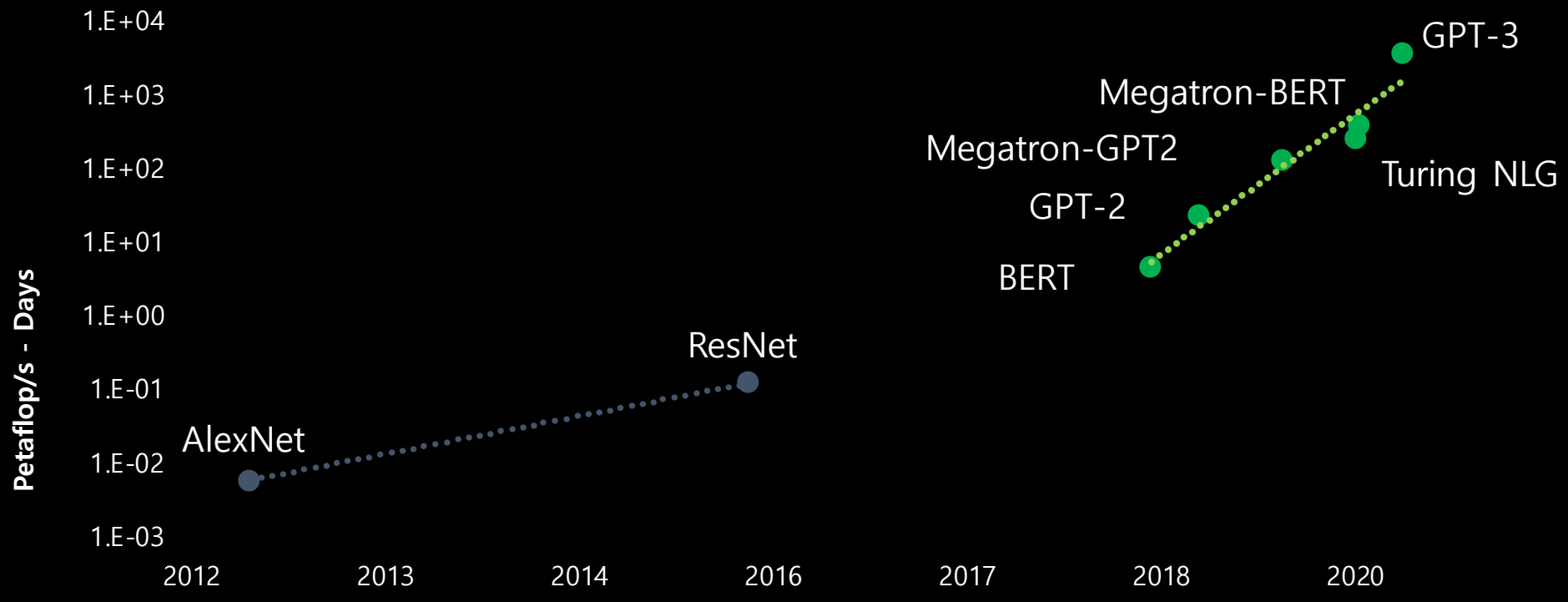
대규모 AI 연구를 위한 효과적인 GPU 데이터 센터 구축 전략

정소영 상무
Solutions Architect Team, Lead
NVIDIA

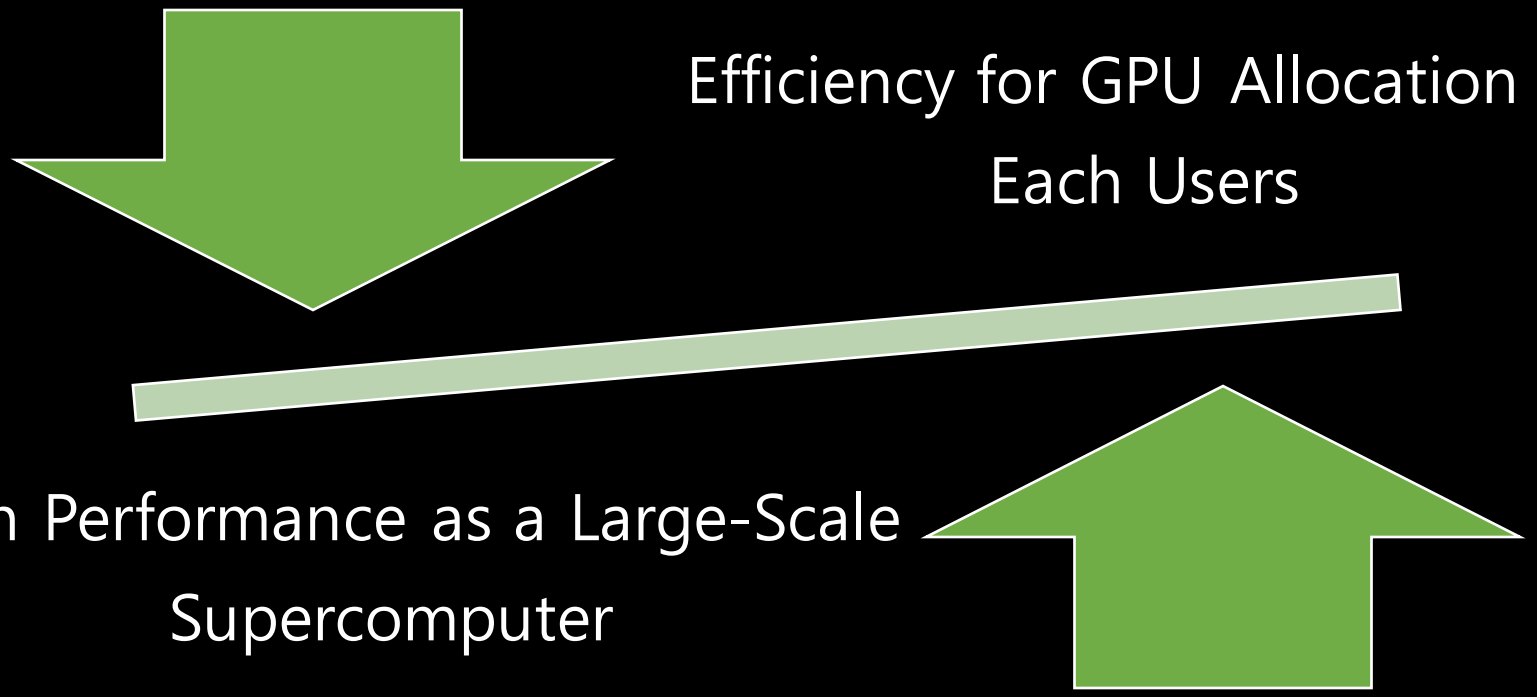
AI TREND

EXPLODING MODEL COMPLEXITY
30,000X in 5 Years | Now Doubling Every 2 Months



화두: DESIGN FOR GPU DATA CENTER

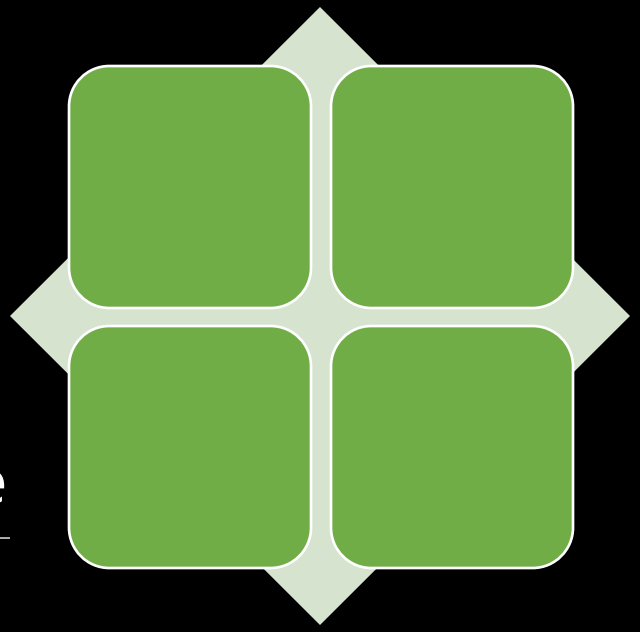
GUIDELINE FOR GPU DATA CENTER DESIGN



COMPONENTS OF GPU DATA CENTER

NVIDIA GPU

Network Architecture

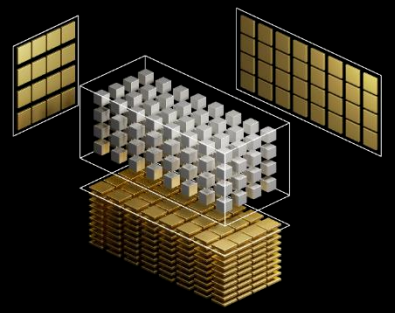


Storage Architecture

Management / Monitoring

NVIDIA GPU

NVIDIA GPU

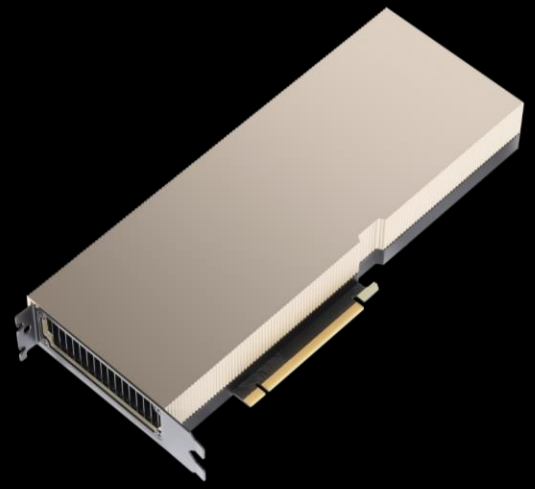


Tensor Core

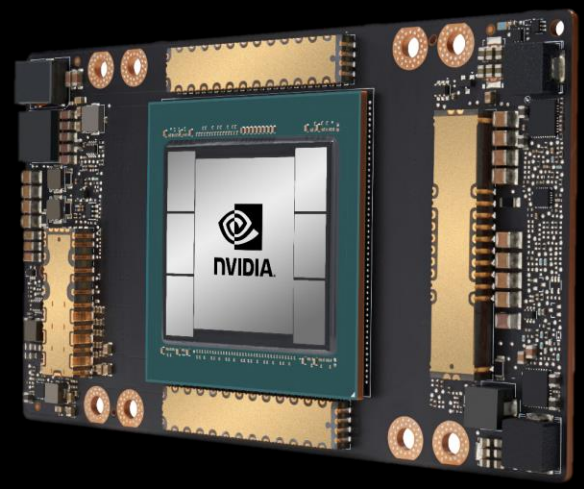
NVLink / NVSwitch

NVIDIA AMPERE A100

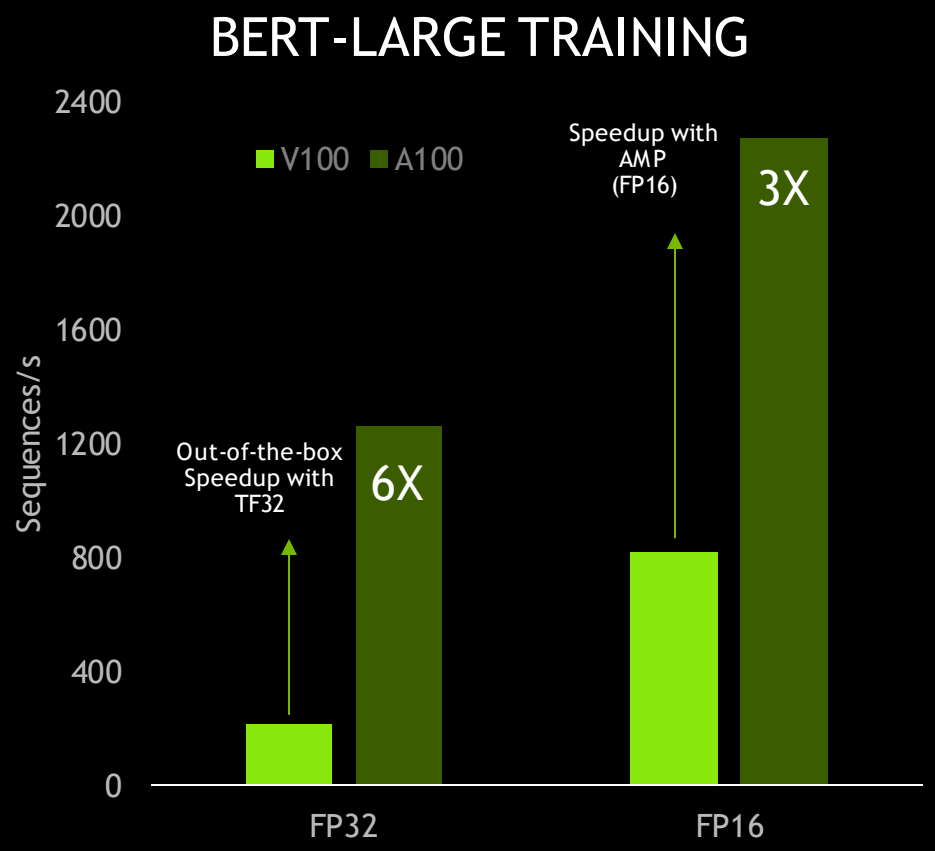
NVIDIA AMPERE A100

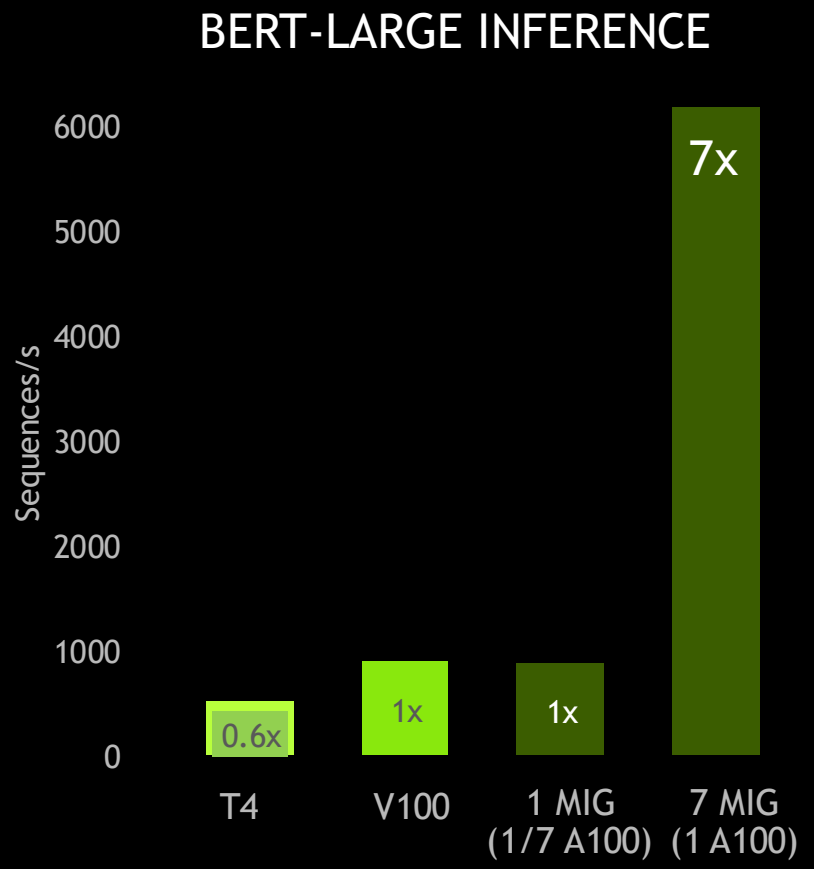


	Peak		Vs Volta
FP32 TRAINING	312	TFLOPS	20X
INT8 INFERENCE	1,248	TOPS	20X
FP64 HPC	19.5	TFLOPS	2.5X
MULTI INSTANCE GPU			7X GPUs



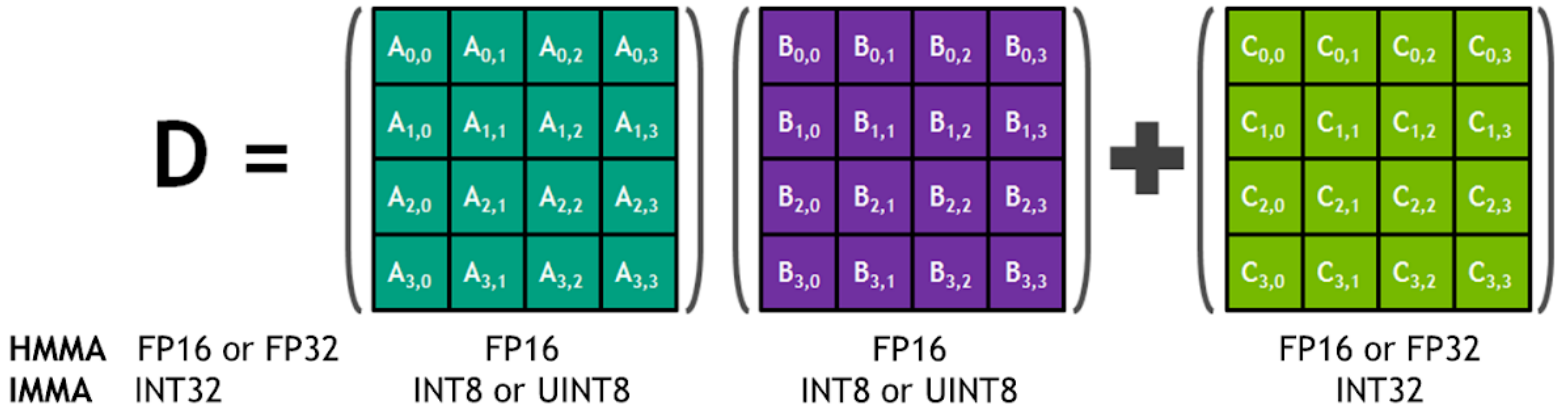
PERFORMANCE





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 and FP16 precision A100: DGX A100 Server with 8xA100 using TF32 precision and FP16 |
BERT Large Inference | T4: TRT 7.1, Precision = INT8, Batch Size =256, V100: TRT 7.1, Precision = FP16, Batch Size =256 | A100 with 7 MIG instances of 1g.5gb : Pre-production TRT, Batch Size =94, Precision = INT8 with Sparsity

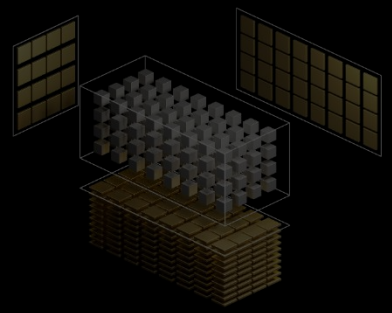
3rd GENERATION TENSOR CORE



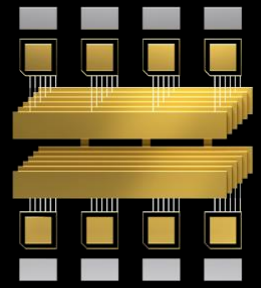
Up to X8 performance than CUDA core in GEMM and Convolution.

Design your DL model to use as many Tensor Core as possible

NVLINK / NVSWITCH

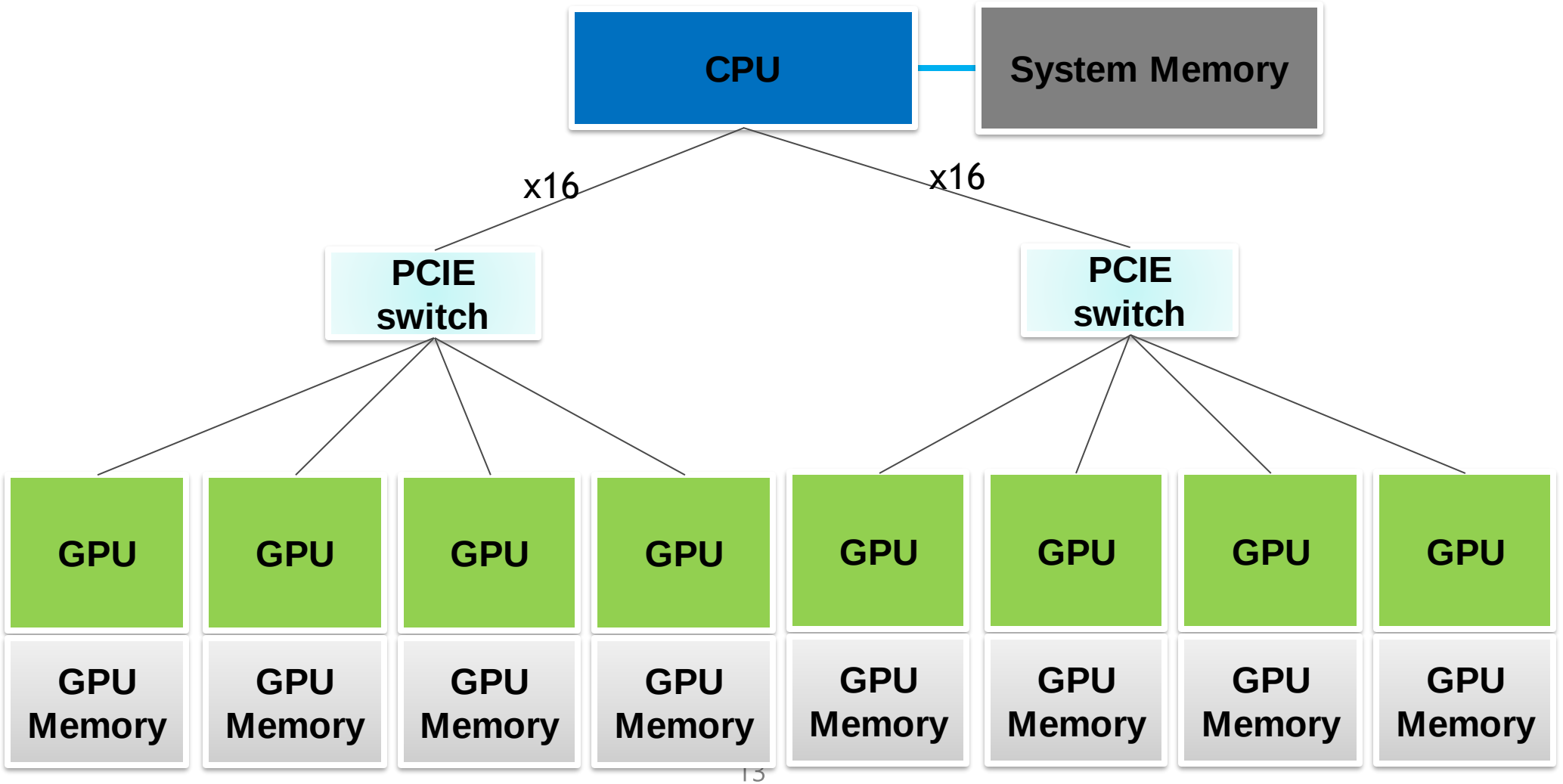


Tensor Core



NVLink / NVSwitch

GPU INTERCONNECT WITH PCIE



PCIE PERFORMANCE

PCI Express version	Introduced	Line code	Transfer rate[i]	Throughput[i]				
				×1	×2	×4	×8	×16
1.0	2003	8b/10b	2.5 GT/s	250 MB/s	0.50 GB/s	1.0 GB/s	2.0 GB/s	4.0 GB/s
2.0	2007	8b/10b	5.0 GT/s	500 MB/s	1.0 GB/s	2.0 GB/s	4.0 GB/s	8.0 GB/s
3.0	2010	128b/130b	8.0 GT/s	984.6 MB/s	1.97 GB/s	3.94 GB/s	7.88 GB/s	15.8 GB/s
4.0	2017	128b/130b	16.0 GT/s	1969 MB/s	3.94 GB/s	7.88 GB/s	15.75 GB/s	31.5 GB/s
5.0[32][33]	*expected in Q2 2019*[34]	128b/130b	32.0 GT/s[ii]	3938 MB/s	7.88 GB/s	15.75 GB/s	31.51 GB/s	63.0 GB/s

https://en.wikipedia.org/wiki/PCI_Express

NVLINK

High-speed communication protocol for near-range semiconductor

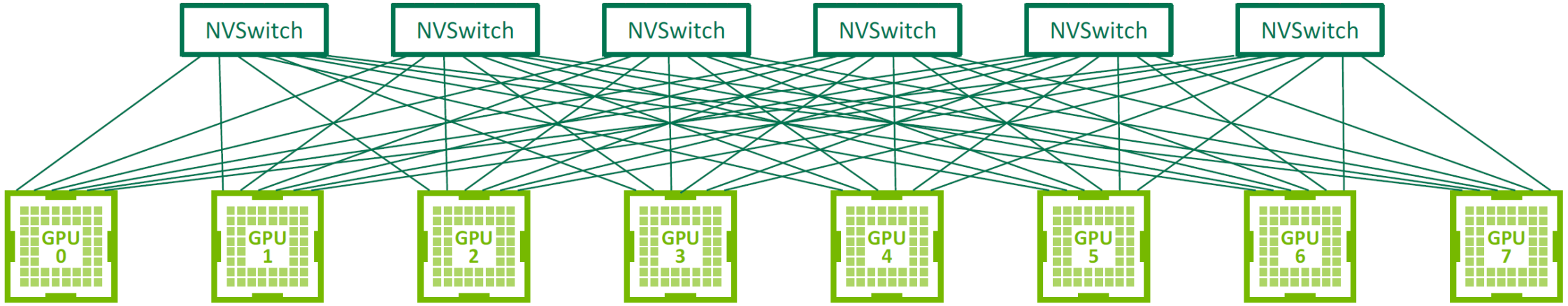
Developed by NVIDIA

GPU-GPU or CPU-GPU

12 Links per A100

50GB/s per NVLink and 600GB/s per A100

NVSWITCH



Provides Full Mesh Topology between GPUs

A100 TYPES

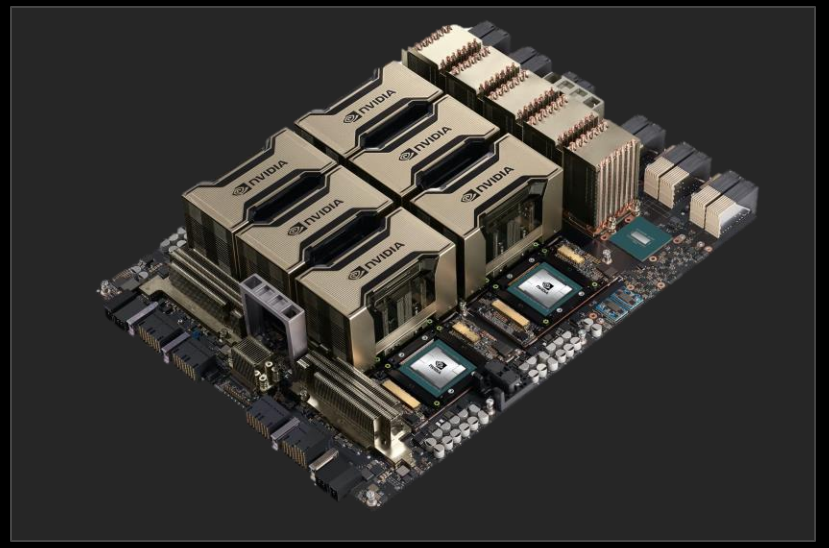
A100 PCIe

NVLINK supported only btw two GPUs

A100 4-GPU SXM

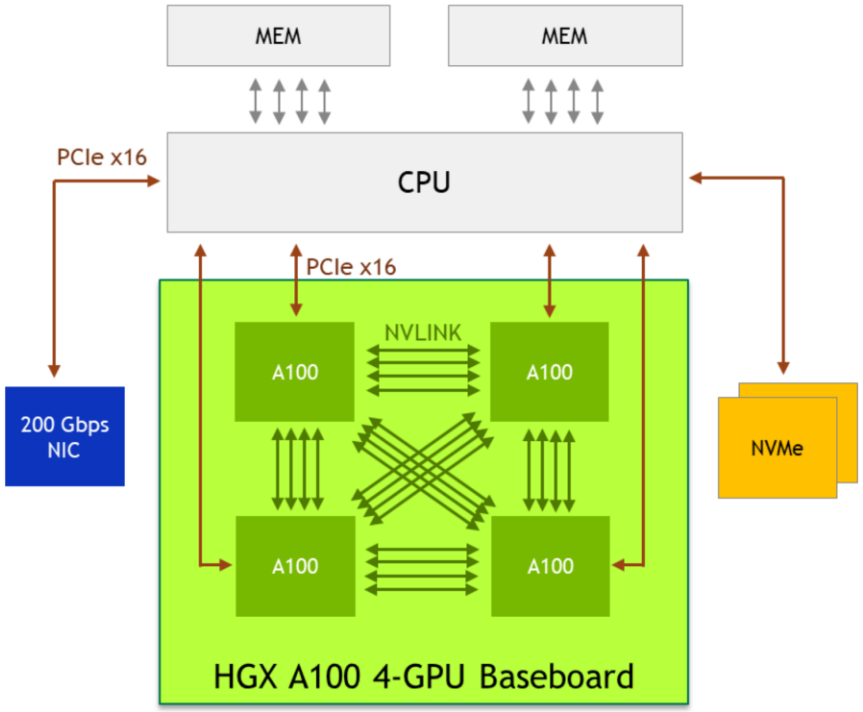
No NVSwitch supported, so limited NVLink throughput

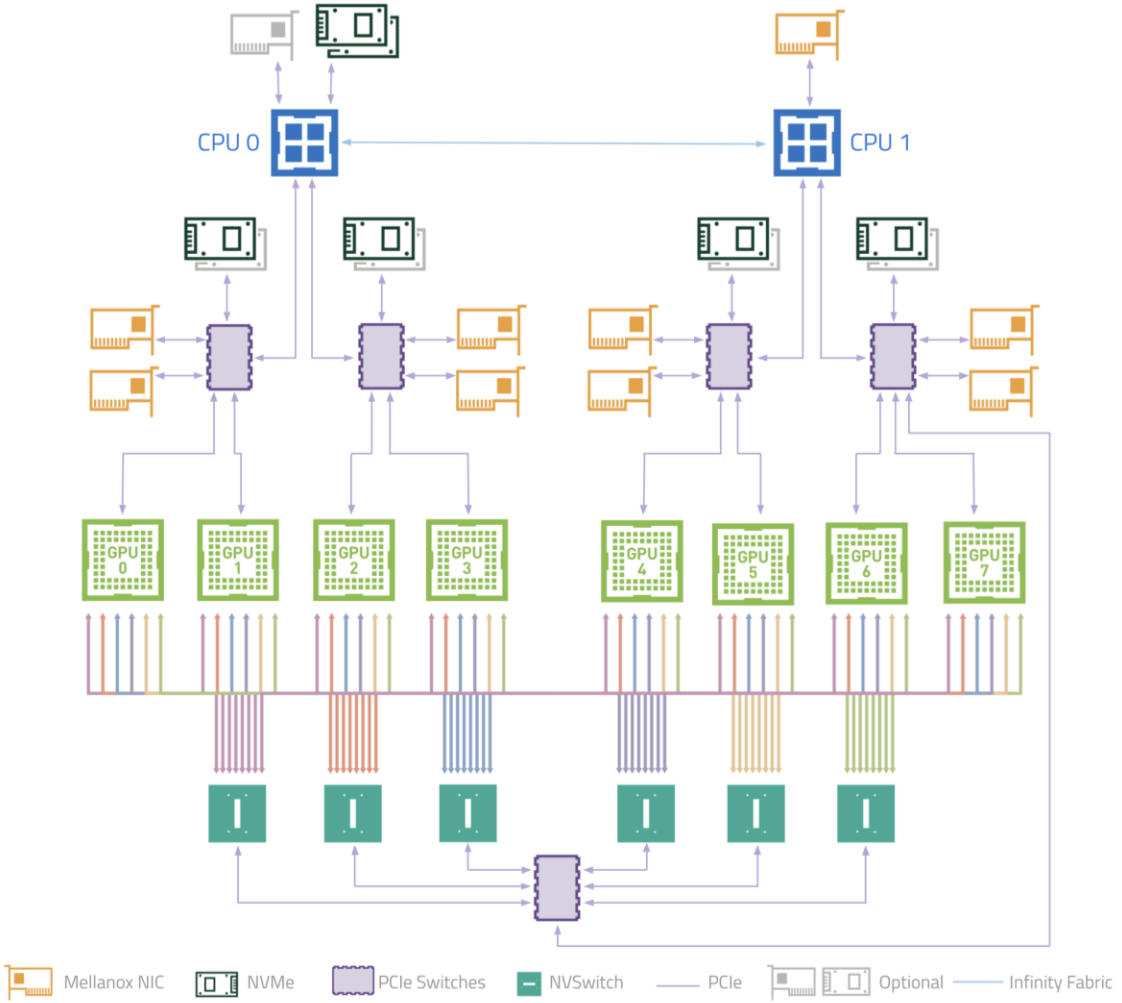
A100 8-GPU SXM



Full-throughput NVLink supported with NVSwitch

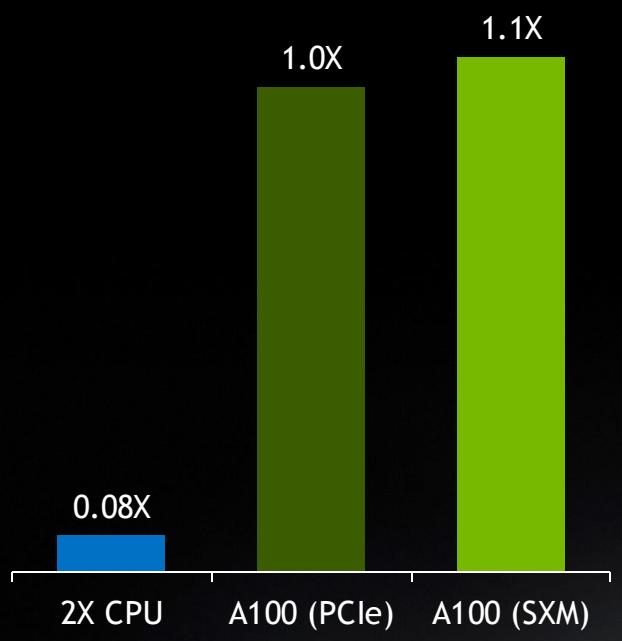
BLOCK DIAGRAM





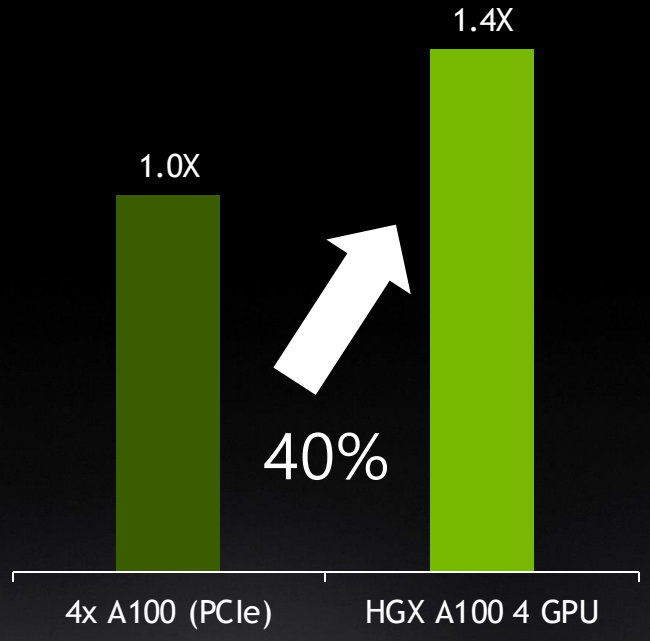
PERFORMANCE

90% Performance with 1x A100 PCIe for Mainstream Servers



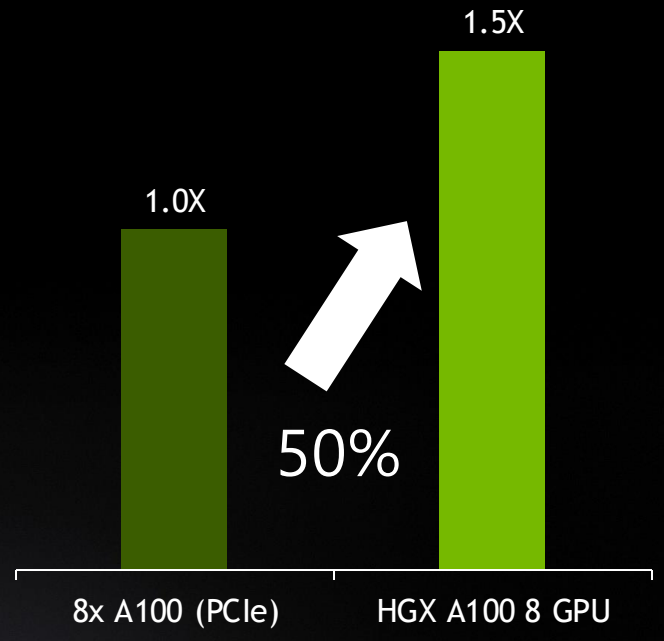
HPC Apps

Up to 1.4x Better Scaling with HGX A100 4 GPU



BERT-LARGE Pre-Training

Up to 1.5x Better Scaling with HGX A100 8GPU



BERT-LARGE Pre-Training

Geometric Mean of application speedups vs. 2x Intel Xeon Gold 6240 : Benchmark Application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT Large Fine Tuner], Quantum Espresso [AUSURF112-jR];, TensorFlow [ResNet-50], VASP 6 [Si Huge], |GPU node: with dual-socket CPUs with 1x A100 SXM and 1xA100 PCIe GPUs.
Scores are based on projection. BERT Pre-Training Throughput using Pytorch Phase 1 Seq Len = 128, PCIe server with 1x A100 (PCIe), 2xA100 (PCIe) and 2x A100 with NVLINK Bridge, 4x A100 (PCIe) 8x A100 (PCIe), HGX A100 4 GPU and DGX A100 8x GPU FP16 |

NETWORK REQUIREMENTS

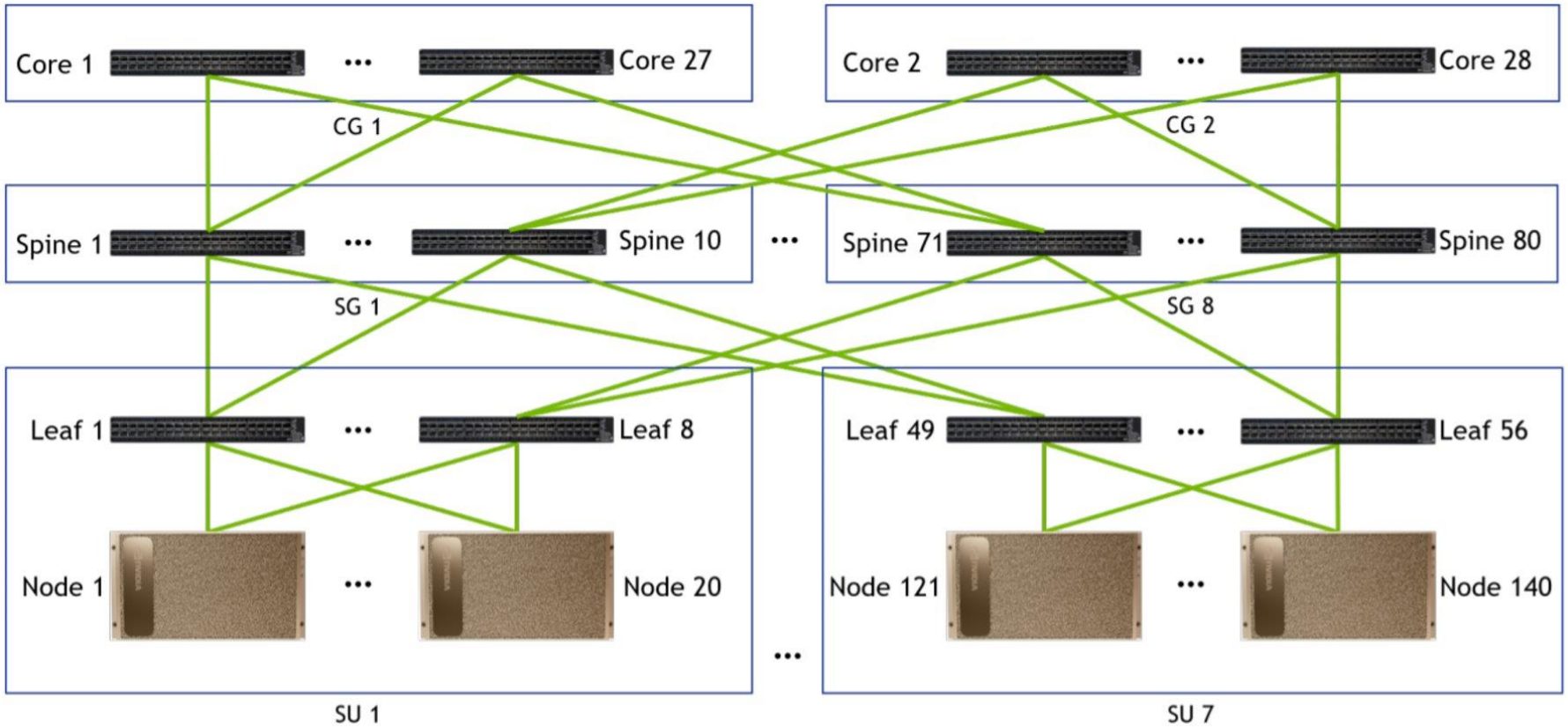
NVIDIA DGX A100 SUPERPOD

Supercomputer with 1,120 * A100 GPUs

https://www.nvidia.com/ko-kr/data-center/dgx-superpod/

NETWORK REQUIREMENTS

Full FAT-tree topology for non-blocking all-reduce operations between GPUs



STORAGE REQUIREMENTS

STORAGE REQUIREMENTS

Read-Intensive or Write-Intensive

- Read data once (DRAM) and copy them to GPU memory for training
- Checkpointing is important to cope with GPU / Node failure
- How frequently checkpointing?

STORAGE REQUIREMENTS

Concurrent IOPS

- Data size: training data volume (read) or checkpointing size (write)
- # of IB NIC per node to storage appliance: 200Gb HDR
- # nodes for concurrent IO: model parallel or data parallel
- Throughput of storage appliance

STORAGE REQUIREMENTS

Ex) GPT-3 with 175B parameters

- Large-scale model parallel + data parallel required for training / inference
- 16 nodes of DGX A100 necessary for model parallel cluster
- So, all GPUs read but GPUs in 16 nodes write
- 100s of GB of training data while several TB of checkpoint for write

STORAGE REQUIREMENTS

Ex) How to calculate IOPS for GPT-3 shared storage?

- Shared volume across all the nodes in SuperPOD
- Basic I/O performance of DGX A100 is 2*200Gbps / 8 = 50GB/s
- Considering 80% utilization, 40GB/s necessary per single DGX A100 node
- 16 * 40GB/s = 720GB/s required for checkpointing throughput.

GPT-3 TRAINING TIME PROJECTION

# of GPUs	100	200	500	1,000	2,000	5,000	**10,000**
Training days (V100)	1730.6	865.3	346.1	173.1	86.5	34.6	**17.3**
Training days (A100)	641	320.5	128.2	64.1	32	12.8	**6.4**

Total compute for training GPT-3: 3.14E+23
Performance projection of A100 from V100: x2.7 (based on BERT perf. projection in MLPerf 0.7)
Model-parallel based training throughput on V100: 2.1E+13 Flops (21 TFlops) (with Deepspeed)
Assumed Mellanox IB Interconnect Technology used with linear-scale performance

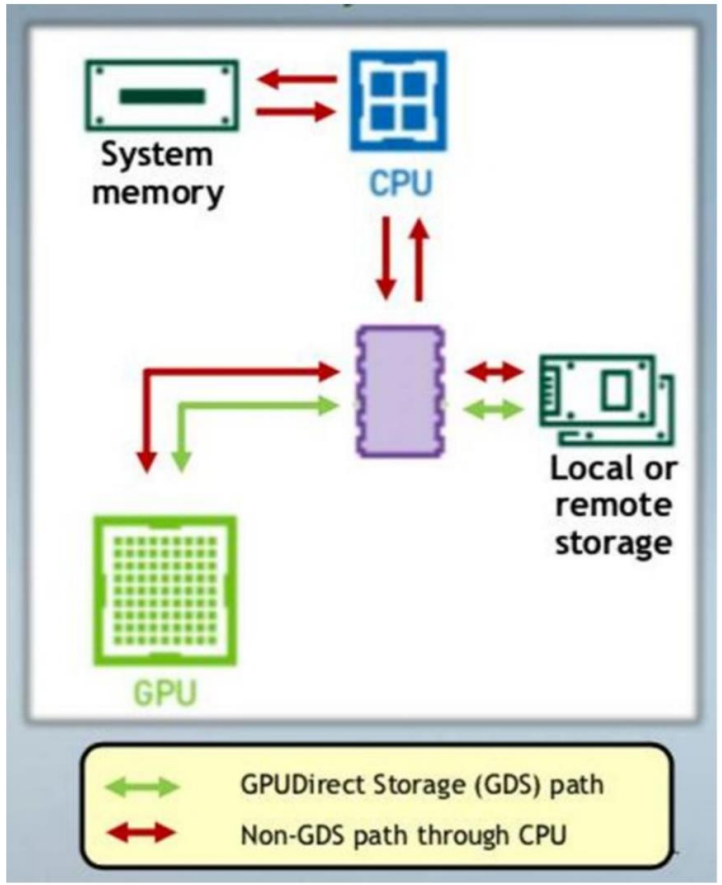
GPUDIRECT STORAGE

The easiest way to get performance for storage-GPU memory transfers

- Skips CPU bounce buffer via DMA
- Works for both local and remote storage
- Accessed via new CUDA cuFile APIs on CPU
- No special HW

Advantages

- Higher peak bandwidth
- Lower latency by avoiding extra copies and dynamic routing that optimizes path, buffers, mechanisms
- Less jitter than fault-based methods
- Greater generality, e.g. alignment



SDS – NVIDIA COLLABORATION

SDS – NVIDIA COLLABORATION

Design for SDSPOD

Tensor Core optimized Deep Learning Model Design

GPU-optimized AI Service Stack Development

Thank you

Q&A