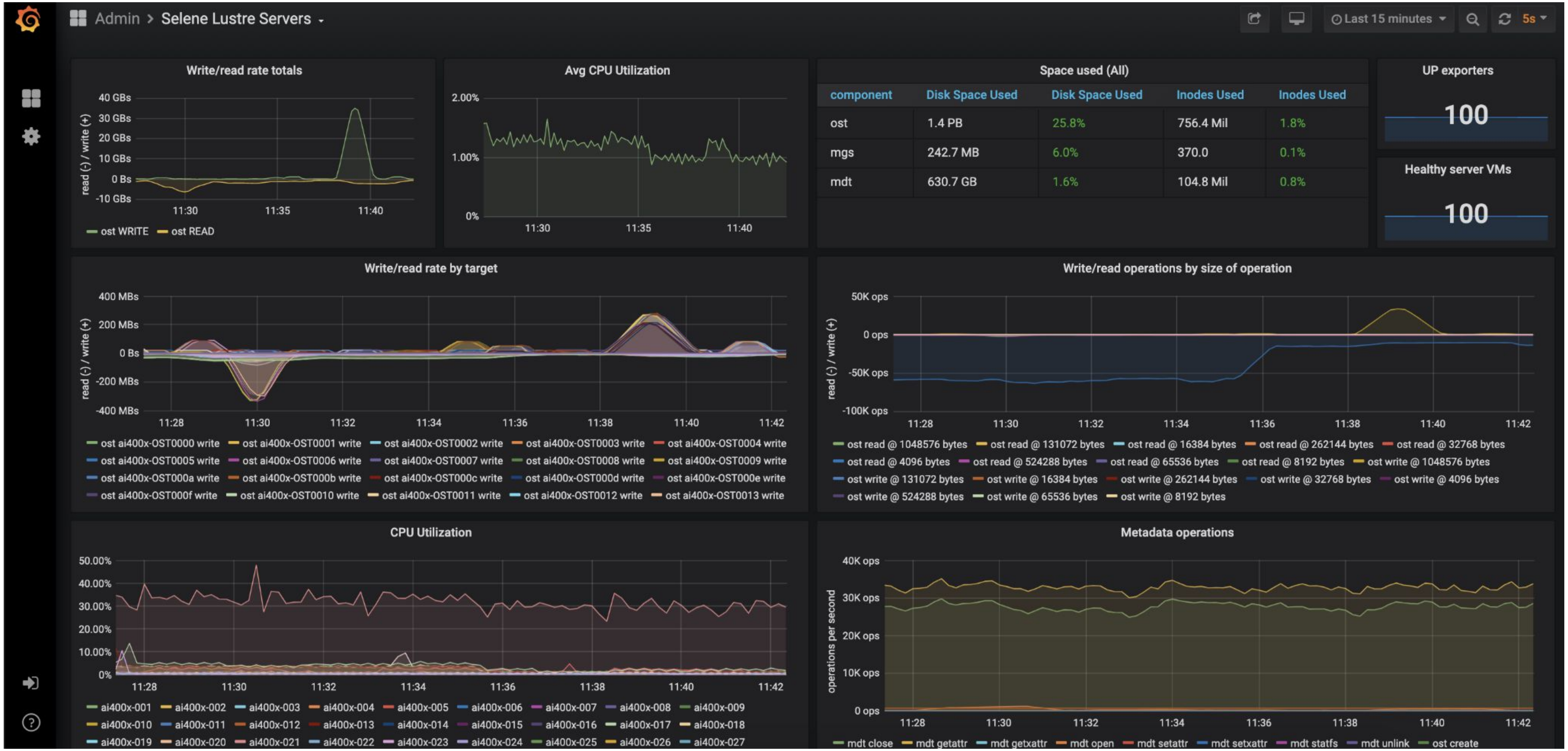
DDN AI400X - Configuration

50 GB/s read, 35 GB/s write, 3M IOPS
250 TB useable all-nvme capacity
8 x HDR100 IB (can also be 100GbE)
Dual controllers, fully-redundant
2 RU, 1.6KW, 5700 BTU/hr

- Utilizes DDN EXAScaler filesystem, based on Lustre
- Each AI400X configured with 4 VMs, presenting 1 MDT and 8 OSTs
  - Total of 20 MDTs and 160 OSTs in Selene production configuration
- Lustre Progressive File Layout (PFL) configured to facilitate efficient striping of 'small', 'medium' and 'large' files
  - `lfs setstripe -E 1G -c 1 -E 128G -c 8 -E eof -c -1 /lustre`
- LNet Multi-Rail utilized by all AI400X VMs and client nodes
  - `# cat /etc/modprobe.d/lustre.conf`
  - `options lnet networks="o2ib0(ib0,ib1)"`
  - `options libcfs cpu_npartitions=20 cpu_pattern=""`
  - `options ko2iblnd peer_credits=32 peer_credits_hiw=16 credits=1024 concurrent_sends=64`
  - `options lnet lnet_transaction_timeout=100`

Monitoring

Full telemetry info for IB and storage



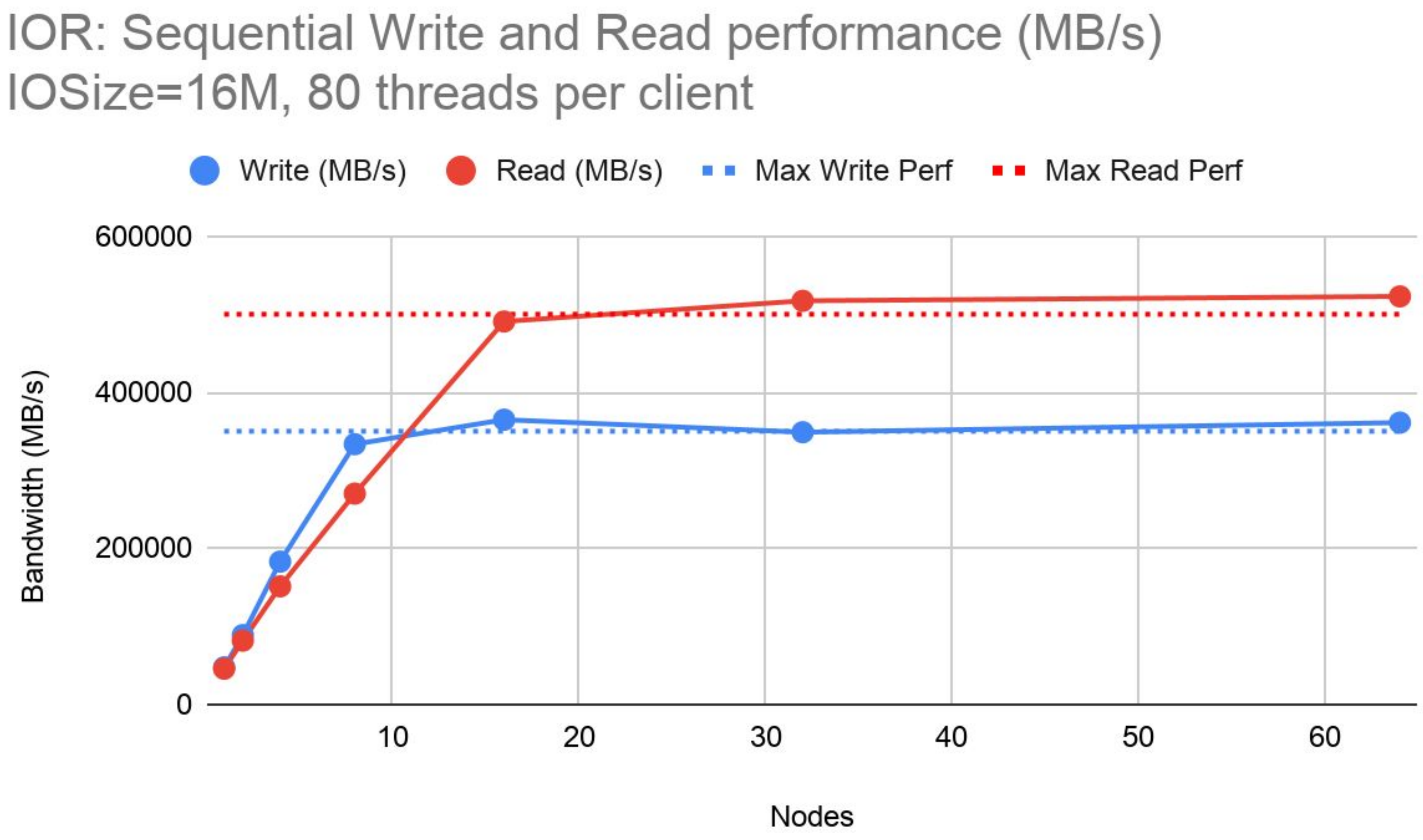
Synthetic Application Perf

Synthetic Benchmarks

10 AI400X Appliances (1 POD)

- From single client, can achieve nearly line rate of 2x200G HDR200 connections
- 16 clients and above can saturate perf of 10 AI400X appliances of 350GB/s write and 500GB/s read

1-64 clients
2-128 rails
8-512 GPUS



NVIDIA

Synthetic Benchmarks

20 AI400X Appliances (Selene Production Configuration)

- Selene production FS composed of 20 AI400X appliances
- Using same IOR test, 64 clients able to achieve 1TB/s read and 700GB/s write
  - Max perf scaling nearly linearly with number of appliances

64 clients
128 rails
512 GPUS

Real Application Perf

Real Selene Workload: MLPerf Training v0.7

Time-to-train: /raid vs /lustre

- Majority of DL workloads follow the same paradigm:
  - Read data, perform computation and all-reduce, read the same data, perform computation and all-reduce, … checkpoint, read data ….

- Select two workloads from the MLPerf Training benchmark as a way to evaluate DL performance
  - BERT: Natural Language Processing model (e.g. text generation, sentiment analysis, question & answer)
  - ResNet50: Image Classification model
  - Run training for both models using datasets stored on local node storage (/raid) and filesystem (/lustre)

BERT 128N

/raid: 114 s

/lustre: 122s (93.4%)

ResNet50 96N (using mmap)

/raid: 96.8 s

/lustre: 99.4 (97.3%)

Source: https://www.nvidia.com/en-us/data-center/resources/ddn-a3i-reference-architecture/

Real Selene Workload: Megatron-LM

Paper: https://arxiv.org/pdf/1909.08053.pdf Repo: https://github.com/NVIDIA/Megatron-LM

- Tools for ongoing research of training large transformer language models at scale by NVIDIA's Applied Deep Learning Research team
  - GPT-2/GPT-3 models and more, scaling from 1B to 1T parameters in size

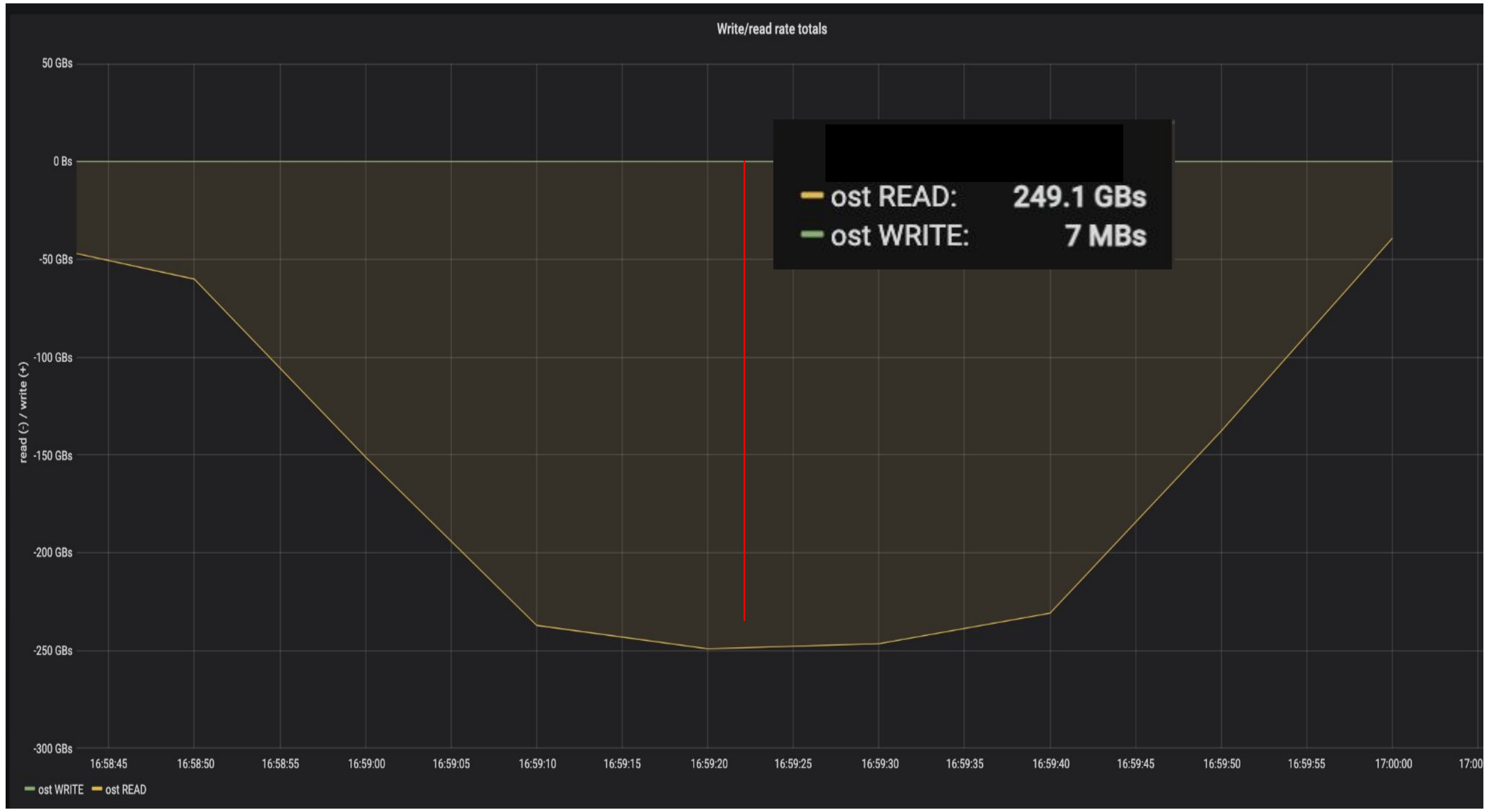
- Interesting use case wrt storage for few reasons:
  - Large models ⇒ large checkpoints
  - Tensor and pipeline parallelism ⇒ multiple checkpoints from different ranks
  - Training on a shared cluster ⇒ single job time limited, need to read/write checkpoints at beginning/end of each job
- Using GPT3 13B as an example model
  - 13B parameters
  - 4 way tensor parallel, 2 way pipeline parallel
  - Total size of checkpoint files == 172GB split across 8 files
  - Distributed training with 128 nodes

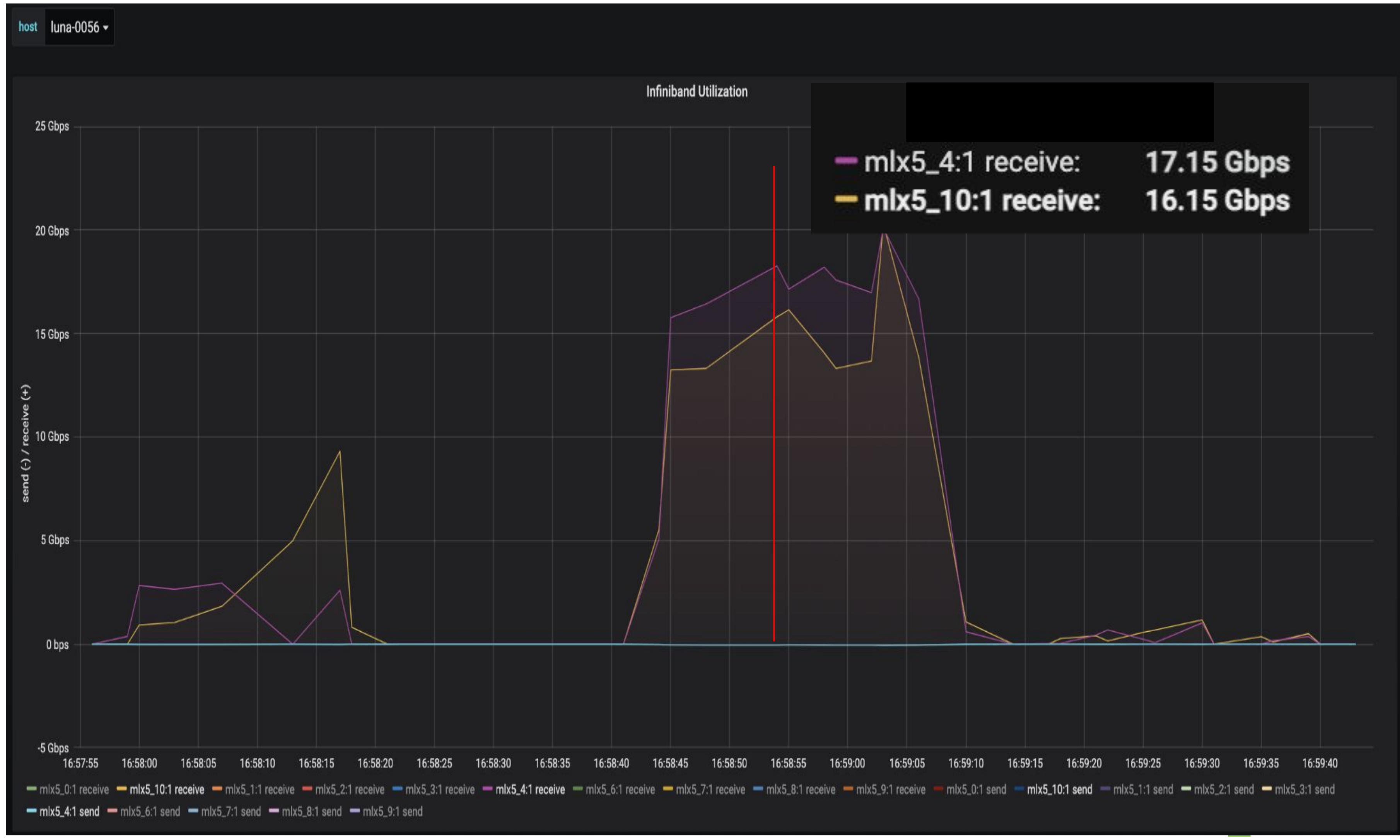
Real Selene Workload: Megatron-LM

GPT3 13B: Initial read of data

128 clients
256 rails
1024 GPUS

- Peak of ~250GB/s data read from FS
  - Each compute node reading shared dataset and model checkpoints to initialize training



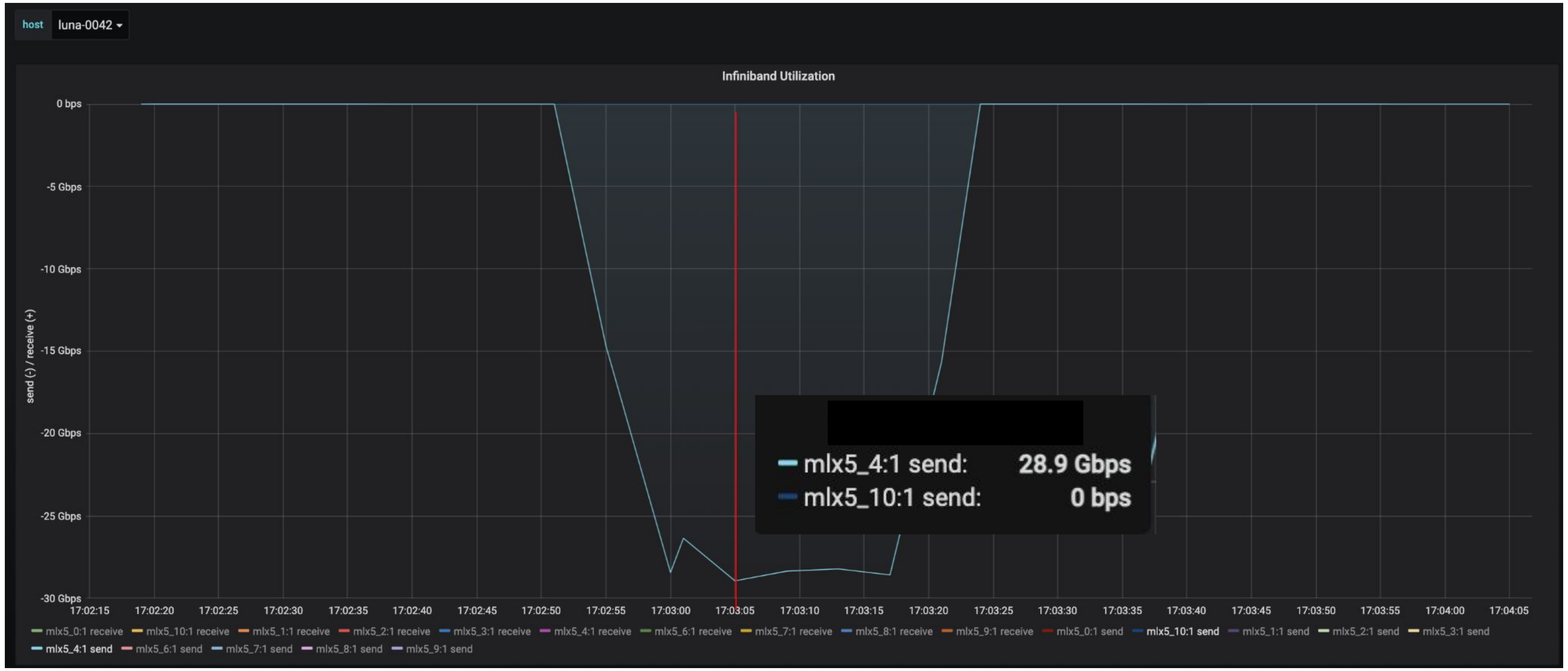


Real Selene Workload: Megatron-LM

GPT3 13B: Checkpointing

- ~7 GB/s data written to FS
  - Checkpoints being written to FS, from rank 0 of each data parallel instance

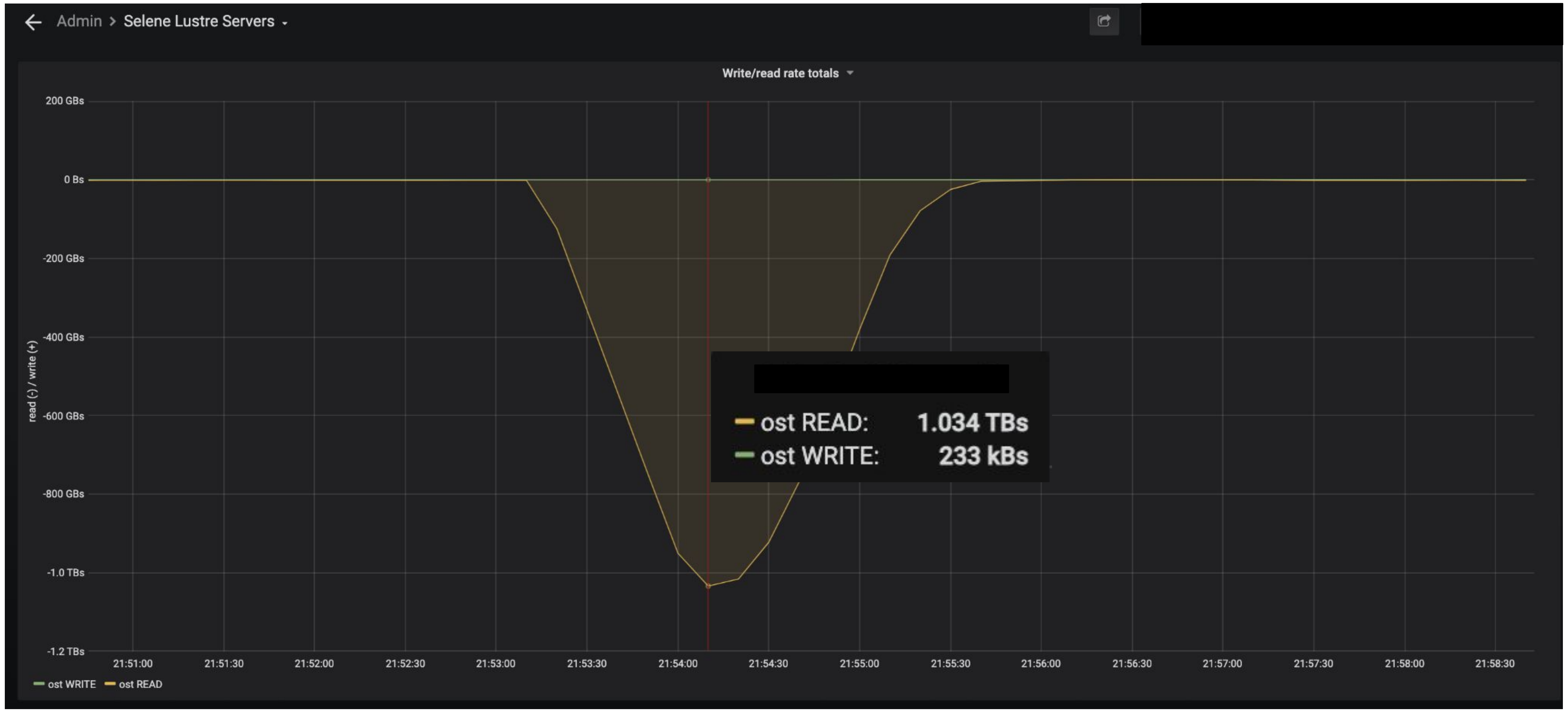
128 clients
256 rails
1024 GPUS



Real Selene Workload: Megatron-LM

Scaling up

Large scale runs of larger model variants can read 1 TB/s under normal production conditions



Client caching: a new feature for workload perf?

Client Caching

Context

The problem:

- DL workloads require reading the same datasets over and over and over again
- Manually copying datasets to local NVMEs (a.k.a. /raid) is a painful process for admins
- Users are not necessarily familiar with data transfer strategies, cost and time, wasting precious compute time

The idea:

- Each compute node can have a directory that can be used as Lustre cache (PCC extension, a.k.a. Hot nodes)
- All necessary datasets would be prepopulated only once in **read-only** /lustre/fsr and nodes would get a copy at first access
- The hit would be minimal at epoch0 and just limited by bandwidth once

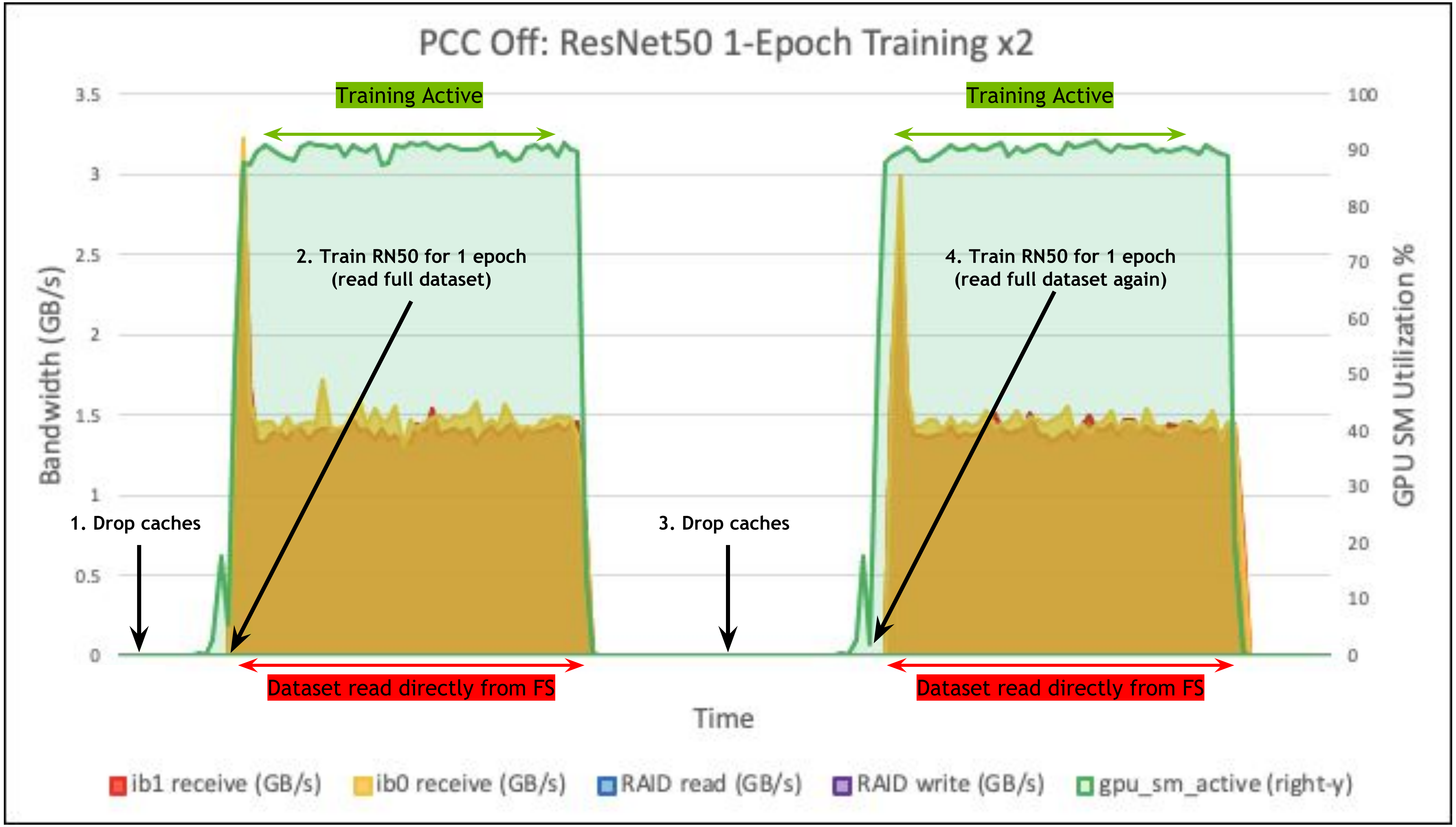
Client Caching DL Experiments

ResNet50 1N Training

- Objective: Run ResNet50 with/without PCC to simulate how it will impact our users (is it transparent?)
- Dataset dimensions: 6 files, 144GB total size
- 1N test methodology
  - Drop caches
  - Train RN50 for 1 epoch (read full dataset)
  - Drop caches
  - Train RN50 for 1 epoch (read full dataset again)
- Repeat the test using two versions of dataset on /lustre
  - Version of dataset not eligible for caching
  - Version of dataset eligible for PCC autocaching

Client Caching DL Experiments: PCC Off

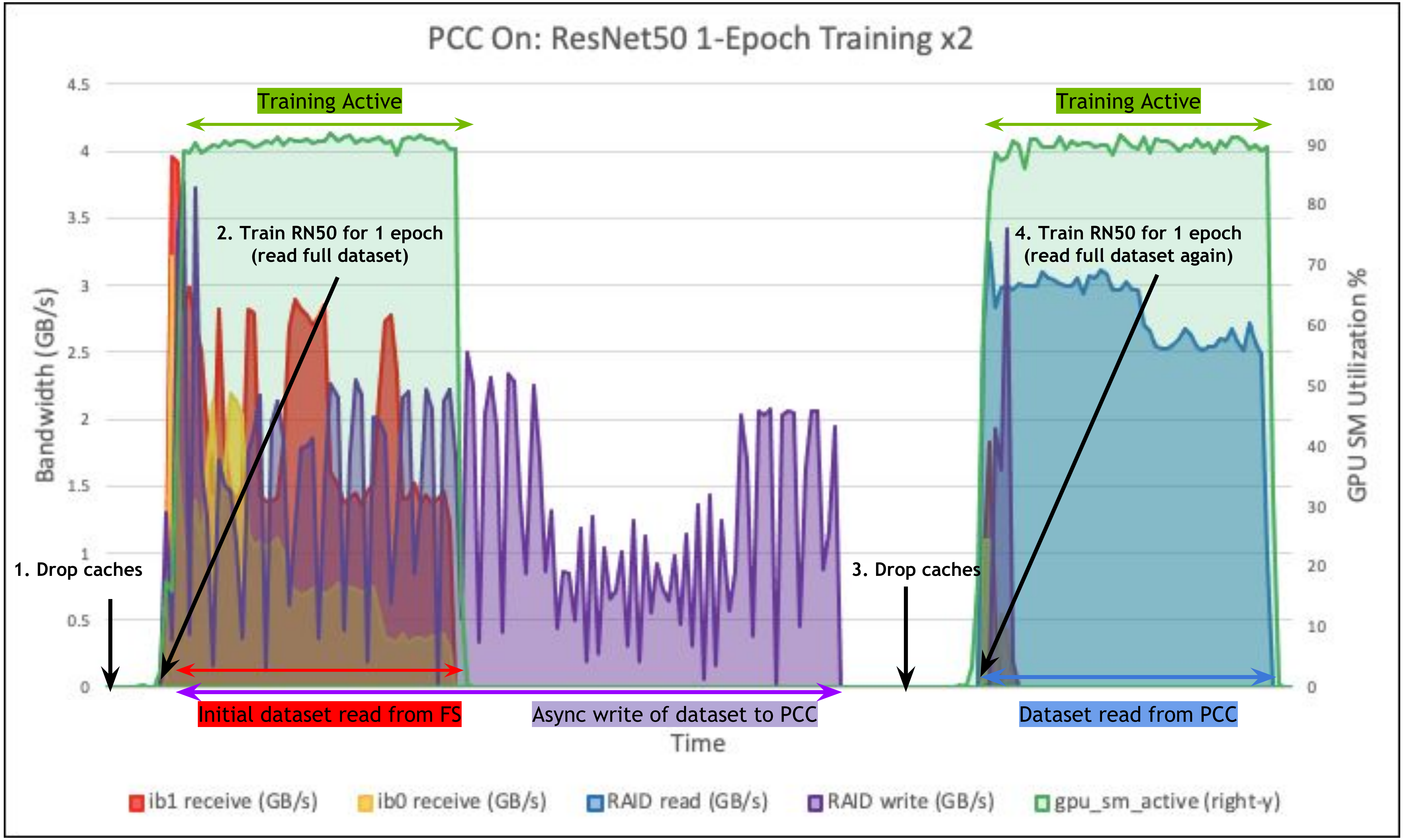
1. Drop caches
2. Train RN50 for 1 epoch (read full dataset)
3. Drop caches
4. Train RN50 again for 1 epoch



1 client
2 rails
8 GPUs

PCC on: Client Caching DL Experiments

1. Drop caches
2. Train RN50 for 1 epoch (read full dataset)
3. Drop caches
4. Train RN50 again for 1 epoch



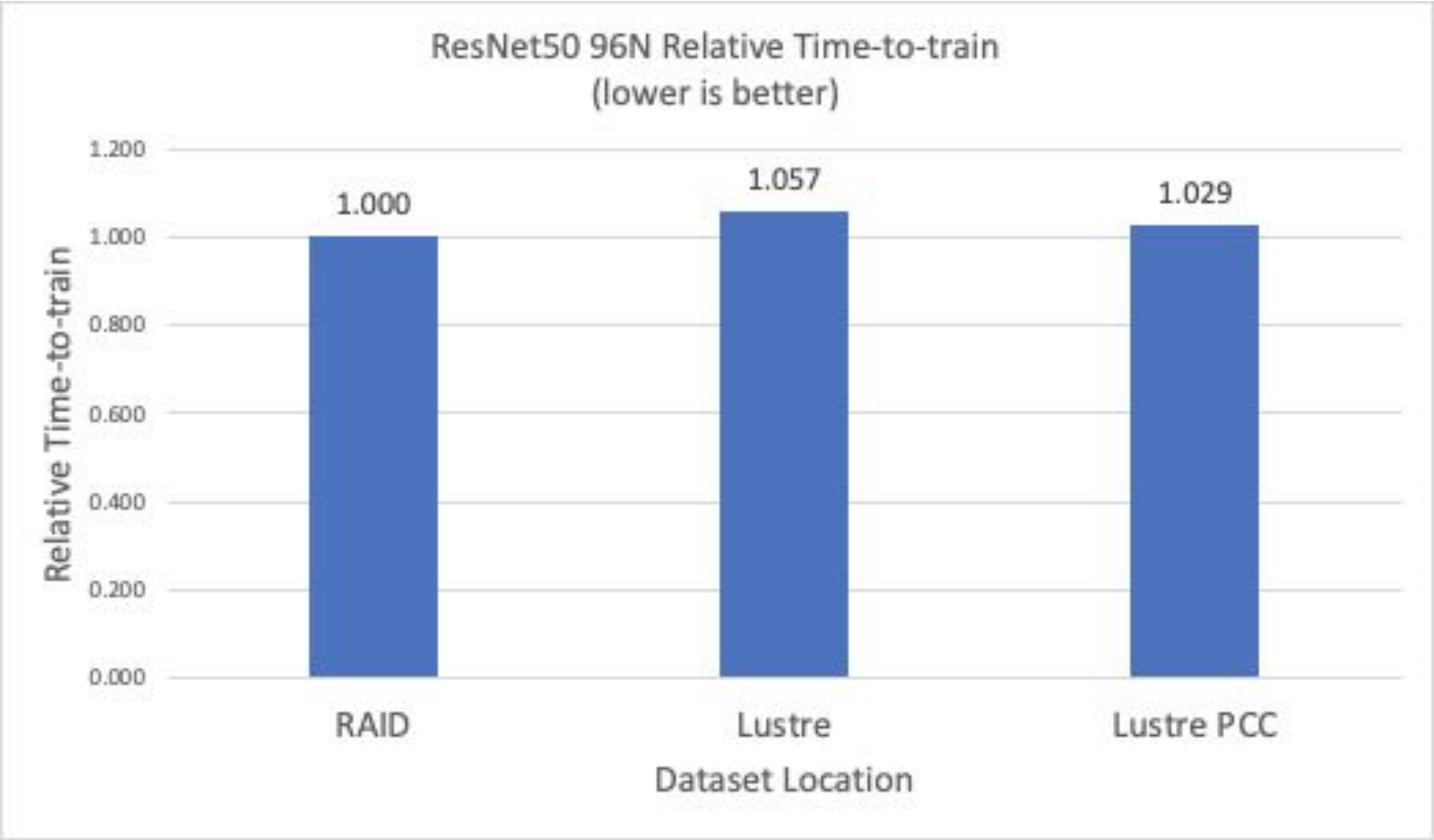
1 client
2 rails
8 GPUs

Multi-node Training with PCC

ResNet50 96N

96 client
192 rails
768 GPUs

- When using 96 clients simultaneously for ResNet50 training, having dataset cached in PCC provides close to 3% performance uplift compared to Lustre alone



NVIDIA.

Conclusion and links

Conclusion and links

Mission accomplished: 1TB/sec!

Solution can be implement on any Lustre setup.

A very flexible and simple solution for both cluster admins and users while providing performance.

Filesystem is reliable and relatively resilient to hardware failures (both fs and fabric with multi-rail).

Newer features are well suited for DL needs at perf.

Links:

- NVIDIA DGX A100 SuperPOD Announcement Blog
  https://blogs.nvidia.com/blog/2020/05/14/dgx-superpod-a100/
- DDN A3I Solutions for NVIDIA DGX A100 SuperPOD Reference Architecture
  https://www.nvidia.com/en-us/data-center/resources/ddn-a3i-reference-architecture/