Science and Technology Facilities Council
Scientific Computing
BASE-II
Blueprinting AI For Science at Exascale
Jan. , 2024

Submitting to MLPerf storage

Understanding Results

Louis Douriez
ldouriez@ddn.com

MLPerf™ Storage v0.5 - Workloads

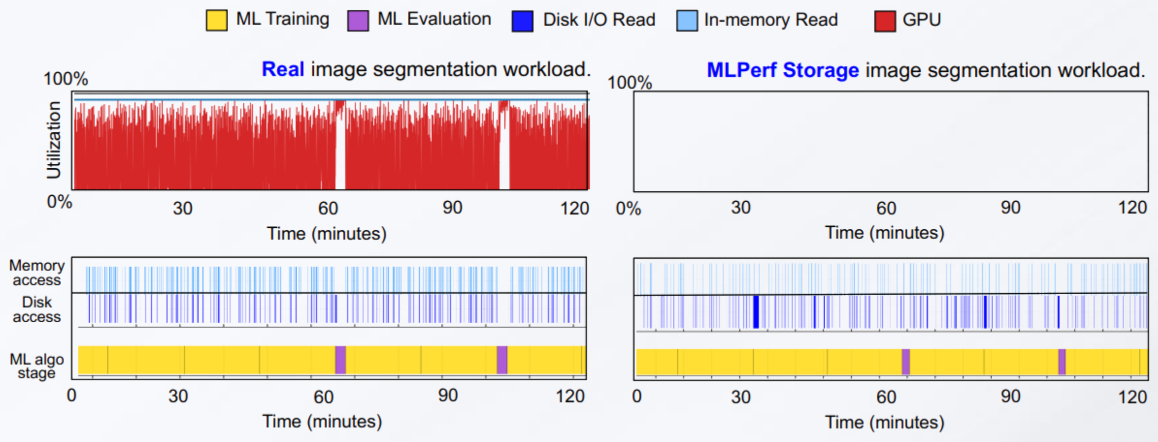
Number of simulated V100 GPUs for AI training

This benchmark suite measures how fast storage systems can supply training data when a model is being trained.

Each workload supported by MLPerf Storage is defined by a corresponding MLPerf Training benchmark. There are two workloads, 3D UNET and BERT-large.

Area	Task	Model	Nominal Dataset (see below)
Vision	Medical image segmentation	3D UNET	KITS 2019 (602x512x512)
Language	Language processing	BERT-large	SQuAD v1.1 (max_seq_len=384)

From Characterizing Machine Learning I/O with MLPerf Storage
Oana Balmau - CHEOPS @ EuroSys, May 8th, 2023

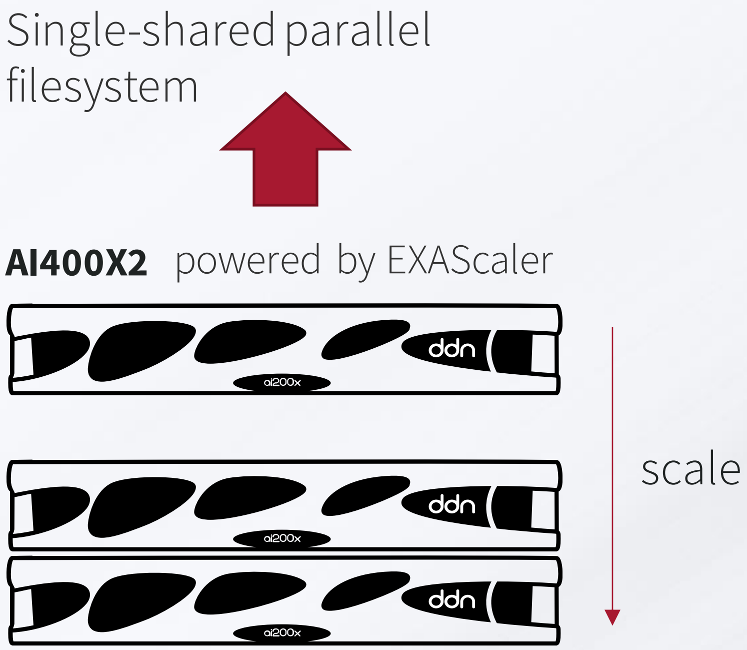


DDN Storage submitted

DDN all-flash appliance

The AI400X2 appliance is a fully integrated and optimized shared data platform with predictable capacity, capability, and performance. The all-NVMe configuration provides optimal performance for a wide variety of workload and data types and ensures that DGX POD operators can achieve the most from at-scale GPU applications, while maintaining a single, shared, centralized data platform.

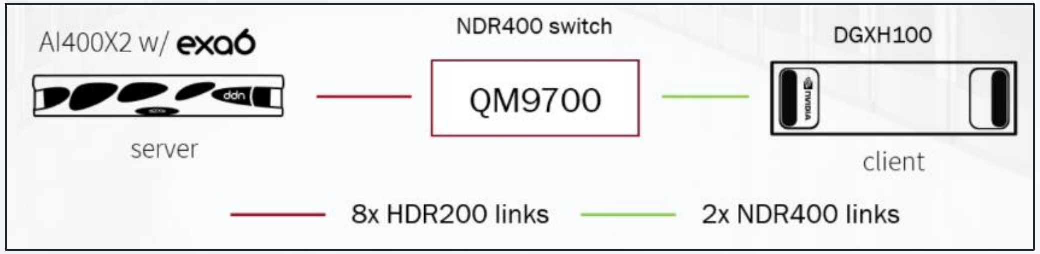
- Every AI400X2 appliance delivers over 90 GB/s and 3M IOPS directly to DGX H100 systems in a DGX SuperPOD.
- Shared performance scales linearly as additional AI400X2 appliances are integrated to the DGX SuperPOD.



MLPerf™ Storage v0.5 - Systems

DDN submitted 2 systems

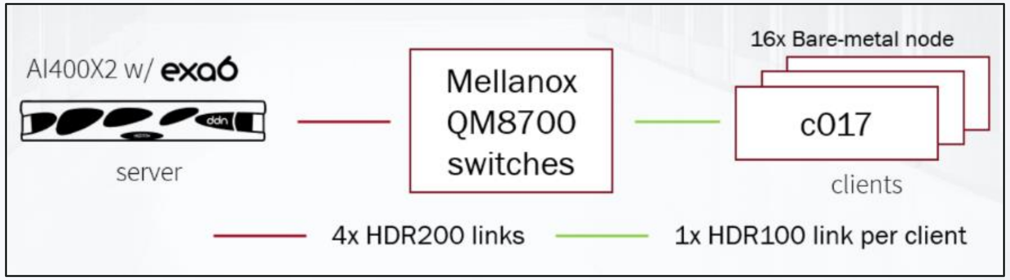
AI400X2_24x13.9TiB_nvme_8xHDR200



Single client system

https://github.com/mlcommons/storage_results_v0.5/blob/main/closed/DDN/systems/AI400X2_24x13.9TiB_nvme_8xHDR200.pdf

AI400X2_24x3.5TiB_nvme_4xHDR200



Multi-clients system

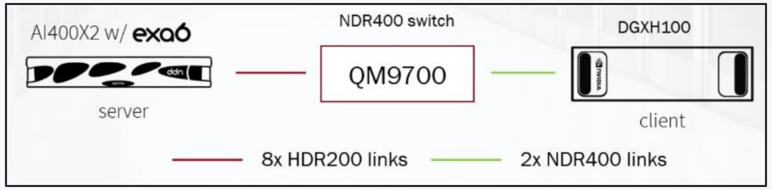
https://github.com/mlcommons/storage_results_v0.5/blob/main/closed/DDN/systems/AI400X2_24x3.5TiB_nvme_4xHDR200.pdf

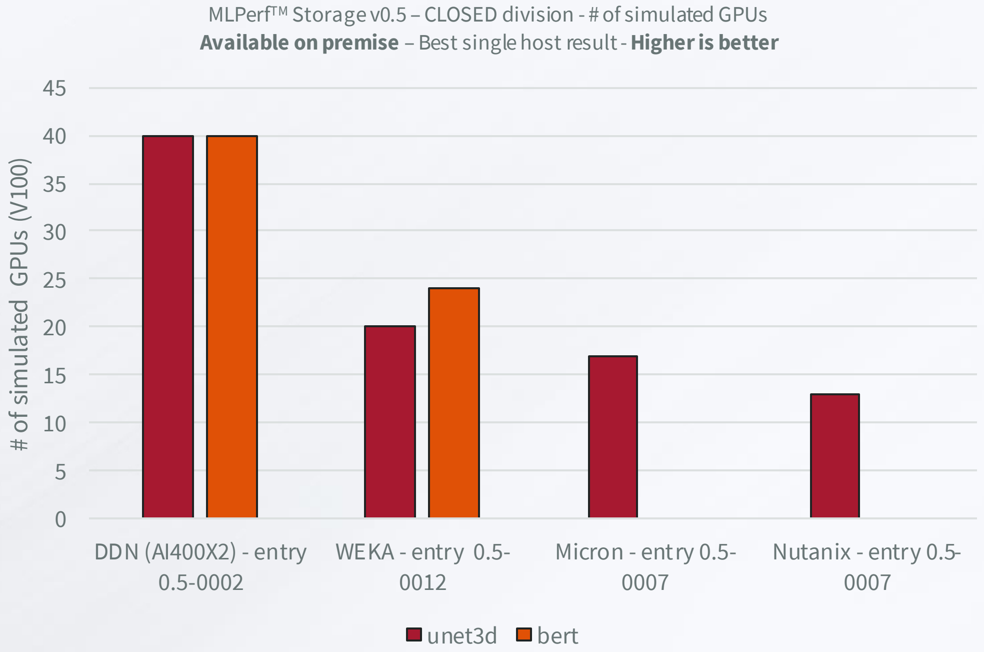
THE AI DATA COMPANY

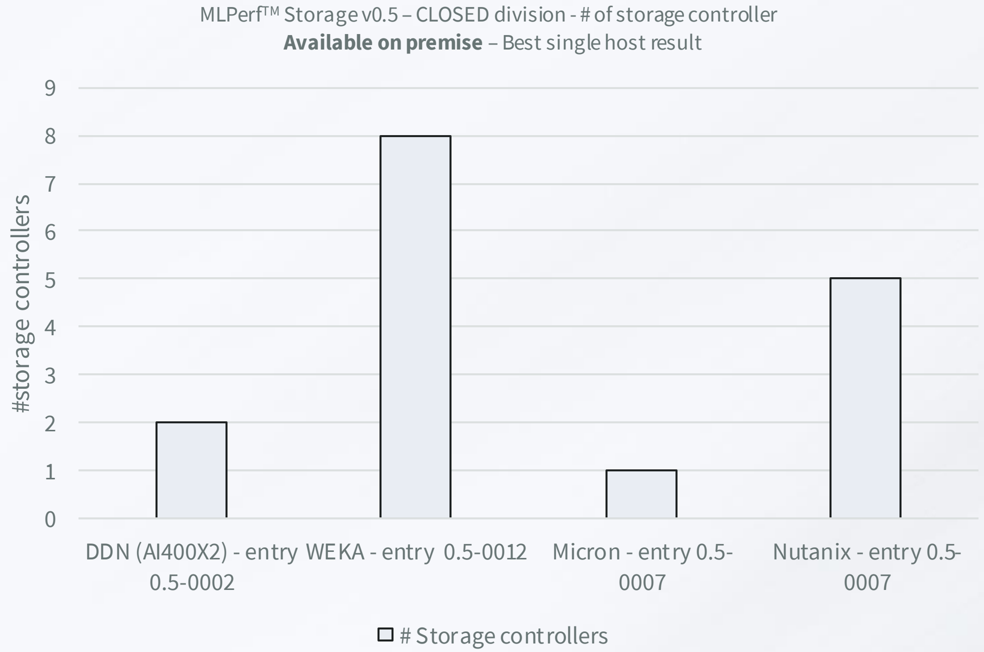
MLPerf™ Storage v0.5 - Results

Training simulation – Single Host - CLOSED

DDN system AI400X2_24x13.9TiB_nvme_8xHDR200[1]
Single client – Single AI400X2





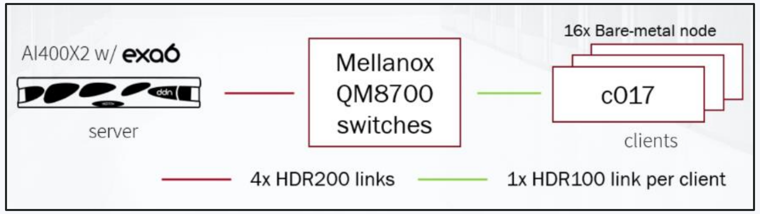


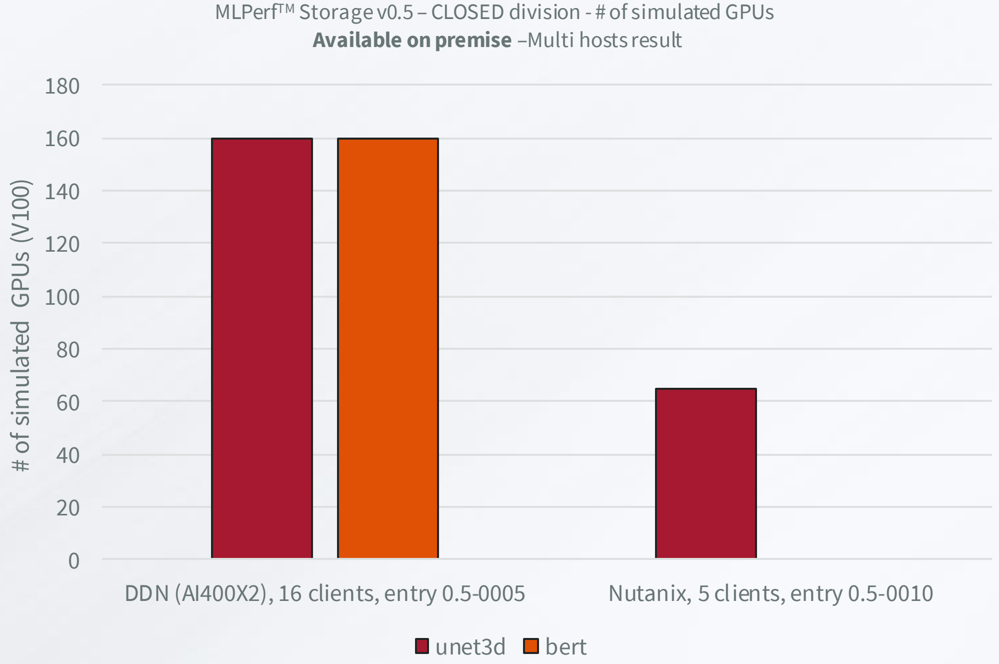
[1] MLPerf™ Storage v0.5 Closed. Retrieved from https://mlcommons.org/en/storage-results-05/ 11 September 2023, entry 0.5-0002. Result verified by MLCommons Association. The MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. See www.mlcommons.org for more information.

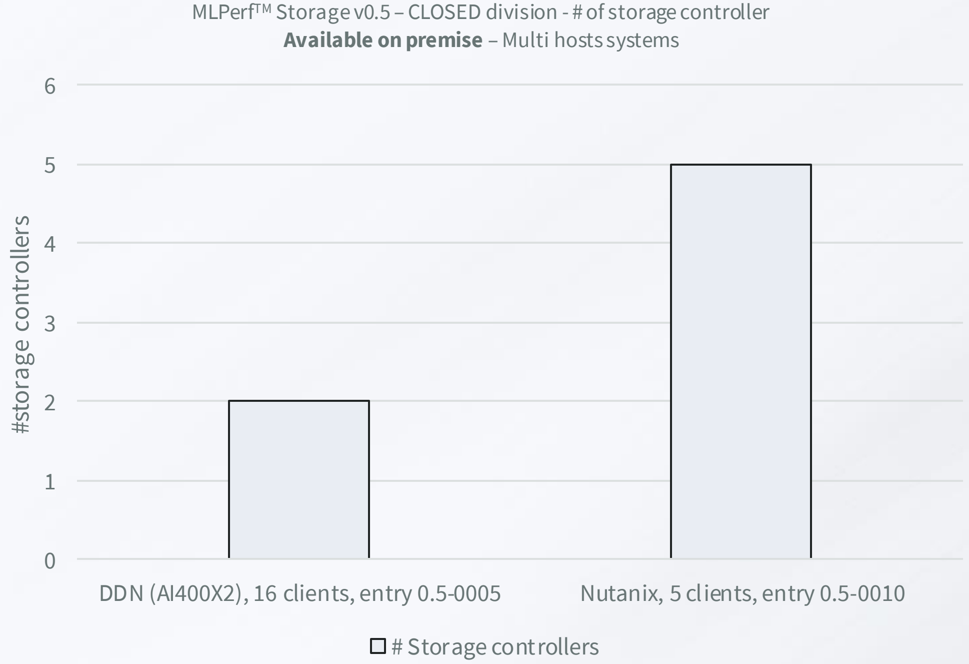
MLPerf™ Storage v0.5 - Results

Training simulation – Multiple hosts -CLOSED

DDN system AI400X2_24x3.5TiB_nvme_4xHDR200[1]
Multiple client – Single AI400X2







[1] MLPerf™ Storage v0.5 Closed. Retrieved from https://mlcommons.org/en/storage-results-05/ 11 September 2023, entry 0.5-0005. Result verified by MLCommons Association. The MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

From LLM Workload to Infrastructure Consideration

The Rise of Data

Jean-Thomas Acquaviva
jtacquaviva@ddn.com

Louis Douriez
ldouriez@ddn

DDN and NVIDIA SuperPOD Top 10!

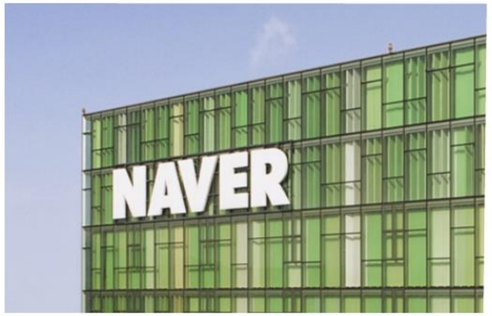
NVIDIA Eos
World's Fastest SuperPOD

NEC AI Research
Japan

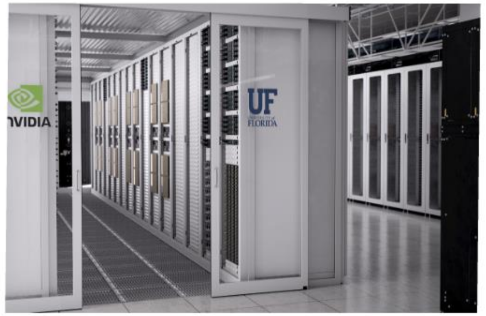
Berzelius
Linköping University

Cambridge-1
UK Life Sciences

PARAM Siddhi AI
India Research and R&D



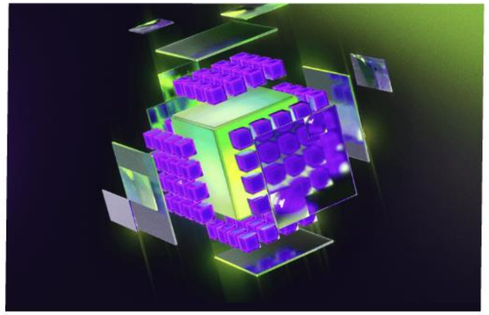
NAVER AI Cloud
South Korea AI Services



HiPerGator
Uni of Florida

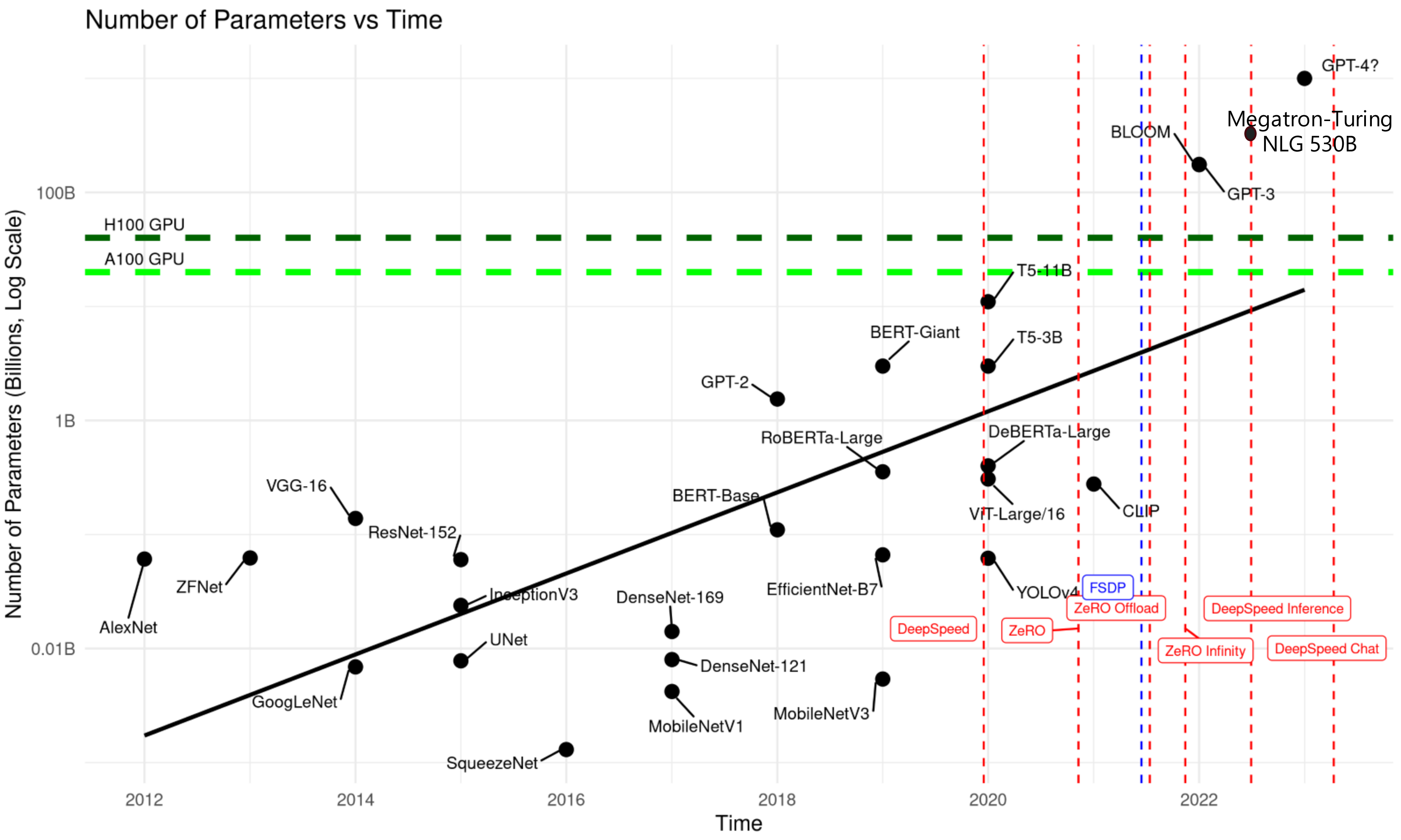
Lambda
US Cloud SuperPOD

NVIDIA Selene
World's First SuperPOD



Scaleway
Europe Cloud SuperPOD

Parameters in LLM over Years: Exponential Growth



In the last 3 years

Model size x1000

GPU memory x5

Log Scale on Y

LLM Memory Consumption: training is demanding

Considering Ψ, the model size expressed in number of parameters

- **For Inference Memory consumption is 2xΨ Byte**
  - **17B model requires 34GB of memory to run**

- **For Training Memory pressure depends on:**
  - **Parameters, half precision, 2xΨ Byte**
  - **Gradient, half precision, 2 x Ψ Byte**
  - **Optimizers states, 3 states single precision 12 x Ψ Byte**

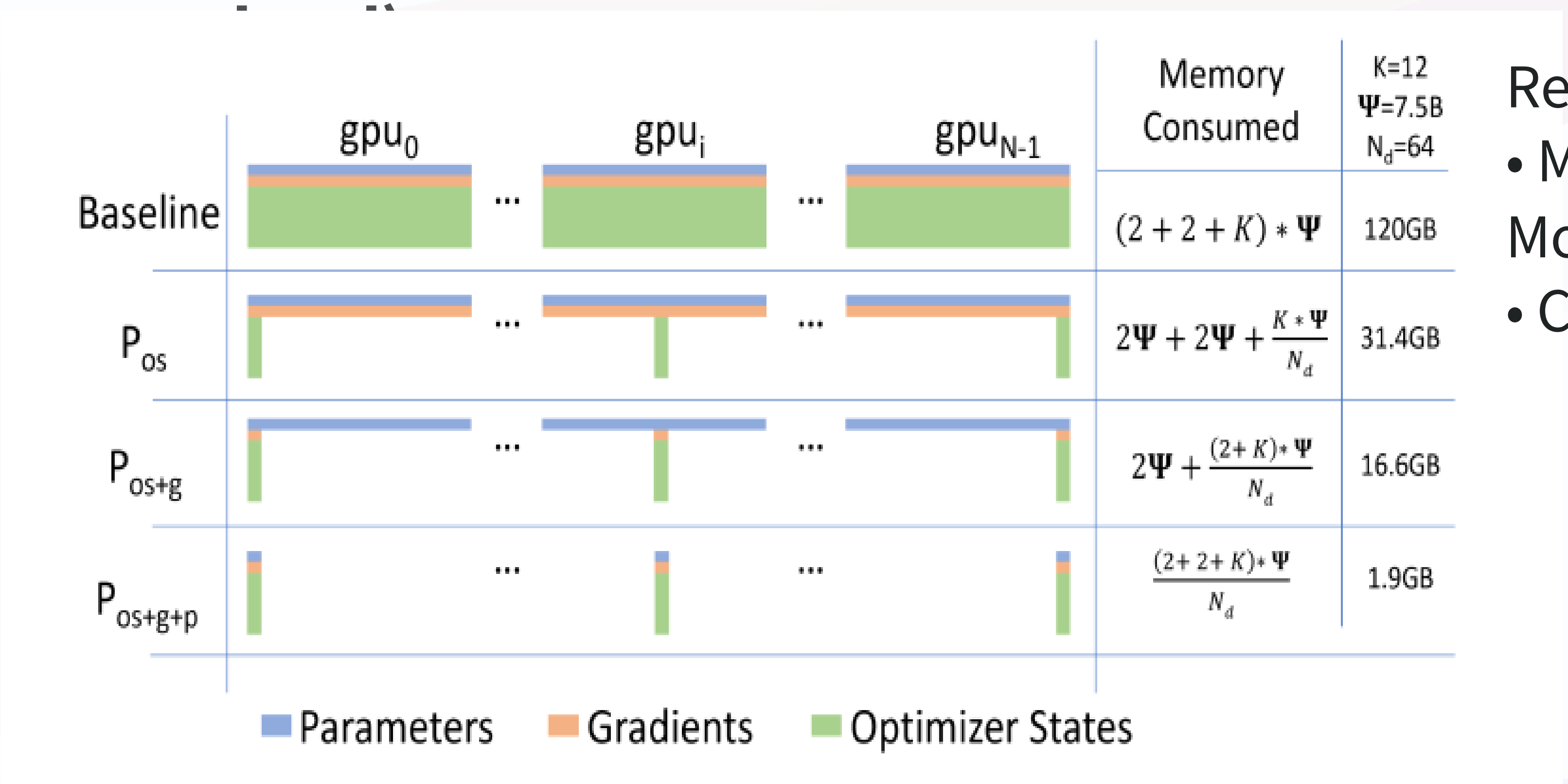
The total amount of memory needed:

Byte needed = 16 x number of parameters

A 17B parameters model = 272 GB of memory: Not available even on latest GPUs

LLM Memory Offloading: Zero [2020]

ZeRO: framework from Microsoft interleaving parallelization schemes to minimize memory footprint (at the cost of some communication



Reduction of memory footprint

- Mixture of Data Parallelism, Model and Pipeline parallelism
- Cap communication overhead

LLM Walking Around the Memory Wall?

Harness multiple GPUs to aggregate their memory

- Efficiency of the transistor budget process?
  - Burning Logic to get Memory
- Require multiple GPUs to perform aggregation
  - Limit investigation on LLM to organizations with consequent infrastructure
- LLMs are large data structure with uneven access, temporal locality exists

LLM Memory Offloading: ZeRO Infinity [2021]

Resurrect Out-of-Core computing

Zero to Infinity, extension of the ZeRO model

Model's parameters, gradients and optimizers states are not offloaded to remote GPUS, but either to CPU memory, local storage and remote storage

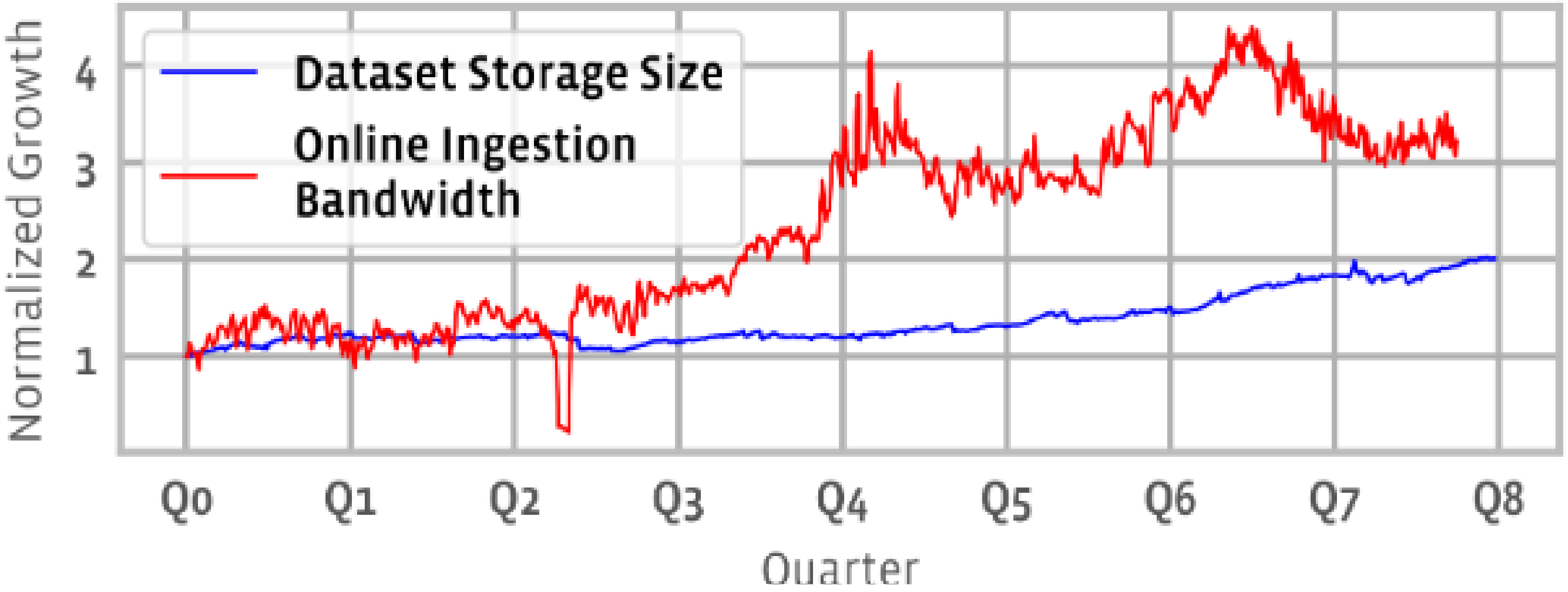
Offloading is an emerging topic: e.g. Hugging Face Accelerate, FlexGen*

*FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU

LLM and Storage: Bandwidth is already the key resource

Bandwidth requirement is growing faster than capacity.

- Bandwidth x4 over 2 years
- Capacity x2 over 2 years



- Number of parameters in model increases x2 than the number of Token in data sets

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

LLaMA: Open and Efficient Foundation Language Models

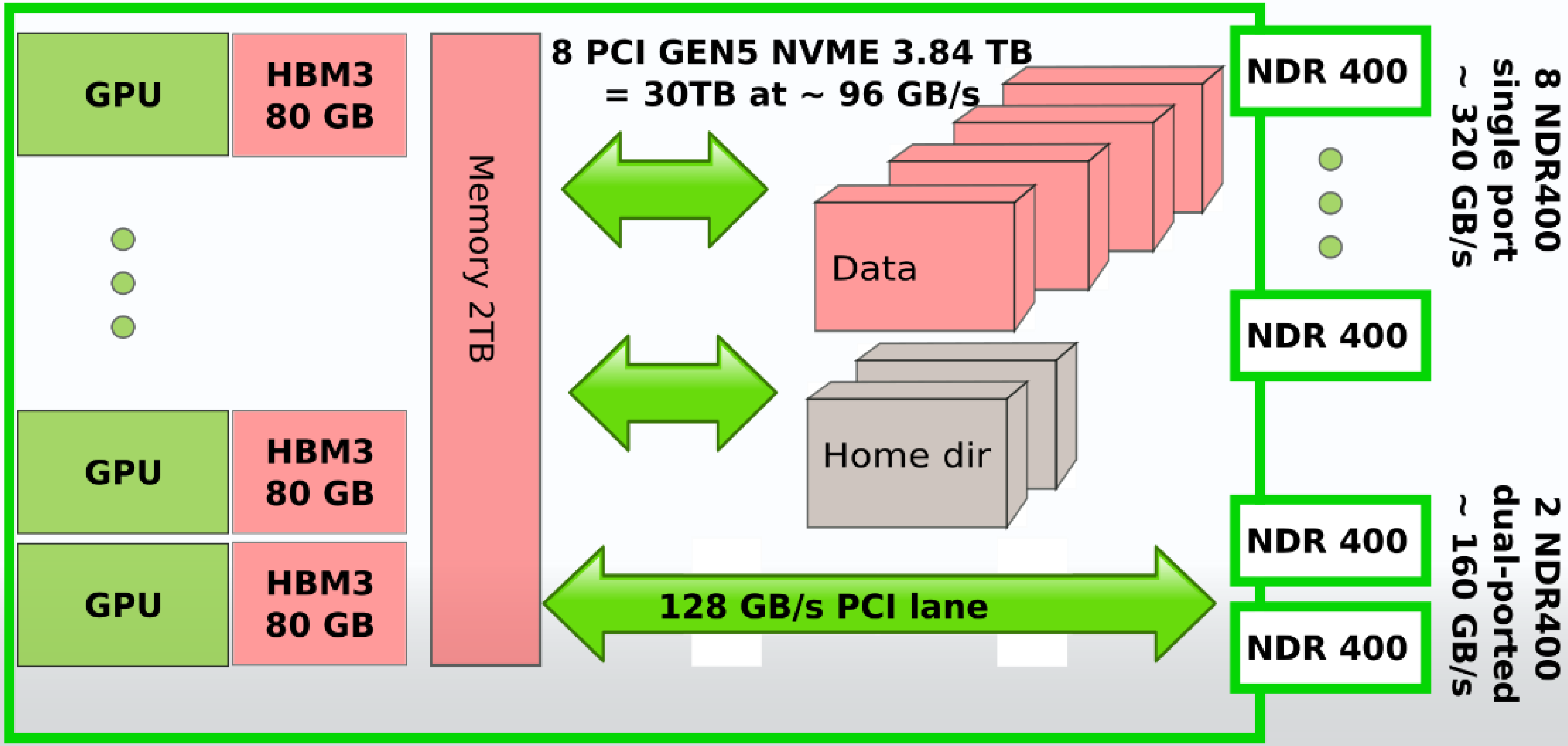
LLM and Storage: Some Subtleties

		GOOD		BETTER		BEST	
		NLP where, data fits in local memory		LLM training with text, audio & image data		LLM training with video, audio, image & text	
#HGX Servers	**# GPUs**	**Read/GPU**	**Write/GPU**	**Read/GPU**	**Write/GPU**	**Read/GPU**	**Write/GPU**
1	8	0.5 GB/s	0.25 GB/s	1.5 GB/s	0.75 GB/s	5 GB/s	2.5 GB/s

NVIDIA Default Recommendation

Additionally, NVIDIA Cloud Reference Architecture recommends **1.3 GB/s to 1.8 GB/s write performance per GPU** for the LLM training use case.

DGX Memory Hierarchy



Two memory levels

- *80 GB per GPU*
- *2TB shared with CPU*

Two storage levels

- *PCI Gen 5 local NVMes*
- *2 NDR400 IB slots for network attached storage.*

Memory requirements per model size

LLM size is getting in parallel filesystem territory

	~LLAMA2	~GPT3	~GPT4	Future
MODEL NAME	BLOOM 7B1	BLOOM 176B	BLOOM-mod-1	BLOOM-mod-2
#Hidden Layers	30	70	960	4800
Hidden Size	4096	14336	10240	20480
# Attention Heads	32	112	16	16
Batch Size Used	32	16	8	2
# Parameters	7.1B	176B	1.2T	24.1T
Memory Needed to Fit the Model for Inference (GB)	3.5	350	2300	$\boxed{44000}$

44TB = 550 H100 for inference
~6000 for training

Inference Experimental Results with BLOOM LLM

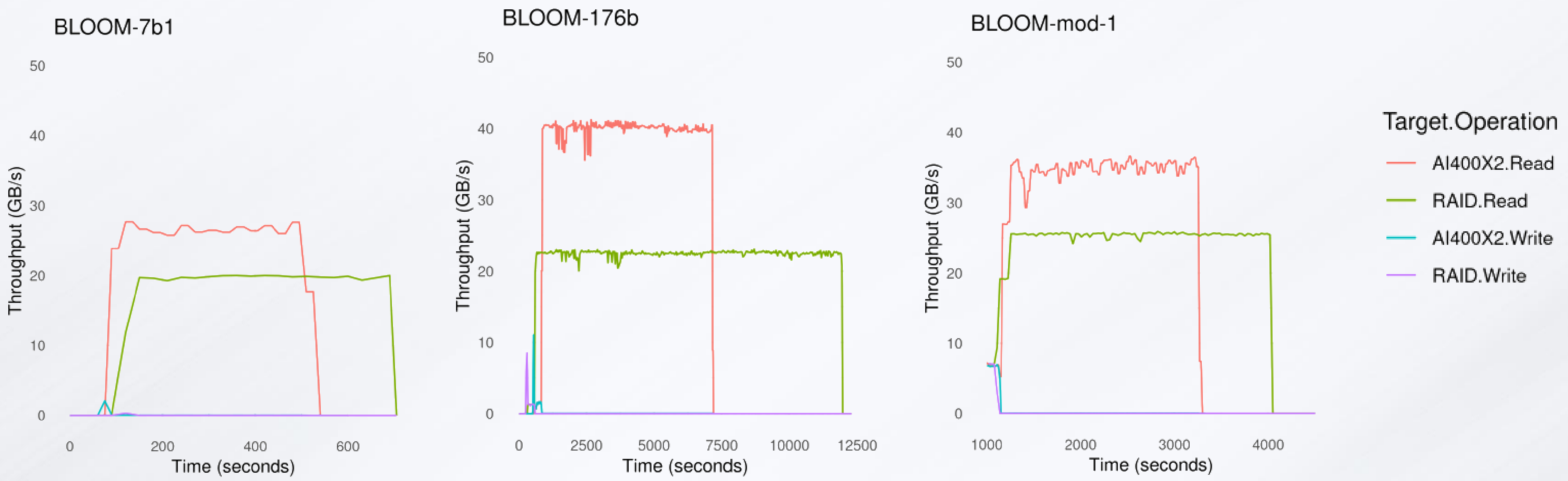
ExaScaler competitive with CPU Offloading

Outperforming consistently Local Storage

IO size? ~ 2GB
Inference offloading is Bandwidth Driven

LLM Offloading performance

A throughput problem



LLM and Storage: Take-Away

- Offloading of models' data to the ExaScaler alleviates complexity and maintains GPU efficiency
  - ExaScaler scales to hundreds of PetaByte, thus removing memory issues from the design consideration and complexity equation.
- Experimentations and measurements for inference are feasible
- Training is extremely expensive to measure: MLPerf Storage

LLM and Storage: foreseeable future

- The coming generation of LLMs will put even more stress on the infrastructure
  - Bandwidth: Training a hundred-trillion parameter LLM is feasible but requires a secondary memory pool up to 1 TiB per GPU with a bandwidth of 100 GB/s bidirectionally [ISCA23]
  - Capacity: AI driven Data-Sets generation will lead to additional capacity pressure
  - Growing models means growing need for Checkpointing: Write importance will rise.
  - Current ML Workload are already 50/50 Read Write [Mascost 21]

Some References

- [BRO20] BROWN, Tom, MANN, Benjamin, RYDER, Nick, *et al.* Language models are few-shot learners. *Advances in neural information processing systems*, 2020, vol. 33, p. 1877-1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
- [SCA22] SCAO, Teven Le, FAN, Angela, AKIKI, Christopher, *et al.* Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. https://arxiv.org/pdf/2211.05100
- [RAJ20] RAJBHANDARI, Samyam, RASLEY, Jeff, RUWASE, Olatunji, *et al.* Zero: Memory optimizations toward training trillion parameter models. In : *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2020. p. 1-16. https://arxiv.org/pdf/1910.02054.pdf%3E
- [ZHA22] ZHAO, Mark, AGARWAL, Niket, BASANT, Aarti, *et al.* Understanding data storage and ingestion for large-scale deep recommendation model training: Industrial product. In : *Proceedings of the 49th Annual International Symposium on Computer Architecture*. 2022. p. 1042-1057. https://arxiv.org/pdf/2108.09373.pdf
- [MLP22] ML Perf Storage, https://mlcommons.org/en/groups/research-storage/
- [BAL23] Characterizing I/O Patterns in Machine Learning ACM Workshop on Challenges and Opportunities of Efficient and Performant Storage Systems. O. Balmau. https://sigmodrecord.org/publications/sigmodRecord/2209/pdfs/09_Dbrainstorming_Blamau.pdf
- [UMO23] UM, Taegeon, OH, Byungsoo, SEO, Byeongchan, *et al.* FastFlow: Accelerating Deep Learning Model Training with Smart Offloading of Input Data Pipeline. *Proceedings of the VLDB Endowment*, 2023, vol. 16, no 5, p. 1086-1099. https://www.vldb.org/pvldb/vol16/p1086-um.pdf
- [RAJ21] RAJBHANDARI, Samyam, RUWASE, Olatunji, RASLEY, Jeff, *et al.* Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In : *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021. p. 1-14.
- [RAJ22] RAJBHANDARI, Samyam, LI, Conglong, YAO, Zhewei, *et al.* Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In : *International Conference on Machine Learning*. PMLR, 2022. p. 18332-18346. https://proceedings.mlr.press/v162/rajbhandari22a/rajbhandari22a.pdf
- [KAP20] KAPLAN, Jared, MCCANDLISH, Sam, HENIGHAN, Tom, *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. https://arxiv.org/pdf/2001.08361.pdf
- [DGX23] NVIDIA DGX SuperPOD: DDN AI400X2 Appliance, Reference Architecture https://www.ddn.com/wp-content/uploads/2023/01/DDN-A3I-AI400X2-NVIDIA-DGX-A100-SuperPOD-Reference-Architecture.pdf
- [HAZ18] HAZELWOOD, Kim, BIRD, Sarah, BROOKS, David, *et al.* Applied machine learning at facebook: A datacenter infrastructure perspective. In : *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2018. p. 620-629. https://research.facebook.com/file/904032783795098/hpca-2018-facebook.pdf
- [ROL23] Is Bare-metal I/O Performance with User-defined Storage Drives Inside VMs Possible? ACM Workshop on Challenges and Opportunities of Efficient and Performant Storage Systems. S. Rolon, O. Balmau https://drive.google.com/file/d/1rnd76S0bttLBc6fWs6NUvR2dpHUVZ-zT/view?usp=share_link
- [ISCA23] ISAEV, Mikhail, MCDONALD, Nic, et VUDUC, Richard. Scaling Infrastructure to Support Multi-Trillion Parameter LLM Training. In : *Architecture and System Support for Transformer Models (ASSYST@ ISCA 2023)*. 2023
- [MASCOTS21] Characterizing Machine Learning I/O Workloads on Leadership Scale HPC Systems Arnab K. Paul† , Ahmad Maroof Karimi† , Feiyi Wang Oak Ridge National Laboratory, USA

Thank You!

Questions?