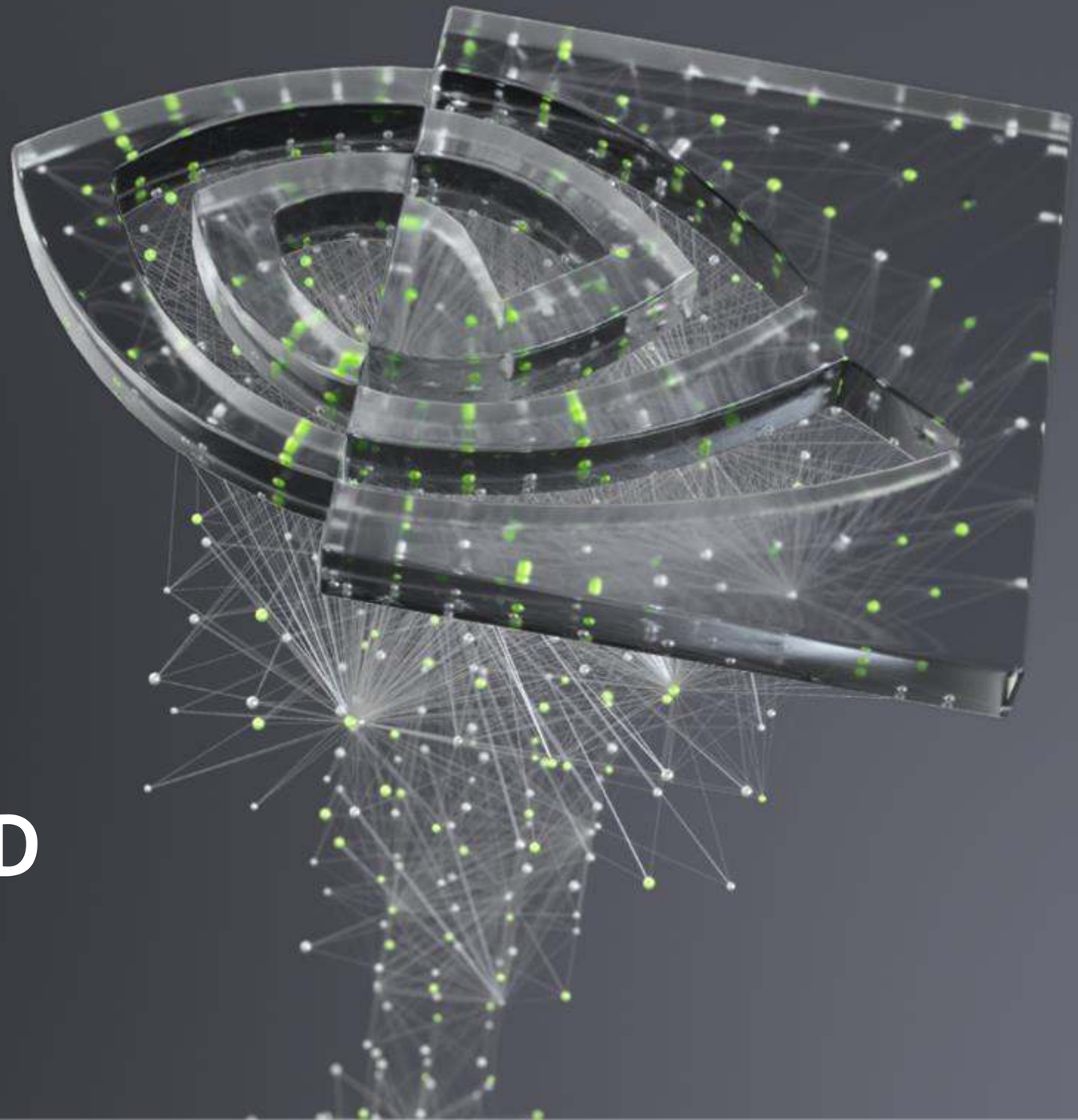


DGX A100 SUPERPOD

Mike Houston, Chief Architect - AI Systems

DATA CENTER ARCHITECTURE

SELENE

DGX A100 SuperPOD Deployment

#1 on MLPerf for commercially available systems

#7 on TOP500 (27.6 PetaFLOPS HPL)

#2 on Green500 (20.5 GigaFLOPS/watt)

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

Built with NVIDIA DGX SuperPOD Arch in 3 Weeks

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

Configuration:

- 2,240 NVIDIA A100 Tensor Core GPUs
- 280 NVIDIA DGX A100 systems
- 494 Mellanox 200G HDR IB switches
- 7 PB of all-flash storage

LESSONS LEARNED

How to Build and Deploy HPC Systems with Hyperscale Sensibilities

Speed and feed matching

Thermal and power design

Interconnect design

Deployability

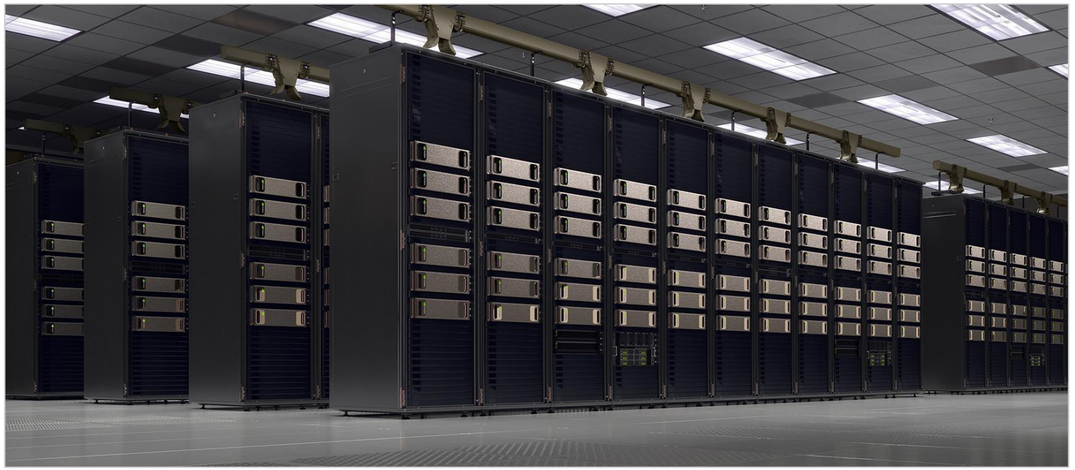
Operability

Flexibility

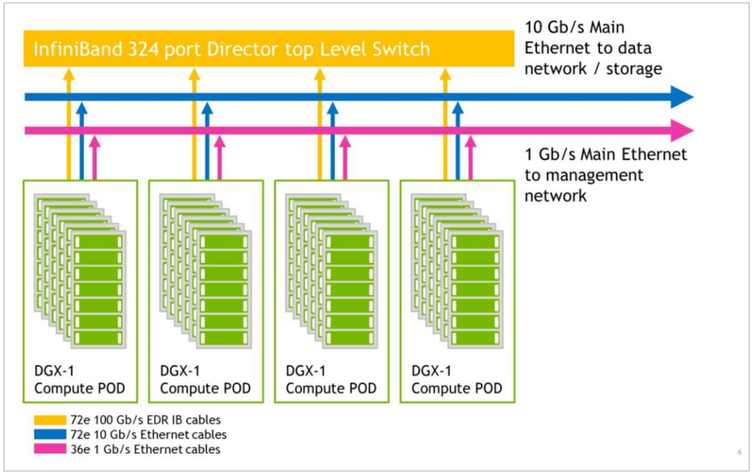
Expandability

DGX-1 PODs

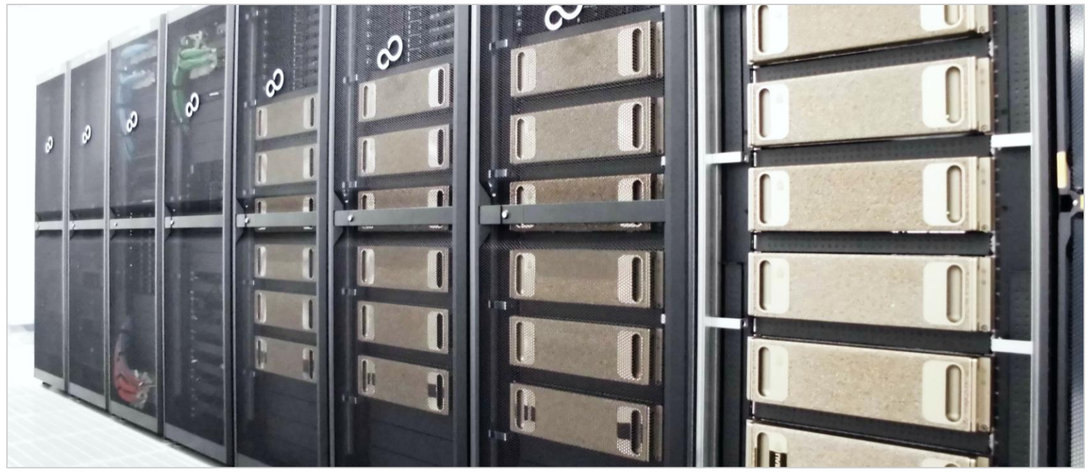
NVIDIA DGX-1 – original layout



DGX-1 Multi-POD



RIKEN RAIDEN



NVIDIA DGX-1 – new layout


NVIDIA
NVIDIA

A NEW DATA CENTER DESIGN

DGX A100 SUPERPOD

Fast Deployment Ready - Cold Aisle Containment Design

A NEW GENERATION OF SYSTEMS

NVIDIA DGX A100

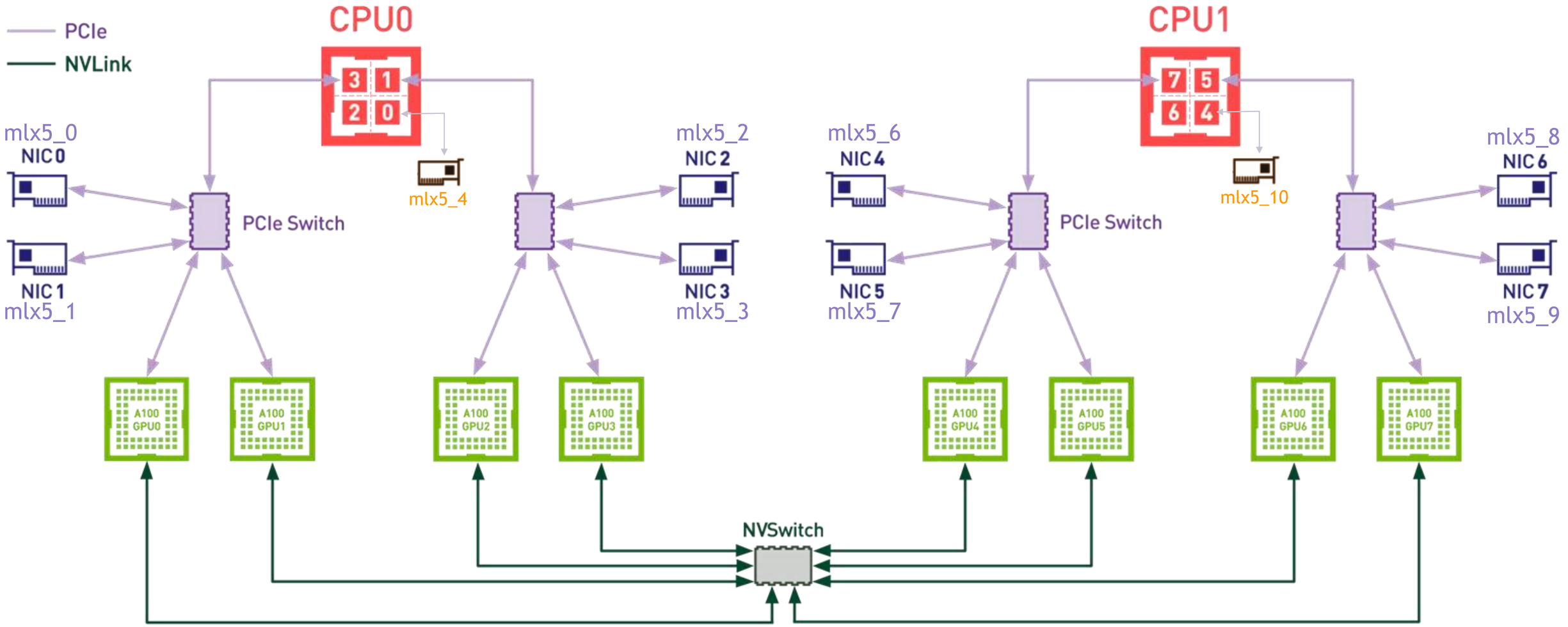
GPUs	**8x NVIDIA A100**
GPU Memory	320 GB total
Peak performance	5 petaFLOPS AI | 10 petaOPS INT8
NVSwitches	6
System Power Usage	**6.5kW max**
CPU	**Dual AMD Rome 7742** 128 cores total, 2.25 GHz(base), 3.4GHz (max boost)
System Memory	**1TB**
Networking	**8x Single-Port Mellanox ConnectX-6** 200Gb/s 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 3.84TB) 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 temp range	**5°C to 30°C** (41°F to 86°F)

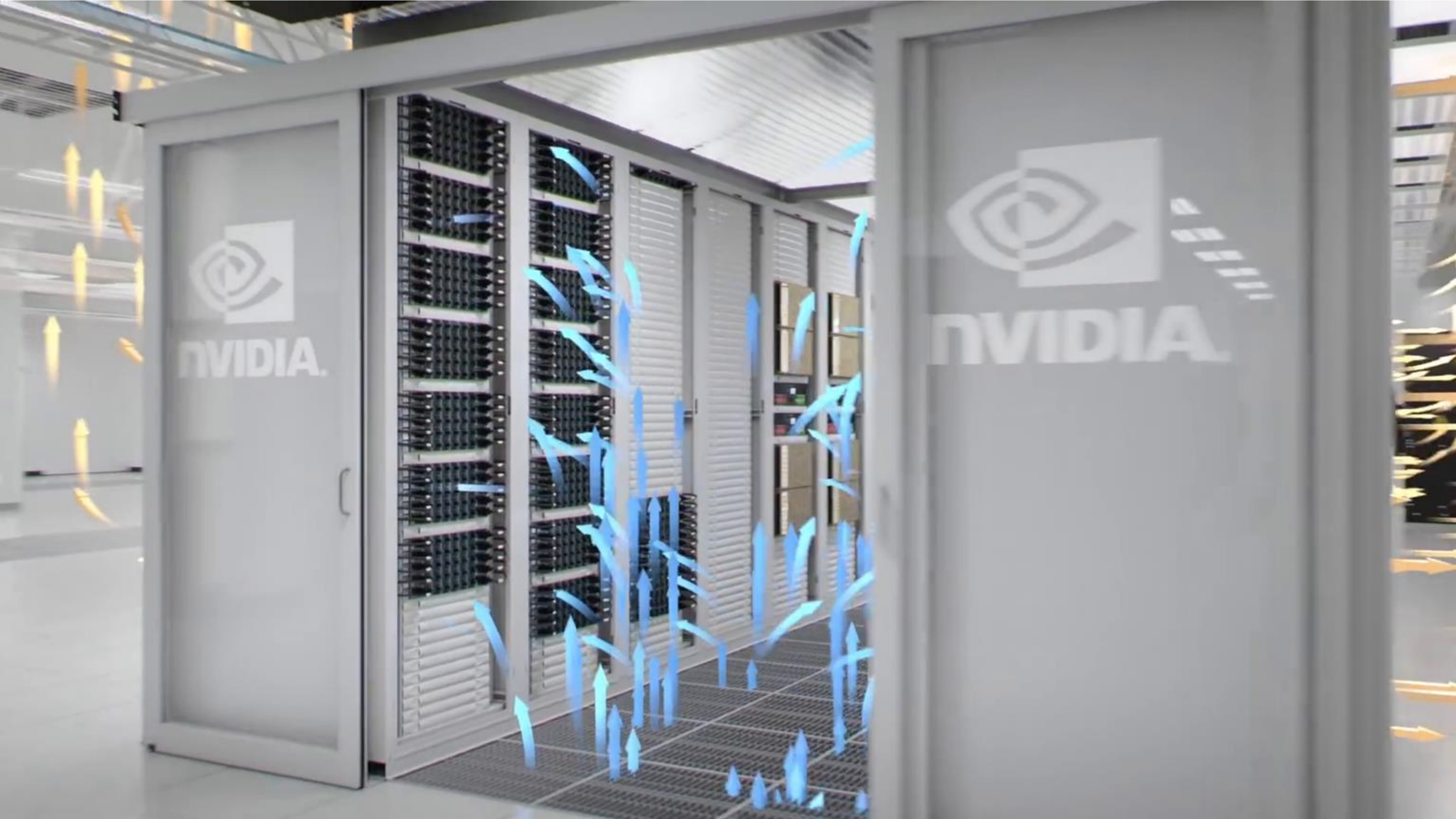
* Optional upgrades

DGX A100

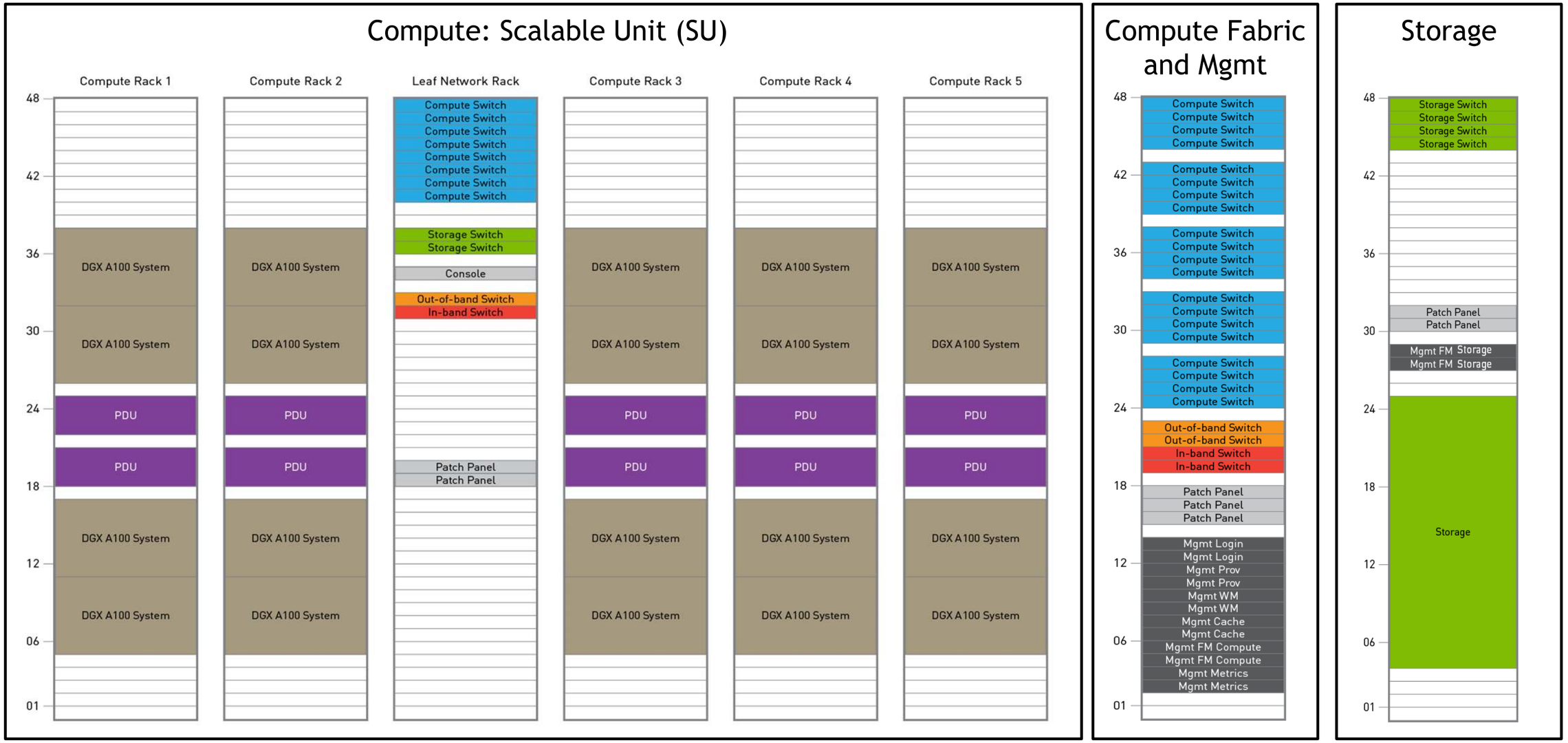
Data plane (can be used as eth or IB)
Compute plane (IB)

High-level Topology Overview (with options)




NVIDIA
NVIDIA

MODULARITY: RAPID DEPLOYMENT



DGX A100 SUPERPOD

A Modular Model

1K GPU SuperPOD Cluster

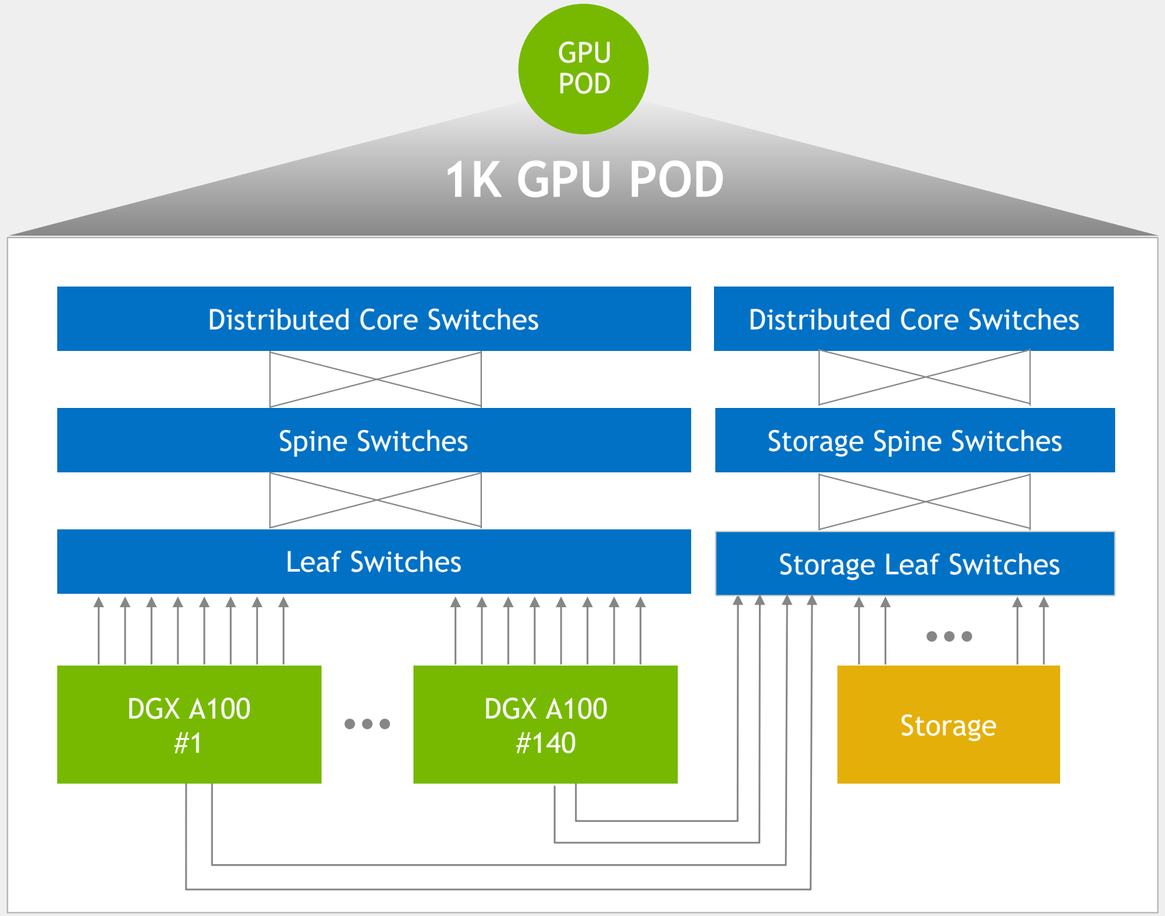
- 140 DGX A100 nodes (1,120 GPUs) in a GPU POD
- 1st tier fast storage - DDN AI400x with Lustre
- Mellanox HDR 200Gb/s InfiniBand - Full Fat-tree
- Network optimized for AI and HPC

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

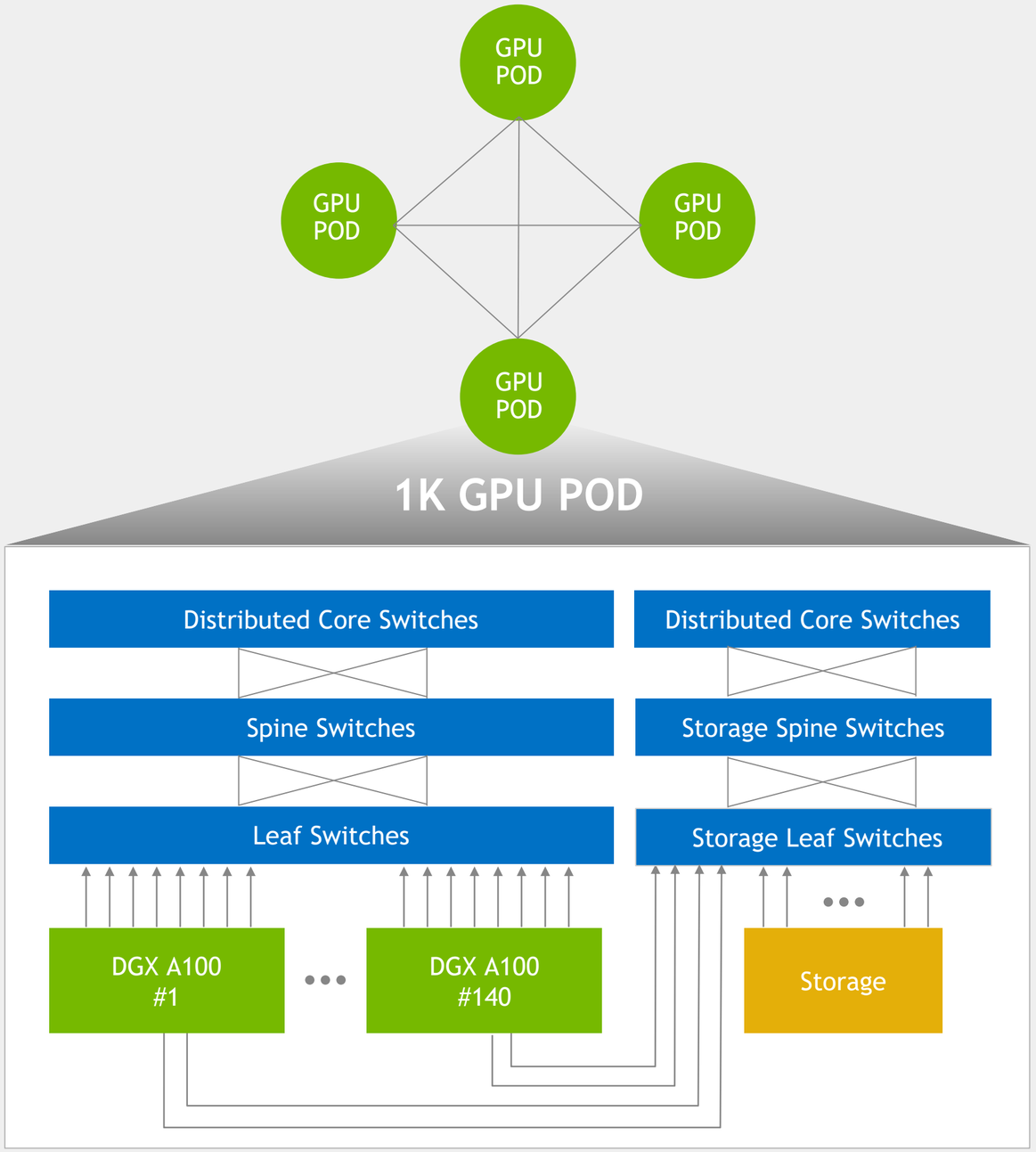


DGX A100 SUPERPOD

An Extensible Model

POD to POD

- Modular IB Fat-tree or DragonFly+
  - Core IB Switches Distributed Between PODs
  - Direct connect POD to POD



MULTI NODE IB COMPUTE

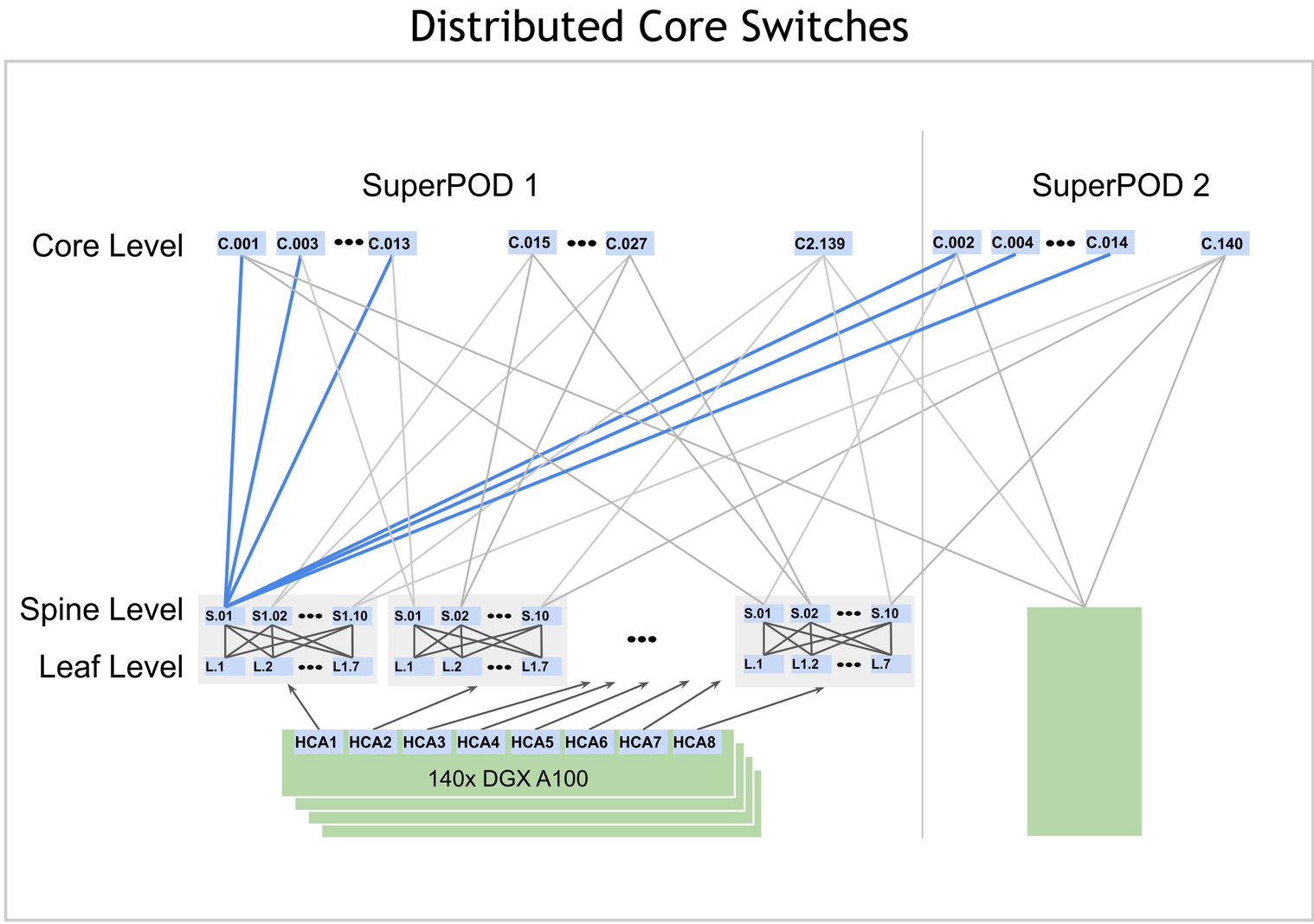
The Details

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 SuperPOD
- SuperPOD contains compute nodes and storage
- All nodes and storage are usable between SuperPODs

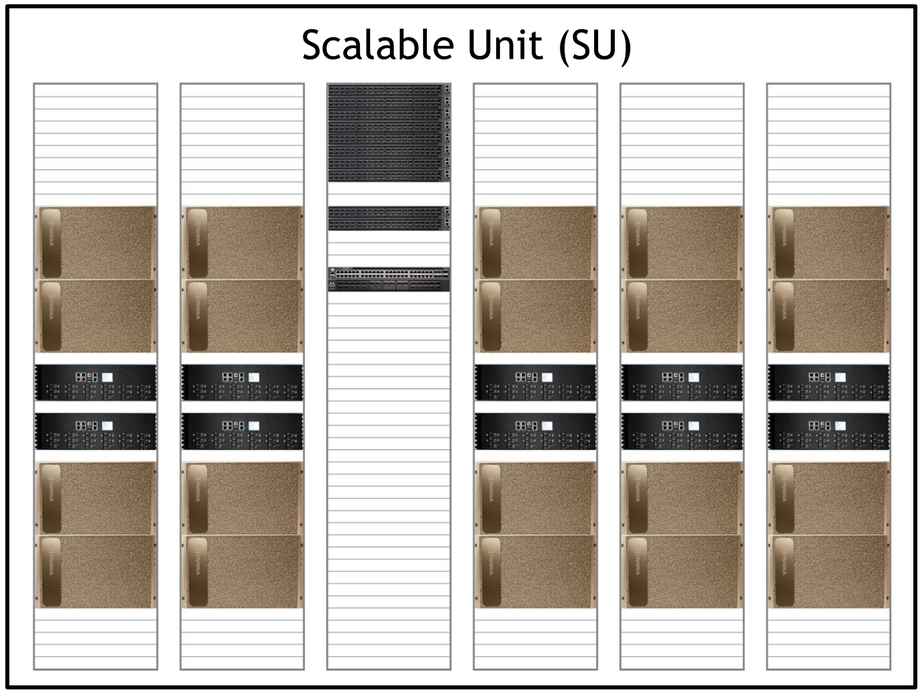
Sharpv2 optimized design
- Leaf and Spines organized in HCA planes
- For a SuperPOD, all HCA1 from 140 DGX-2 connect to a HCA1 Plane fat-tree network
- Traffic from HCA1 to HCA1 between any two nodes in a POD stay either at the Leaf or Spine level
- Only use core switches when
  - Moving data between HCA planes (e.g. mlx5_0 to mlx5_1 in another system)
  - Moving any data between SuperPODs

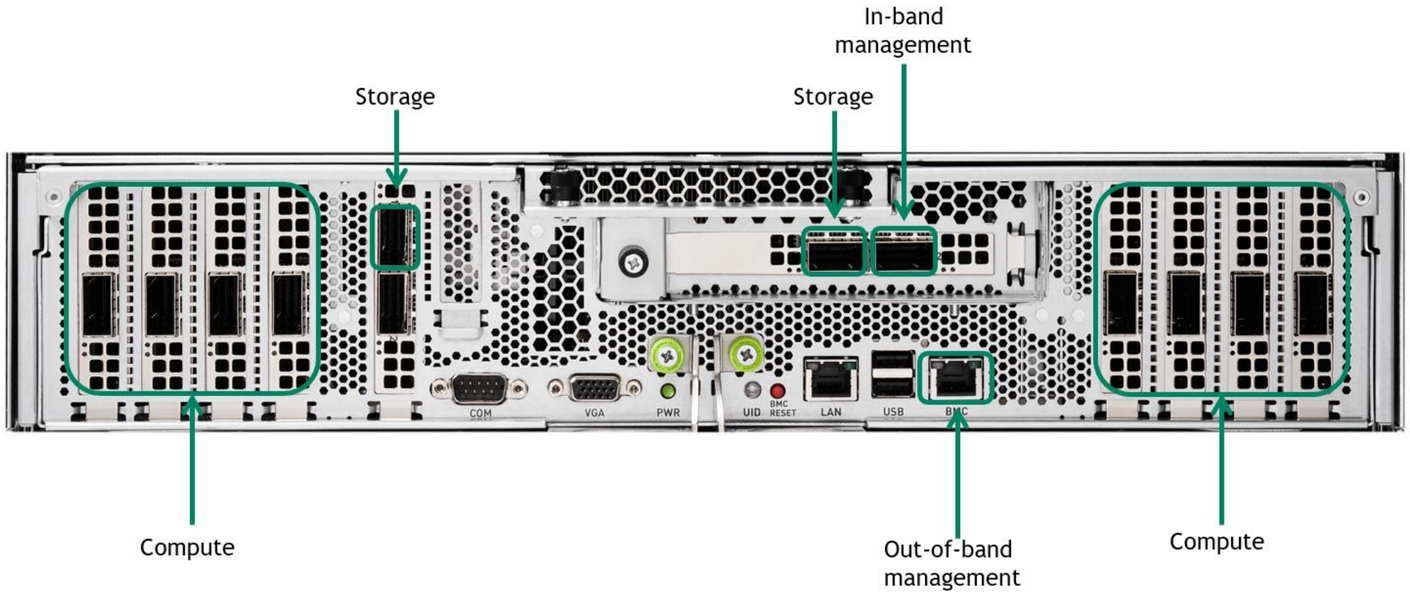


DESIGNING FOR PERFORMANCE

In the Data Center

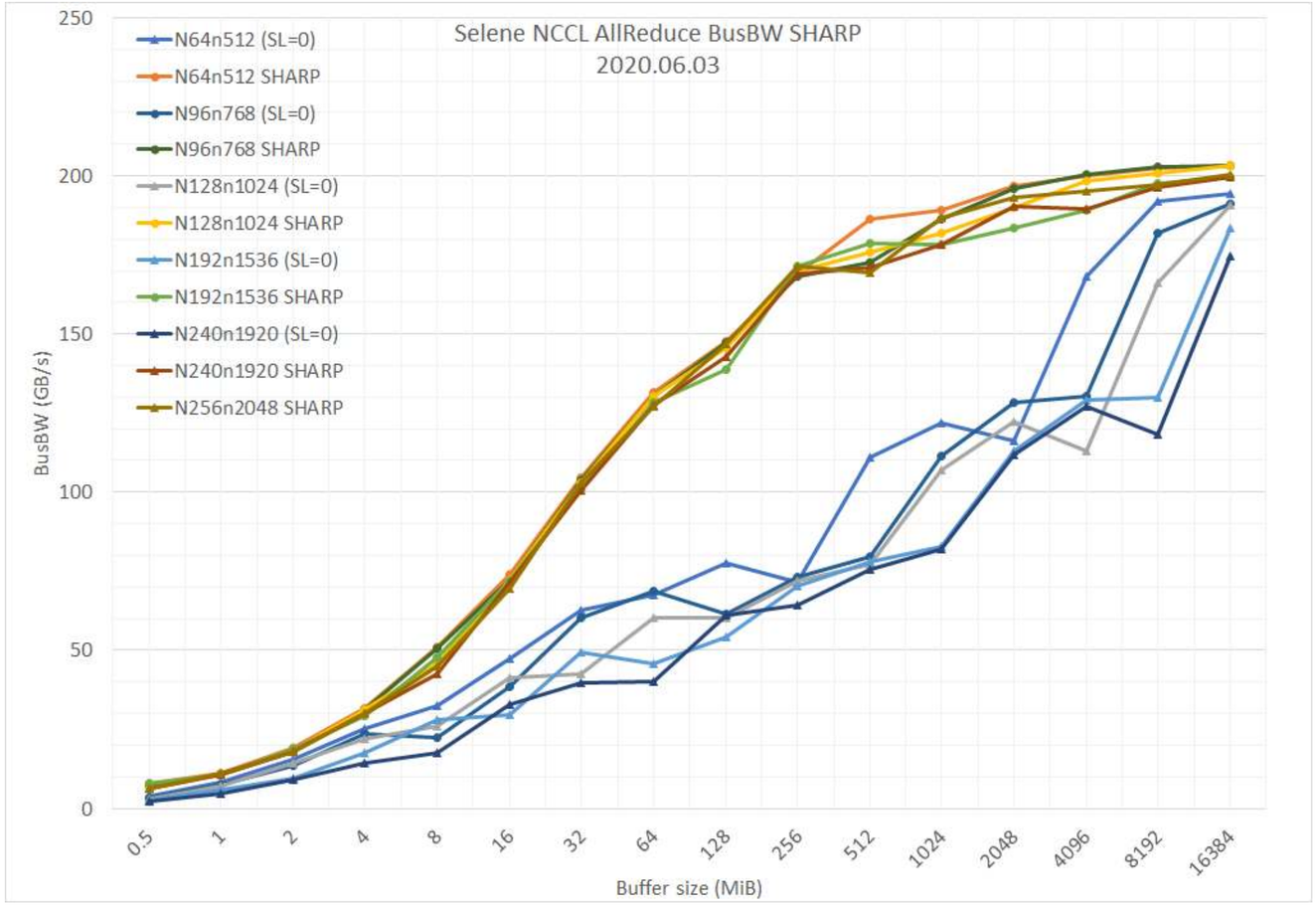
All design is based on a radix optimized approach for Sharpv2 support and fabric performance and to align with design of Mellanox Quantum switches.

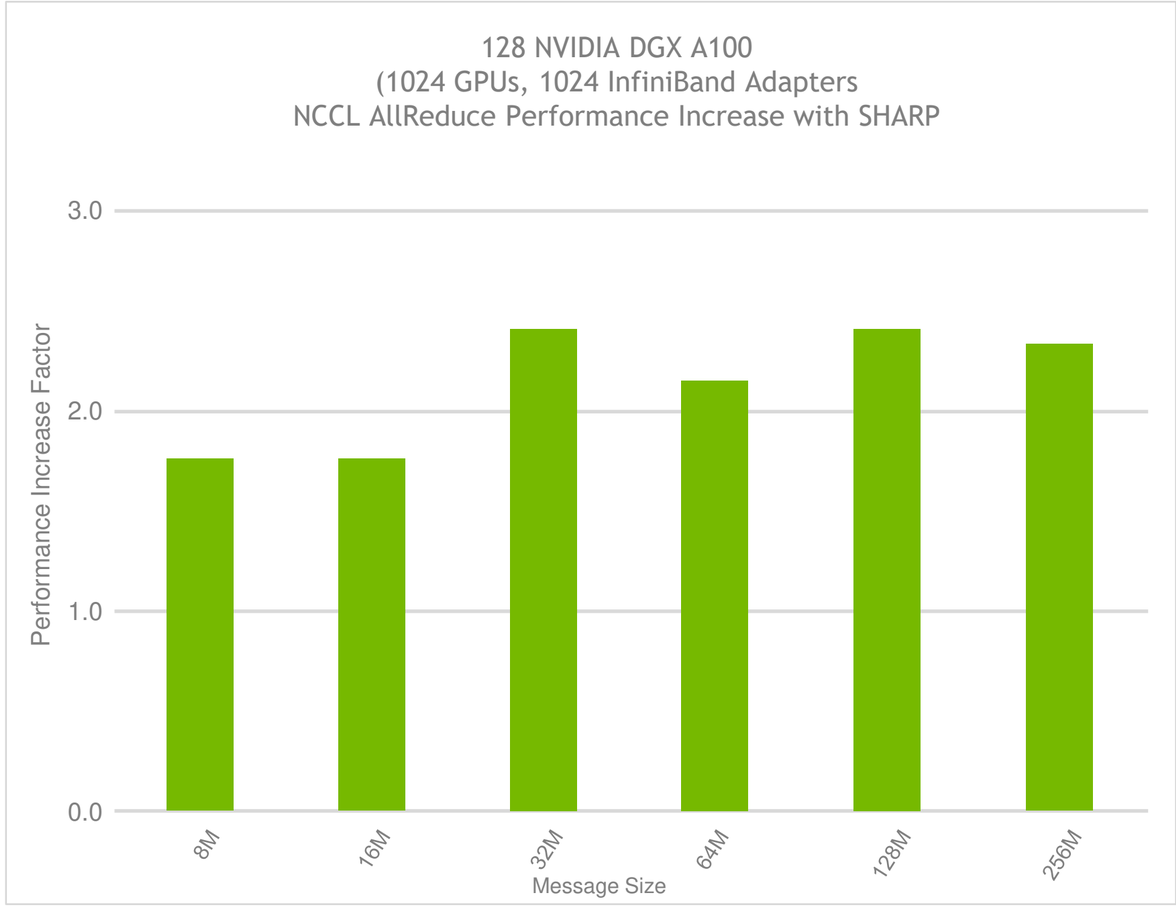




SHARP

HDR200 Selene Early Results





STORAGE

Parallel filesystem for perf and NFS for home directories

Per SuperPOD:

Fast Parallel FS: Lustre (DDN)
- 10 DDN AI400X Units
- Total Capacity: 2.5 PB
- Max Perf Read/Write: 490/250 GB/s
- 80 HDR-100 cables required
- 16.6KW

Shared FS: Oracle ZFS5-2
- HA Controller Pair/768GB total
- 8U Total Space (4U per Disk Shelf, 2U per controller)
- 76.8 TB Raw - 24x3.2TB SSD
- 16x40GbE
- Key features: NFS, HA, snapshots, dedupe
- 2kW

STORAGE HIERARCHY

- Memory (file) cache (aggregate): 112TB/sec - 0.5PB (2TB/node)
- NVMe cache (aggregate): 14TB/Sec - 8.4PB (30TB/node)
- Network filesystem (cache - Lustre): 1TB/sec - 5PB
- Object storage: 100GB/sec - 100+PB

BUILDING DURING COVID-19

BUILDING AT SPEED OF LIGHT

Challenges We Had

- Bring-up + build
- Building at scale right away
- Very tight schedules

What We Built

- Procedures for rapid deployment
- Scalable unit design
- DCOps training material

What We Got

- Top 10 supercomputer
- Fastest available MLPerf machine

MODULAR DEPLOYMENT SPEED

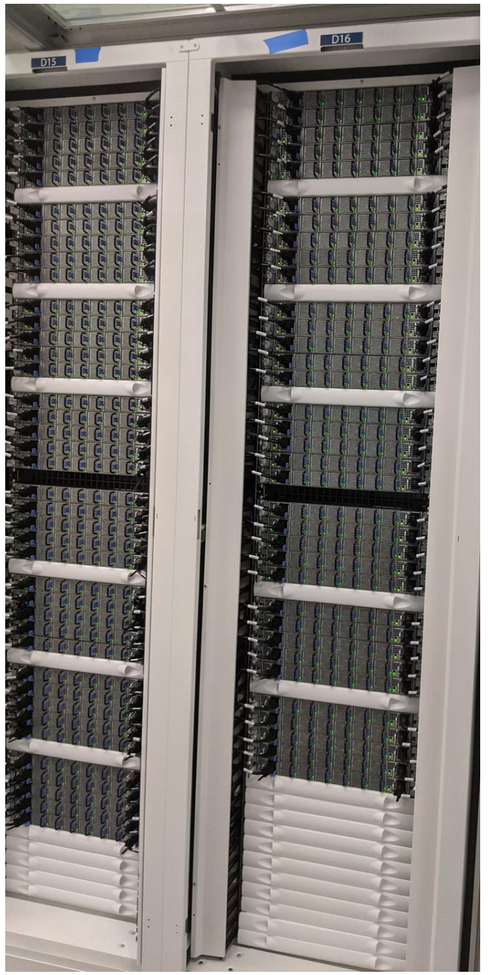
Each SuperPOD (140 systems) < 10 days

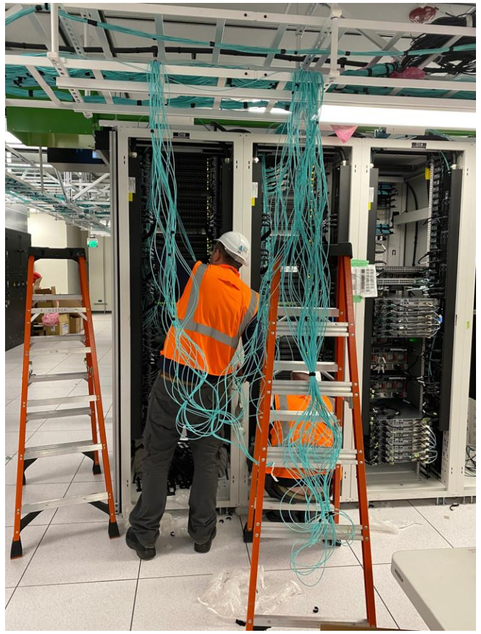
- Designed to be deployable by 2 DCOps engineers doing 20 systems per shift
- Maximum we deployed was 60 systems in one day across 2 shifts (loading dock limited)
- Rack, connectivity check, automated provisioning, burn-in, identify issues, fix, hand off
- Average time from racked to user running is 4 hours


CAUTION
DANGER


WARNING
TIPPING HAZARD. DO NOT LEAN.

BUNDLING OF CABLES OFFSITE



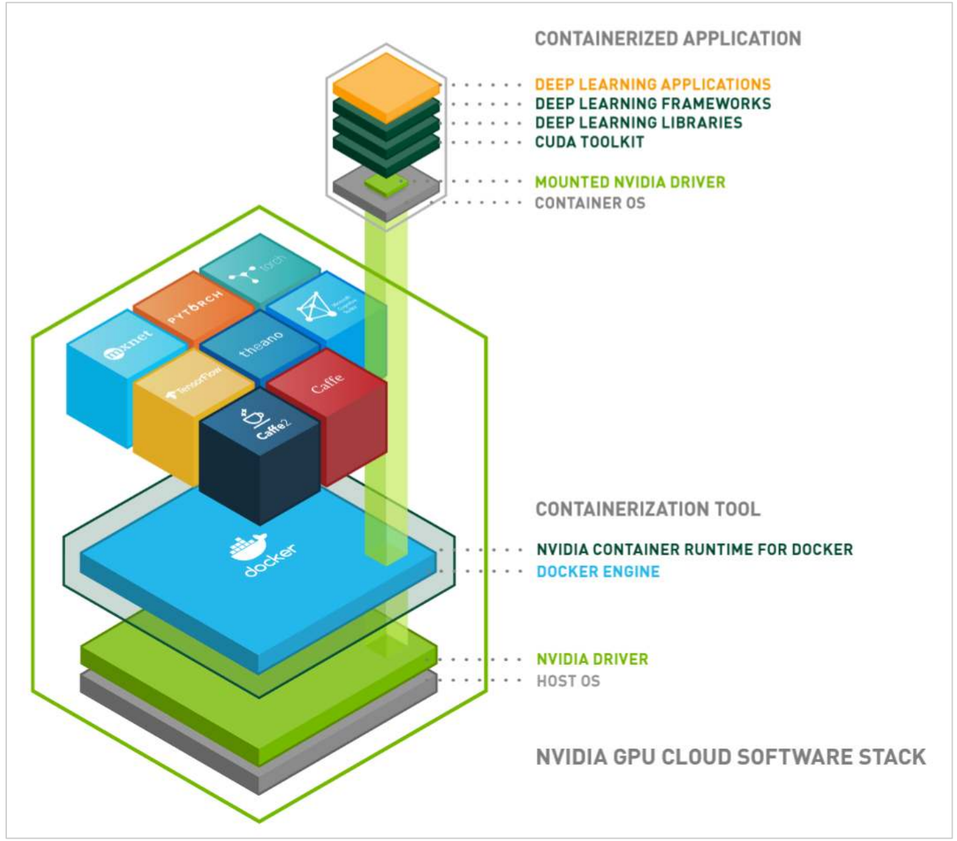


SW OPERATIONS

SCALE TO MULTIPLE NODES

Software Stack - Application

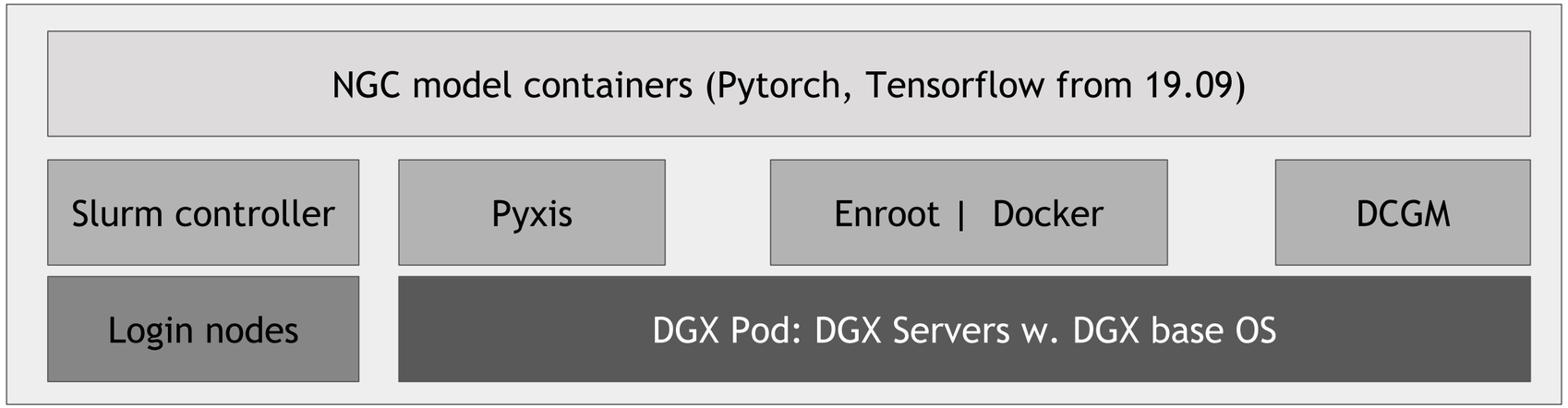
- Deep Learning Model:
  - Hyperparameters tuned for multi-node scaling
  - Multi-node launcher scripts
- Deep Learning Container:
  - Optimized TensorFlow, GPU libraries, and multi-node software
- Host:
  - Host OS, GPU driver, IB driver, container runtime engine (docker, enroot)



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
- DeepOps: NVIDIA open-source toolbox for GPU cluster management w/Ansible playbooks



INTEGRATING CLUSTERS IN THE DEVELOPMENT WORKFLOW

Supercomputer-scale CI (Continuous integration internal at NVIDIA)

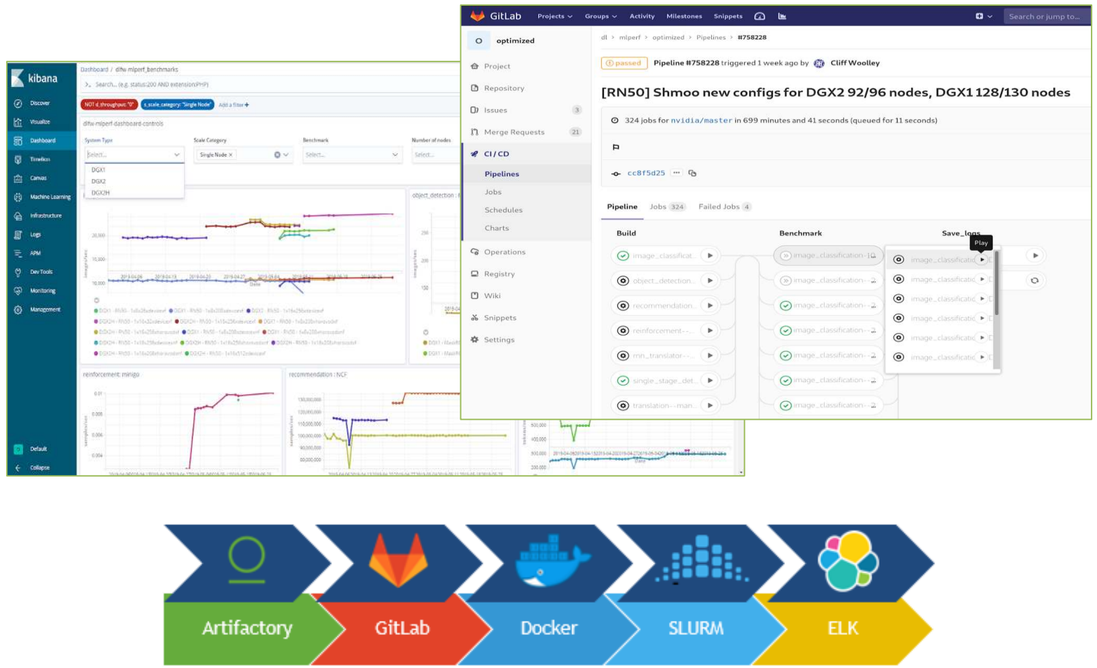
- Integrating DL-friendly tools like GitLab, Docker w/ HPC systems

  Kick off 10000's of GPU hours of tests with a single button click in GitLab

  … build and package with Docker
  … schedule and prioritize with SLURM
  … on demand or on a schedule
  … reporting via GitLab, ELK stack, Slack, email

  Emphasis on keeping things simple for users while hiding integration complexity

  Ensure reproducibility and rapid triage



LINKS

RESOURCES

Presentations

GTC Sessions (https://www.nvidia.com/en-us/gtc/session-catalog/) :

- Under the Hood of the new DGX A100 System Architecture [S21884]
- Inside the NVIDIA Ampere Architecture [S21730]
- CUDA New Features And Beyond [S21760]
- Inside the NVIDIA HPC SDK: the Compilers, Libraries and Tools for Accelerated Computing [S21766]
- Introducing NVIDIA DGX A100: the Universal AI System for Enterprise [S21702]
- Mixed-Precision Training of Neural Networks [S22082]
- Tensor Core Performance on NVIDIA GPUs: The Ultimate Guide [S21929]
- Developing CUDA kernels to push Tensor Cores to the Absolute Limit on NVIDIA A100 [S21745]

HotChips:

- Hot Chips Tutorial - Scale Out Training Experiences – Megatron Language Model
- Hot Chips Session - NVIDIA’s A100 GPU: Performance and Innovation for GPU Computing

Pyxis/Enroot https://fosdem.org/2020/schedule/event/containers_hpc_unprivileged/

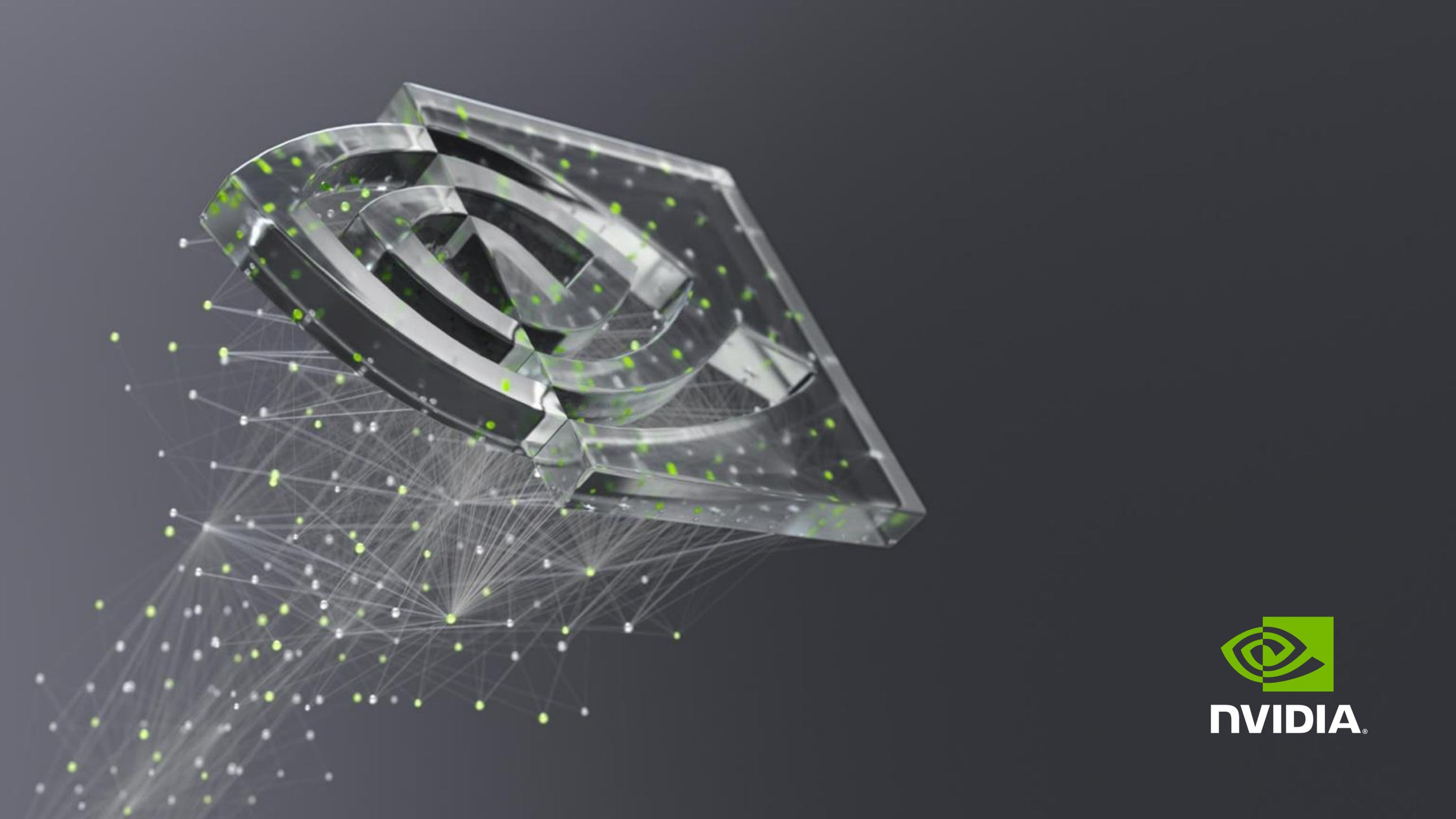
RESOURCES

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/
- 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/


NVIDIA