디지털 트랜스포메이션을 통한
비즈니스 혁신과 4차 산업혁명 대응을 위한

GIT 솔루션즈 데이

TESLA PLATFORM

2018.10.17 / baynex

BayNex

GiT 굿모닝아이텍(주)

목차

nVidia 제품군

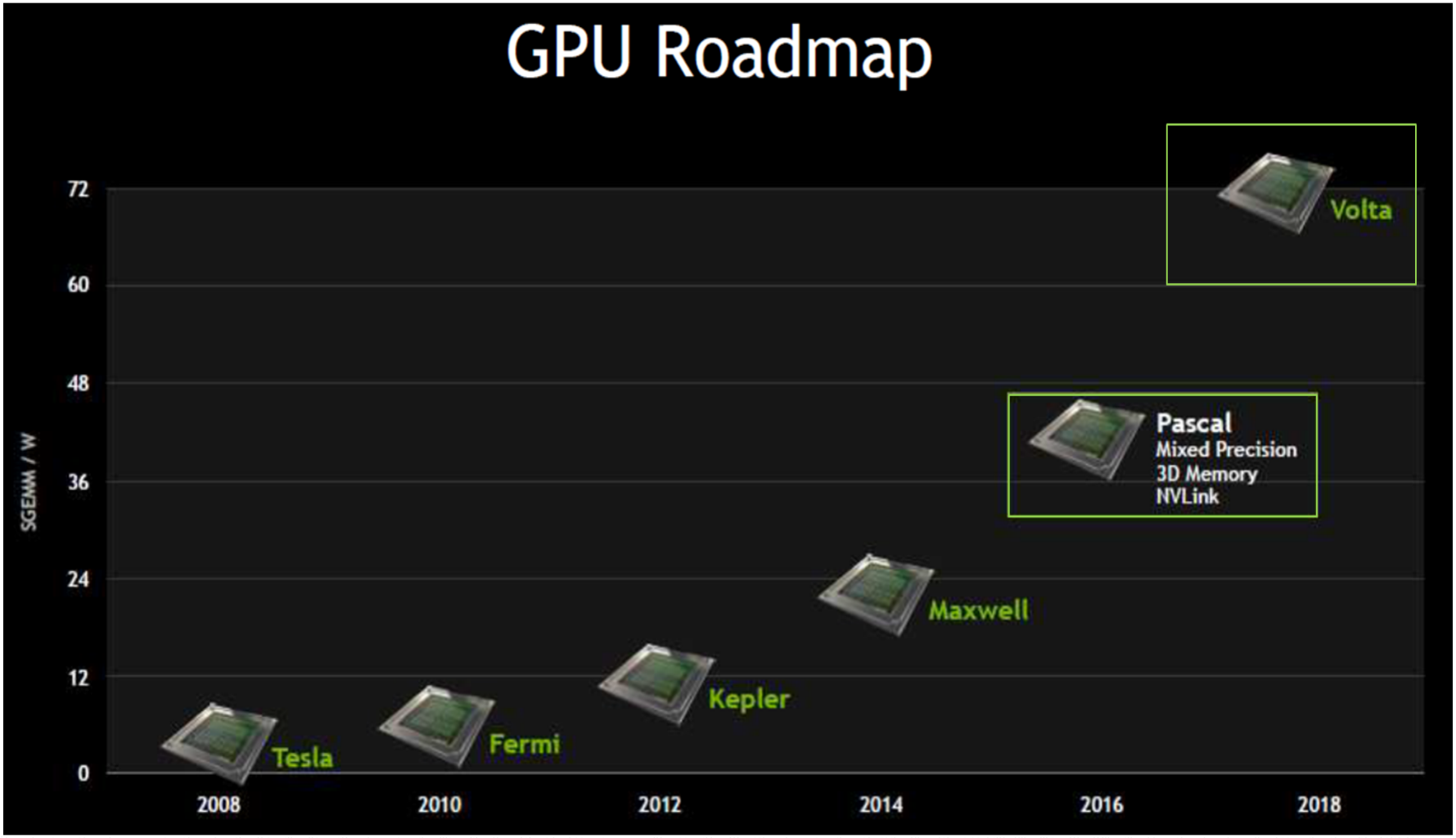
GeForce

Quadro

TESLA

TEGRA

GPU Roadmap


GPU Roadmap
SGEMM / W
72
60
48
36
24
12
0
2008
2010
2012
2014
2016
2018
Tesla
Fermi
Kepler
Maxwell
Pascal
Mixed Precision
3D Memory
NVLink
Volta

Tesla GPU Type

PCIe Type

SXM2 Type

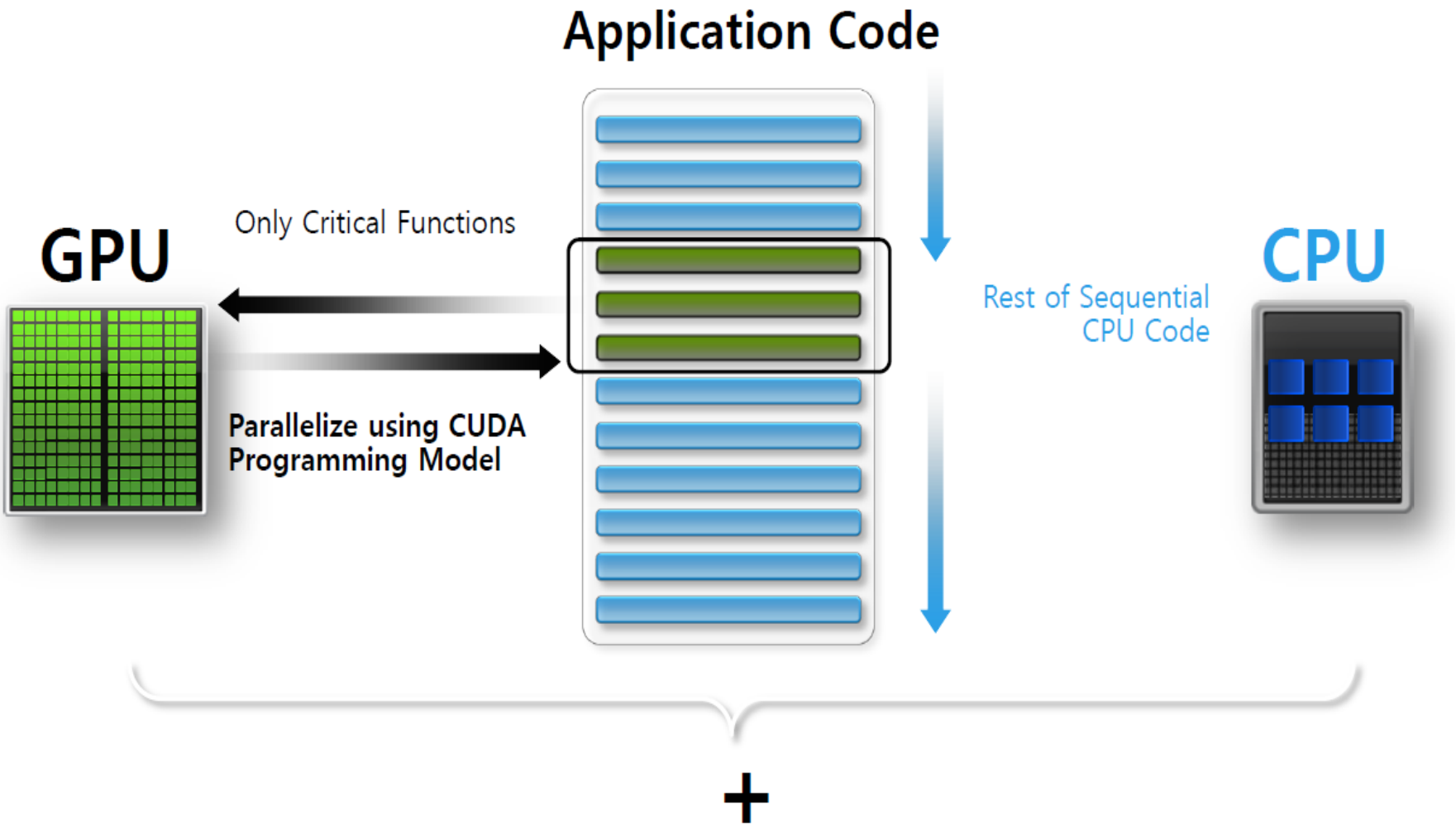
Mezzanine Type

일반 랙 서버

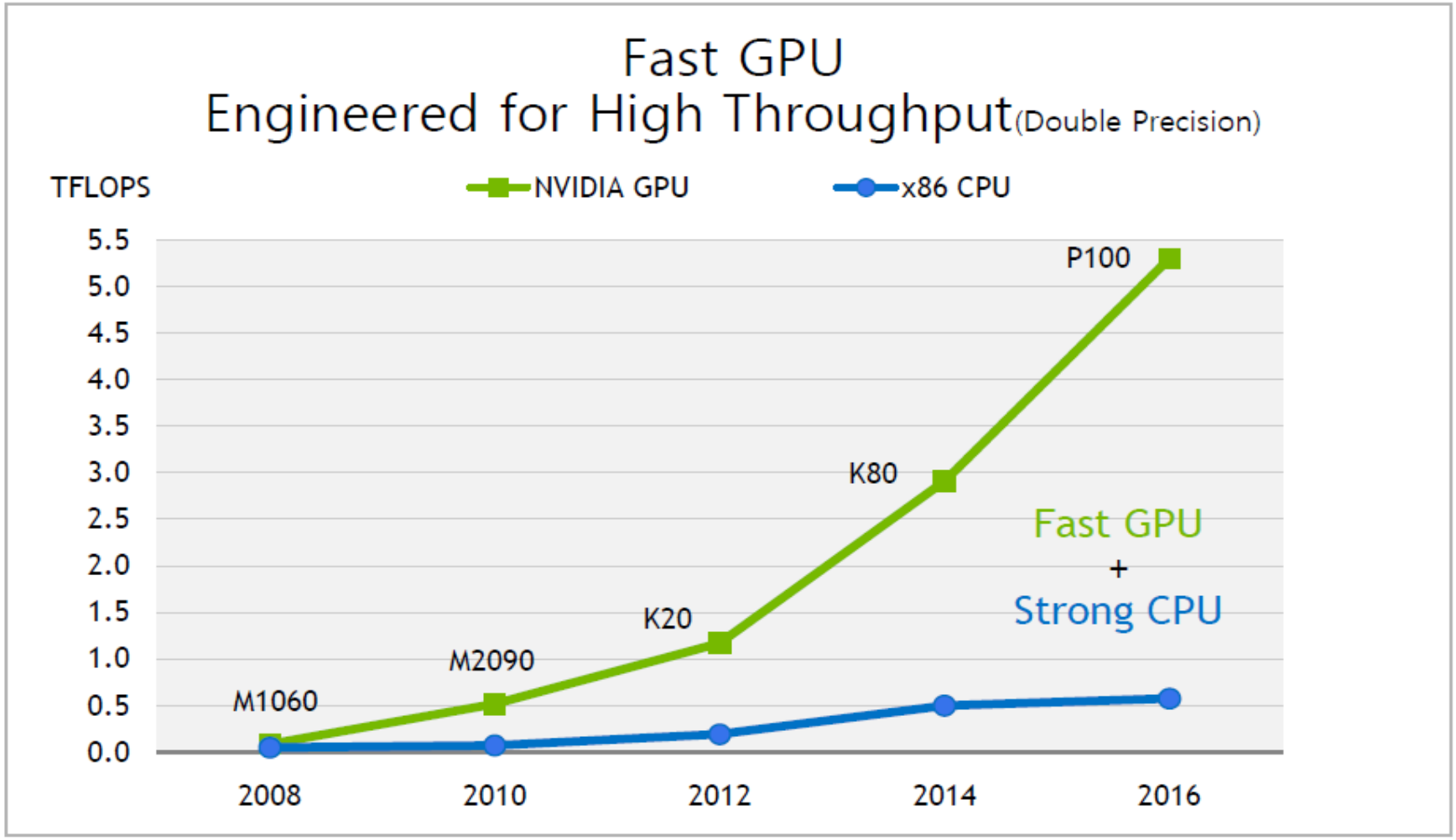
전용 서버

블레이드서버

실행구조


Application Code
GPU
Only Critical Functions
Parallelize using CUDA
Programming Model
Rest of Sequential
CPU Code
CPU
+

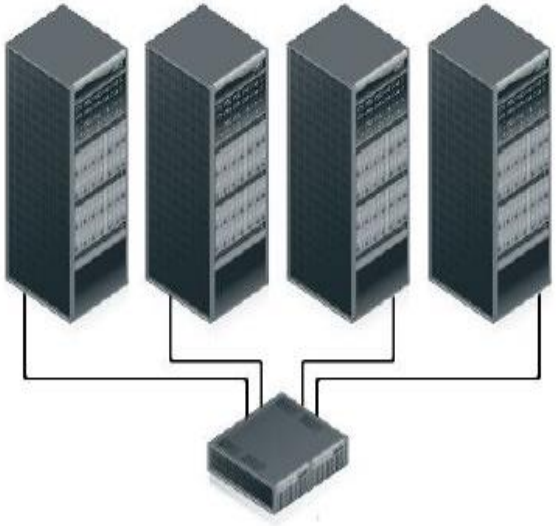
GPU 성능


Fast GPU
Engineered for High Throughput(Double Precision)
TFLOPS
NVIDIA GPU
x86 CPU
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
M1060
M2090
K20
K80
P100
Fast GPU
+
Strong CPU
2008
2010
2012
2014
2016

GPU 주요 용도

HPC

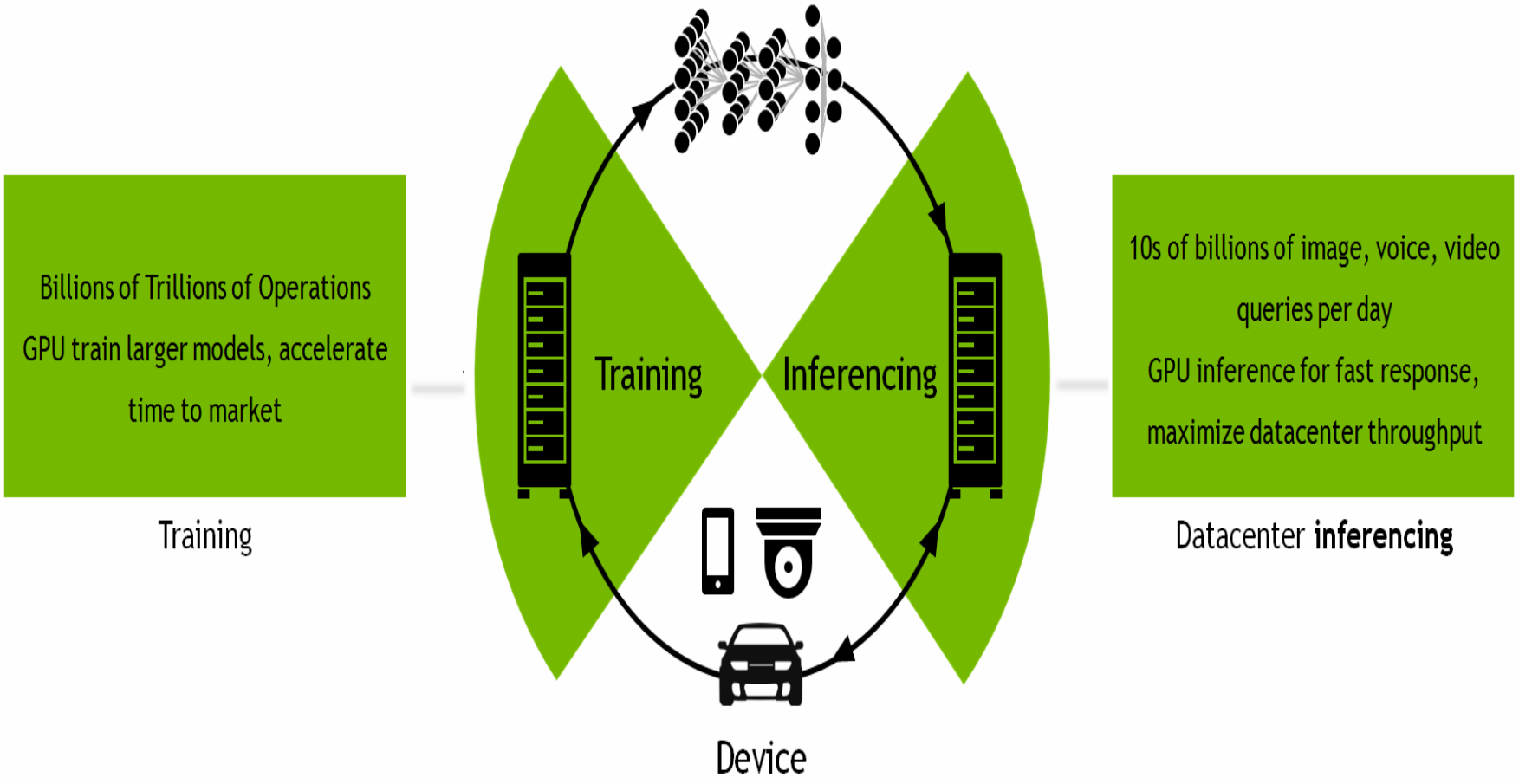
High Performance Computing



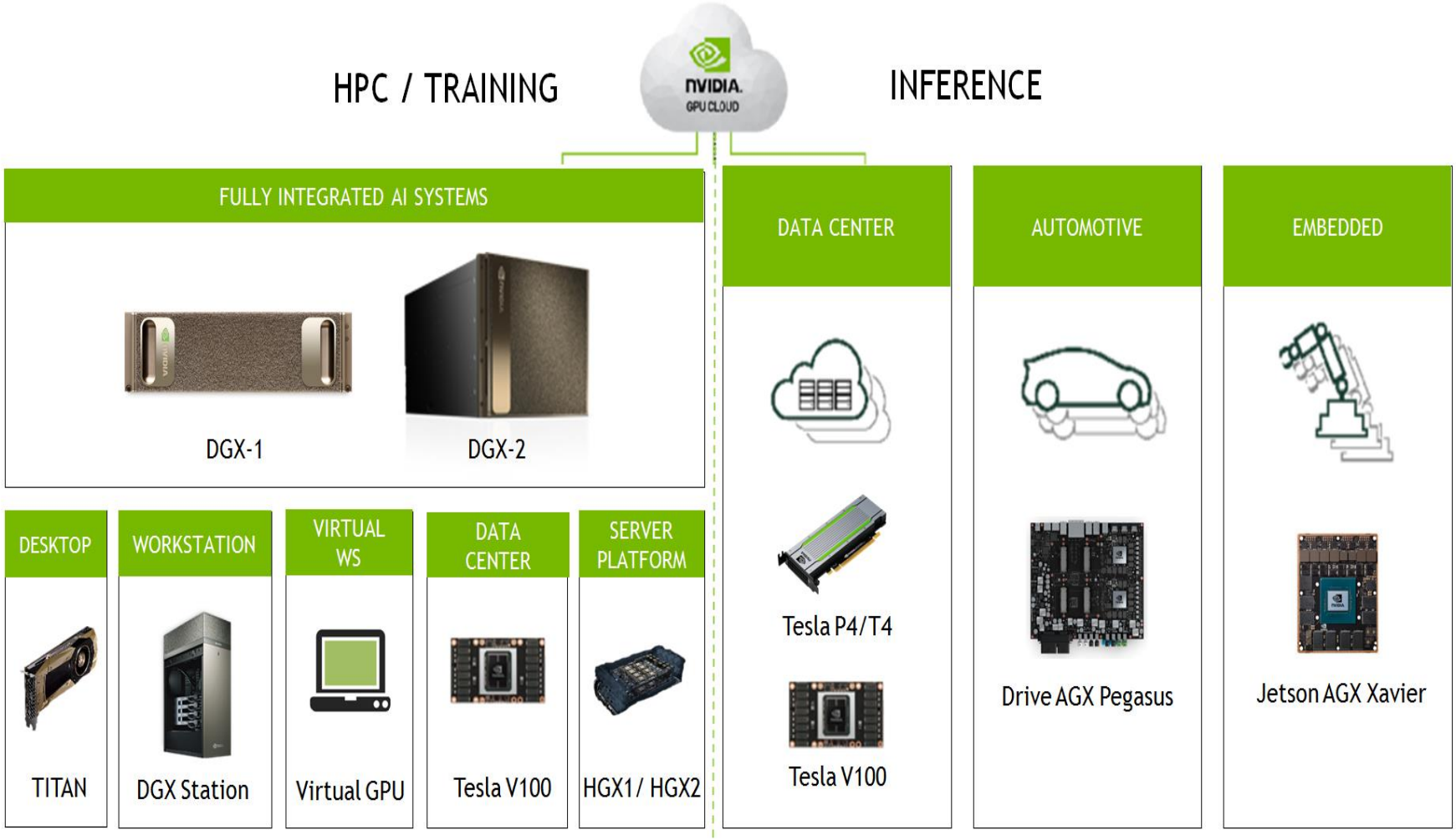
AI

artificial intelligence

AI (Training, Inference)



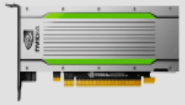
END-TO-END PRODUCT FAMILY



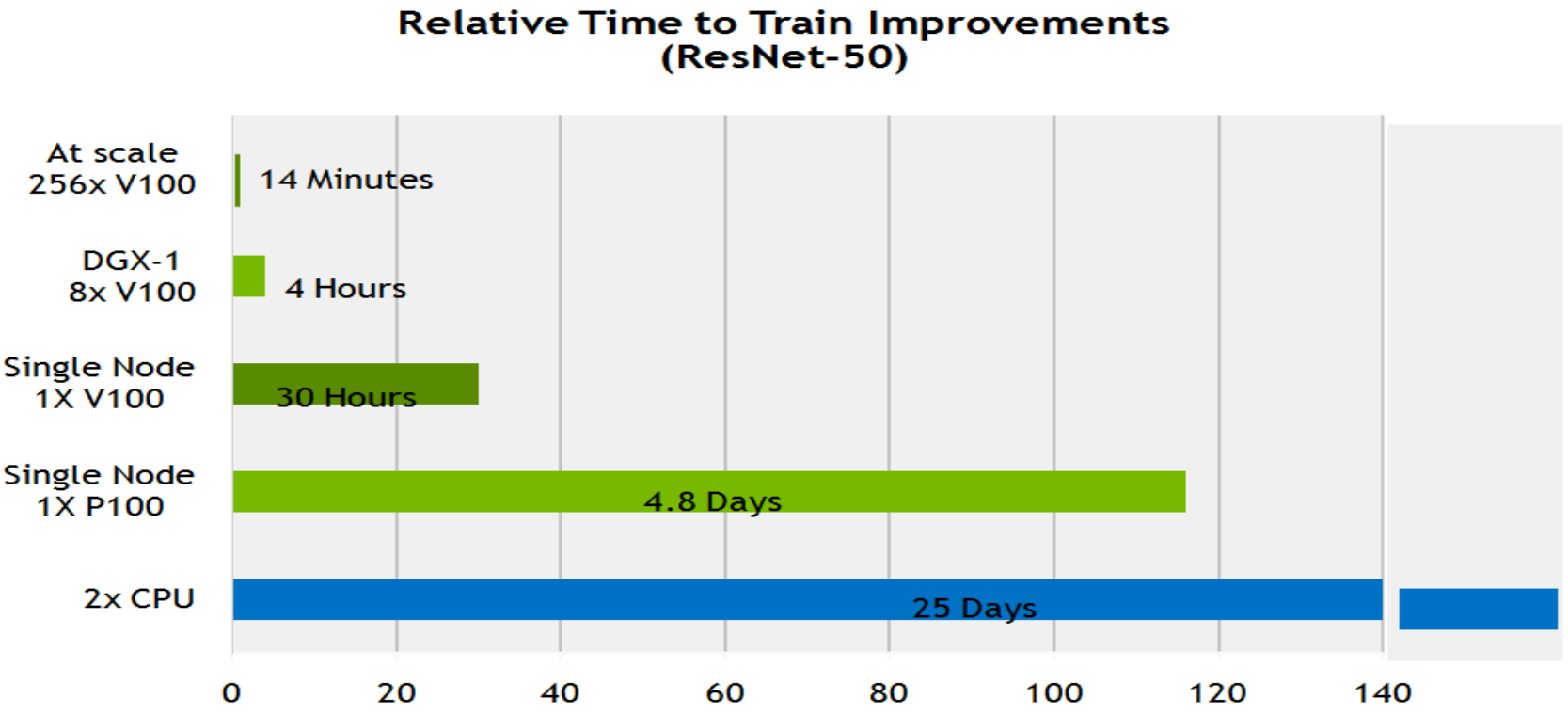
TESLA SPEC

	P100 (SXM2)	P100 (PCIE)	P40	P4	T4	V100 (PCIE)	V100 (SXM2)	V100 (FHHL)
GPU CHIP	GP100	GP100	GP102	GP104	TU104	GV100	GV100	GV100
PEAK FP64 (TFLOPs)	5.3	4.7	NA	NA	NA	7	7.8	6.5
PEAK FP32 (TFLOPs)	10.6	9.3	12	5.5	8.1	14	15.7	13
PEAK FP16 (TFLOPs)	21.2	18.7	NA	NA	65	112	125	105
PEAK TOPs	NA	NA	47	22	260	NA	NA	NA
Memory Size	16 GB HBM2	16/12 GB HBM2	24 GB GDDR5	8 GB GDDR5	16 GB HBM2	32 GB HBM2	32 GB HBM2	16GB HBM2
Memory BW	732 GB/s	732/549 GB/s	346 GB/s	192 GB/s	320GB/s	900 GB/s	900 GB/s	900 GB/s
Interconnect	NVLINK + PCIe Gen3	PCIe Gen3	PCIe Gen3	PCIe Gen3	PCIe Gen3	PCIe Gen3	NVLINK + PCIe Gen3	PCIe Gen3
ECC	Internal + HBM2	Internal + HBM2	GDDR5	GDDR5	GDDR6	Internal + HBM2	Internal + HBM2	Internal + HBM2
Form Factor	SXM2	PCIE Dual Slot	PCIE Dual Slot	PCIE LP	PCIE LP	PCIE Dual Slot	SXM2	PCIE Single Slot Full Height Half Length
Power	300 W	250 W	250 W	50-75 W	70 W	250W	300W	150W

TESLA PRODUCTS RECOMMENDATION

PRODUCT	V100	T4	P4
	 V100 for PCIe  V100 for NVLink	 T4 for PCIe	 P4 for PCIe
Target Use Cases	• Universal GPU for accelerating HPC and AI Workloads • Ultra-high end GPU-accelerated virtualization for 3D professional applications	• Most efficient platform for both real-time and large-batch inference. • Low power, low profile optimized for scale out DL inference deployment	• Low power, low profile optimized for scale out DL inference deployment • Efficient inference and video processing • Optimal performance/$ for most 3D professional apps deployed in VDI
Form Factors	• Tesla V100 for NVLink: Ultimate DL performance Tesla V100 for PCIe: Highest versatility for all workloads	• PCIe	• PCIe
Best Configs.	• PCIe: 2-4 GPU/node • NVLink: 8 way Hybrid Cube Mesh	• 1-2 GPU/node	• 1-2 GPU/node
1st Server Ship	• Available Now	• Available Q4 2018	• Available Now

성능 비교



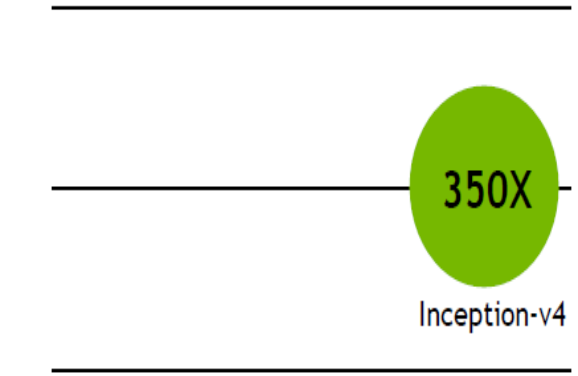
ResNet-50, 90 epochs to solution | CPU Server: dual socket Intel Xeon Gold 6140

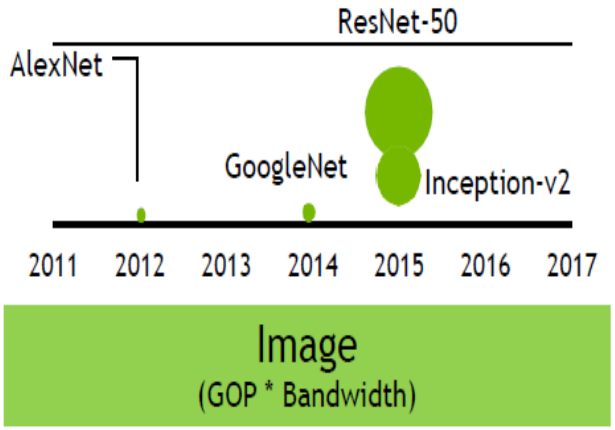
최근 딥러닝의 추세

딥 러닝의 폭발적인 성장:

- Data증가, 계산과 복잡도에 대한 요구증가
- GPU 1개의 메모리 용량 초과
- GPU 1개의 계산성능 초과

멀티 GPU의 scale out 필요성 증대





DGX STATION

DGX STATION SPEC

GPUs	4x NVIDIA® Tesla® V100
TFLOPS (GPU FP16)	480
GPU Memory	16 GB per GPU
NVIDIA Tensor Cores	2,560 (total)
NVIDIA CUDA Cores	20,480 (total)
CPU	Intel Xeon E5-2698 v4 2.2 GHz (20-core)
System Memory	256 GB LRDIMM DDR4
Storage	Data: 3 x 1.92 TB SSD RAID 0 OS: 1 x 1.92 TB SSD
Network	Dual 10 Gb LAN
Display	3x DisplayPort, 4K Resolution
Acoustics	< 35 dB
Maximum Power Requirements	1500 W
Operating Temperature Range	10 - 30 ℃
Software	Ubuntu Desktop Linux OS DGX Recommended GPU Driver CUDA Toolkit

DEEP LEARNING CONTAINERS ON NGC (NVIDIA GPU CLOUD)

Caffe	ONNX
Caffe2	PYTORCH
Microsoft Cognitive Toolkit	TensorFlow
mxnet	theano
NVIDIA DIGITS	torch

DGX-1V

8 V100 GPUs

6 NVLinks per GPU

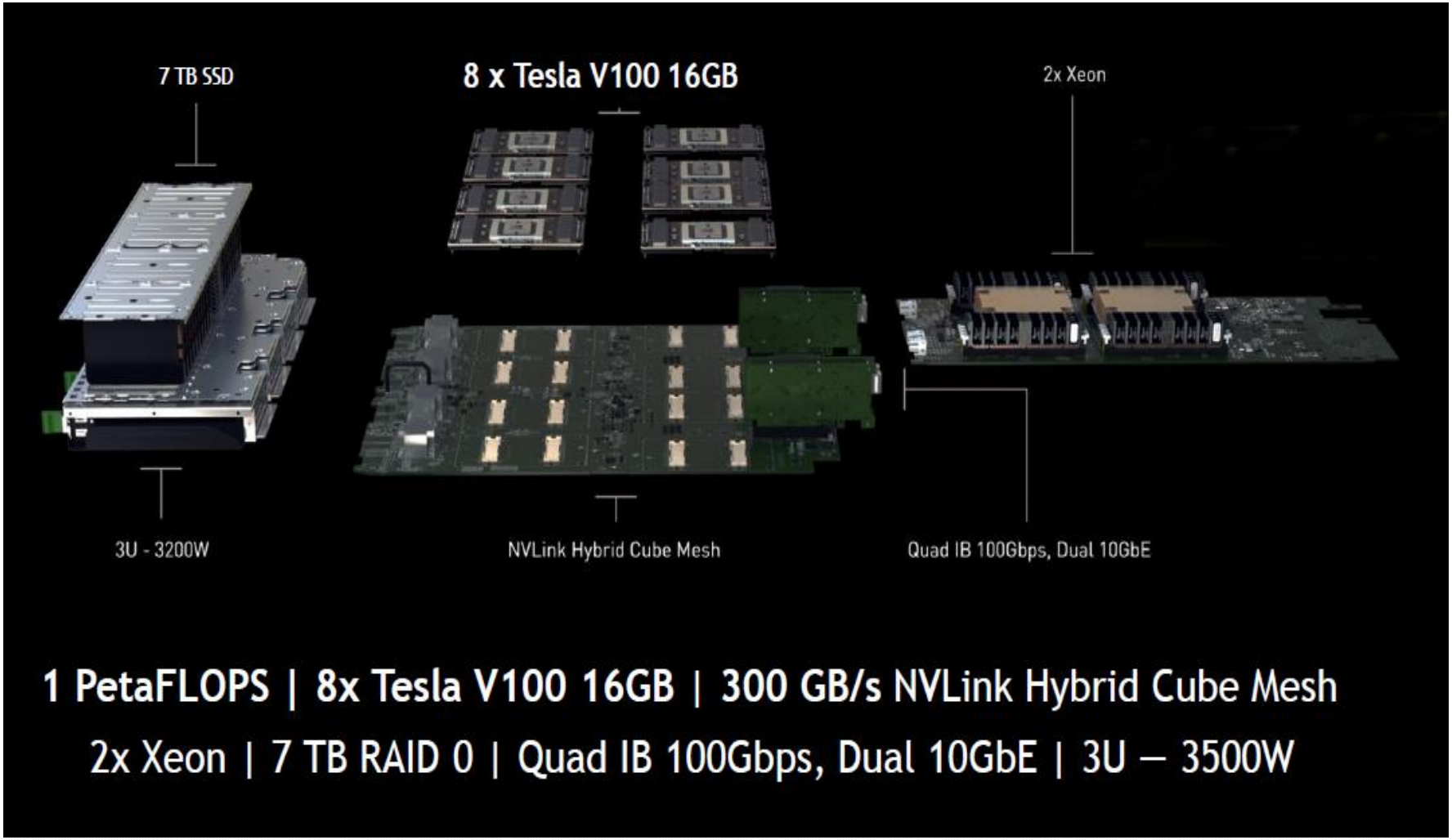
Each link is 50GB/s (bidirectional)

300GB/s GPU 양방향 대역폭

DGX-1은 Hybrid Cube Mesh 구조

데이터 병렬 트레이닝 구조에 최적화된 NCCL 지원

DGX-1V 내부


7 TB SSD
8 x Tesla V100 16GB
2x Xeon
3U - 3200W
NVLink Hybrid Cube Mesh
Quad IB 100Gbps, Dual 10GbE
1 PetaFLOPS | 8x Tesla V100 16GB | 300 GB/s NVLink Hybrid Cube Mesh
2x Xeon | 7 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3500W

DGX-1V TOPOLOGY

GPU - CPU link:

PCIe

12.5+12.5 GB/s eff BW

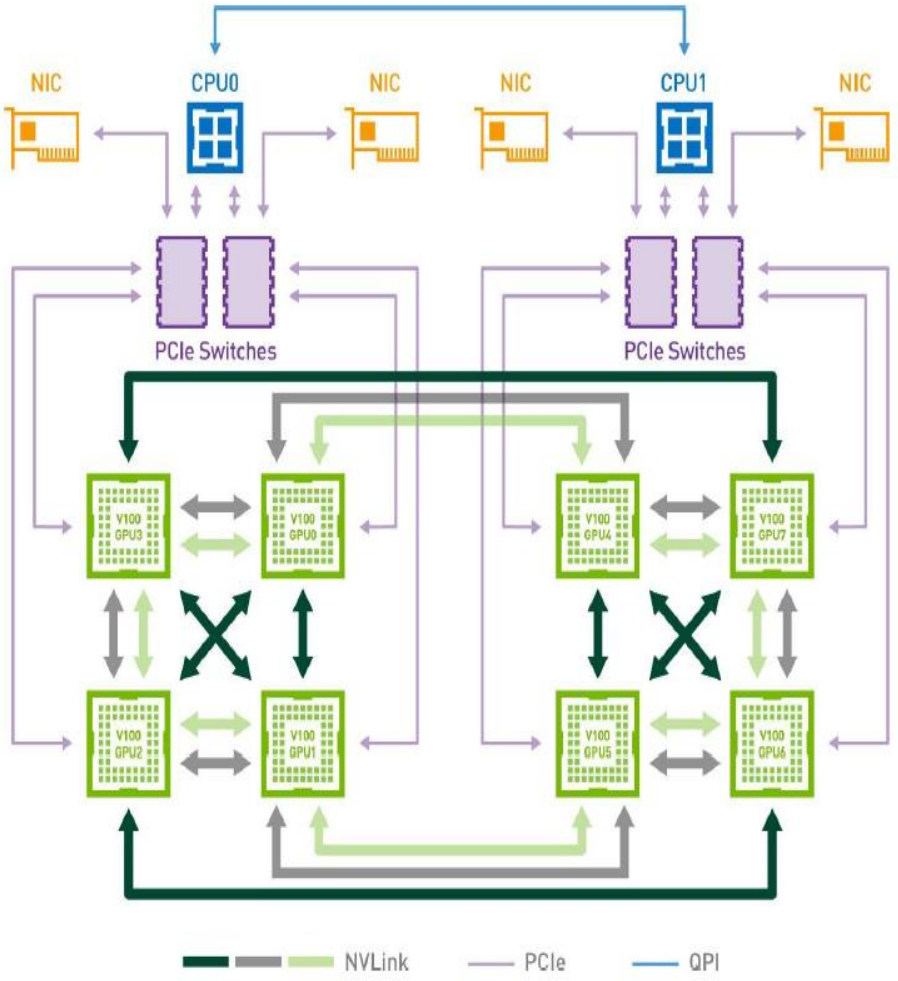
GPUDirect P2P:

GPU - GPU link is NVLink

Hybrid Cube mesh topology

GPUDirect RDMA:

GPU - NIC link is PCIe



DGX-1V TOPOLOGY

$ nvidia-smi topo -m

	GPU0	GPU1	GPU2	GPU3	GPU4	GPU5	GPU6	GPU7	mlx5_0	mlx5_2	mlx5_1	mlx5_3	CPU Affinity
GPU0	X	NV1	NV1	NV2	NV2	SYS	SYS	SYS	PIX	SYS	PHB	SYS	0-19,40-59
GPU1	NV1	X	NV2	NV1	SYS	NV2	SYS	SYS	PIX	SYS	PHB	SYS	0-19,40-59
GPU2	NV1	NV2	X	NV2	SYS	SYS	NV1	SYS	PHB	SYS	PIX	SYS	0-19,40-59
GPU3	NV2	NV1	NV2	X	SYS	SYS	SYS	NV1	PHB	SYS	PIX	SYS	0-19,40-59
GPU4	NV2	SYS	SYS	SYS	X	NV1	NV1	NV2	SYS	PIX	SYS	PHB	20-39,60-79
GPU5	SYS	NV2	SYS	SYS	NV1	X	NV2	NV1	SYS	PIX	SYS	PHB	20-39,60-79
GPU6	SYS	SYS	NV1	SYS	NV1	NV2	X	NV2	SYS	PHB	SYS	PIX	20-39,60-79
GPU7	SYS	SYS	SYS	NV1	NV2	NV1	NV2	X	SYS	PHB	SYS	PIX	20-39,60-79
mlx5_0	PIX	PIX	PHB	PHB	SYS	SYS	SYS	SYS	X	SYS	PHB	SYS	
mlx5_2	SYS	SYS	SYS	SYS	PIX	PIX	PHB	PHB	SYS	X	SYS	PHB	
mlx5_1	PHB	PHB	PIX	PIX	SYS	SYS	SYS	SYS	PHB	SYS	X	SYS	
mlx5_3	SYS	SYS	SYS	SYS	PHB	PHB	PIX	PIX	SYS	PHB	SYS	X	

Legend:

  X   = Self
  SYS = Connection traversing PCIe as well as the SMP interconnect between NUMA nodes (e.g., QPI/UPI)
  NODE = Connection traversing PCIe as well as the interconnect between PCIe Host Bridges within a NUMA node
  PHB = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX = Connection traversing a single PCIe switch
  NV# = Connection traversing a bonded set of # NVLinks

SCALE-OUT 요구 증가

- Scale up to 16 GPUs
- 직접적인 peer GPU memory 접근
- Full non-blocking bandwidth
- 메모리 access시 모든 GPU links를 활용
- Multi-GPU 프로그래밍의 단순화

DGX-2

DGX-2

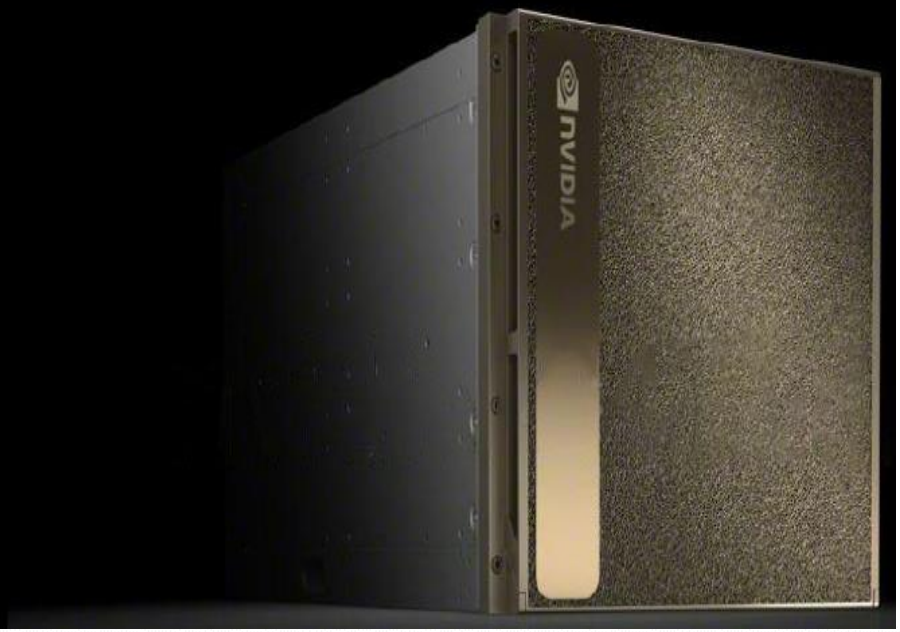
THE WORLD'S MOST POWERFUL AI SYSTEM FOR THE MOST COMPLEX AI CHALLENGES

16 V100 32GB GPUs Fully Interconnected

NVSwitch fabric: 2.4 TB/s Connectivity between GPUs

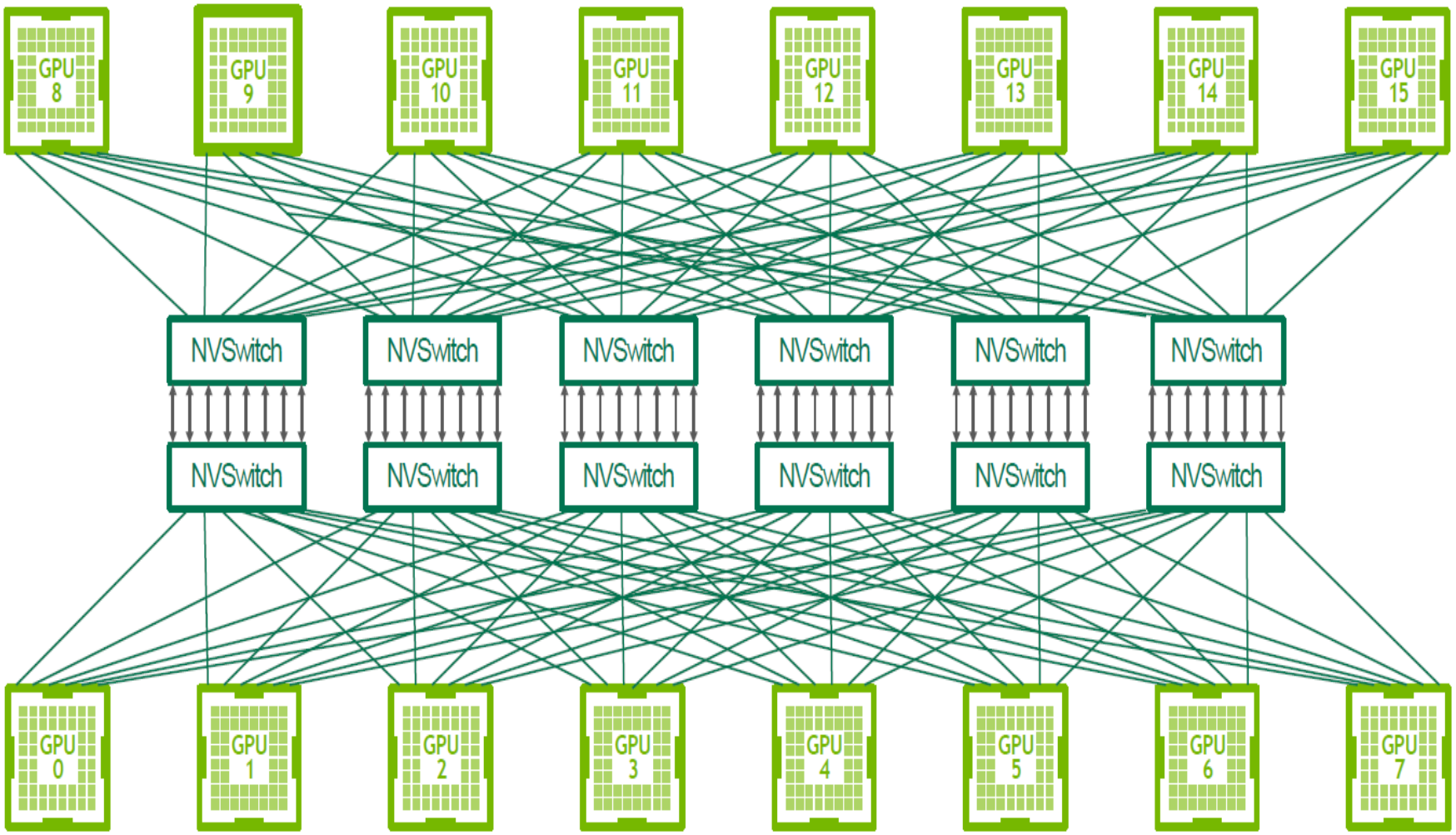
12X GPU-GPU Bandwidth

0.5 TB of Unified GPU Memory

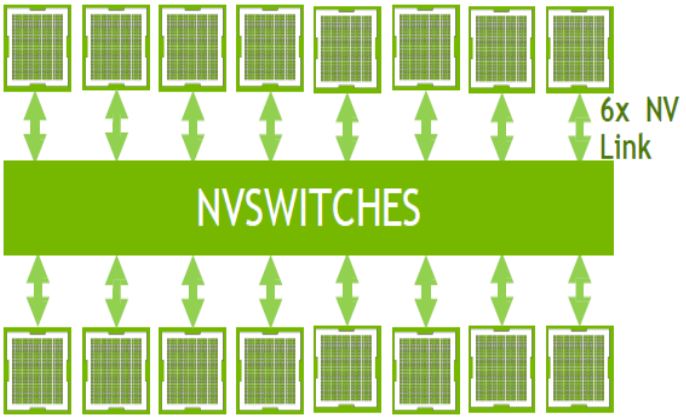


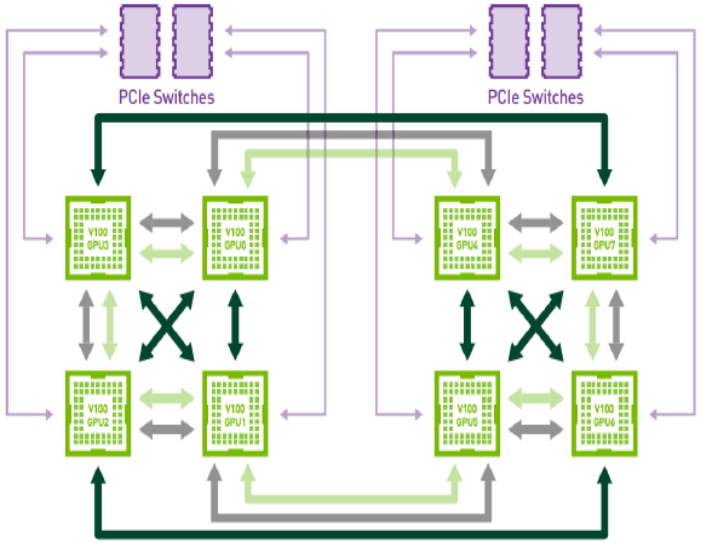
DIRECT PEER MEMORY ACCESS

FULL NON-BLOCKING BANDWIDTH



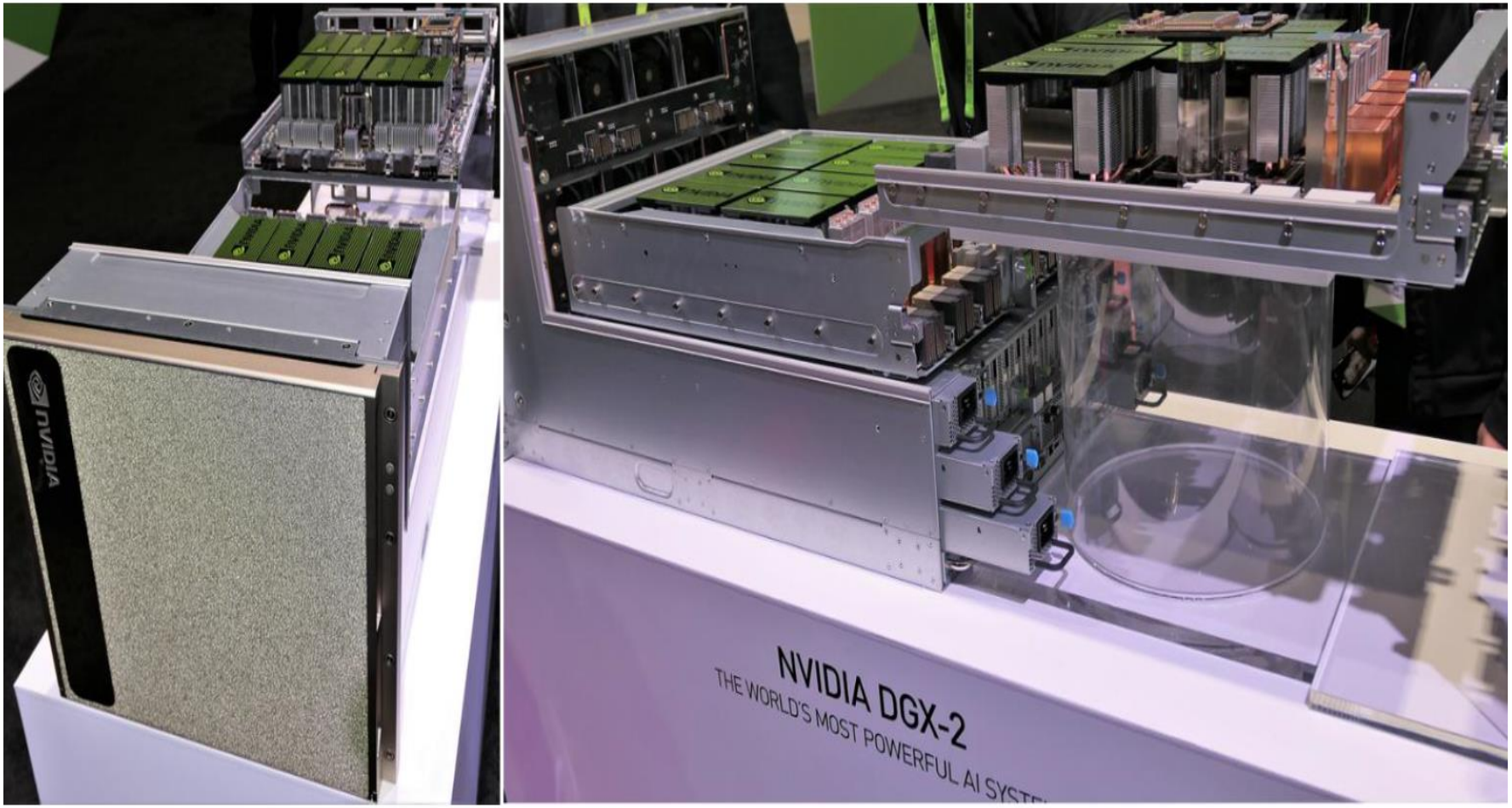
DGX-2 vs DGX-1 구조 비교





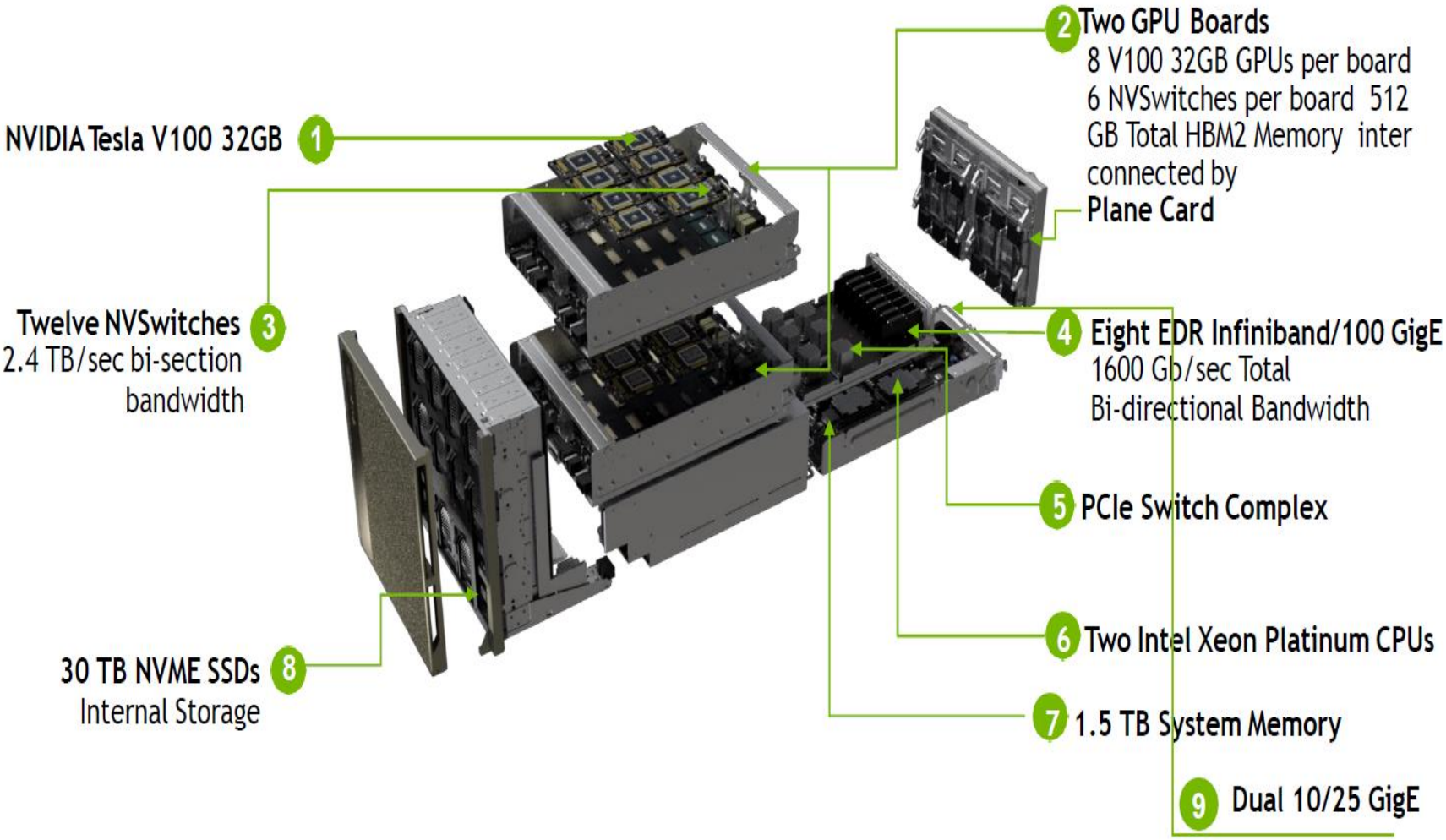
	DGX-2	DGX-1
GPU 개수	16	8
1GPU to 1GPU	Always 6 NVLink	1 NVLink or 2 NVLink
연결방식	Fully Connected	4 GPUs connected directly
구성방식	Symmetric	Hybrid cube mesh

DGX-2



10U - 10000W

DGX-2 내부

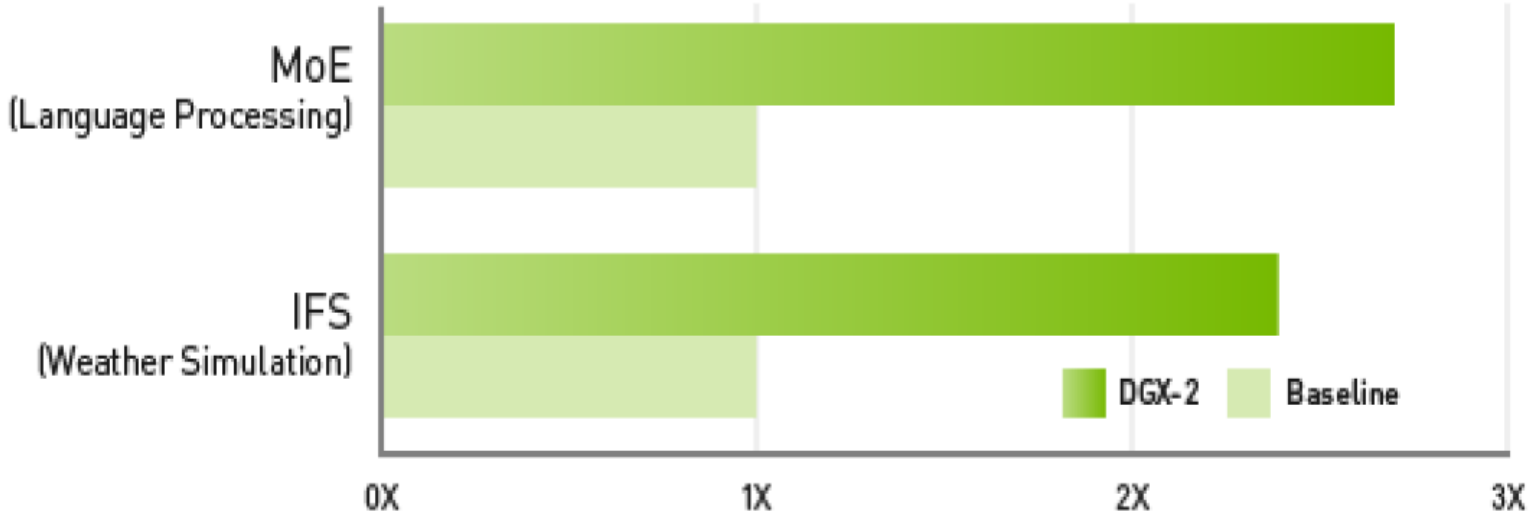

NVIDIA Tesla V100 32GB 1
2 Two GPU Boards
8 V100 32GB GPUs per board
6 NVSwitches per board 512 GB Total HBM2 Memory inter connected by
Plane Card
Twelve NVSwitches 3
2.4 TB/sec bi-section bandwidth
4 Eight EDR Infiniband/100 GigE
1600 Gb/sec Total
Bi-directional Bandwidth
5 PCIe Switch Complex
6 Two Intel Xeon Platinum CPUs
30 TB NVME SSDs 8
Internal Storage
7 1.5 TB System Memory
9 Dual 10/25 GigE

SPEC

DGX-2 Specs		DGX-1 Specs
SYSTEM SPECIFICATIONS		SYSTEM SPECIFICATIONS
GPUs	16X NVIDIA® Tesla V100 32GB	8X NVIDIA® Tesla V100
GPU Memory	512GB total	128GB/ 256GB total
Performance	2 petaFLOPS	1 peta FLOPS
CUDA Cores	81920	40960
Tensor Cores	10240	5120
NVSwitches	12	N/A
CPU	Dual Intel Xeon Platinum 8168, 2.7 GHz, 24-cores	Dual Intel Xeon E5-2698v4 20 cores
System Memory	1.5TB	512 GB
Network	8X 100Gb/sec	4X 100Gb/sec
	Infiniband/100GigE	Infiniband/100GigE
	Dual 10/25Gb/sec Ethernet	Dual 10Gb/sec Ethernet
Storage	OS: 2X 960GB NVME SSDs Internal Storage: 30TB (8X 3.84TB) NVME SSDs	OS: 480GB SSD / Storage : 4X 1.92TB SSD Raid 0
Software	Ubuntu Linux OS See Software stack for details	Ubuntu Linux OS See Software stack for details
System Weight	340 lbs (154.2 kgs)	134 lbs (60.8 kgs)
System Dimensions	Height: 17.3 in (440.0 mm) Width: 19.0 in (482.3 mm) Length: 31.3 in (795.4 mm)	Height: 131.0 mm Width: 444 mm Length: 866 mm
	No Front Bezel 32.8 in (834.0 mm)	
	With Front Bezel	With Front Bezel
Temp Range	5°C to 35°C	10°C to 35°C

성능 비교

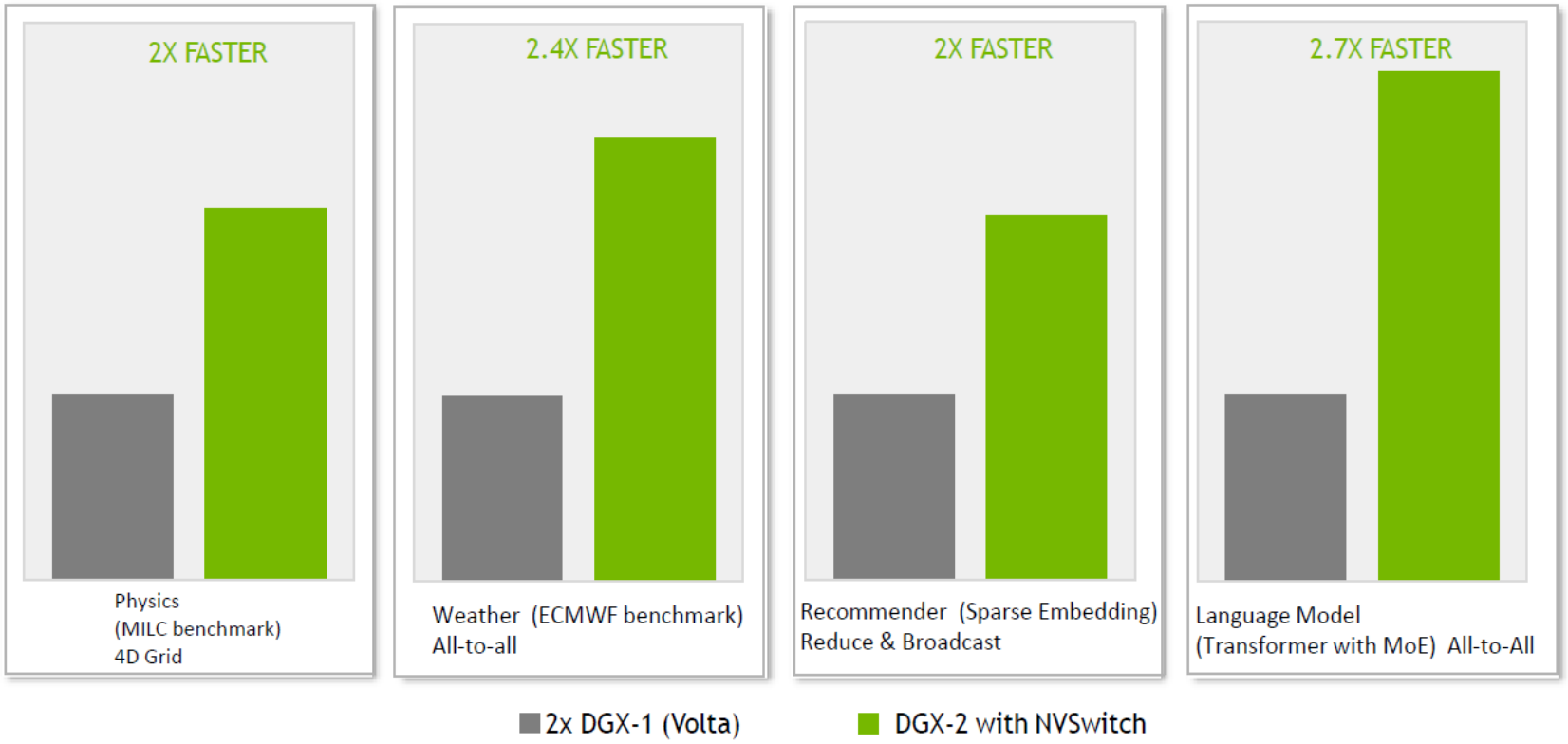
NVSwitch Delivers a >2X Speedup for Deep Learning and HPC*



System Configs: Each of the two DGX-1 servers have dual-socket Xeon E5 2690v4 Processor, 8 x V100 GPUs; servers connected via a 4 EDR (100Gb) InfiniBand connections. DGX-2 server has dual-socket Xeon Scalable Processor Platinum 8168 Processors, 16 x Tesla V100 GPUs.

성능 비교

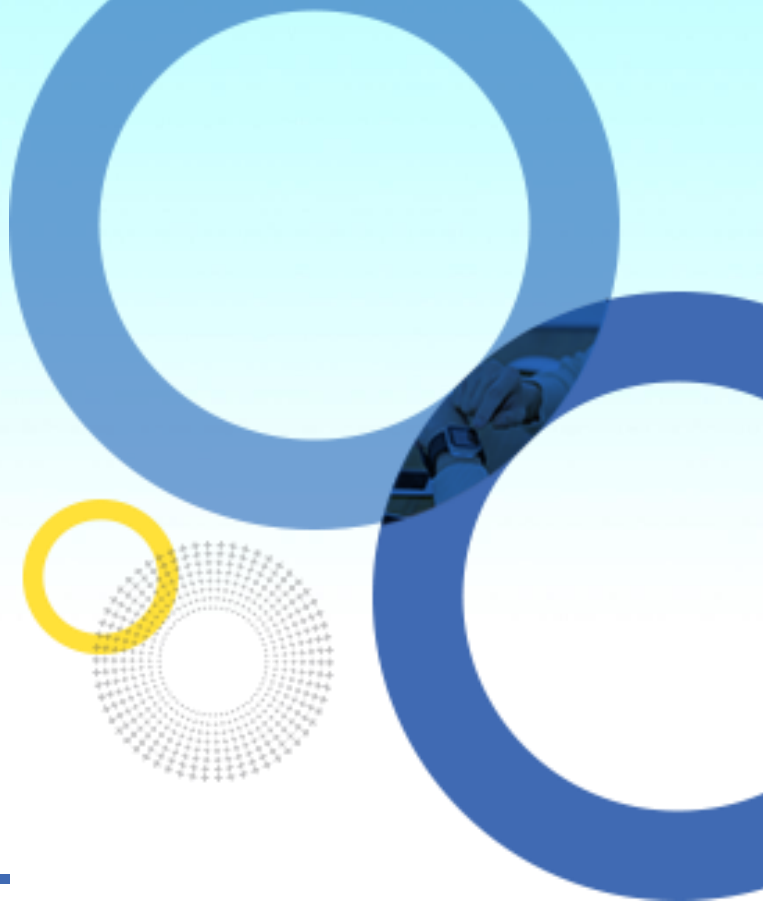
2X HIGHER PERFORMANCE WITH NVSWITCH



2 DGX-1V servers have dual socket Xeon E5 2698v4 Processor. 8 x V100 GPUs. Servers connected via 4X 100Gb IB ports | DGX-2 server has dual-socket Xeon Platinum 8168 Processor. 16 V100 GPUs

디지털 트랜스포메이션을 통한
비즈니스 혁신과 4차 산업혁명 대응을 위한

GIT 솔루션즈 데이



감사합니다.

BayNex

GiT 굿모닝아이텍(주)