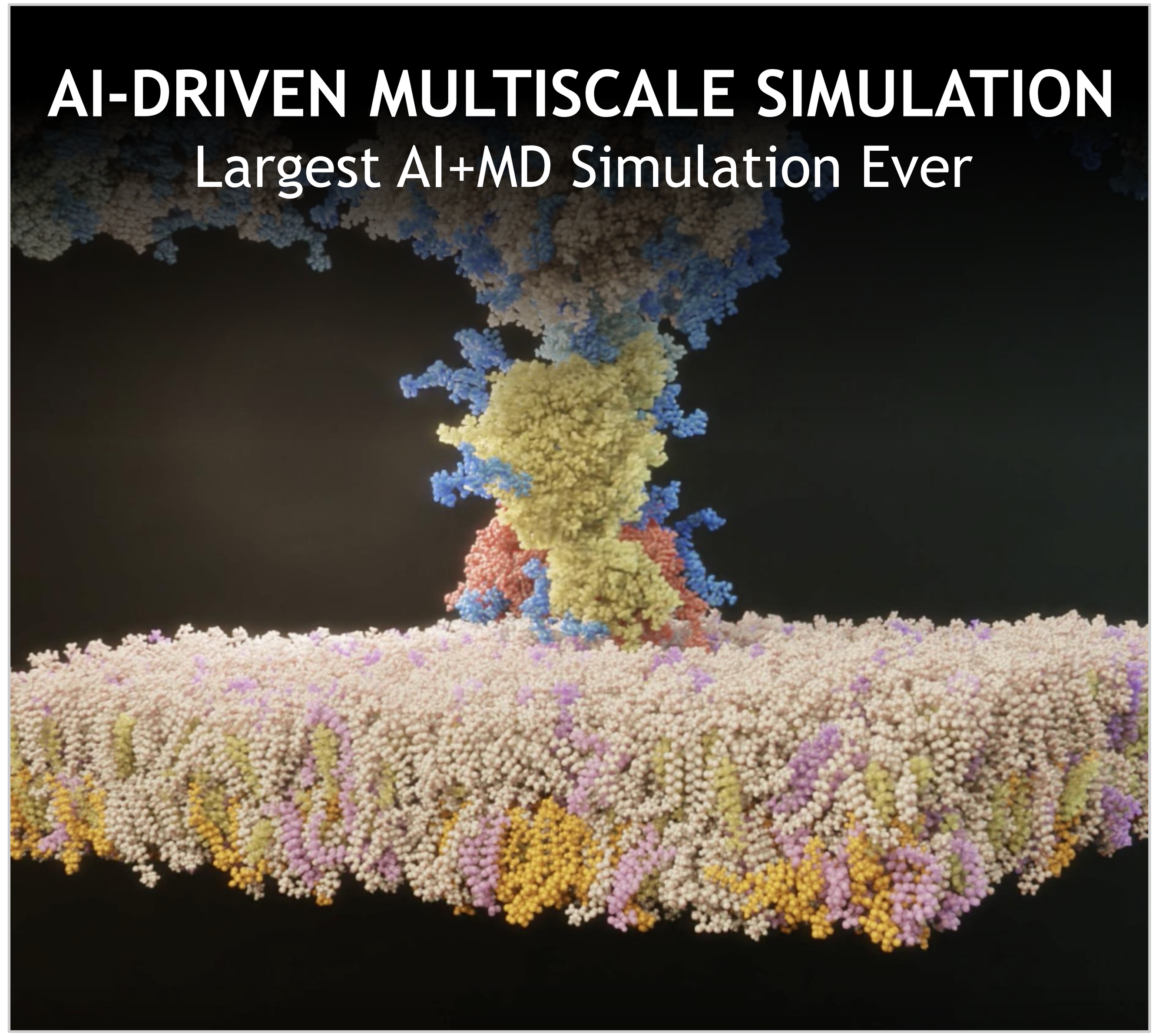
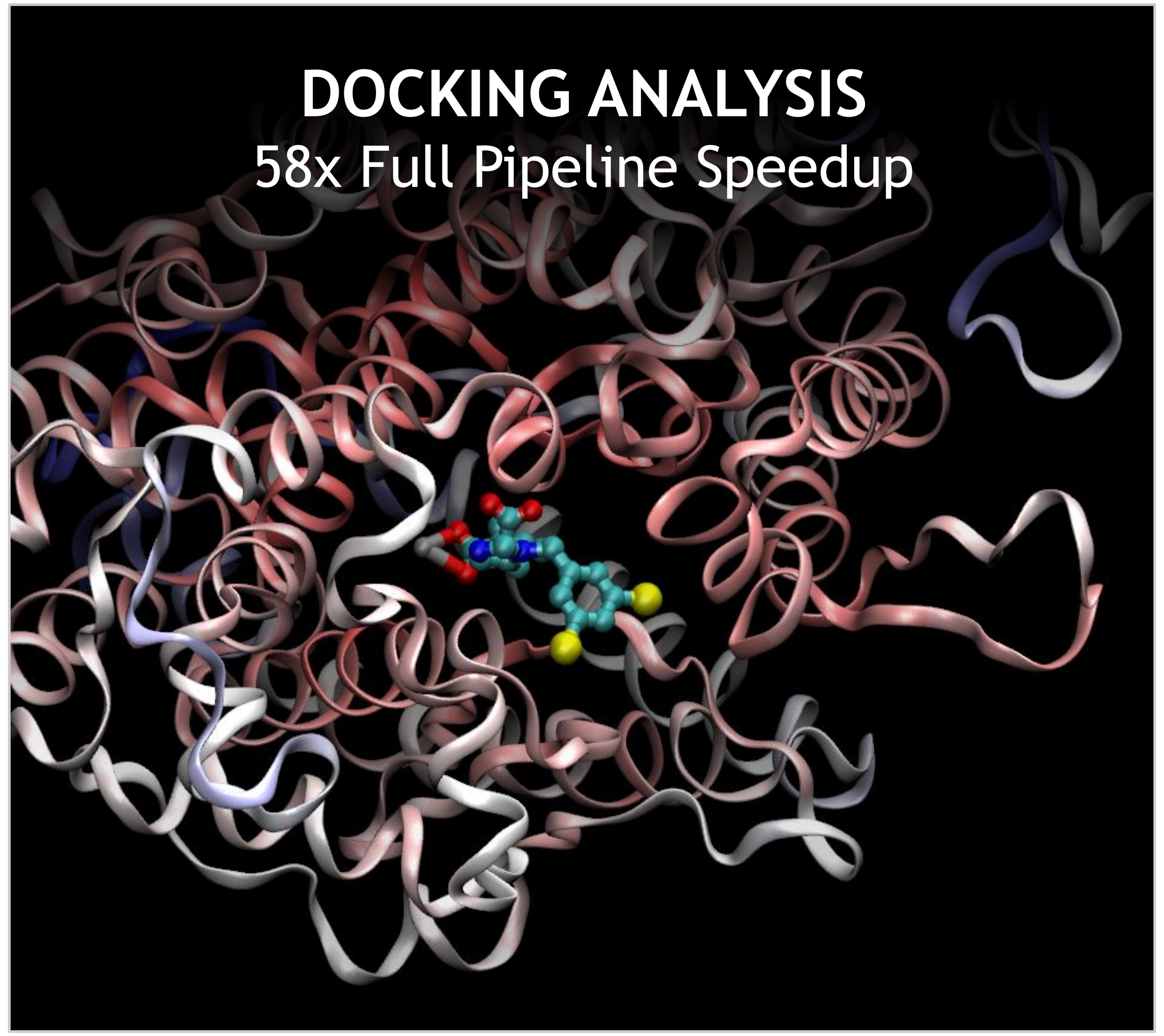
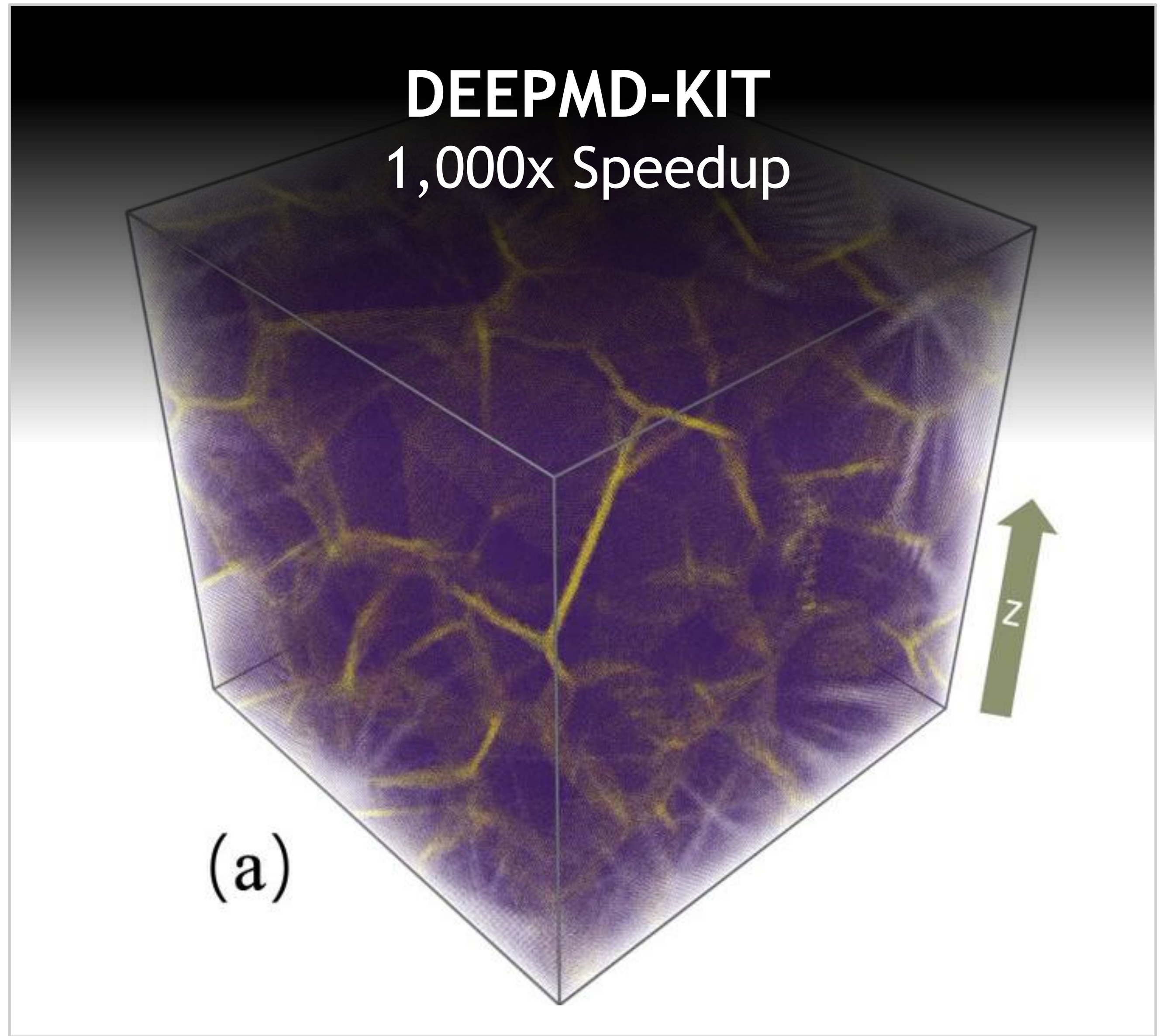
“BIG IRON” AI SYSTEMS

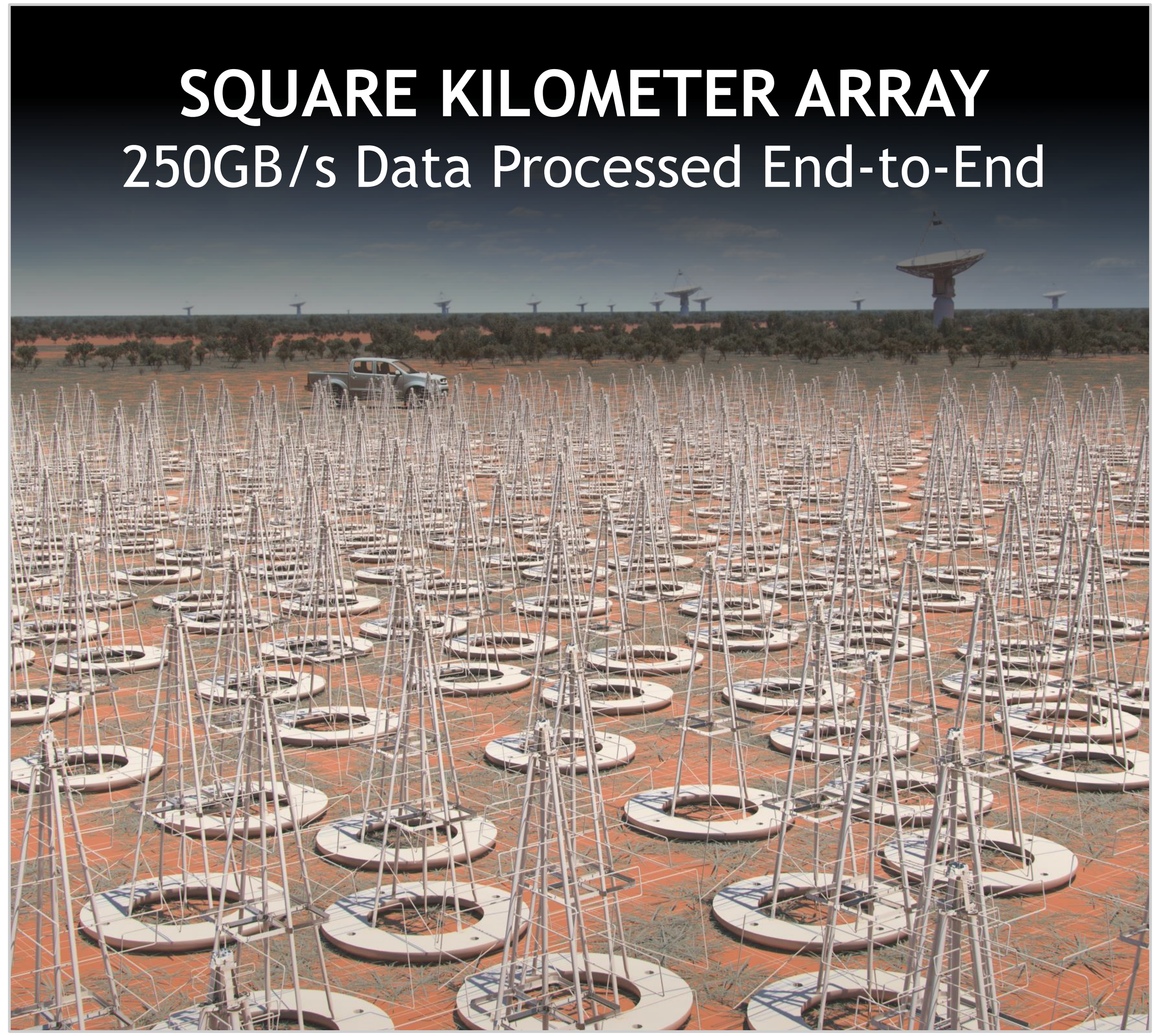
MIKE HOUSTON, VP AND CHIEF ARCHITECT OF AI SYSTEMS | FEB. 28, 2022

AI SUPERCOMPUTING DELIVERING SCIENTIFIC BREAKTHROUGHS

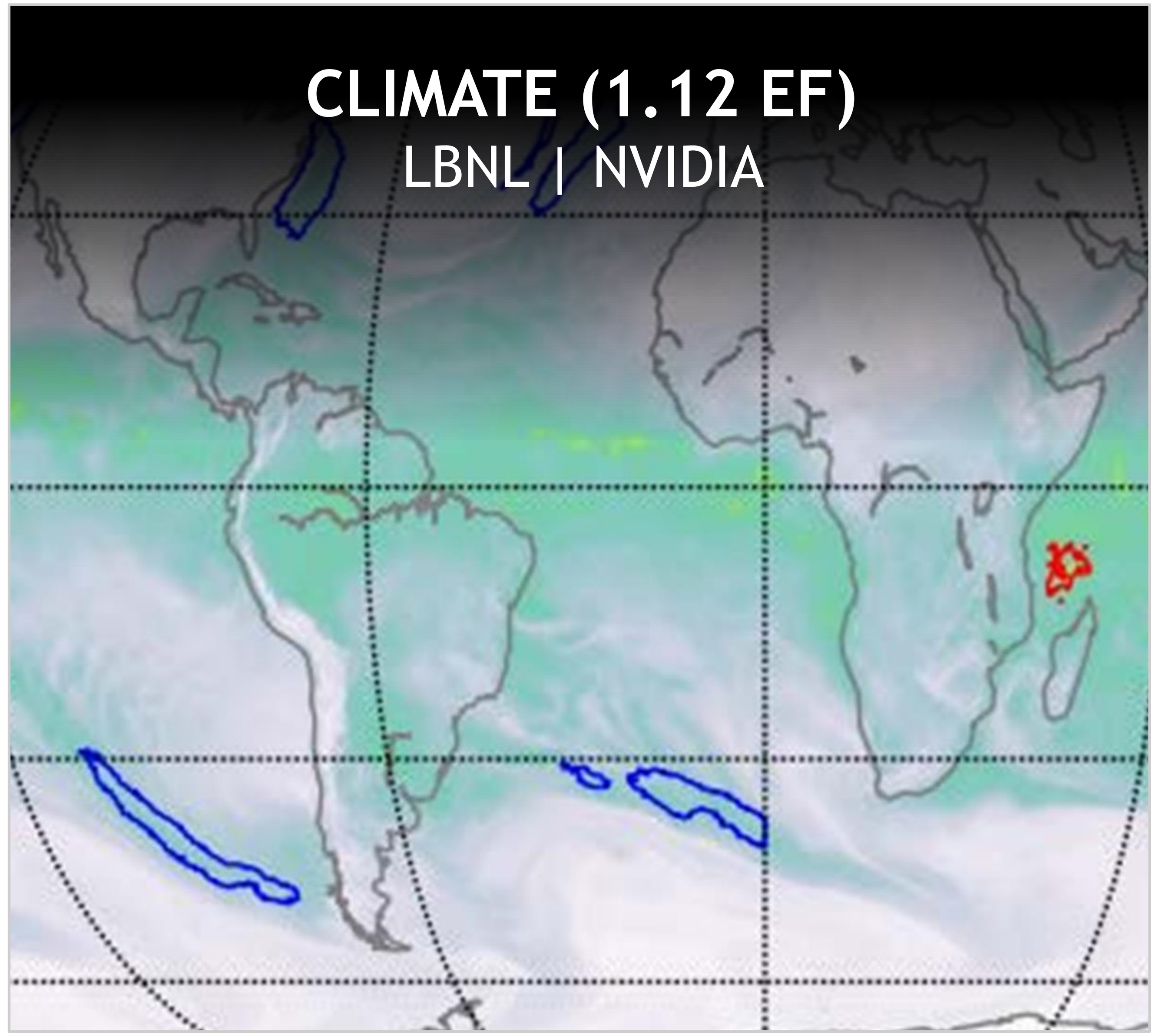


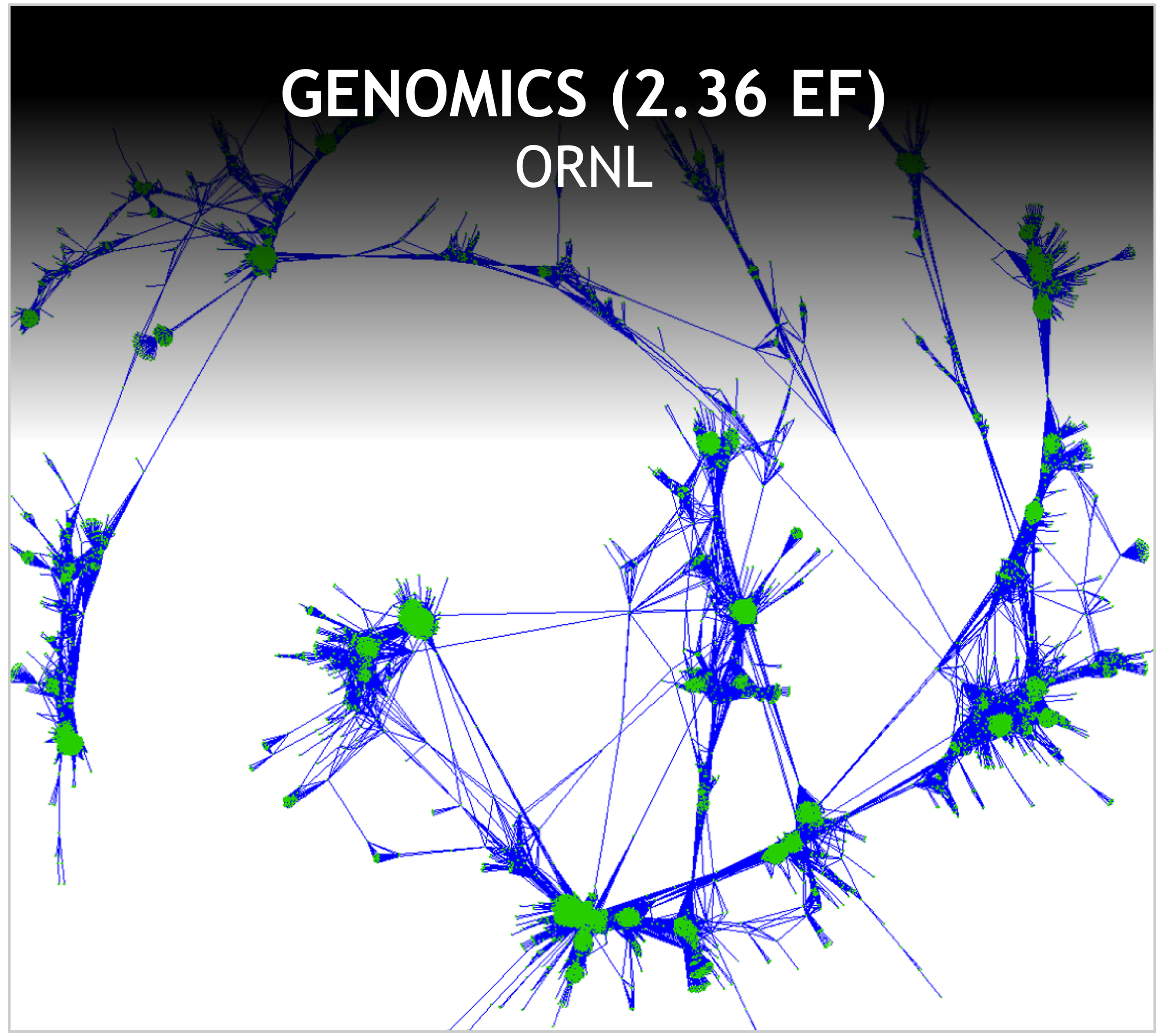


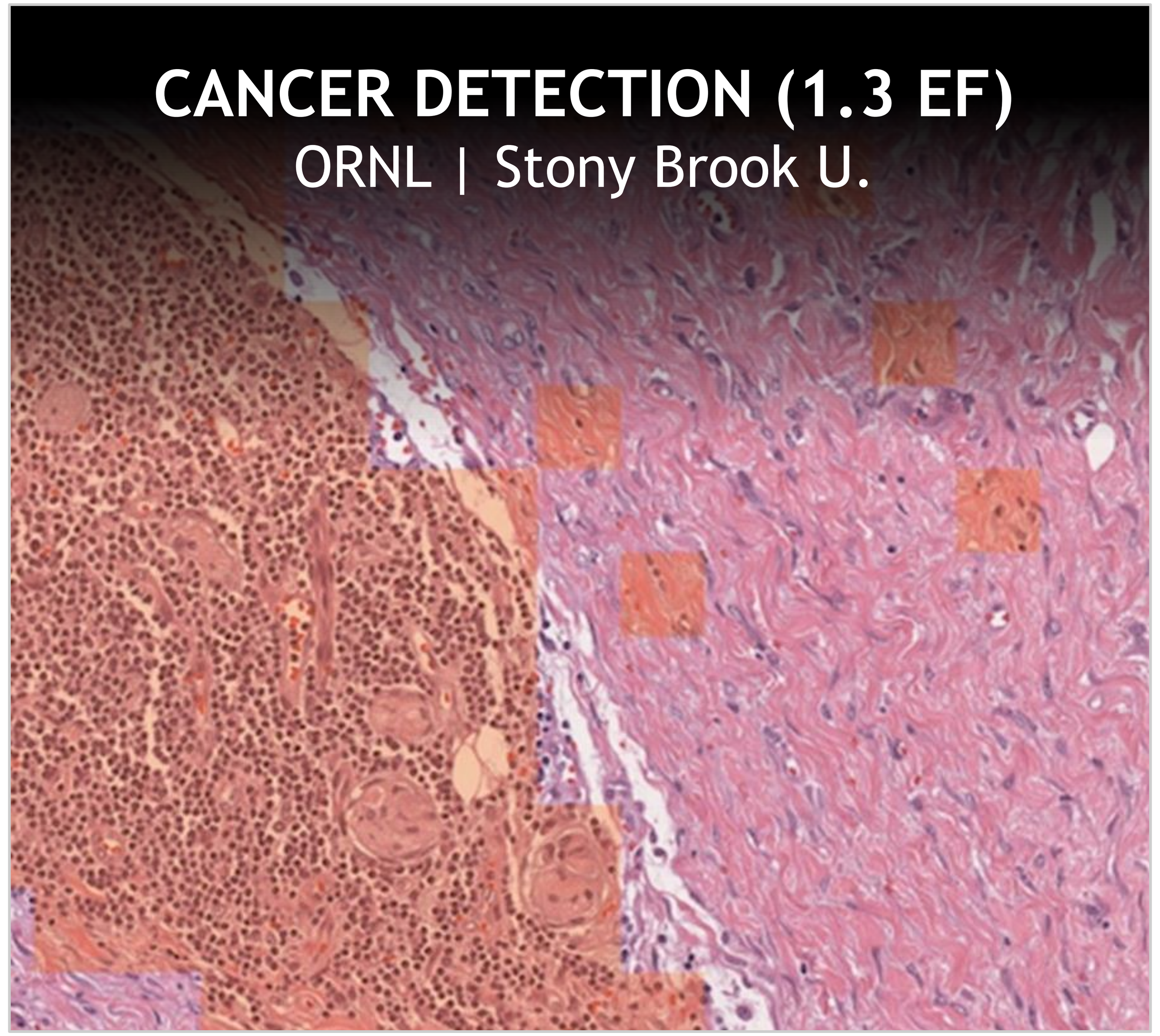




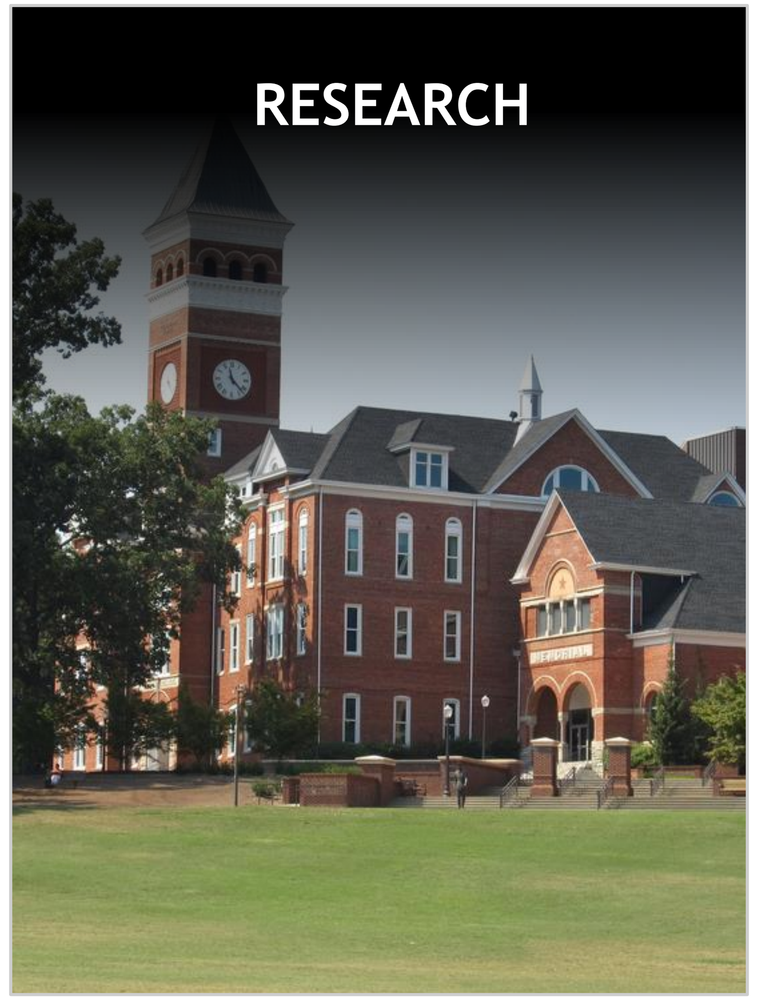
EXASCALE AI SCIENCE

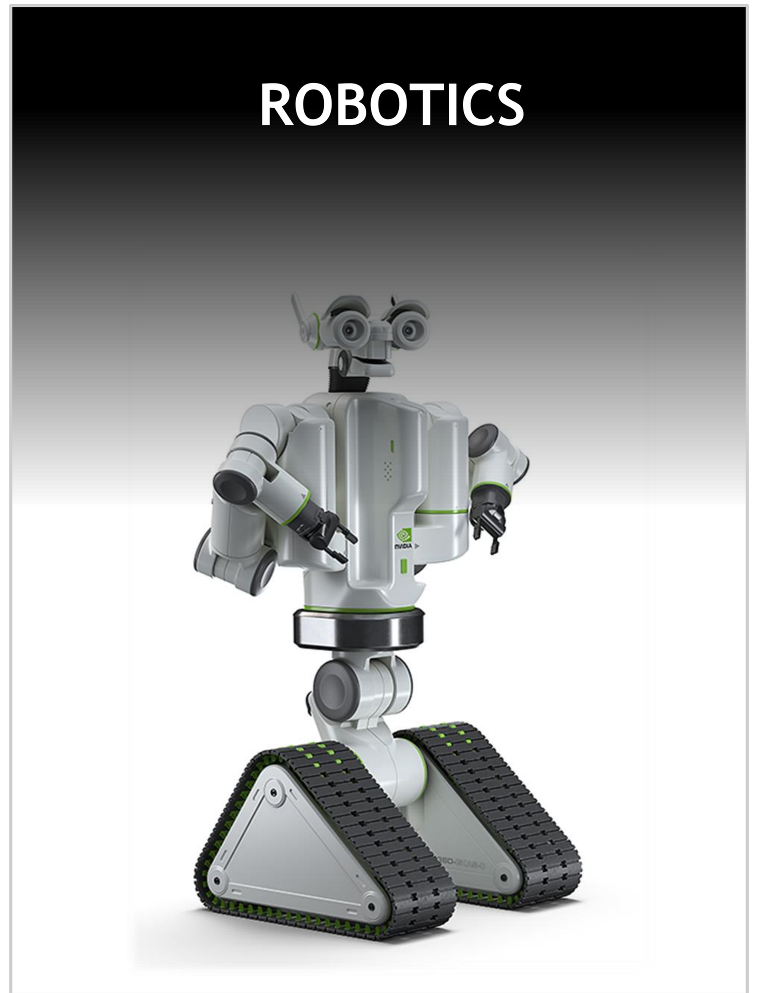


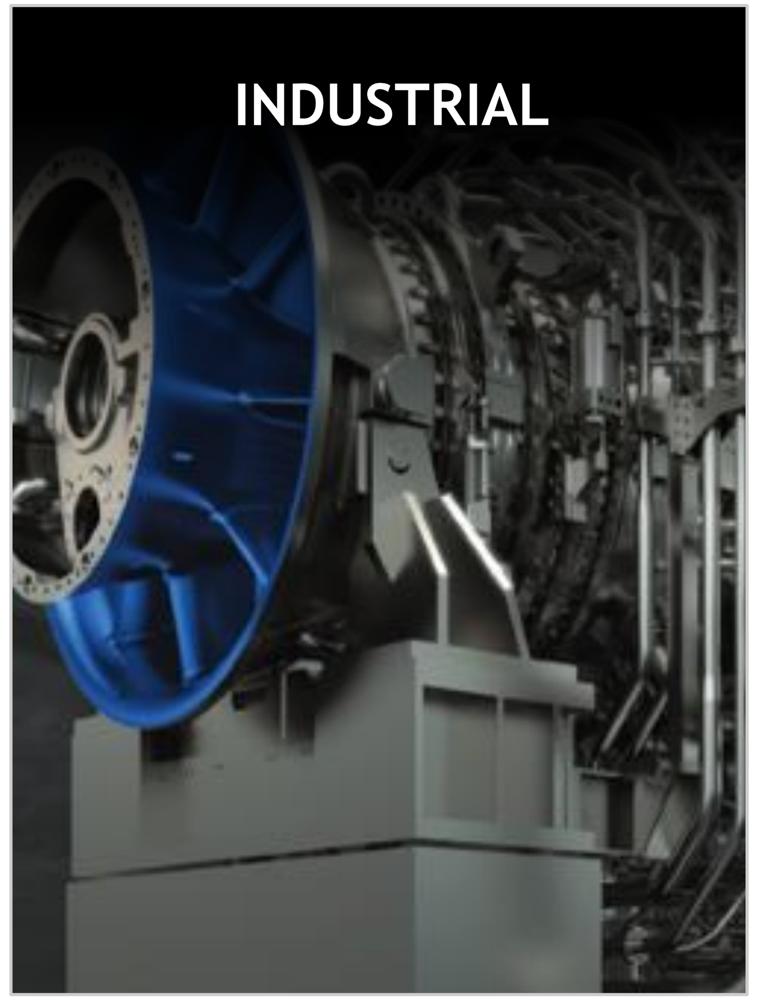


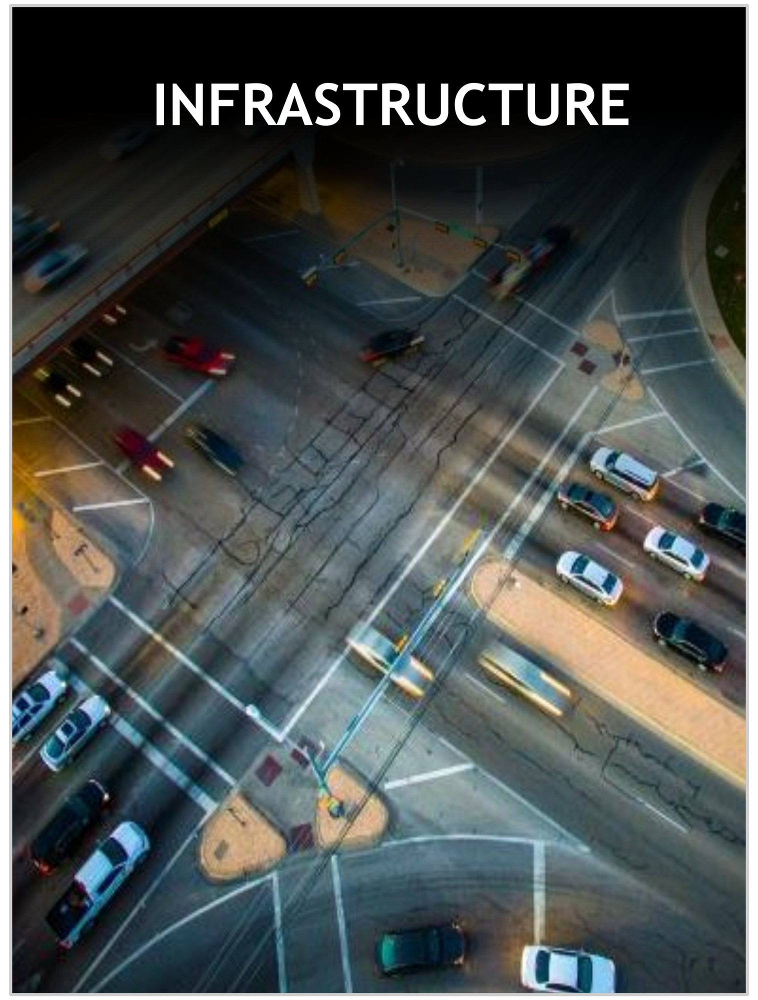


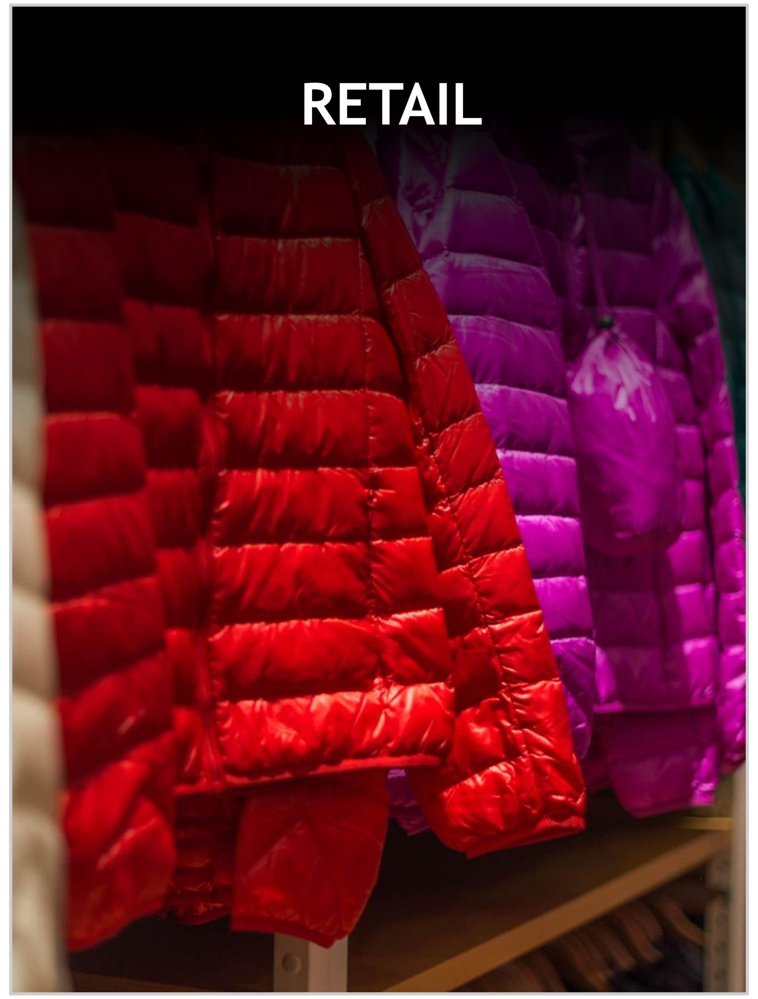
AI ADOPTION ACROSS EVERY INDUSTRY

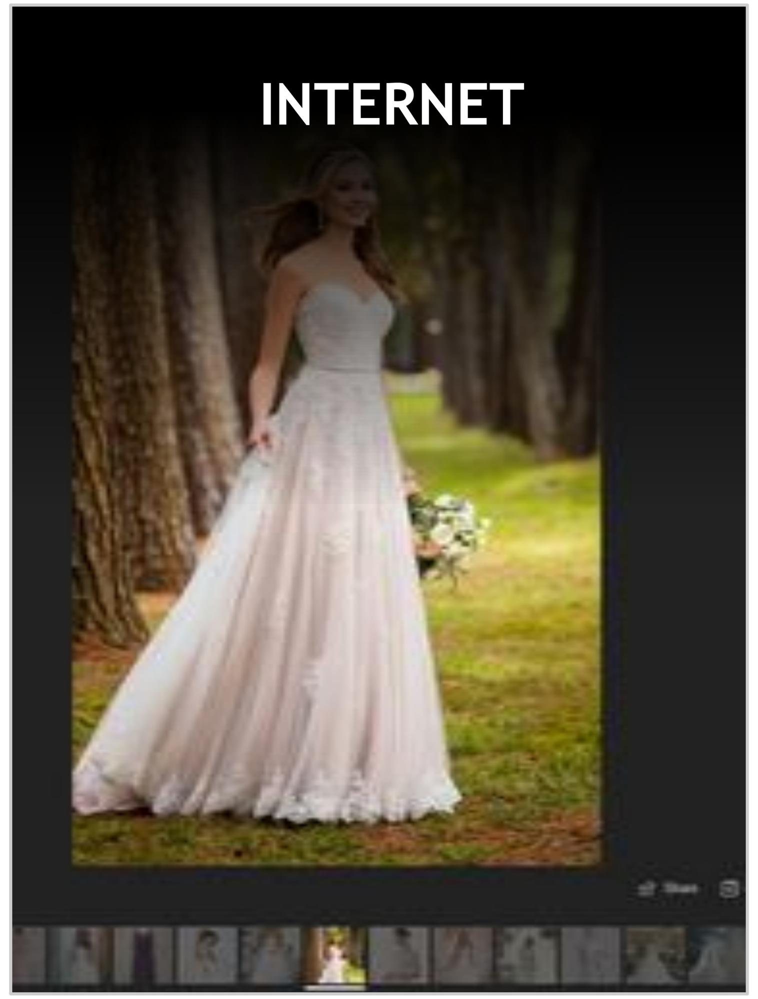


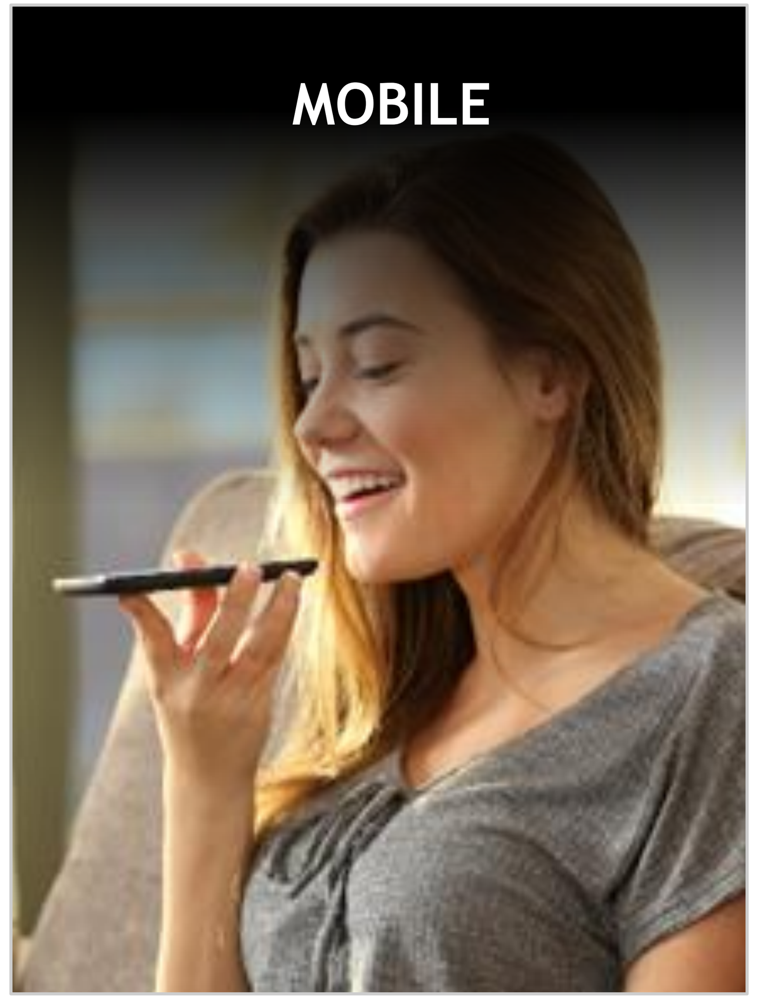


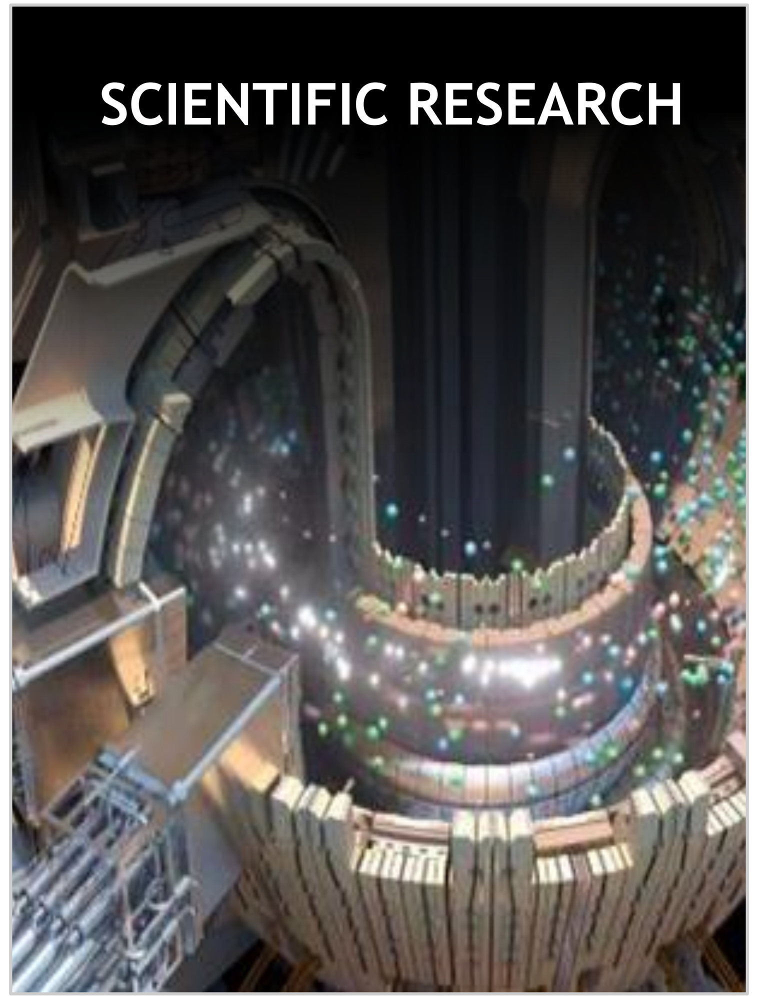






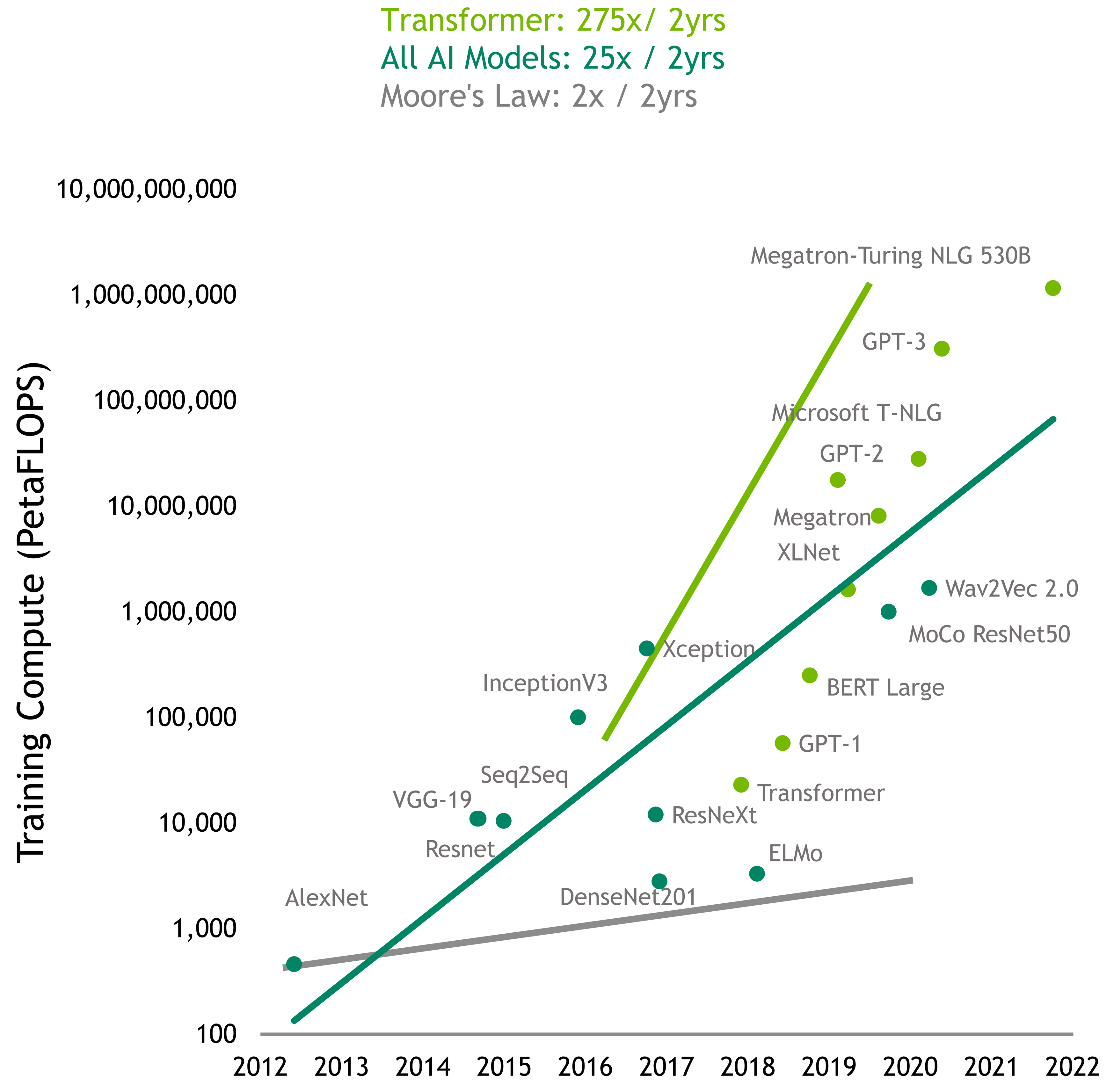




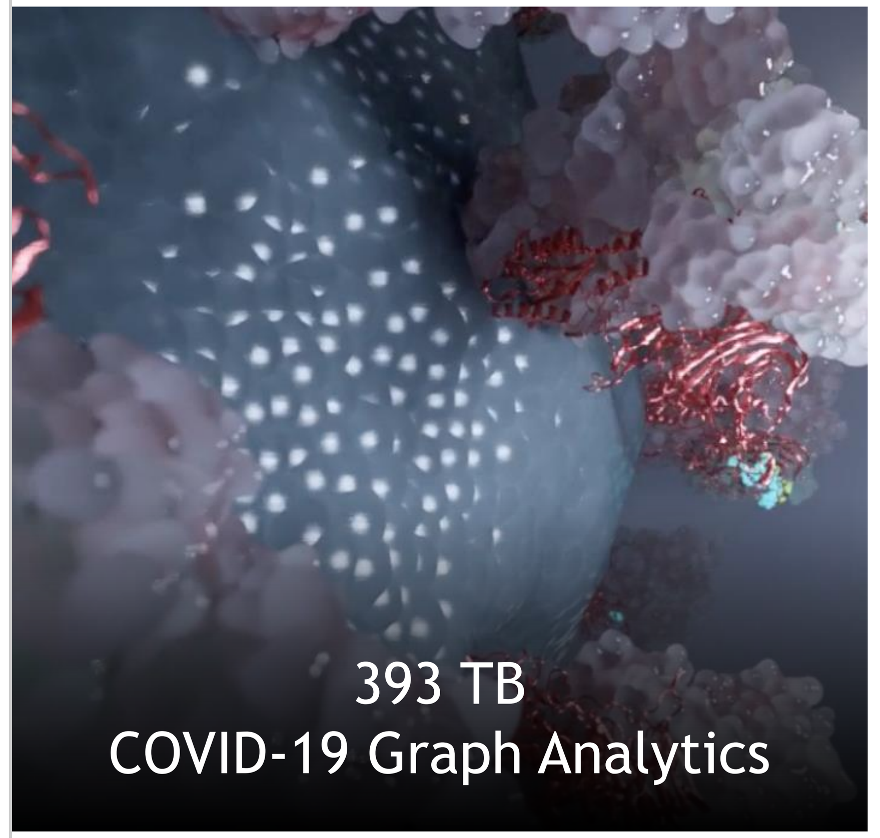


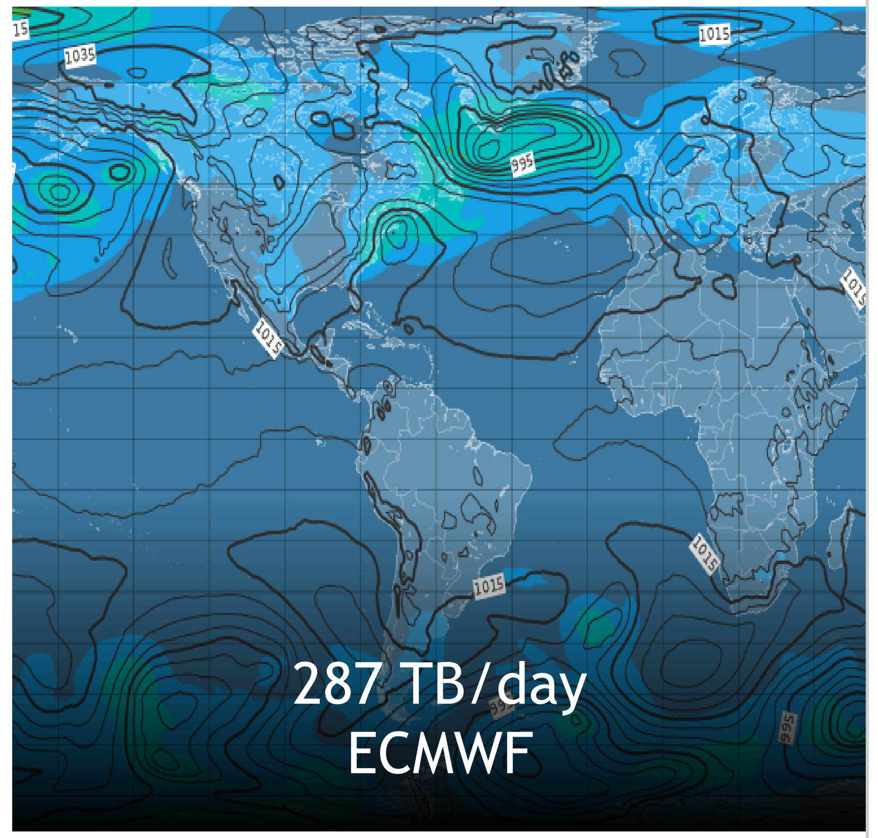
EXPLODING DATA AND MODEL SIZE

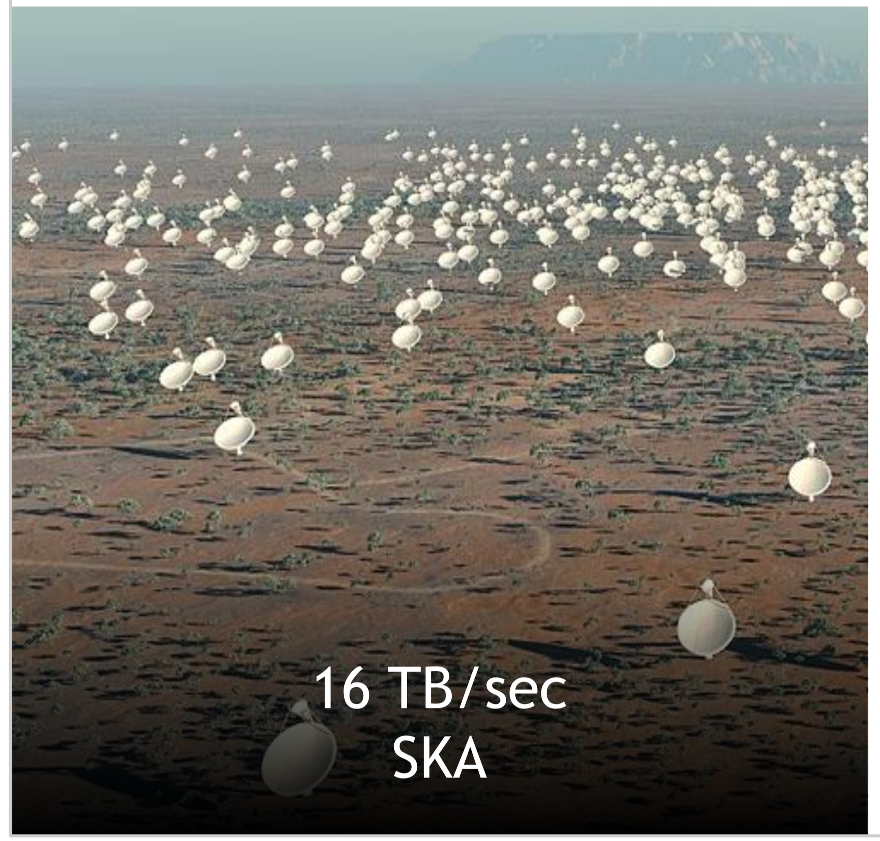
EXPLODING MODEL SIZE
Driving Superhuman Capabilities



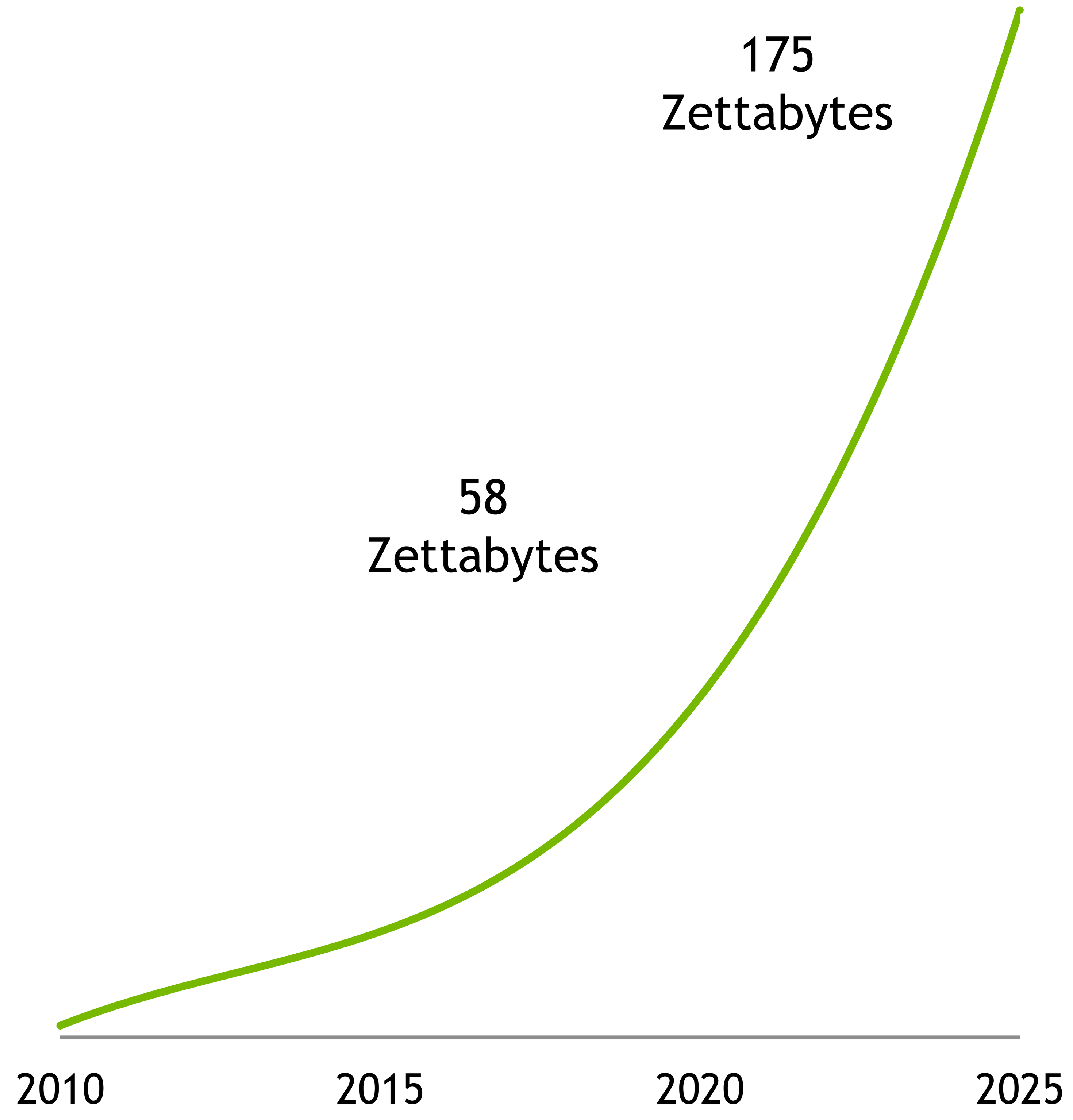
BIG DATA GROWTH
90% of the World's Data in Last 2 Years







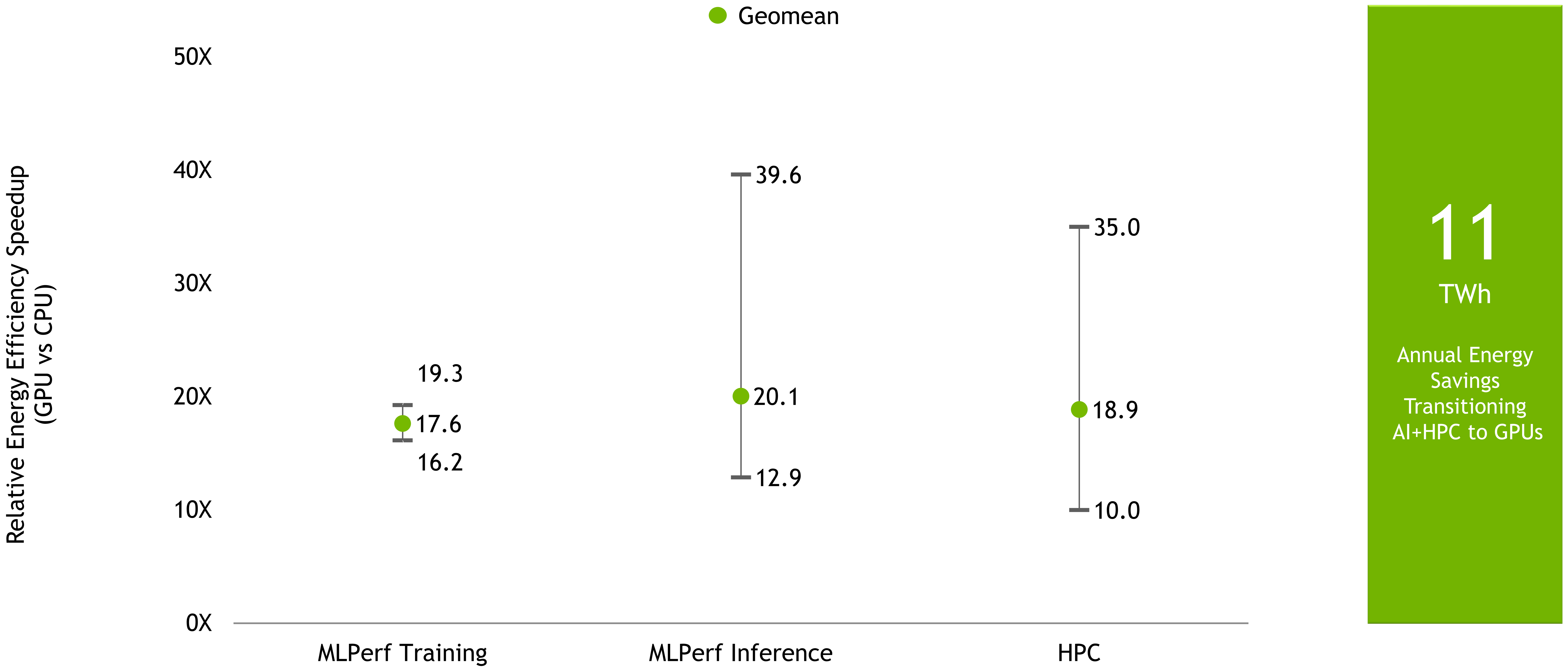
GROWTH IN SCIENTIFIC DATA
Fueled by Accurate Sensors & Simulations



Source for Big Data Growth chart: IDC — The Digitization of the World (May, 2020)

ACCELERATED COMPUTING DELIVERS GLOBAL ENERGY SAVINGS ON AI AND HPC

Up to 40X Energy Savings



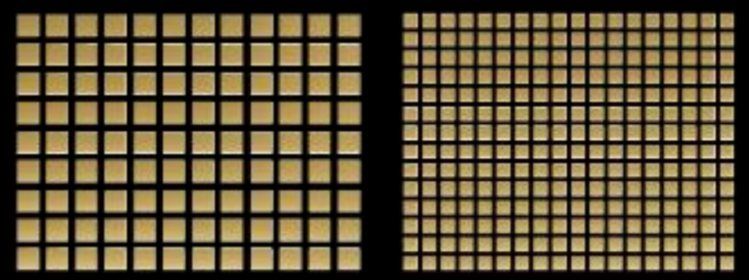
MLPerf AI Training: 1.0-1058, 1.0-1059 | 1.0-1042, 1.0-1043
MLPerf AI Inference: 1.0-30 | 1.0-19, 1.0-20, 1.0-22
HPC Appps: AMBER, Chroma, GTC, LAMMPS, NAMD, SPECFEM3D | DGX A100 (4 GPU) vs Dual Socket Platinum 8280

ENERGY EFFICIENT SYSTEMS FOR MASSIVE WORKLOADS WITH HYPERSCALE SENSIBILITIES

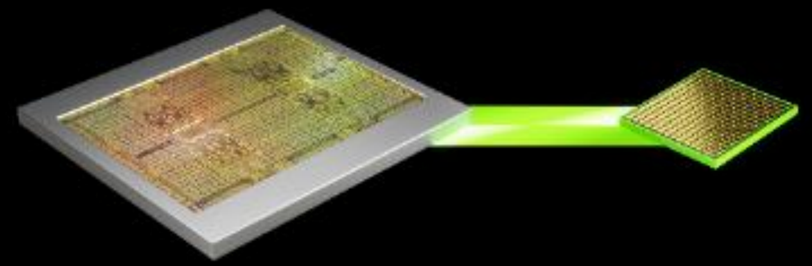
- Speed and feed matching
- Thermal and power design
- Interconnect design
- Deployability
- Operability
- Flexibility
- Expandability

NVIDIA A100 80GB

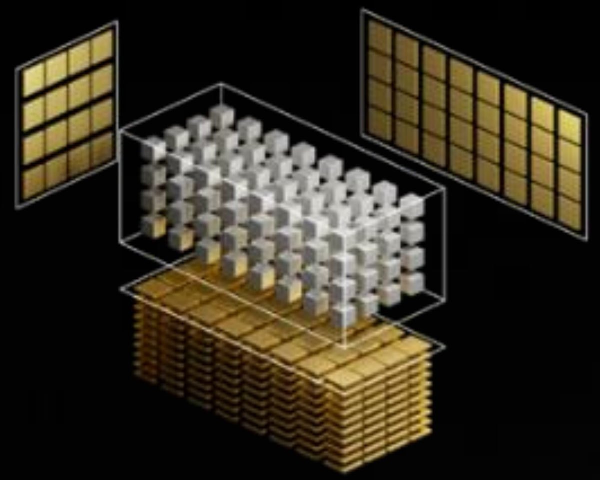
Supercharging The World’s Highest Performing AI Supercomputing GPU



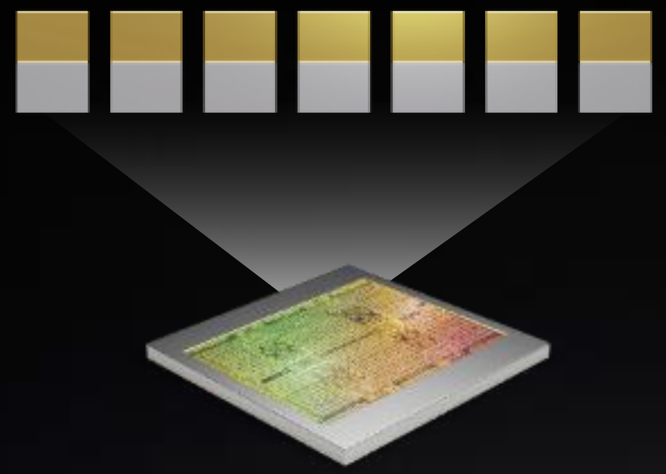
80GB HBM2e
For largest datasets and models



2TB/s +
World’s highest memory bandwidth to feed the world’s fastest GPU

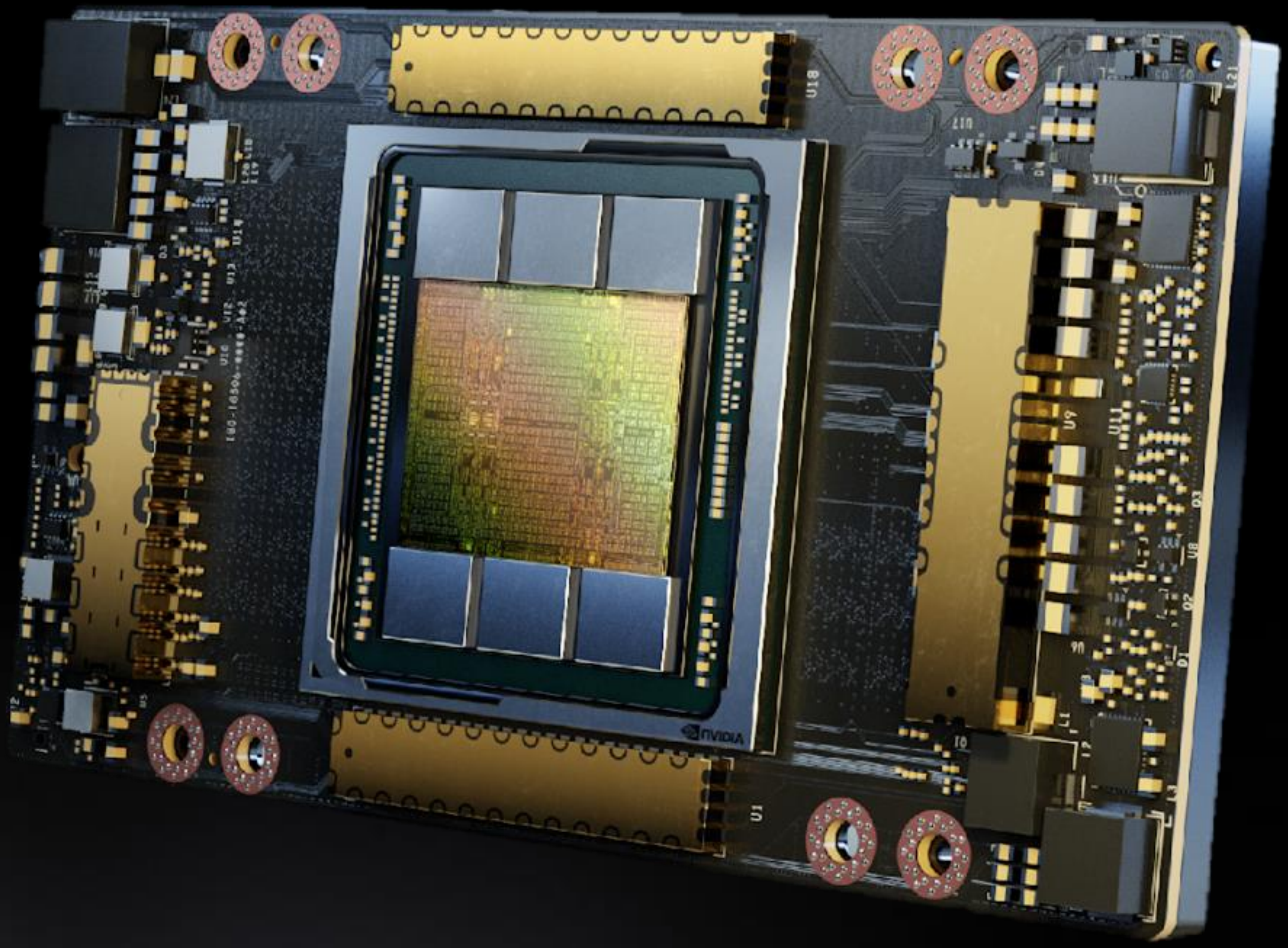


3rd Gen Tensor Core

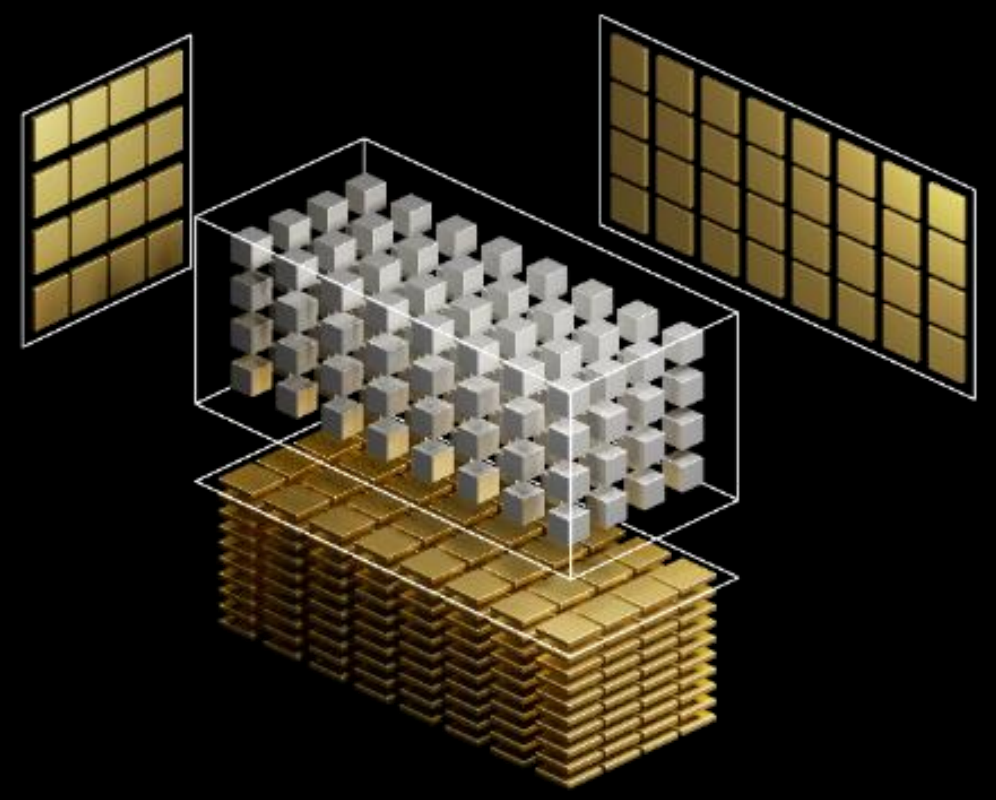


Multi-Instance GPU

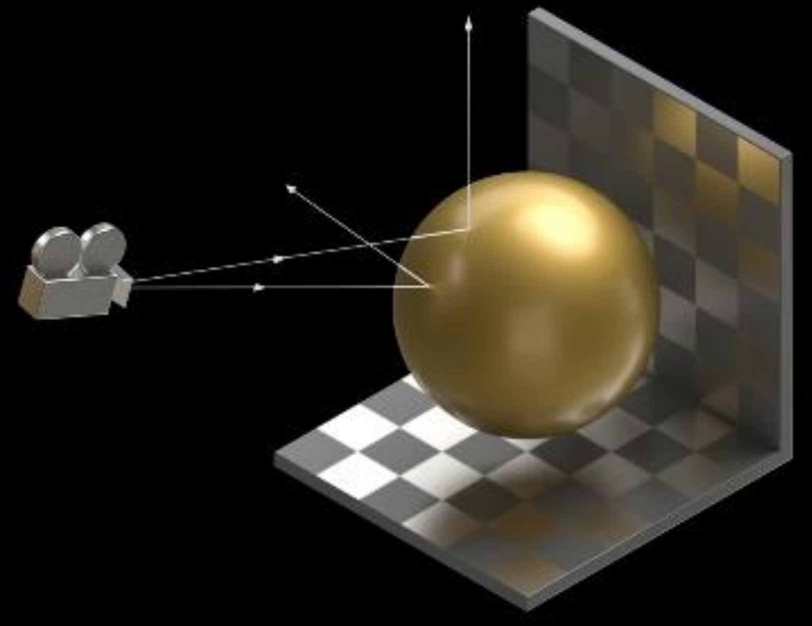
3rd Gen NVLink



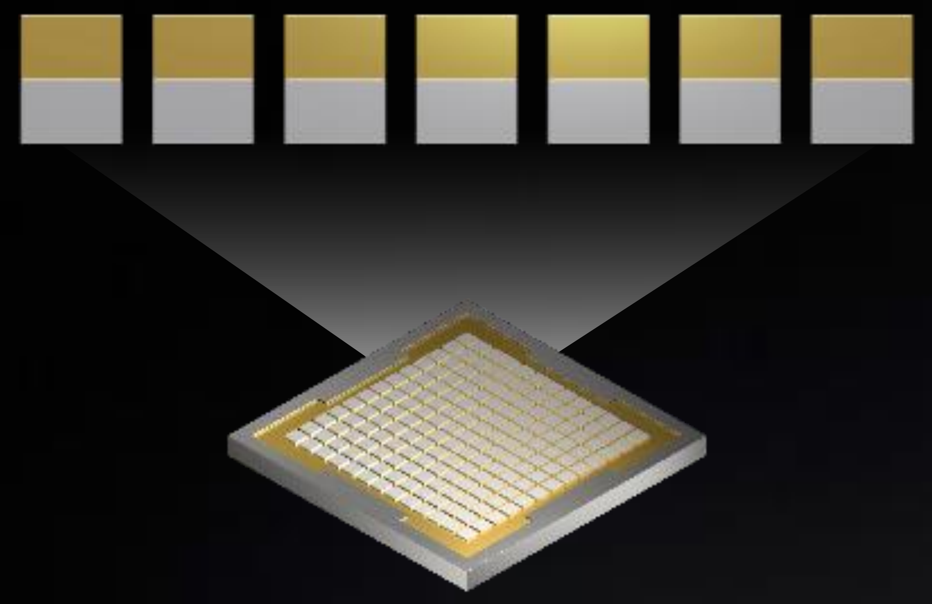
AMPERE ARCHITECTURE



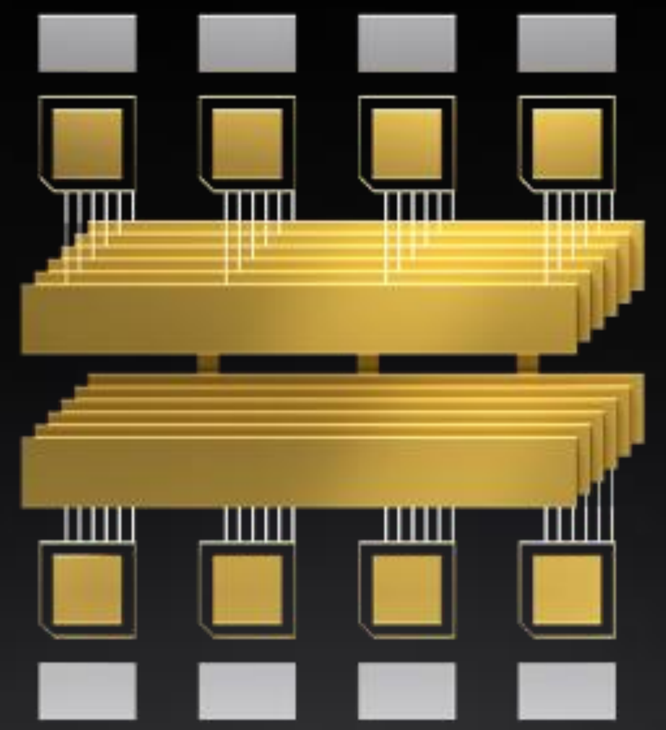
3rd Gen Tensor Cores
Faster, Flexible, Easier to use
20x AI Perf with TF32
2.5x HPC Perf



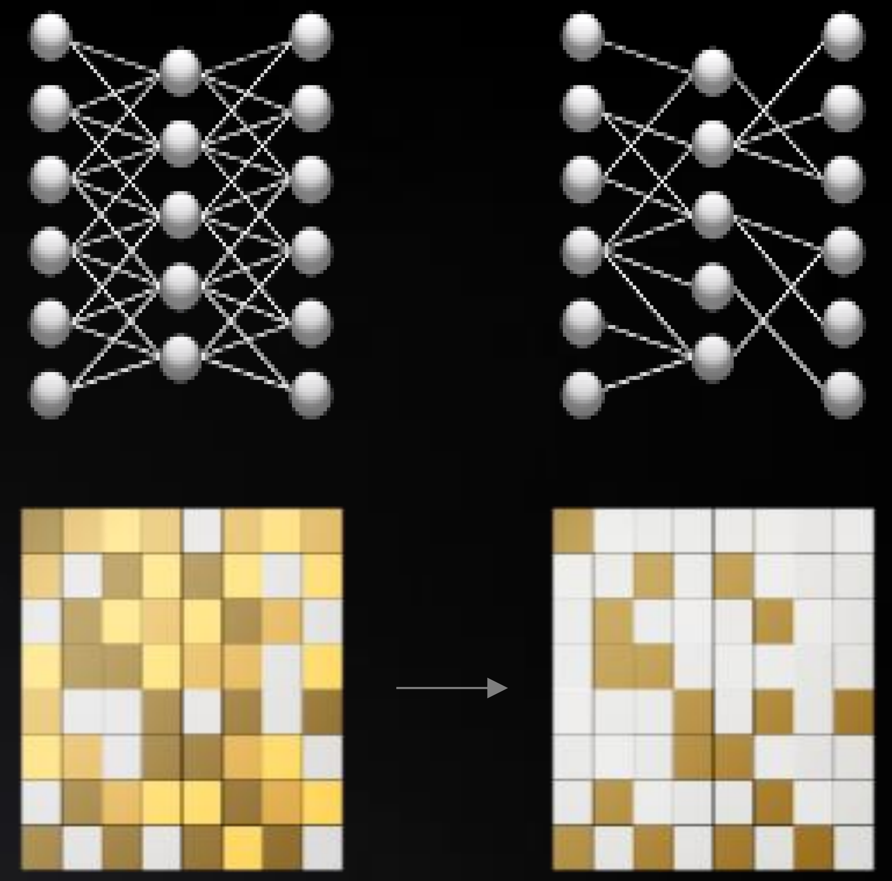
2nd Gen RT Cores
Up to 2X throughput of previous generation



New Multi-Instance GPU
Optimal utilization with right sized GPU
7x Simultaneous Instances per GPU

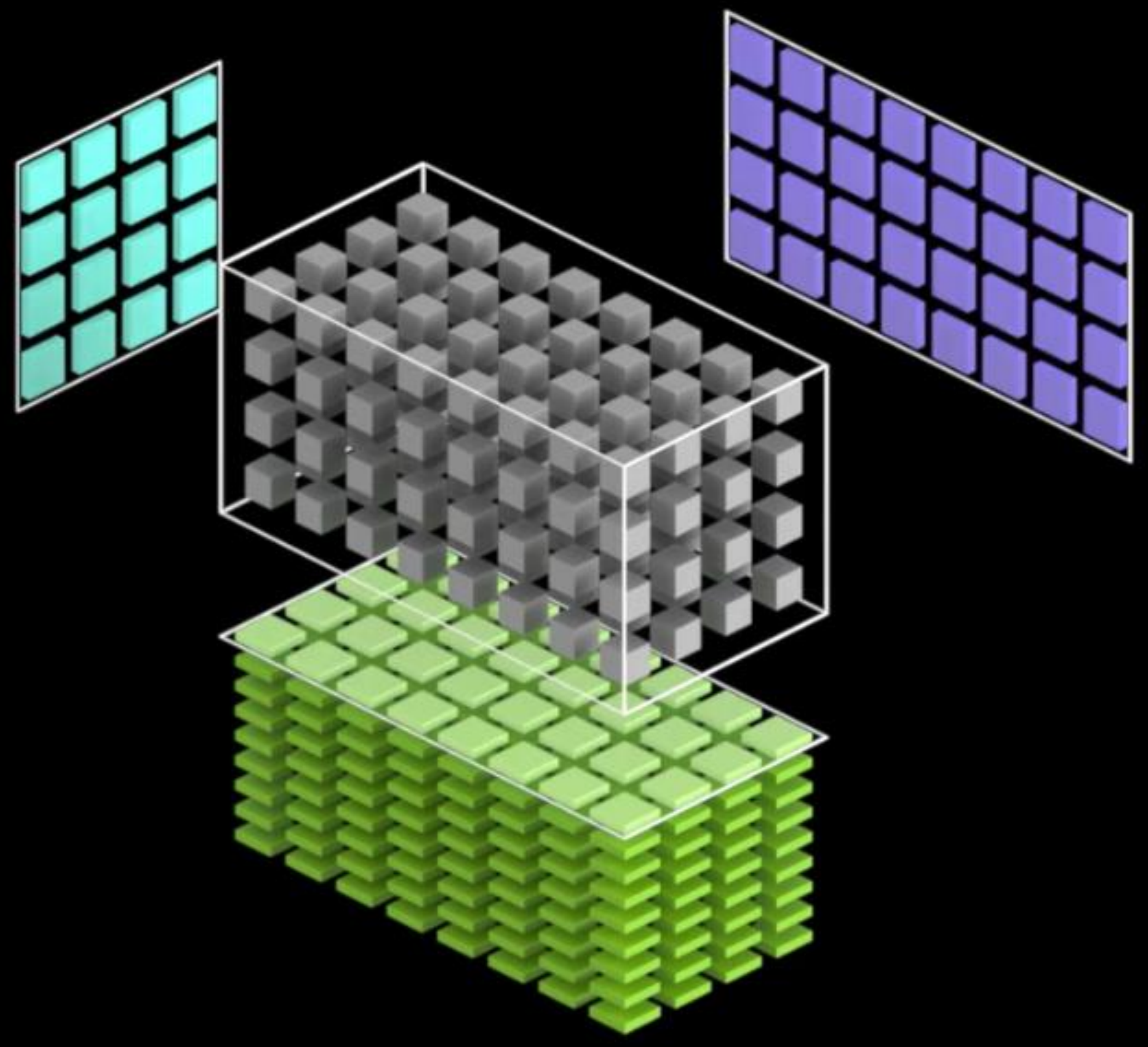


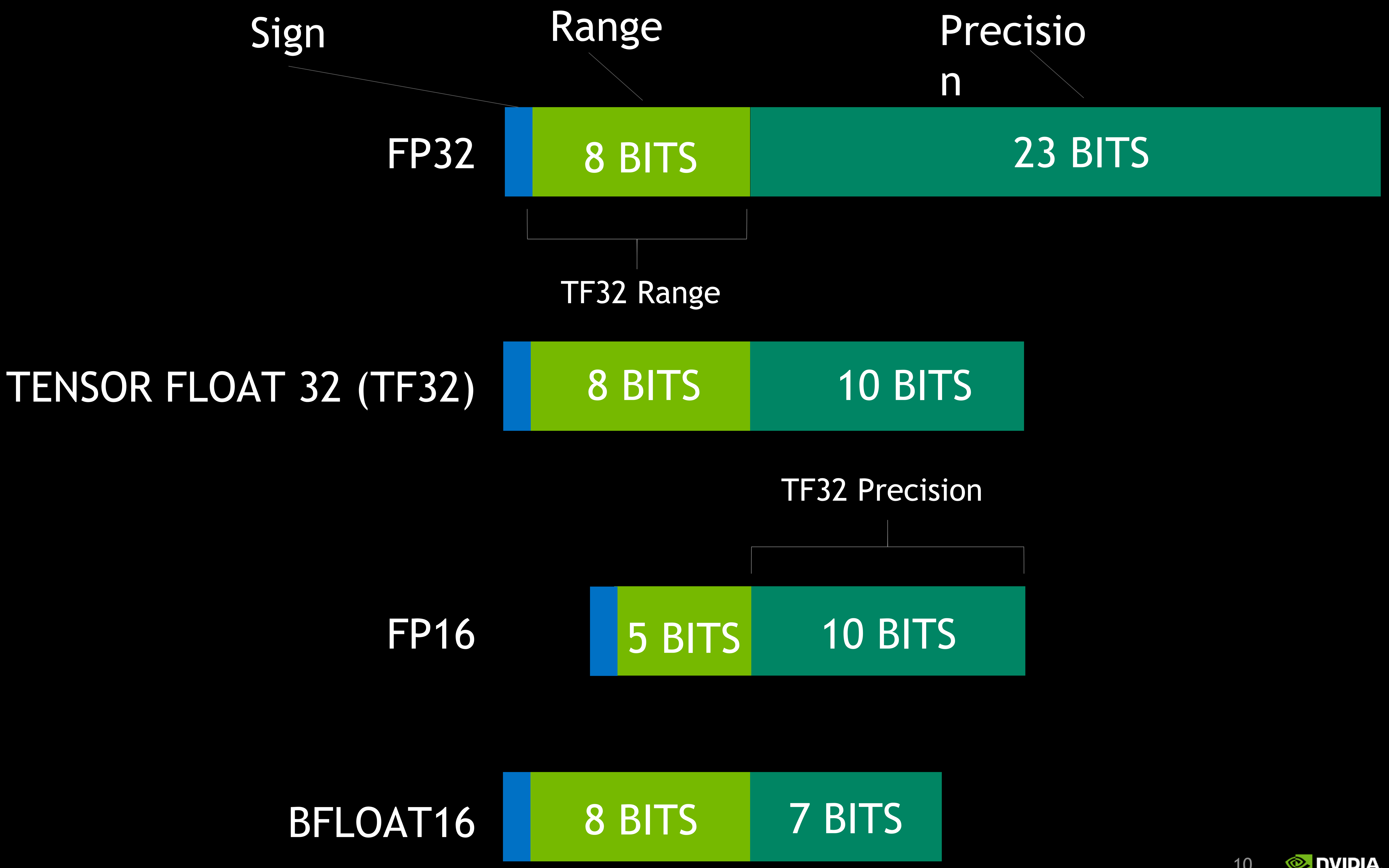
3rd Gen NVLINK and NVSWITCH
Efficient Scaling to Enable Super GPU
2X More Bandwidth



New Sparsity Acceleration
Harness Sparsity in AI Models
2x AI Performance

NEW TF32 TENSOR CORES



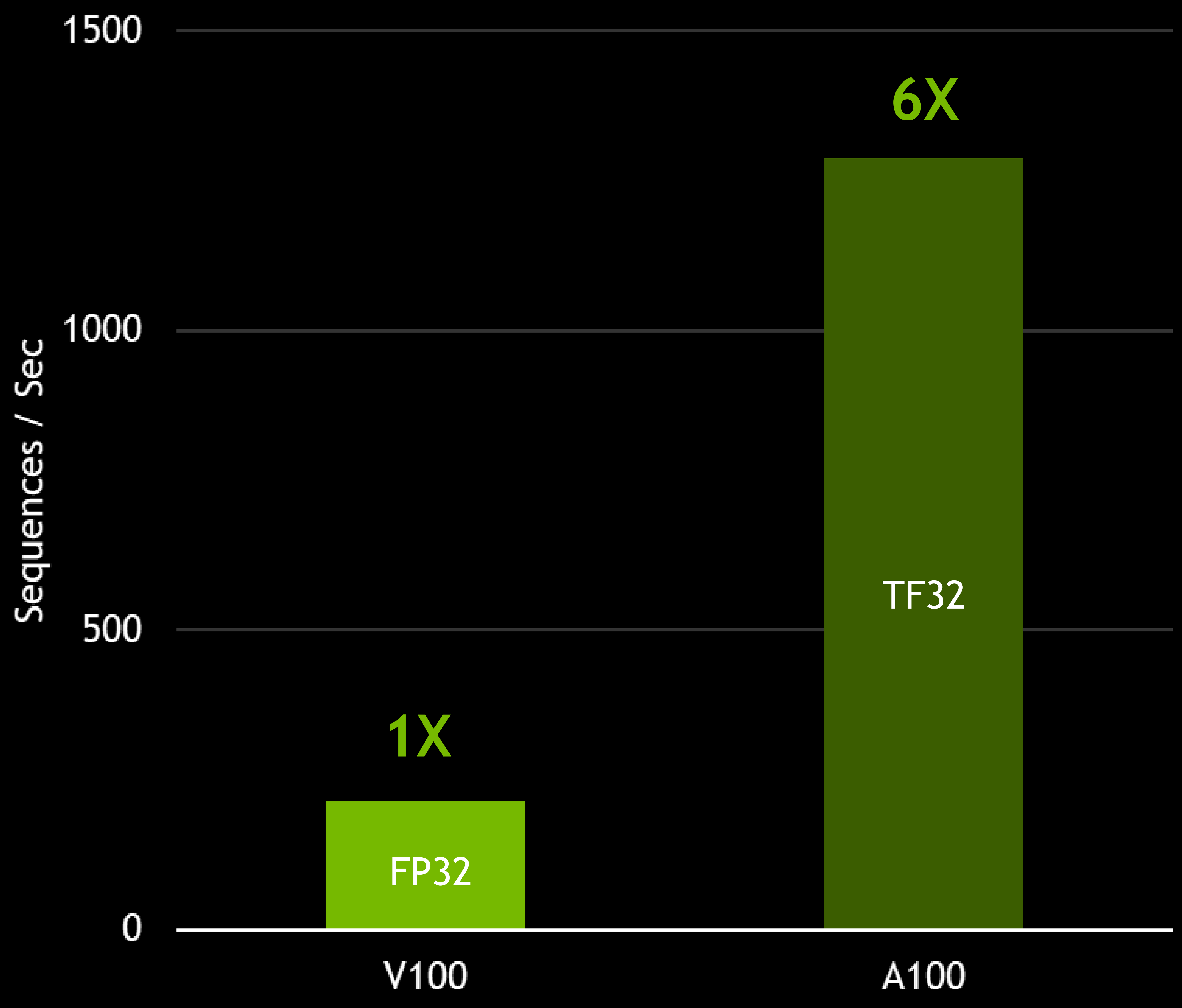


Range of FP32 and Precision of FP16

Input in FP32 and Accumulation in FP32

No Code Change Speed-up for Training

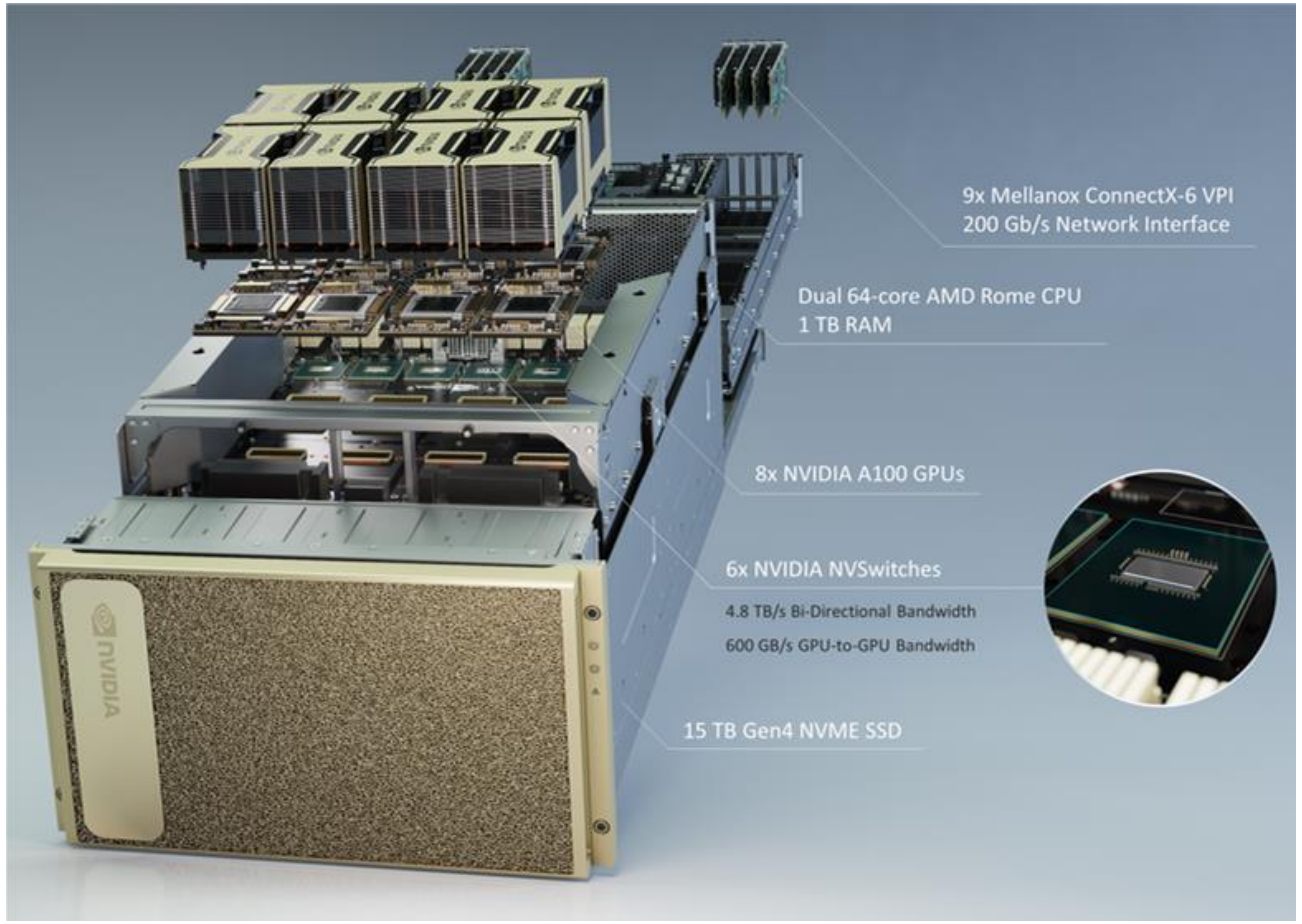
TF32 FOR AI TRAINING - BERT



BERT Pre-Training Throughput using Pytorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 V100: DGX-1 Server with 8xV100 using FP32 precision
A100: DGX A100 Server with 8xA100 using TF32 precision |

NVIDIA DGX A100

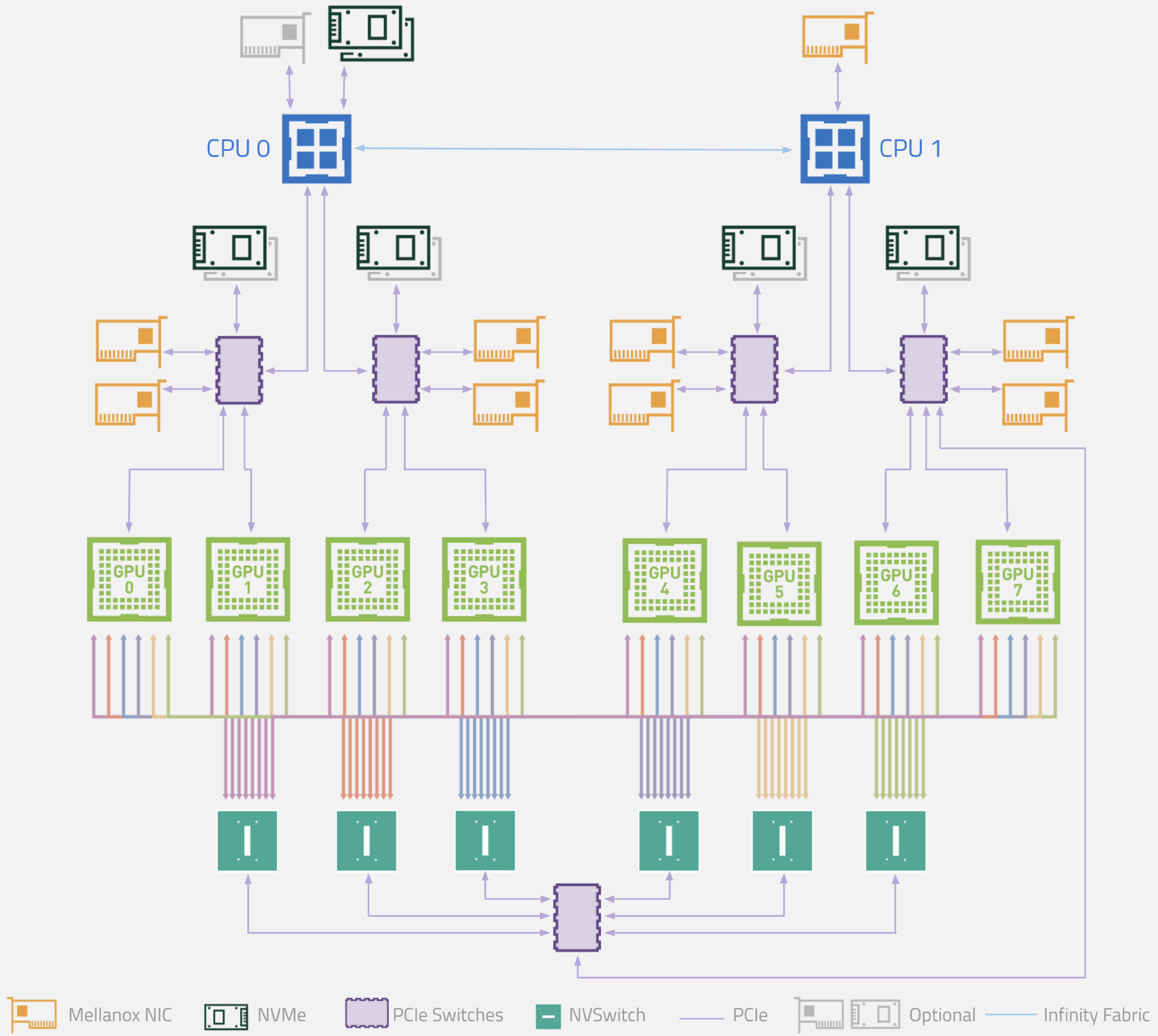
GPUs	**8x NVIDIA A100**
GPU Memory	320 GB total
Peak performance	5 petaFLOPS AI 10 petaOPS INT8
NVSwitches	6
System Power Usage	**6.5kW max**
CPU	**Dual AMD Rome 7742** 128 cores total, 2.25 GHz(base), 3.4GHz (max boost)
System Memory	**1TB**
Networking	**8x Single-Port Mellanox ConnectX-6** 200Gb/s HDR Infiniband (**Compute Network**) 1x (or **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: 15TB (4x 3.86TB) U.2 NVME drives
Software	Ubuntu Linux OS (**5.3+ kernel**)
System Weight	271 lbs (123 kgs)
Packaged System Weight	315 lbs (143 kgs)
Height	**6U**
Operating temperature range	**5C to 30C** (41F to 86F)



* Optional upgrades

Many NICs on compute plane to try to get closer to NVLink performance. Heavy focus on IO capabilities to "feed the beast".

NVIDIA DGX A100 INTERNAL I/O



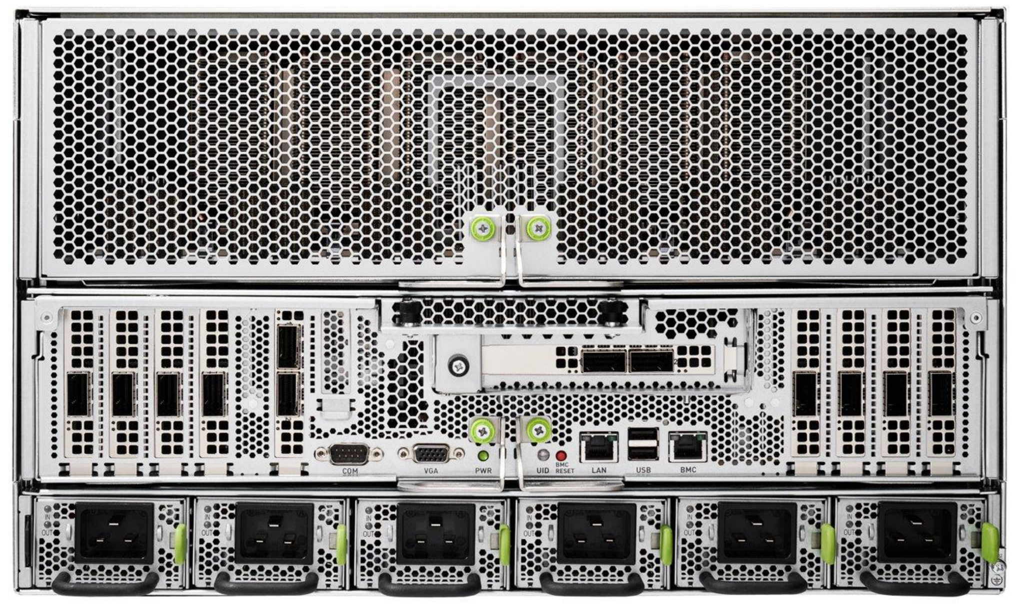
EFFICIENT BUILDING BLOCKS

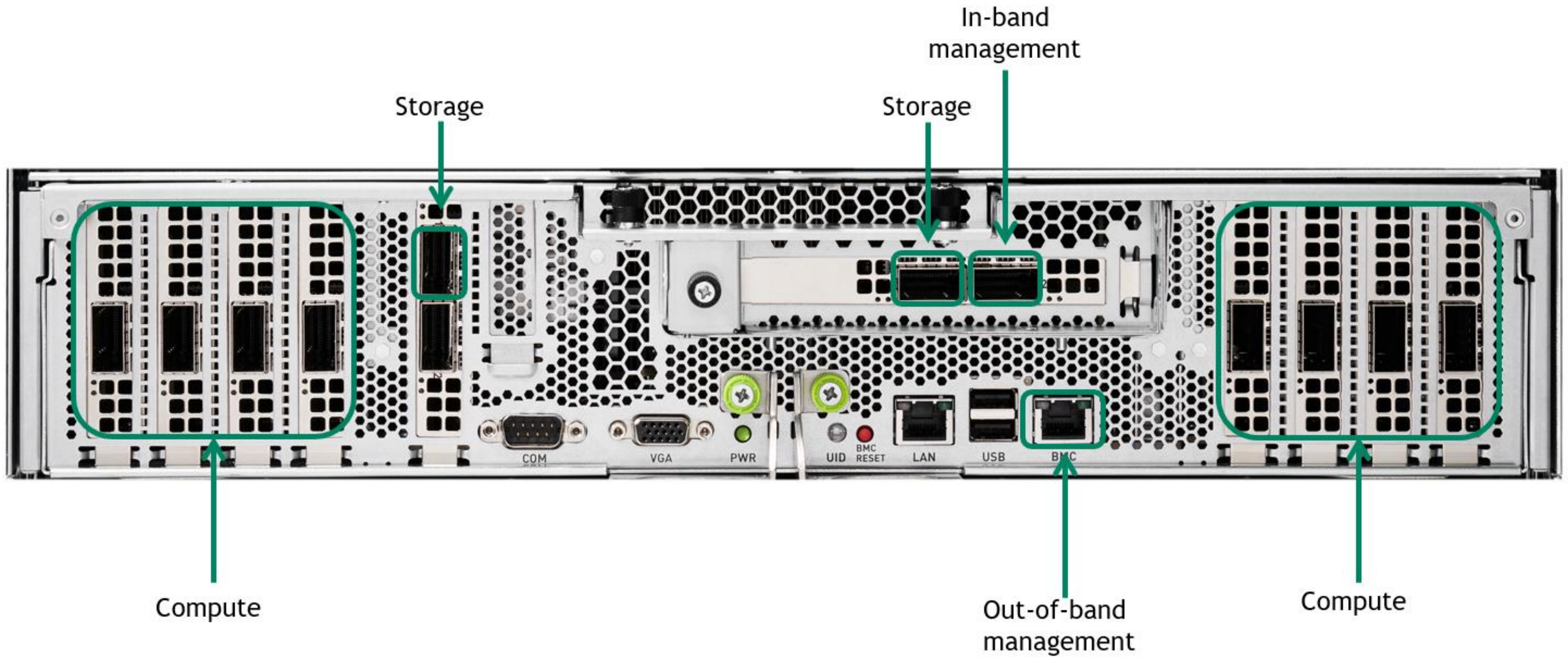
Scalable Units

NVIDIA

DESIGNING FOR PERFORMANCE

In the datacenter





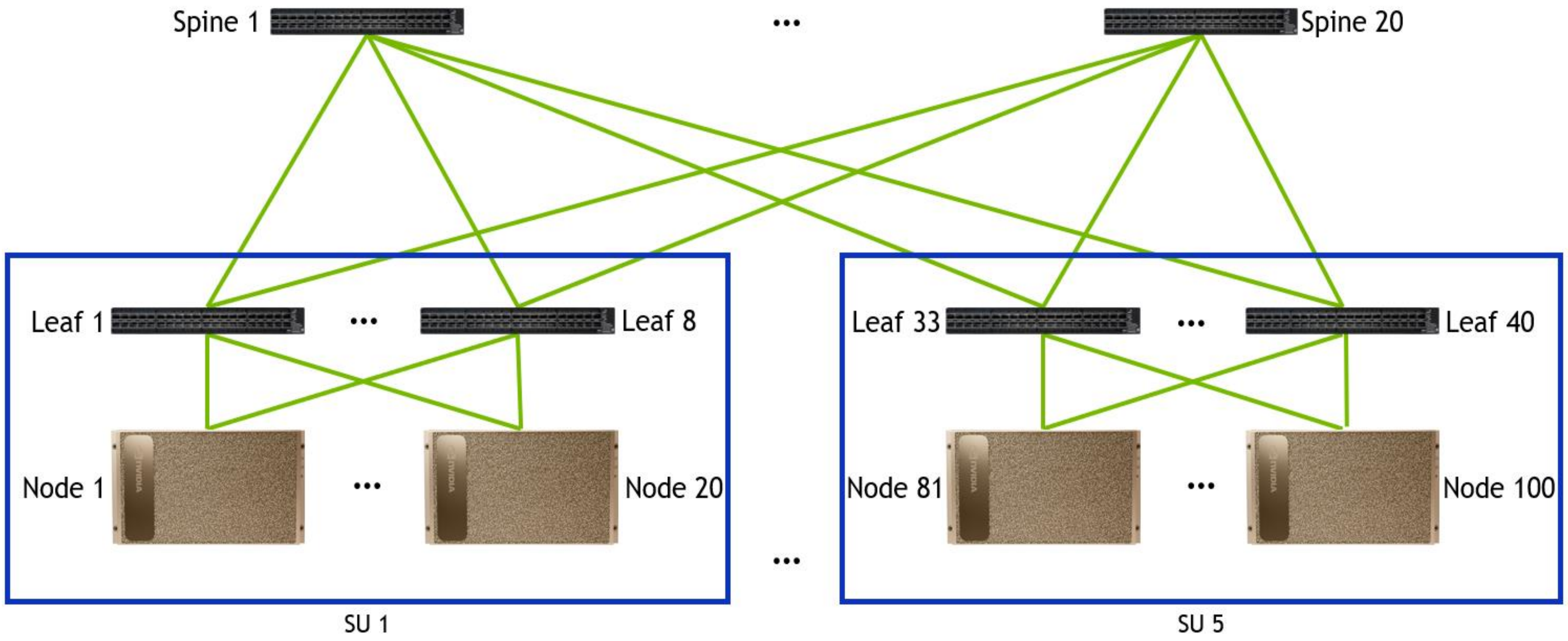
A POD AT ANY SCALE

Growing with Scalable Units (SU)

Nodes	SUs	QM8790 Switches			Cables		
		Leaf	Spine	Core	Leaf	Spine	Core
10	½	8	2		80	80	
20 (Single SU)	1	8	4		160	160	
40	2	16	10		320	320	
80	4	32	20		640	640	
100	5	40	20		800	800	
140 (DGX A100 SuperPOD)	7	56	80	28	1120	1120	560

Full fat tree compute fabric

100 node example

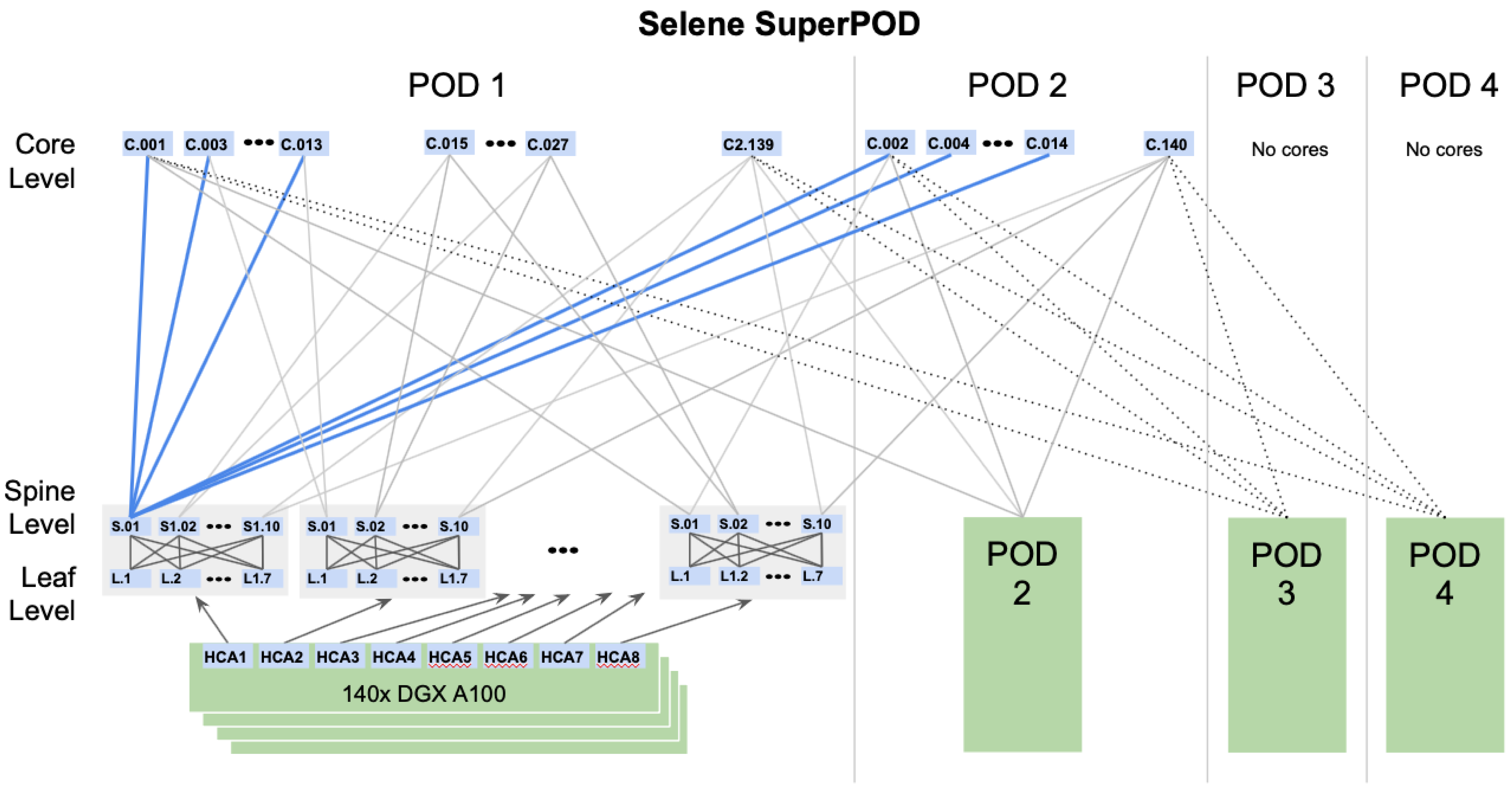


Multi-node IB compute

Designed with Mellanox 200Gb HDR IB network

Separate compute and storage fabric
- 8 Links for compute
- 2 Links for storage (Lustre)
- Both networks share a similar fat-tree design

Modular POD design
- 140 DGX A100 nodes are fully connected in a POD
- POD contains compute nodes and storage
- All nodes and storage are usable between PODs
- resilient to failures at the spine level



Sharp optimized design
- Leaf and Spines organized in HCA planes
- For a POD, all HCA1 from 140 DGX-2 connect to a HCA1 Plane fat-tree network
- Traffic from HCA1 to HCA1 between any two nodes in a POD stay either at the Leaf or Spine level
- Only use core switches when
  - 1. Moving data between HCA planes (e.g. mlx5_0 to mlx5_1 in another system)
  - 2. Moving any data between PODs

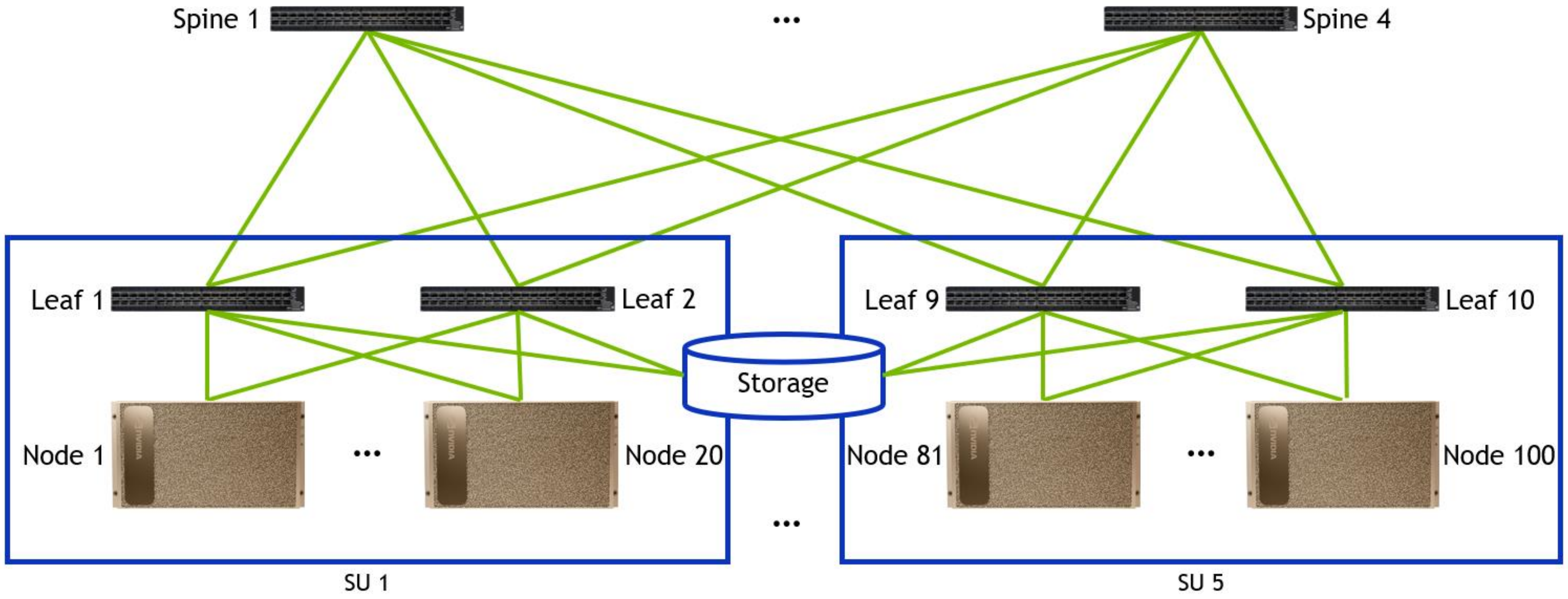
A POD AT ANY SCALE

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



DGX A100 SuperPOD

A modular model

1K GPU SuperPOD Cluster

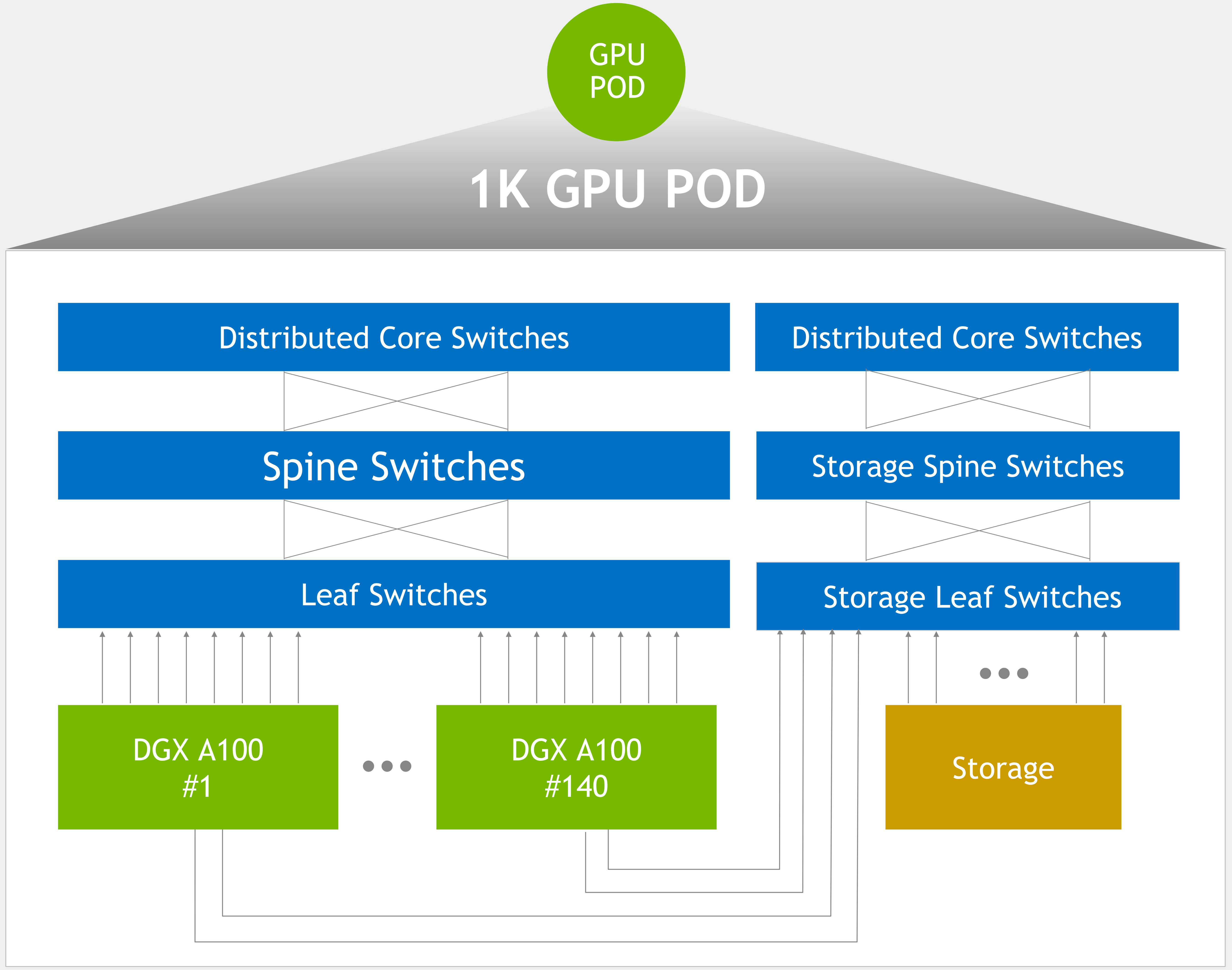
- 140 DGX A100 nodes (1120 GPUs) in a GPU POD
- 1st tier fast storage - DDN AI400x with Lustre
- Mellanox HDR 200Gb/s InfiniBand - Full Fat-tree
- Network optimized for AI and HPC

DGX A100 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 vs Storage
- Adaptive routing and SharpV2 support for offload

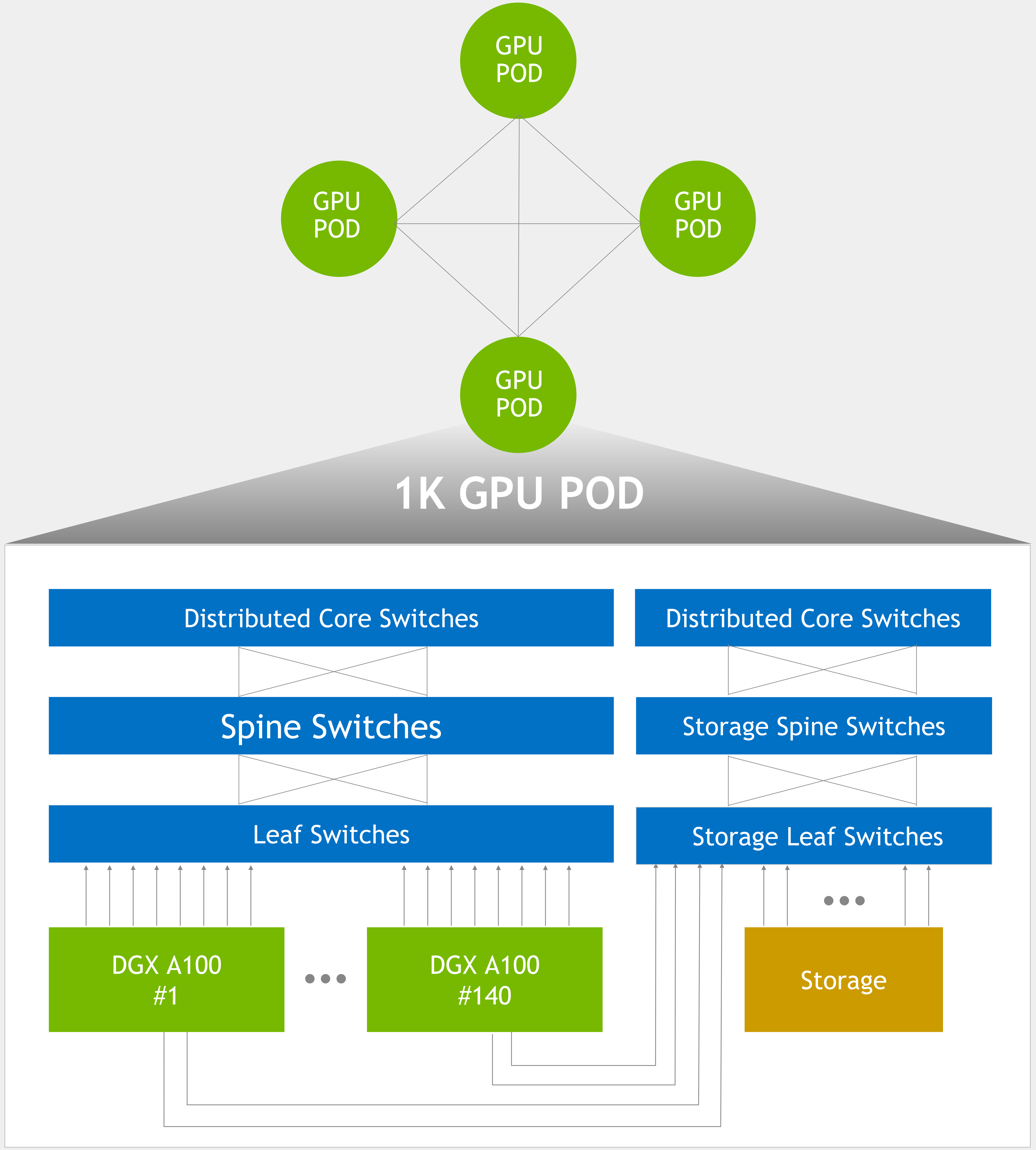


The DGXA100 Superpod

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 vs Storage
- Adaptive routing and SharpV2 support for offload

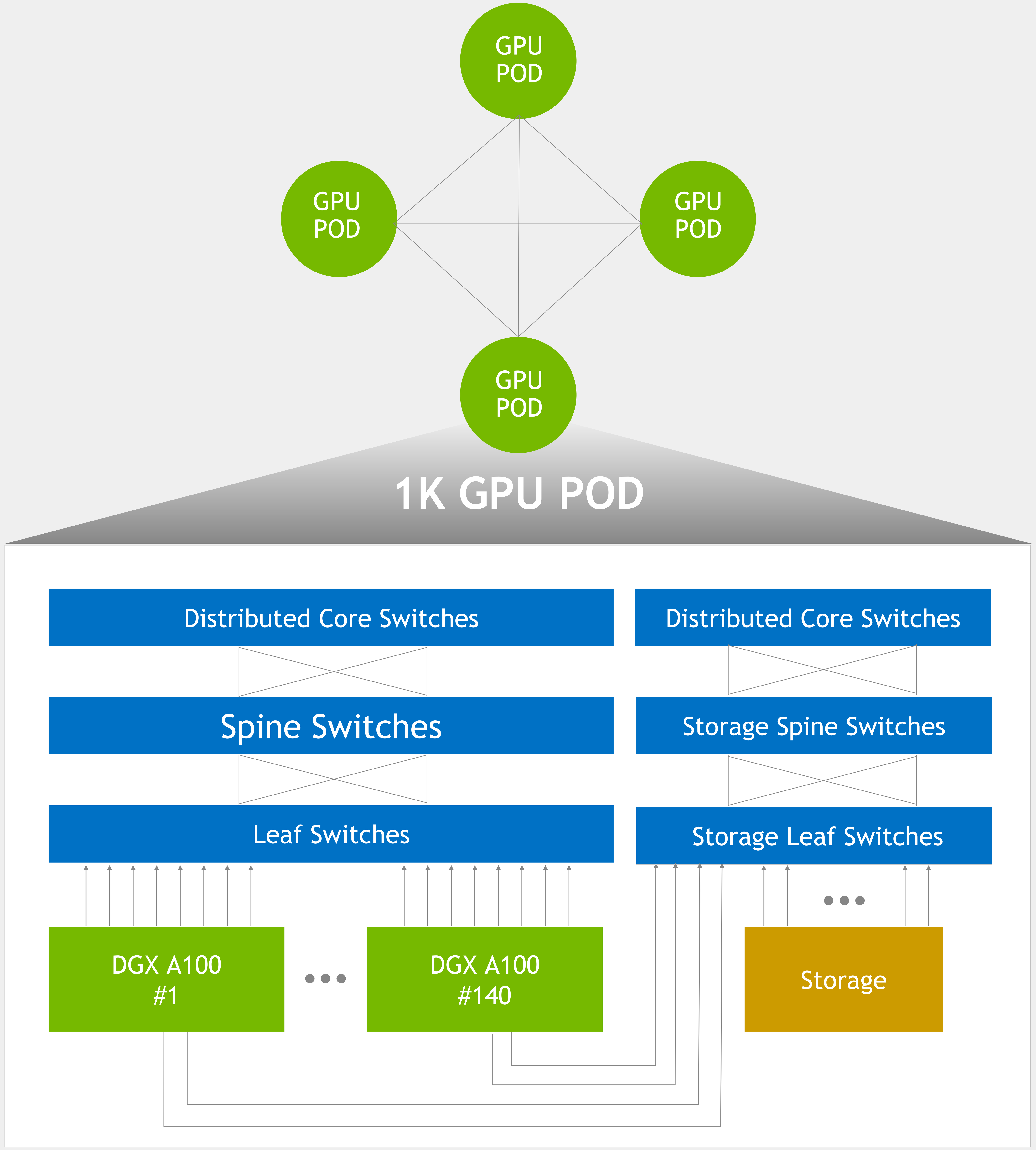


The DGXA100 Superpod

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 vs Storage
- Adaptive routing and SharpV2 support for offload

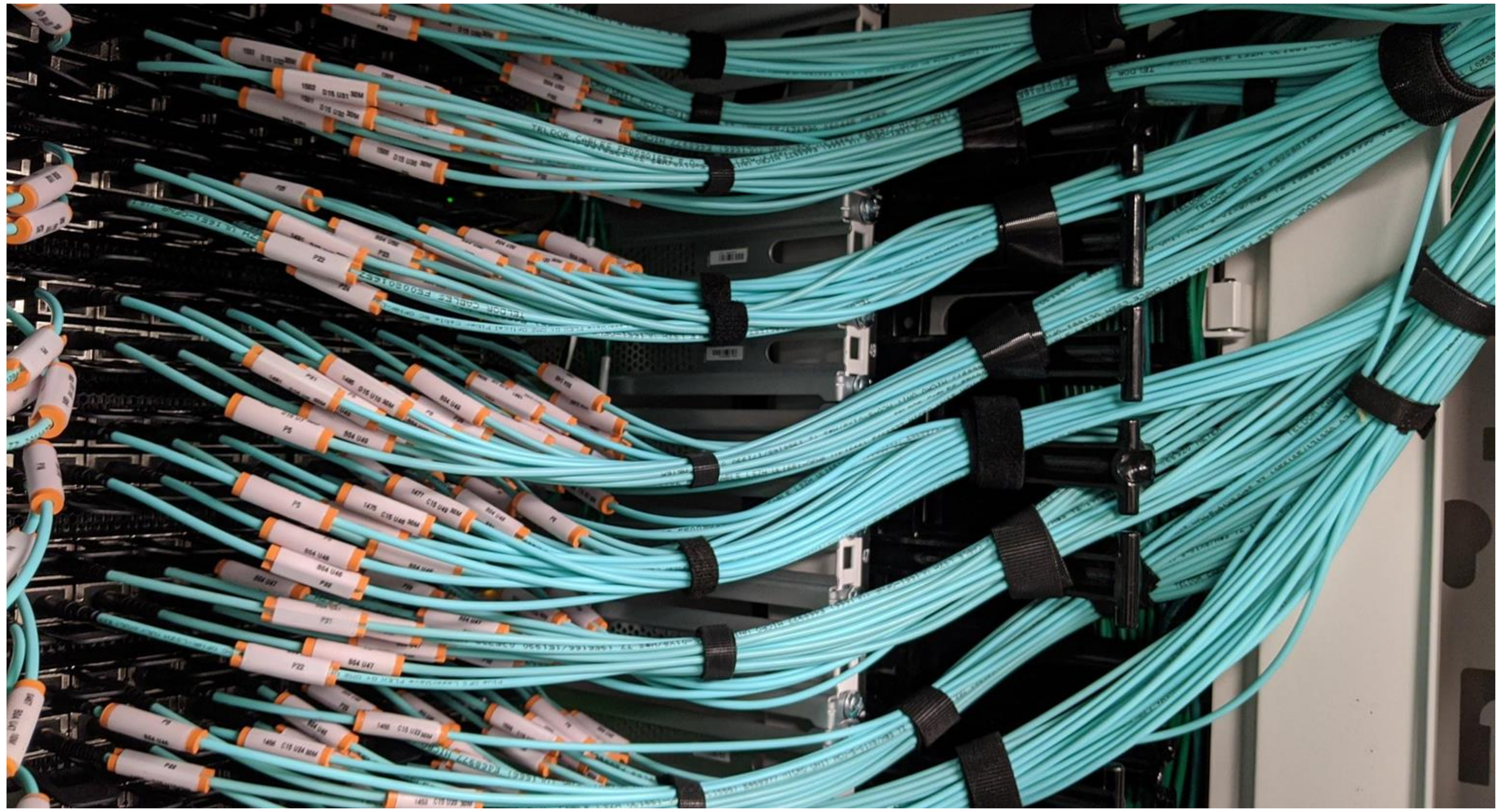


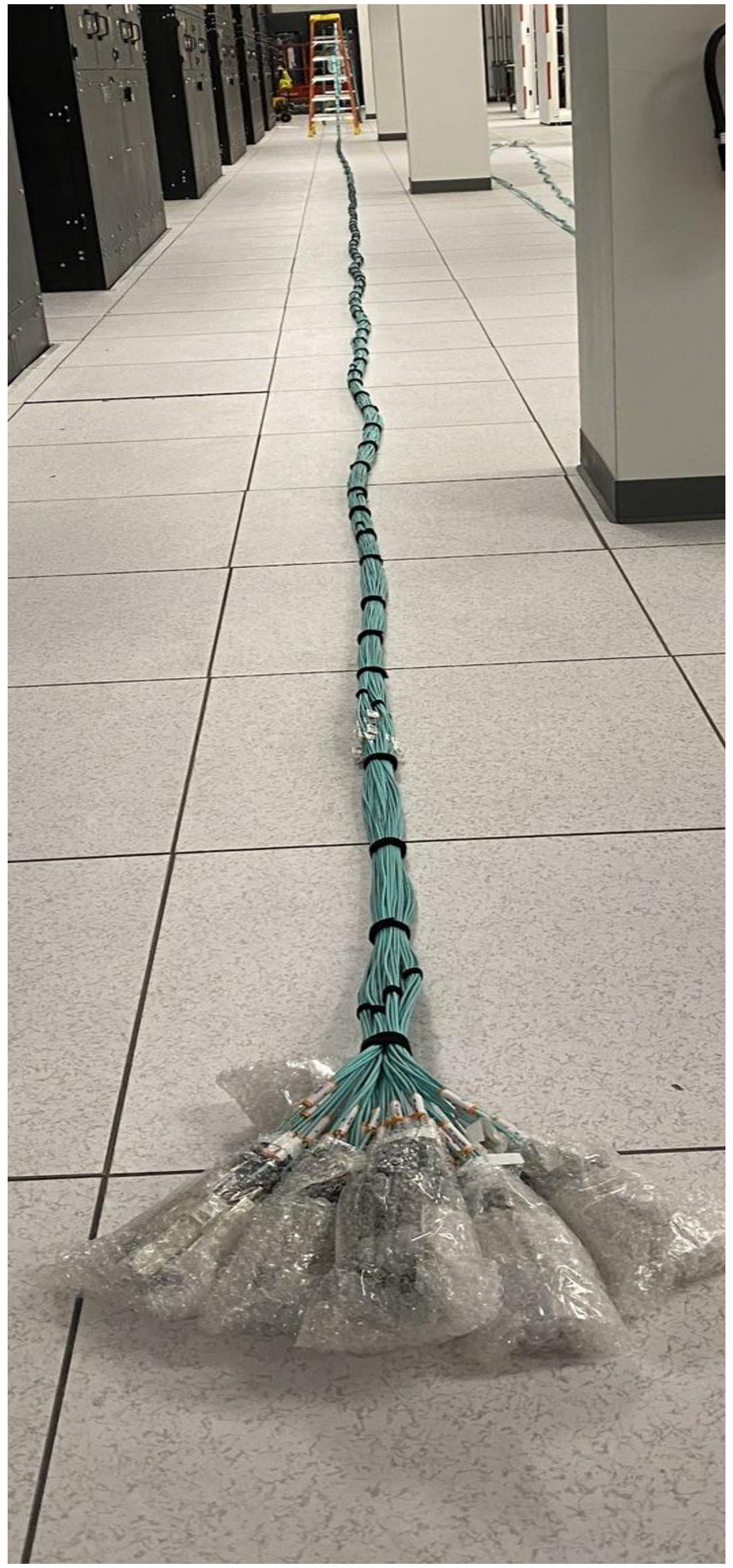
STORAGE HIERARCHY

- Memory (file) cache (aggregate): 224TB/sec — 1.1PB (2TB/node)
- NVMe cache (aggregate): 28TB/Sec — 16.8PB (30TB/node)
- Network filesystem (cache — Lustre): 2TB/sec — 10PB
- Object storage: 100GB/sec — 100+PB

60KM OF IB CABLES

a.k.a. why good dressing is crucial





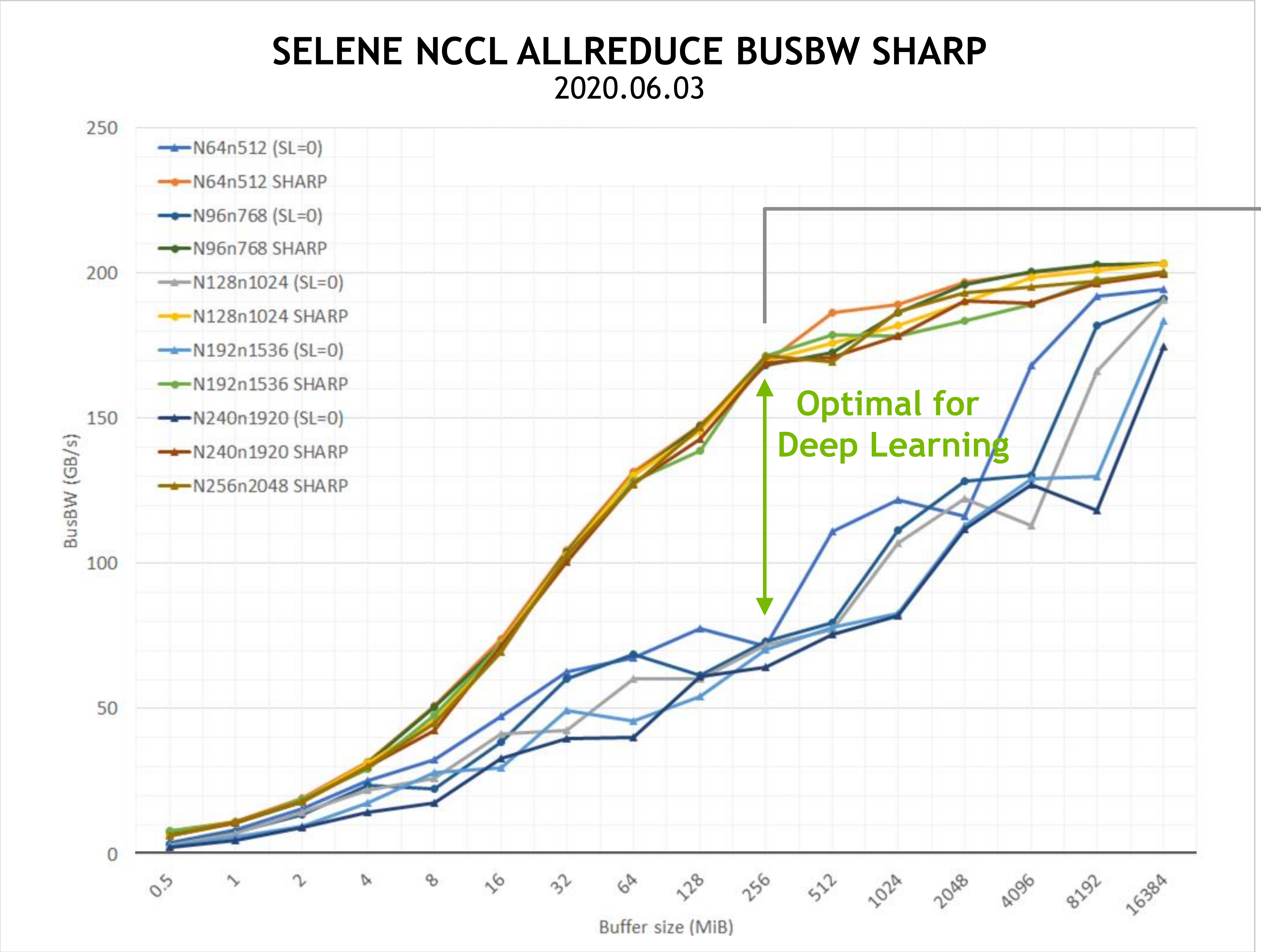
SELENE
DGX A100 SuperPOD Deployment

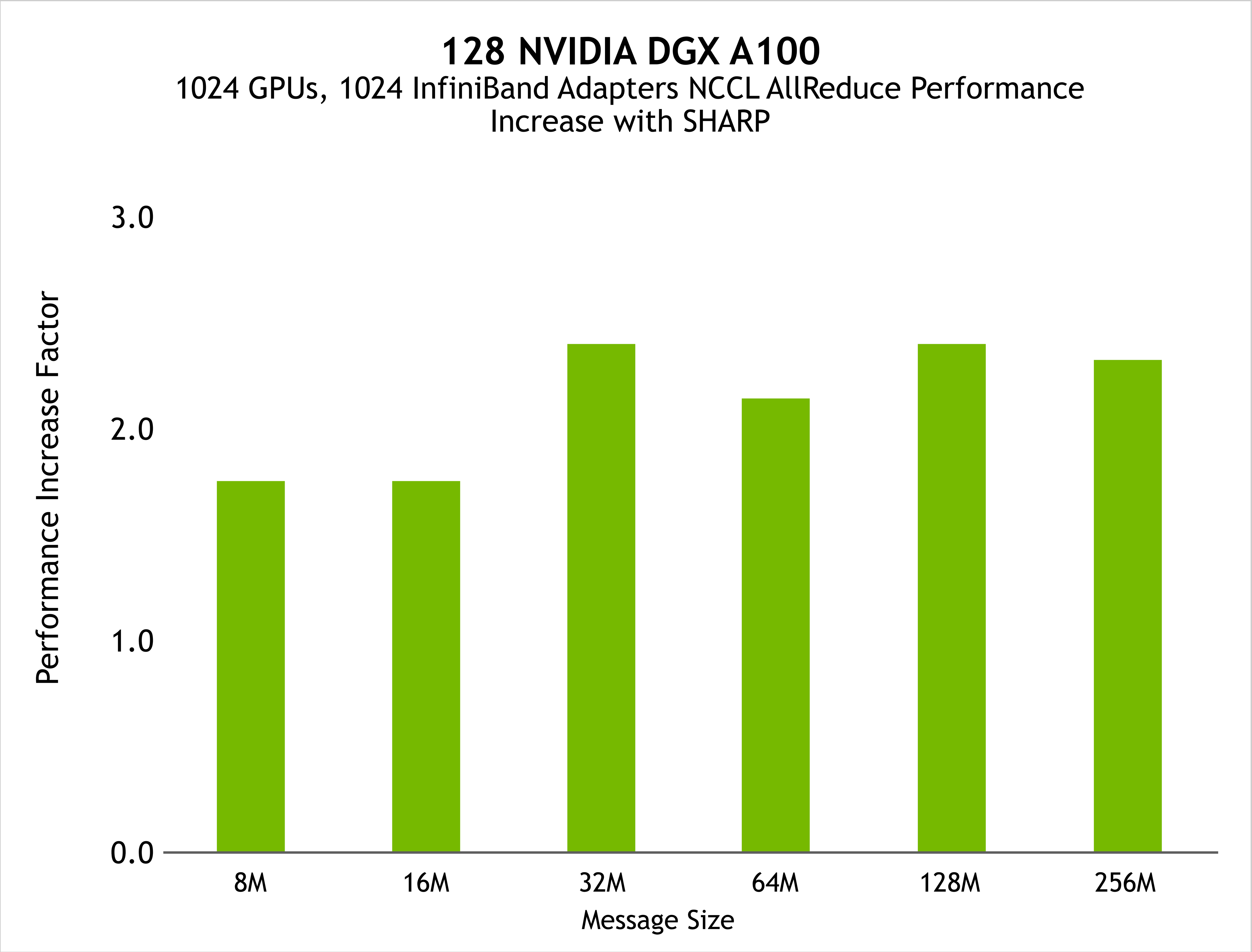
- #6 on TOP500 (63.46 PetaFLOPS HPL)
- #10 on Green500 (26.2 GF/W) — single scalable unit
- #5 on HPCG (1.6 PetaFLOPS)
- #3 on HPL-AI (556 PetaFLOPS)
- Fastest Industrial System in U.S. — 1+ ExaFLOPS AI
- Built with NVIDIA DGX SuperPOD Architecture
  - NVIDIA DGX A100 and NVIDIA Mellanox IB
  - NVIDIA's 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

SHARP OFFLOADS COMPUTE INTO NETWORK

HDR200 Selene Results





AIR COOLED DGX SUPERPOD

Airflow

NVIDIA

LIQUID COOLING PROTOTYPE

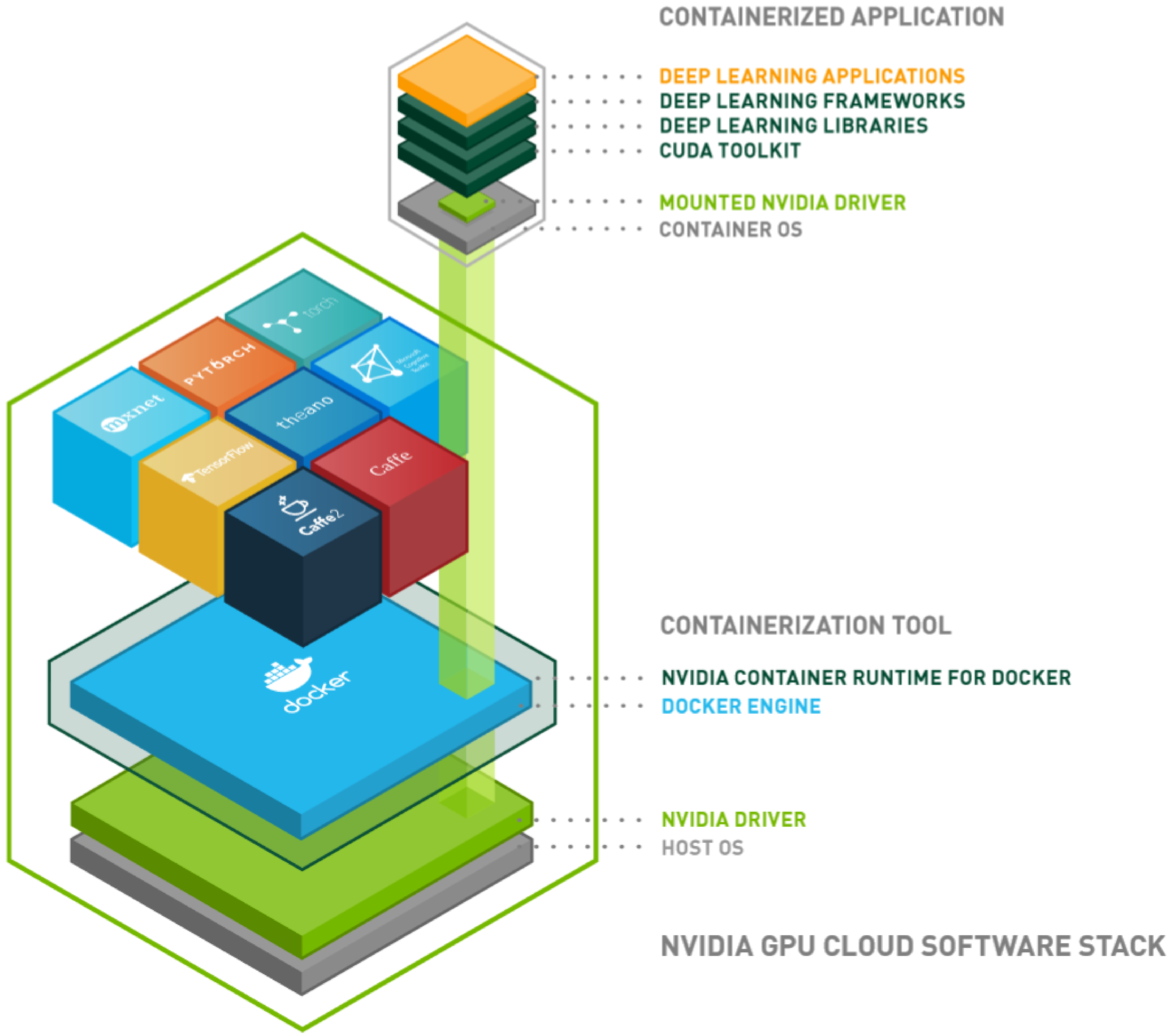
~15% Energy Efficiency Improvement

SOFTWARE OPERATIONS

SCALE TO MULTIPLE NODES

Software Stack — Application

- Deep Learning Model:
  - Hyperparameters tuned for multi-node scaling
  - Multi-node launcher scripts
- Deep Learning Container:
  - Optimized TensorFlow, GPU libraries, and multi-node software
- Host:
  - Host OS, GPU driver, IB driver, container runtime engine (docker, enroot)

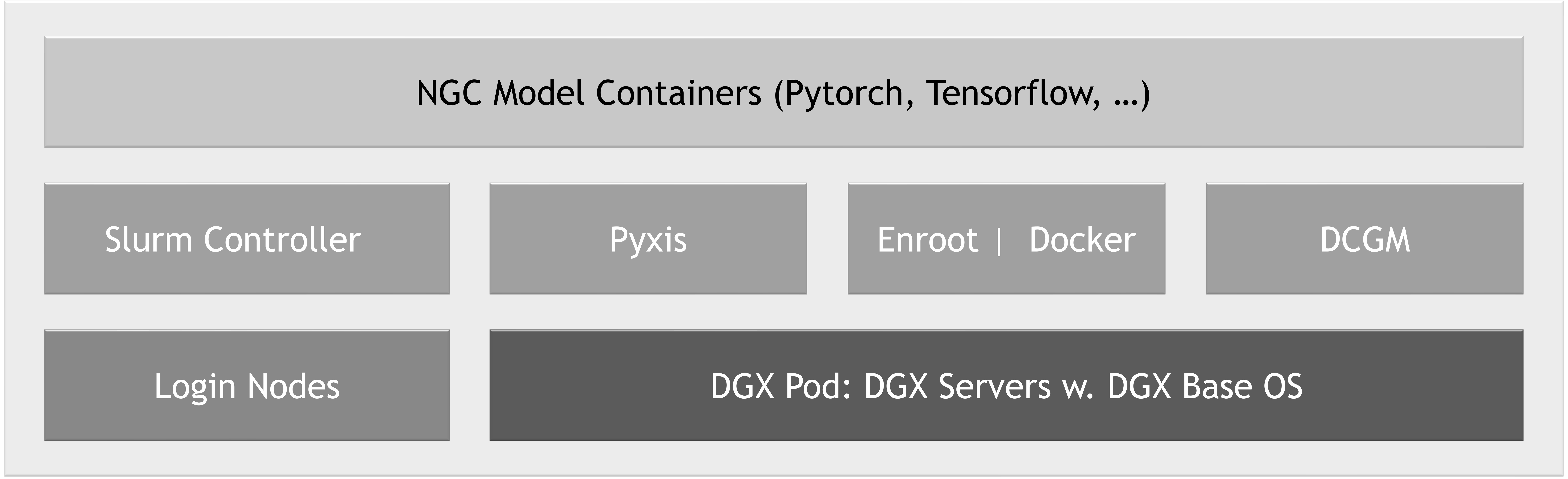


NVIDIA

SCALE TO MULTIPLE NODES

Software Stack — System

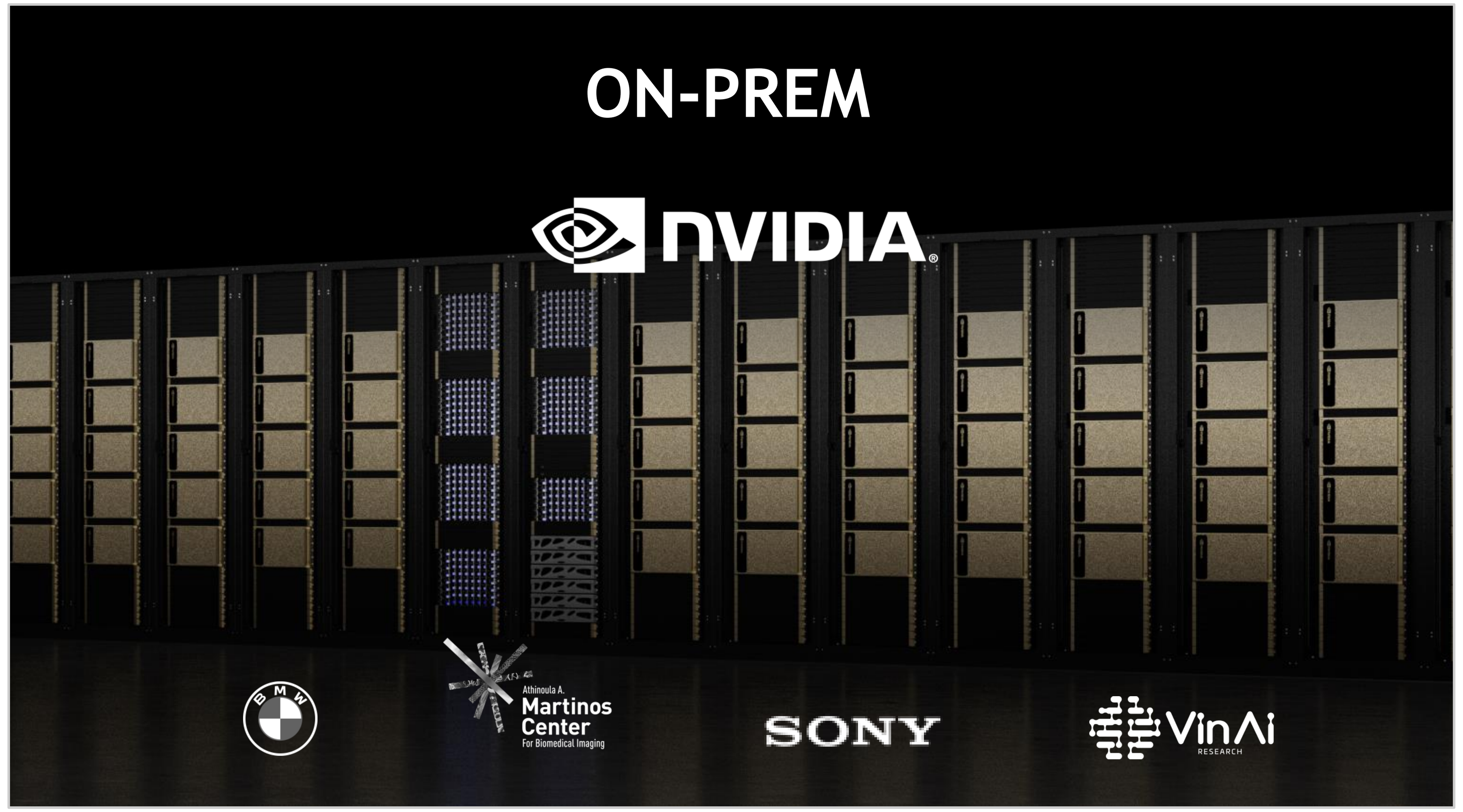
- Slurm: User job scheduling & management
- Enroot: NVIDIA open-source tool to convert traditional container/OS images into unprivileged sandboxes
- Pyxis: NVIDIA open-source plugin integrating Enroot with Slurm
- Base Command: NVIDIA services for GPU cluster management

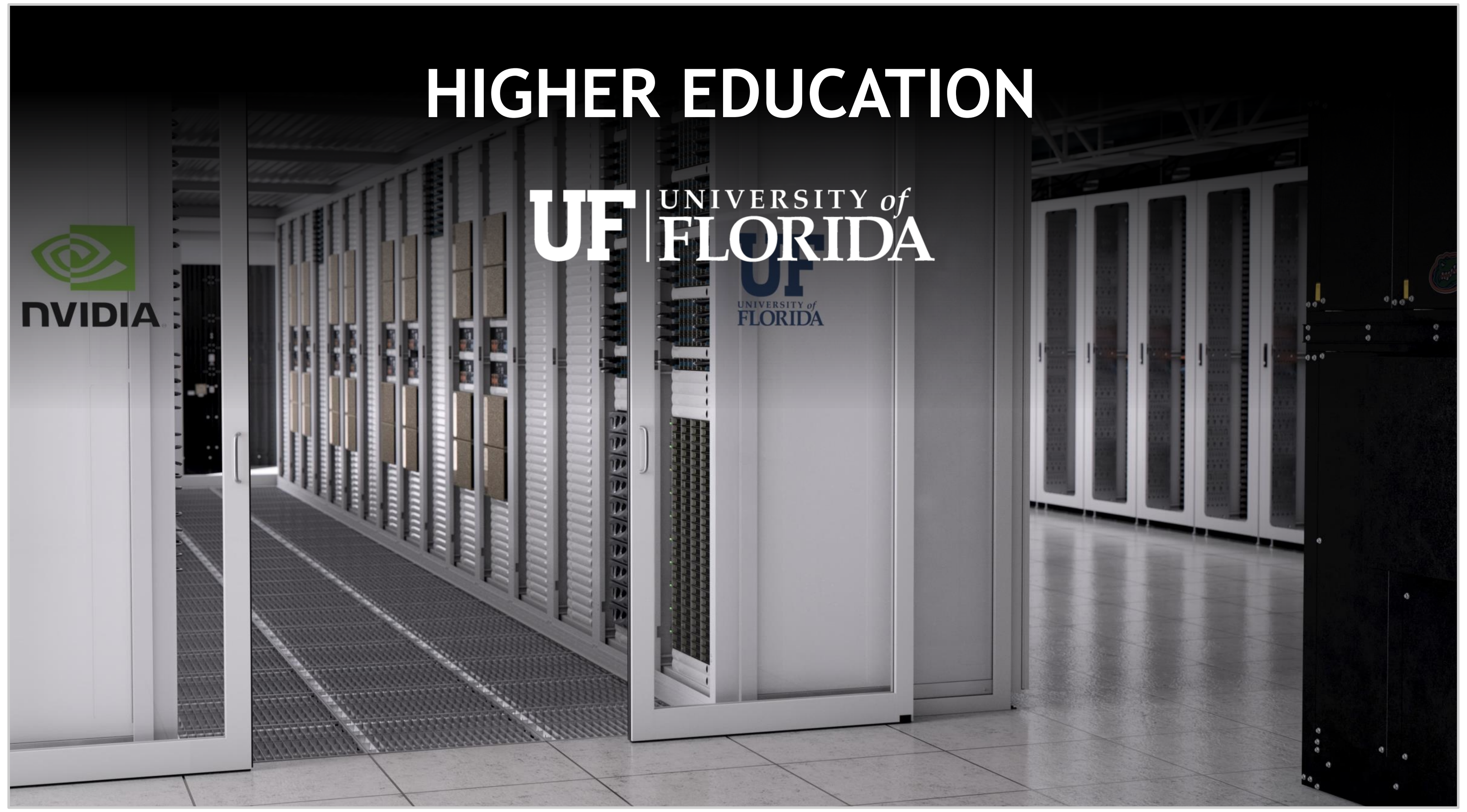


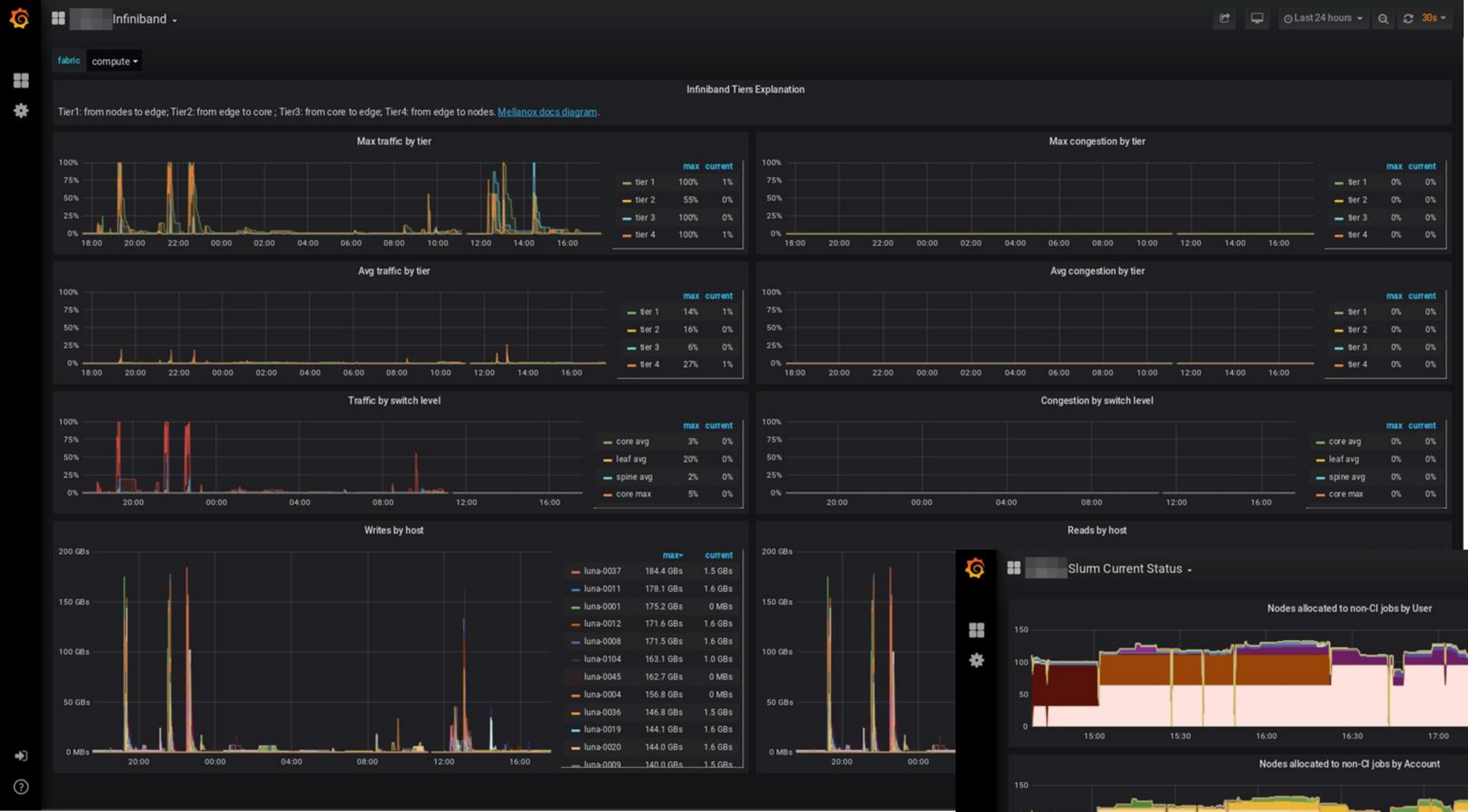
MICROSOFT AZURE NDm A100 v4 — THE WORLD’S FASTEST CLOUD INSTANCE ON DEMAND

Based on SuperPOD Design – Set Records on MLPerf v1.1 Training

NVIDIA SUPERPODS — EFFICIENT COMPUTE AT SCALE REPLICATING REFERENCE DESIGN

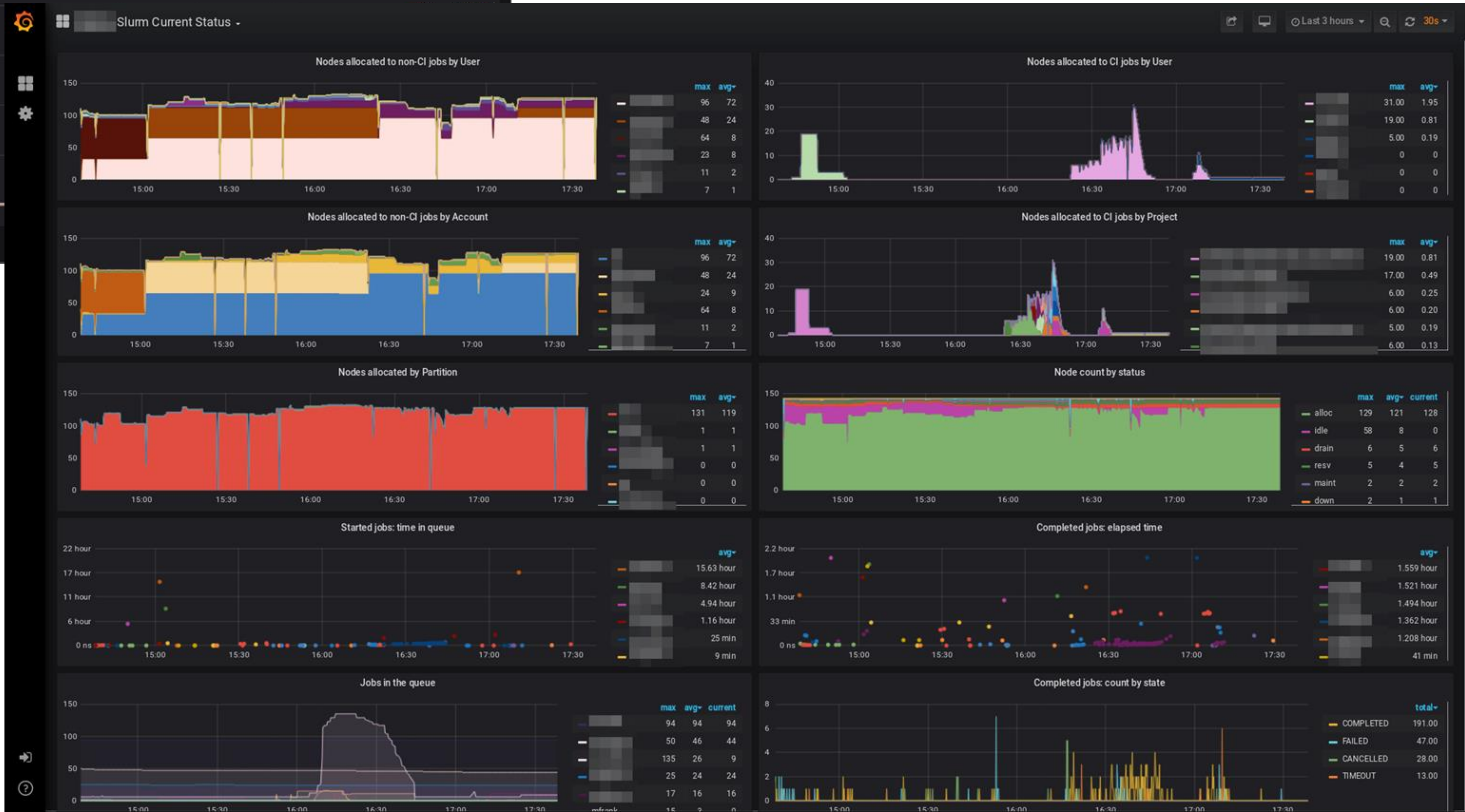






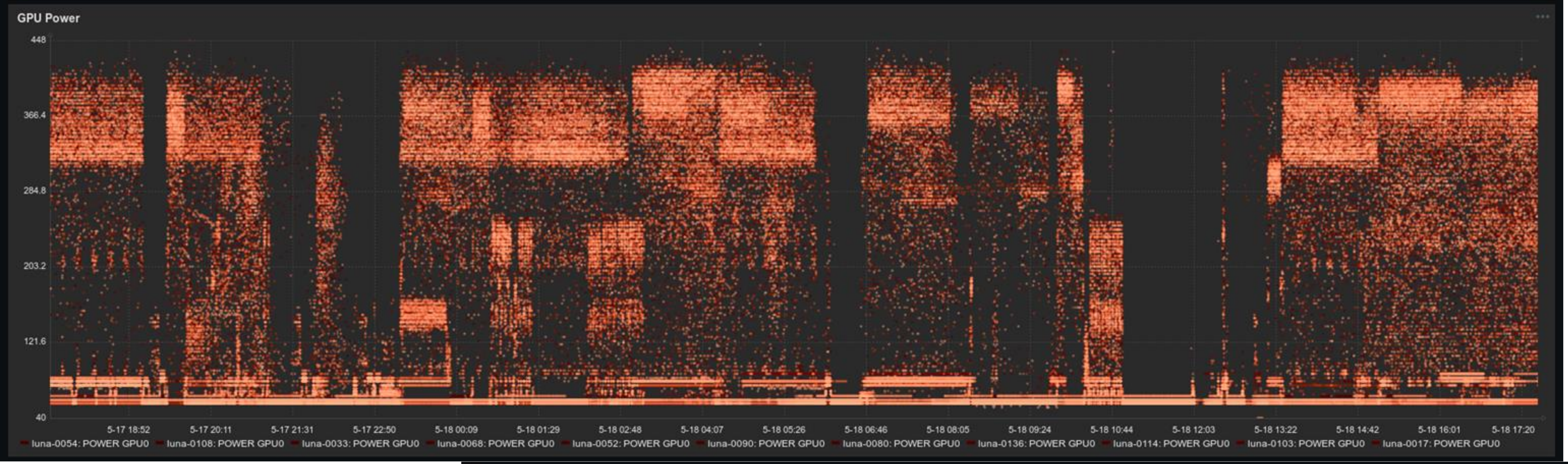
MONITORING

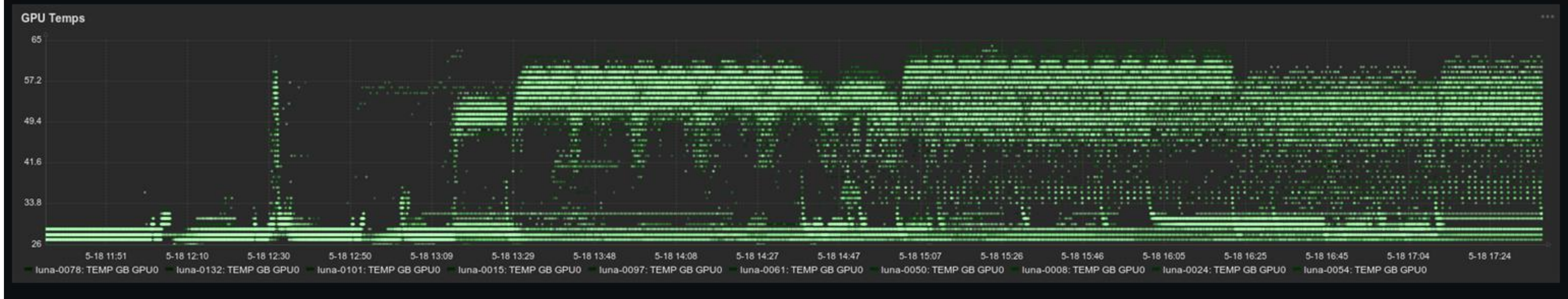
Infiniband, Power, Nodes, etc



MONITORING

Power, Thermals





ENERGY EFFICIENCY — MATCHING HIGH THROUGHPUT AND HIGH UTILIZATION

NVIDIA.

ENERGY EFFICIENCY — MATCHING HIGH THROUGHPUT AND HIGH UTILIZATION

Resources

Links and other doc

DGX A100 Page https://www.nvidia.com/en-us/data-center/dgx-a100/

Blogs

- DGX A100 SuperPOD https://blogs.nvidia.com/blog/2020/05/14/dgx-superpod-a100/
- https://blogs.nvidia.com/blog/2020/08/14/making-selene-pandemic-ai/
- DDN Blog for DGX A100 Storage https://www.ddn.com/press-releases/ddn-a3i-nvidia-dgx-a100/
- Kitchen Keynote summary https://blogs.nvidia.com/blog/2020/05/14/gtc-2020-keynote/
- Double Precision Tensor Cores https://blogs.nvidia.com/blog/2020/05/14/double-precision-tensor-cores/

Want to join us?

If you are interested in exploring roles at NVIDIA:

https://www.nvidia.com/en-us/about-nvidia/careers/


NVIDIA