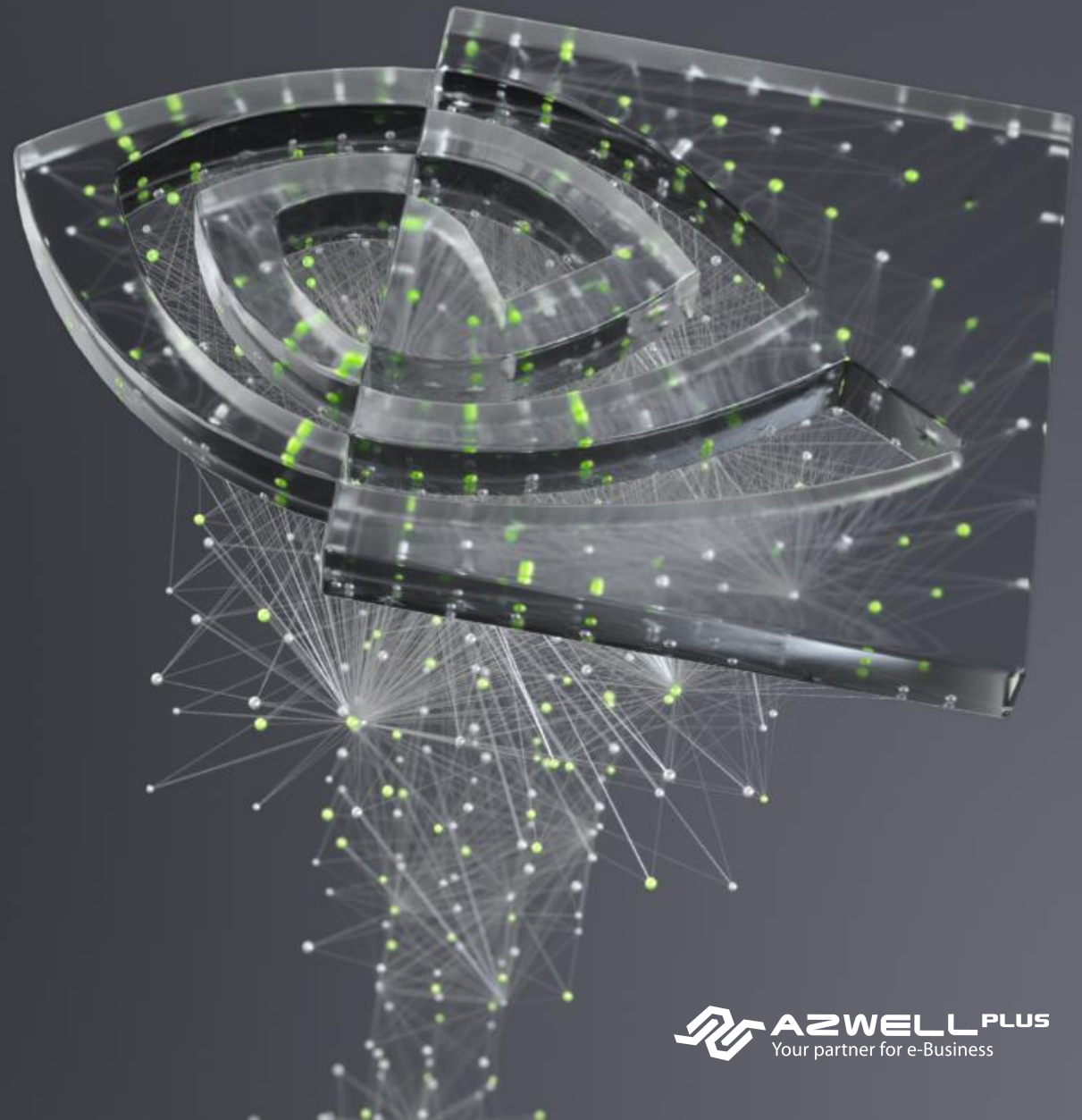
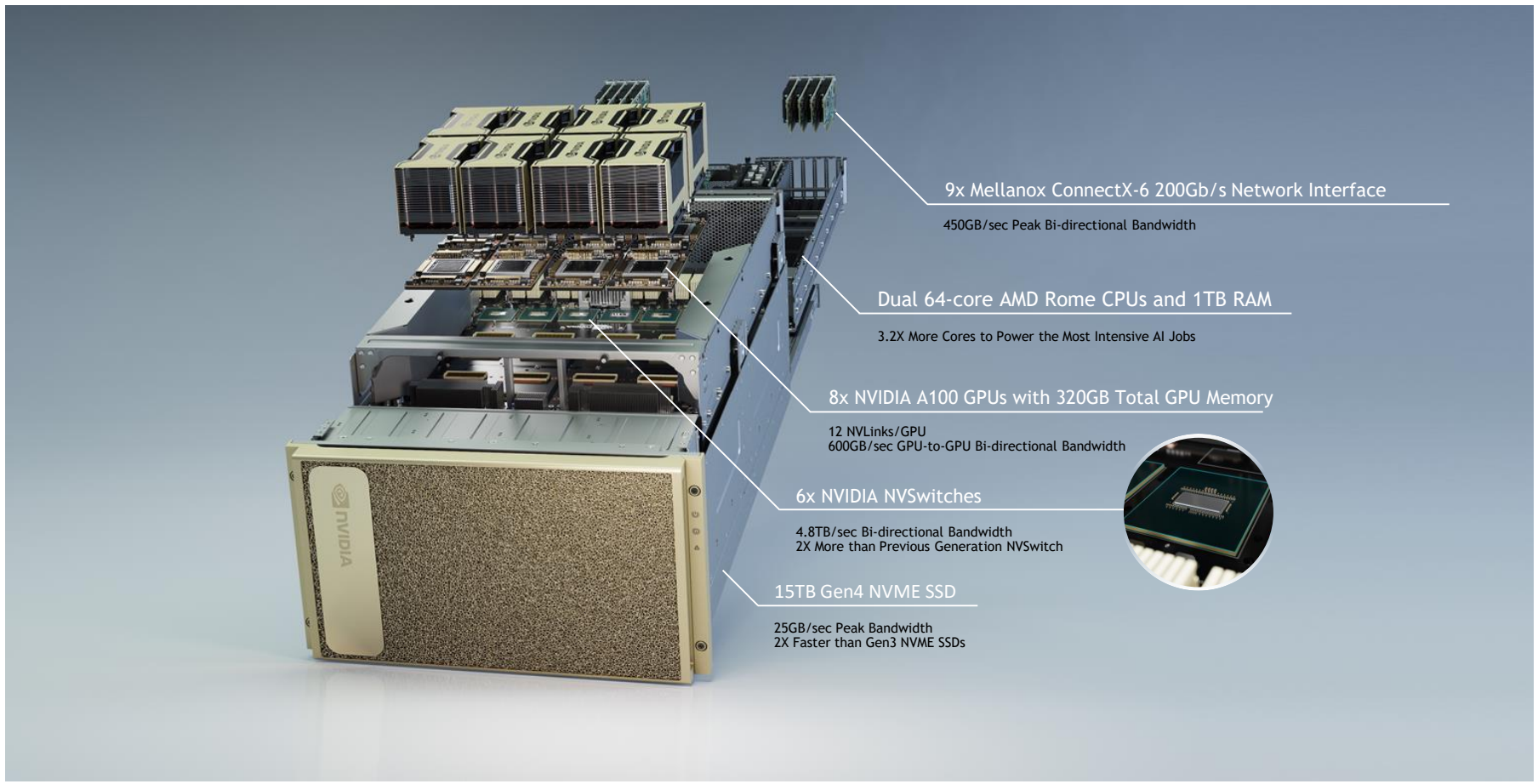
NVIDIA

DGX A100



AZWELL PLUS
Your partner for e-Business

GAME-CHANGING PERFORMANCE FOR INNOVATORS

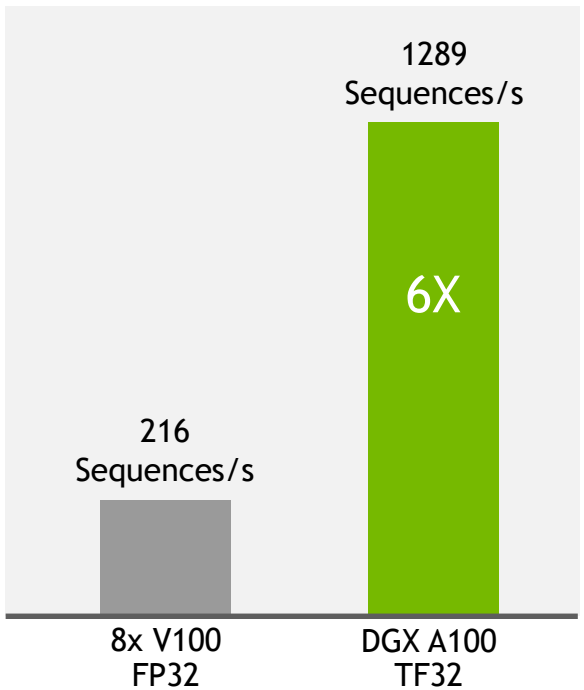


NVIDIA DGX A100 SYSTEM SPECS

App Focus Components	
GPUs	8x NVIDIA A100 Tensor Core GPUs
GPU Memory	320GB Total
NVIDIA NVSwitch	6
Performance	5 petaFLOPS AI 10 petaOPS INT8
CPU	Dual AMD Rome, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)
System Memory	1TB
Networking	9x Mellanox ConnectX-6 VPI HDR InfiniBand/200GigE 10th Dual-port ConnectX-6 optional
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives

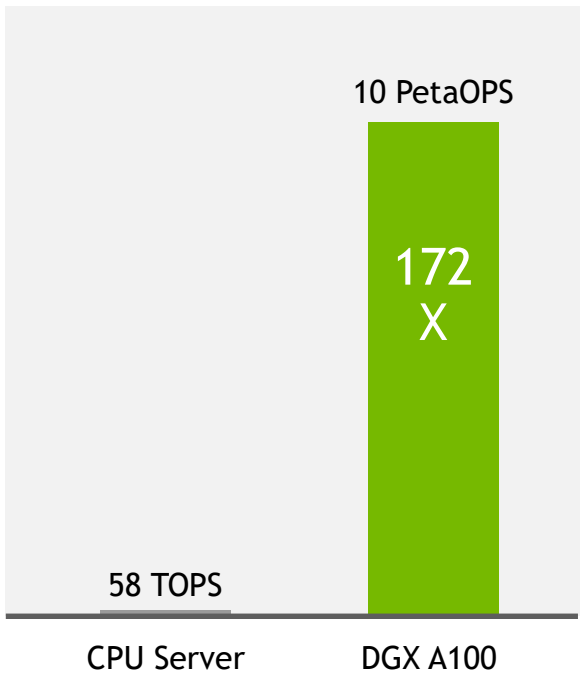
Power and Physical Dimensions	
System Power Usage	6.5 kW Max
System Weight	271 lbs (123 kgs)
System Dimensions	6 Rack Units (RU) Height: 10.4 in (264.0 mm) Width: 19.0 in (482.3 mm) Max Length: 35.3 in (897.1 mm) Max
Operating Temperature	5°C to 30°C (41°F to 86°F)
Cooling	Air

DGX A100 PERFORMANCE



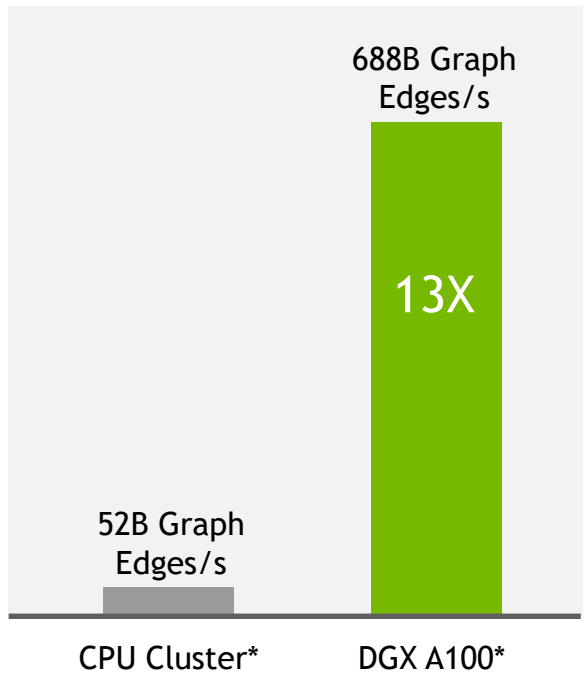
Training
NLP: BERT-Large

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 8x V100 using FP32 precision
DGX A100: DGX A100 with 8x A100 using TF32 precision



Inference
Peak Compute

CPU Server: 2x Intel Platinum 8280 using INT8
DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity



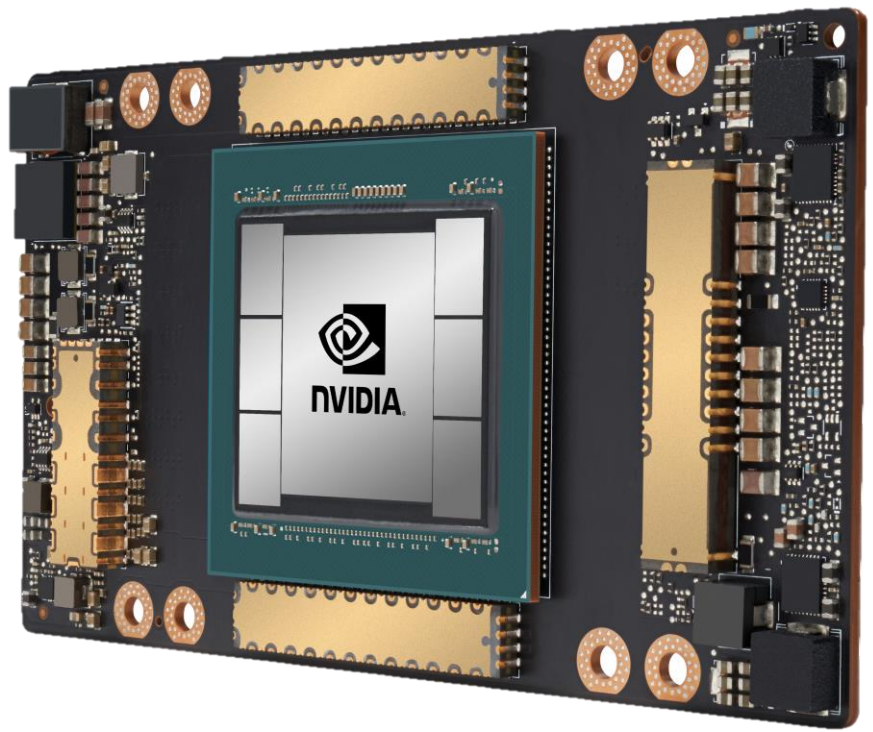
Analytics
PageRank

3000x CPU Servers vs. 4x DGX A100 Published Common Crawl Data Set: 128B Edges, 2.6TB Graph

ANNOUNCING NVIDIA A100

Greatest Generational Leap – 20X Volta

	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

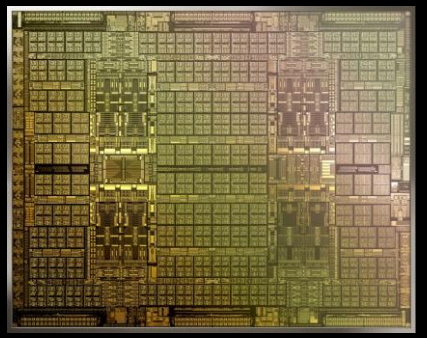


54B XTOR | 826mm2 | TSMC 7N | 40GB Samsung HBM2 | 600 GB/s NVLink

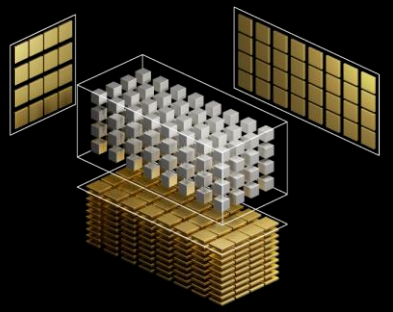
NVIDIA A100 SPECS TABLE

	Peak Performance
Transistor Count	54 billion
Die Size	826 mm^2
FP64 CUDA Cores	3,456
FP32 CUDA Cores	6,912
Tensor Cores	432
Streaming Multiprocessors	108
FP64	9.7 teraFLOPS
FP64 Tensor Core	19.5 teraFLOPS
FP32	19.5 teraFLOPS
TF32 Tensor Core	156 teraFLOPS | 312 teraFLOPS*
BFLOAT16 Tensor Core	312 teraFLOPS | 624 teraFLOPS*
FP16 Tensor Core	312 teraFLOPS | 624 teraFLOPS*
INT8 Tensor Core	624 TOPS | 1,248 TOPS*
INT4 Tensor Core	1,248 TOPS | 2,496 TOPS*
GPU Memory	40 GB
Interconnect	NVLink 600 GB/s PCIe Gen4 64 GB/s
Multi-Instance GPUs	Various Instance sizes with up to 7MIGs @5GB
Form Factor	4/8 SXM GPUs in HGX A100
Max Power	400W (SXM)

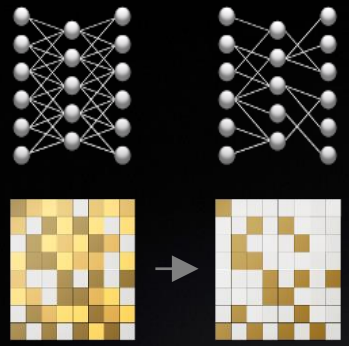
5 MIRACLES OF A100



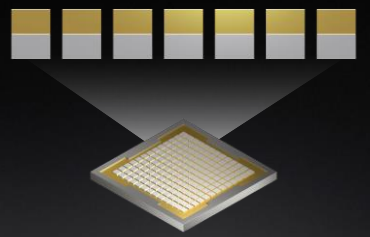
Ampere
World's Largest 7nm chip
54B XTORS, HBM2



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



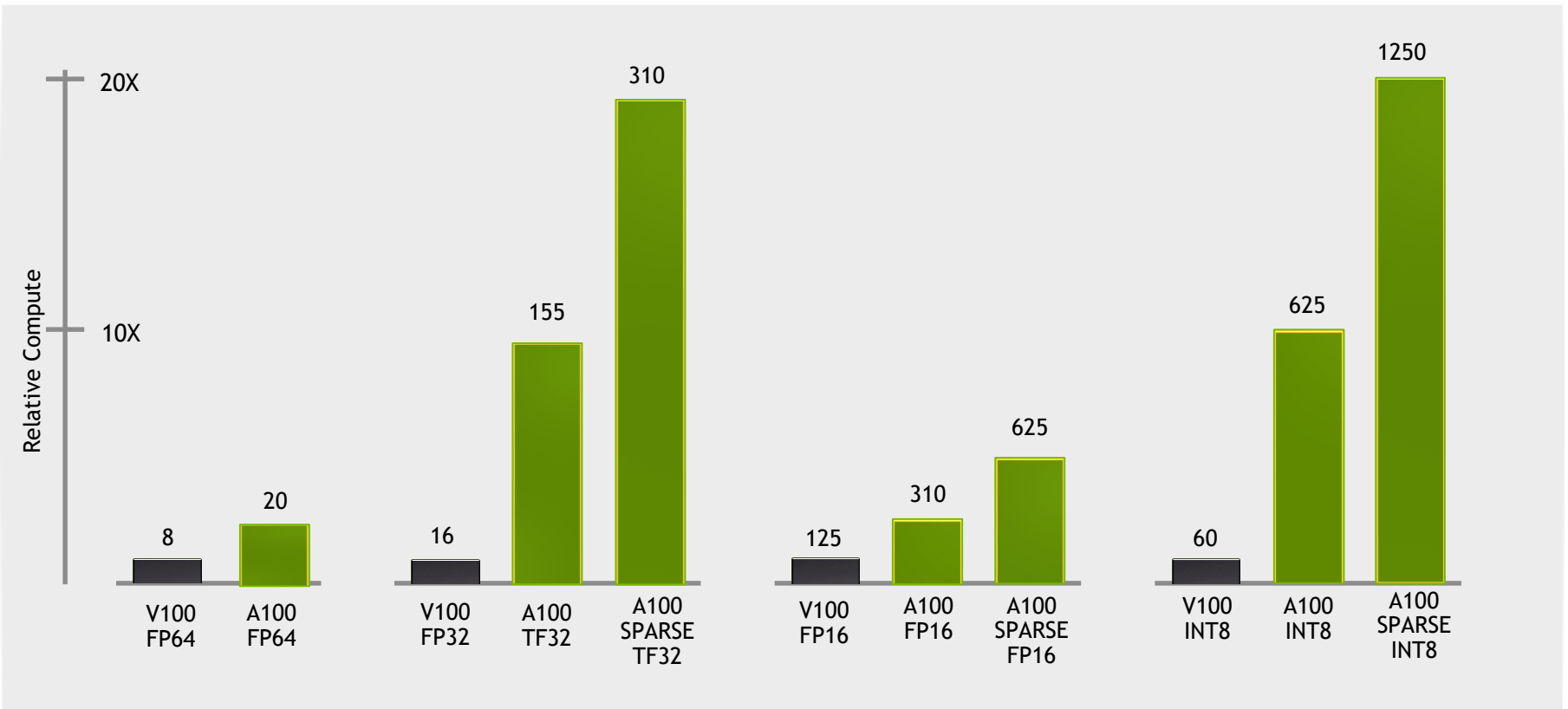
New Sparsity Acceleration
Harness Sparsity in AI Models
2x AI Performance



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

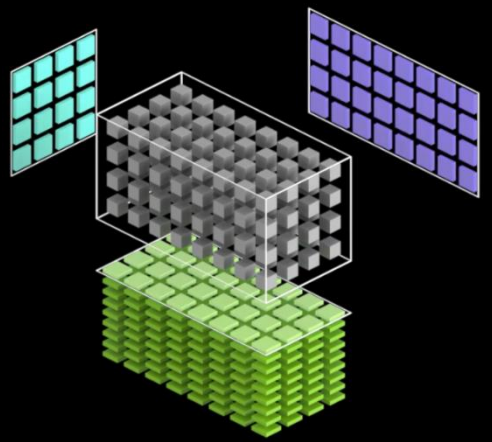
3rd Gen NVLINK and NVSWITCH
Efficient Scaling to Enable Super GPU
2X More Bandwidth

NVIDIA A100
GREATEST GENERATIONAL LEAP – 20X VOLTA

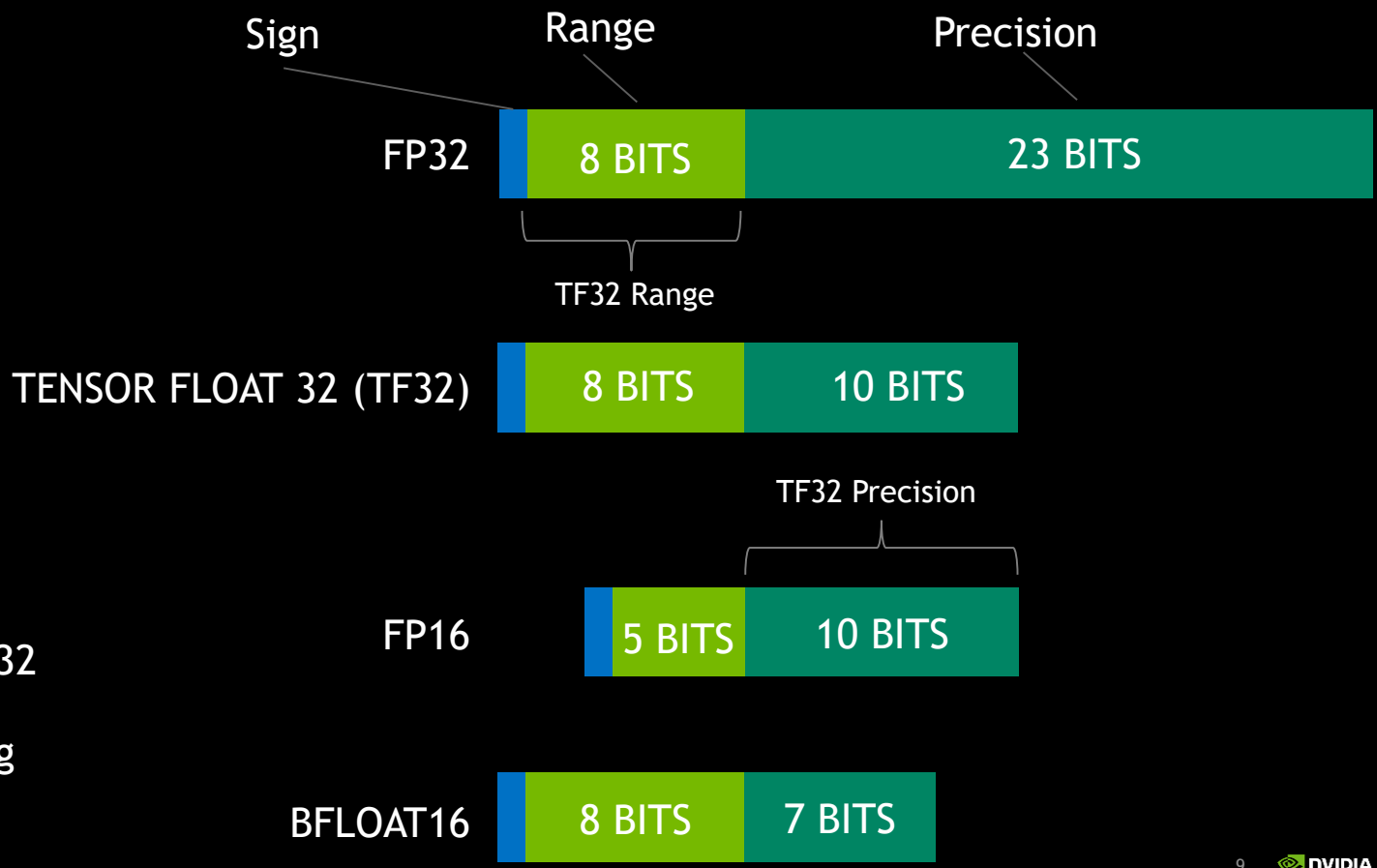


Peak Performance in Trillion Operations Per Second (TOPS) of A100 Compared to V100 | V100 Rounded off to the nearest whole number | A100 rounded off to the nearest 5.

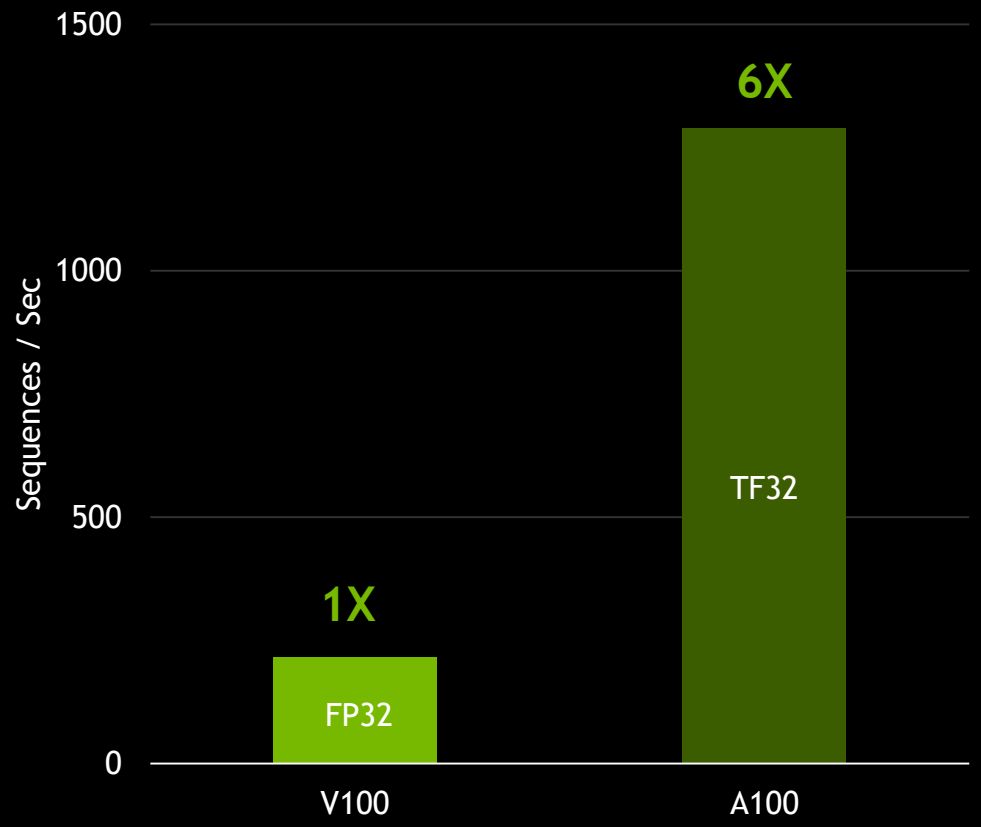
NEW TF32 TENSOR CORES



- Range of FP32 and Precision of FP16
- Input in FP32 and Accumulation in FP32
- No Code Change Speed-up for Training



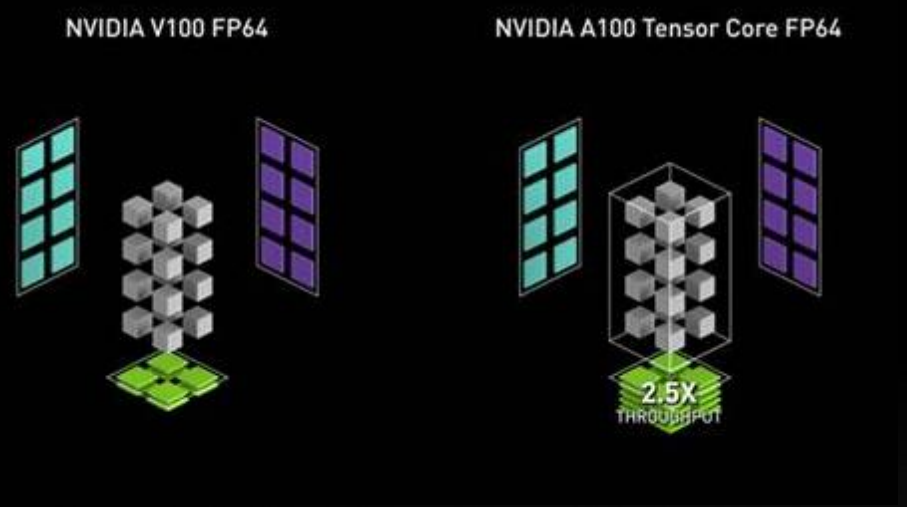
6X OUT OF THE BOX SPEEDUP WITH TF32 FOR AI TRAINING



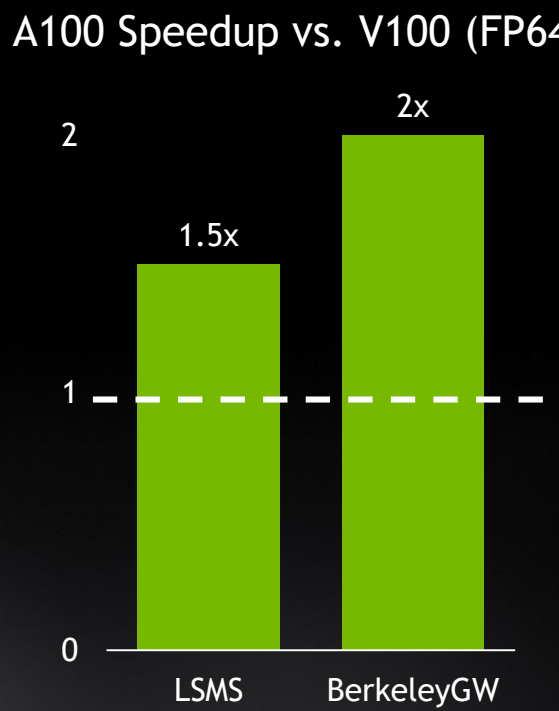
BERT Pre-Training Throughput using Pytorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 V100: DGX-1 Server with 8xV100 using FP32 precision A100: DGX A100 Server with 8xA100 using TF32 precision |

UP TO 2X MORE PERFORMANCE

Leveraging New A100 FP64 Tensor Cores

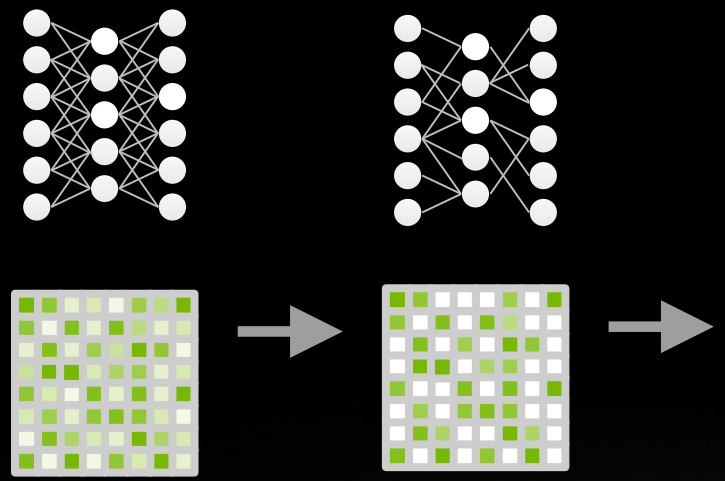


- IEEE FP64
- Drop-in acceleration via cuBLAS, cuTensor, and cuSolver



Application [Benchmarks]: BerkeleyGW [Chi Sum + MTXEL] using DGX-1V (8xV100) and DGX-A100 (8xA100) | LSMS [Fe128] single V100 SXM2 vs. A100 SXM4

STRUCTURAL SPARSITY BRINGS ADDITIONAL SPEEDUPS



Dense Matrix

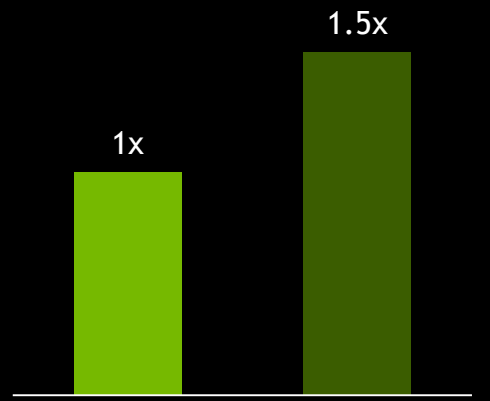
Sparse Matrix

2X

Faster Execution

A100 Tensor Core

BERT Large Inference

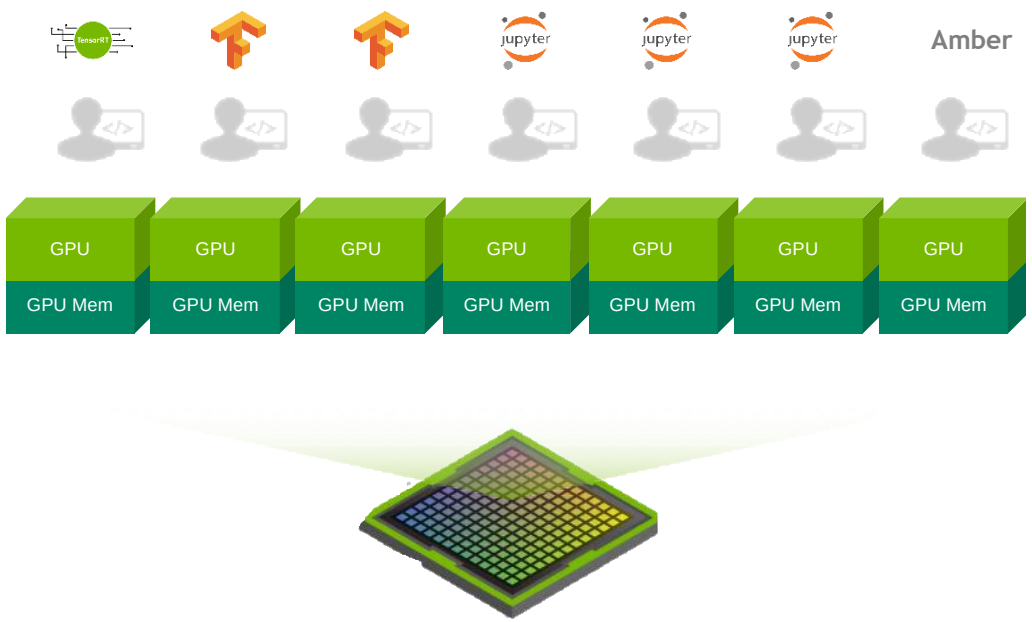


- Structured sparsity: Half the values are zero
- Skip half of the compute and mem fetches
- Compute up to 2x rate vs non-sparse

BERT Large Inference | precision = INT8 with and without sparsity, large Batch inference, A100 with 7 MIGS

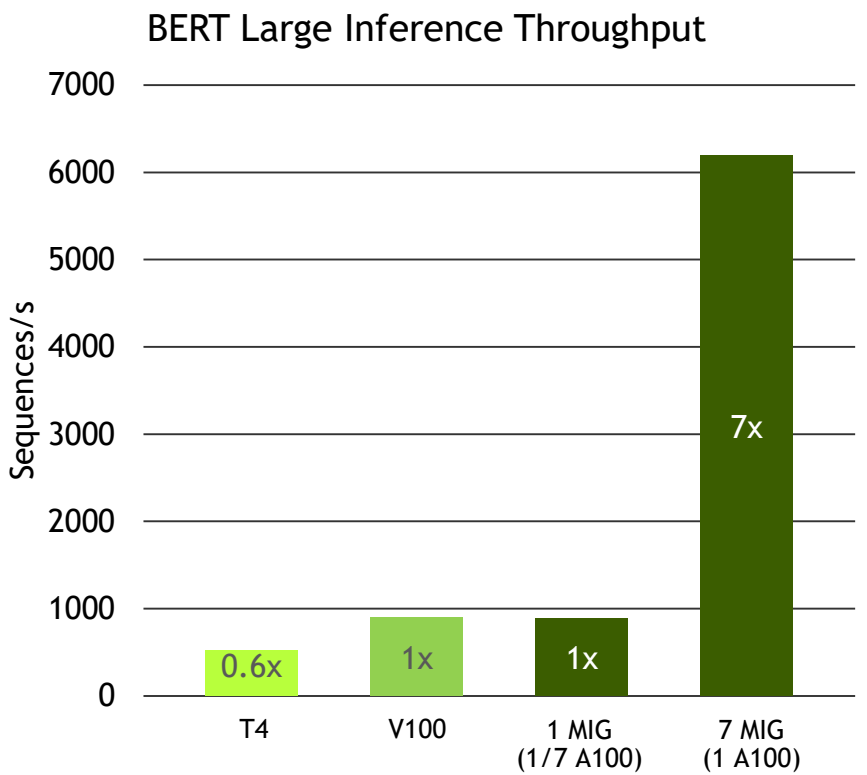
NEW MULTI-INSTANCE GPU (MIG)

Optimize GPU Utilization, Expand Access to More Users with Guaranteed Quality of Service



- Up To 7 GPU Instances In a Single A100: Dedicated SM, Memory, L2 cache, Bandwidth for hardware QoS & isolation
- Simultaneous Workload Execution With Guaranteed Quality Of Service: All MIG instances run in parallel with predictable throughput & latency
- Right Sized GPU Allocation: Different sized MIG instances based on target workloads
- Flexibility to run any type of workload on a MIG instance
- Diverse Deployment Environments: Supported with Bare metal, Docker, Kubernetes, Virtualized Env.

7X HIGHER INFERENCE THROUGHPUT WITH MIG



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

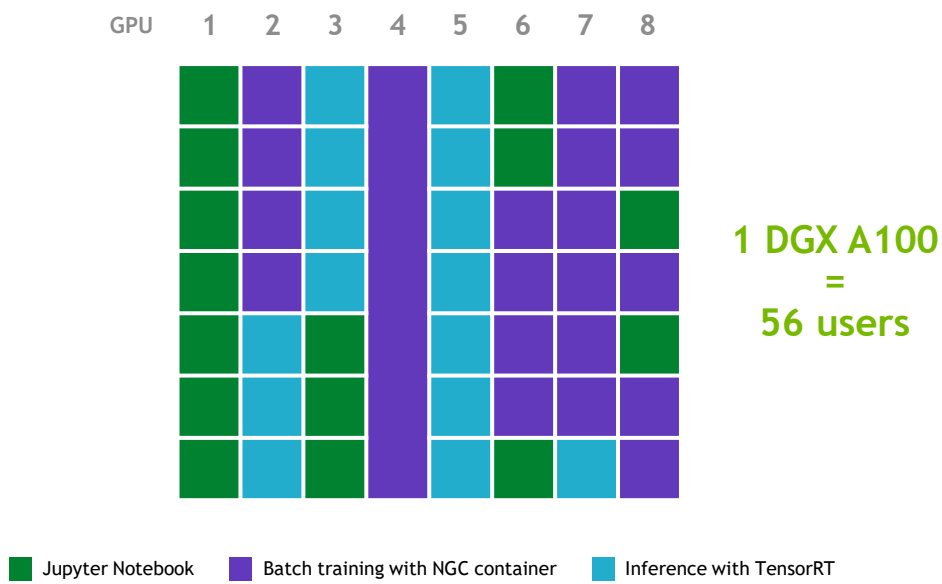
MULTI-INSTANCE GPU (MIG) ON DGX A100

More Users and Better GPU Utilization

GPU Instance Size	Number of GPU Instances Available	GPU Memory
1 GPU Slice	7	5 GB
2 GPU Slice	3	10 GB
3 GPU Slice	2	20 GB
4 GPU Slice	1	20 GB
7 GPU Slice	1	40 GB

Flexible Utilization

Configure GPUs for vastly different workloads
with GPU instances that are fault-isolated



1 DGX A100
=
56 users

MOST POWERFUL TOOL FOR A DATA SCIENCE TEAM

Using DGX A100 with MIG to Give Every Developer Power to Explore

One DGX A100 delivers:

- 5 petaFLOPS of AI training power, or
- 10 petaOPS of AI inference power
- With MIG, a team of 25 developers can share a DGX A100

Each developer gets:

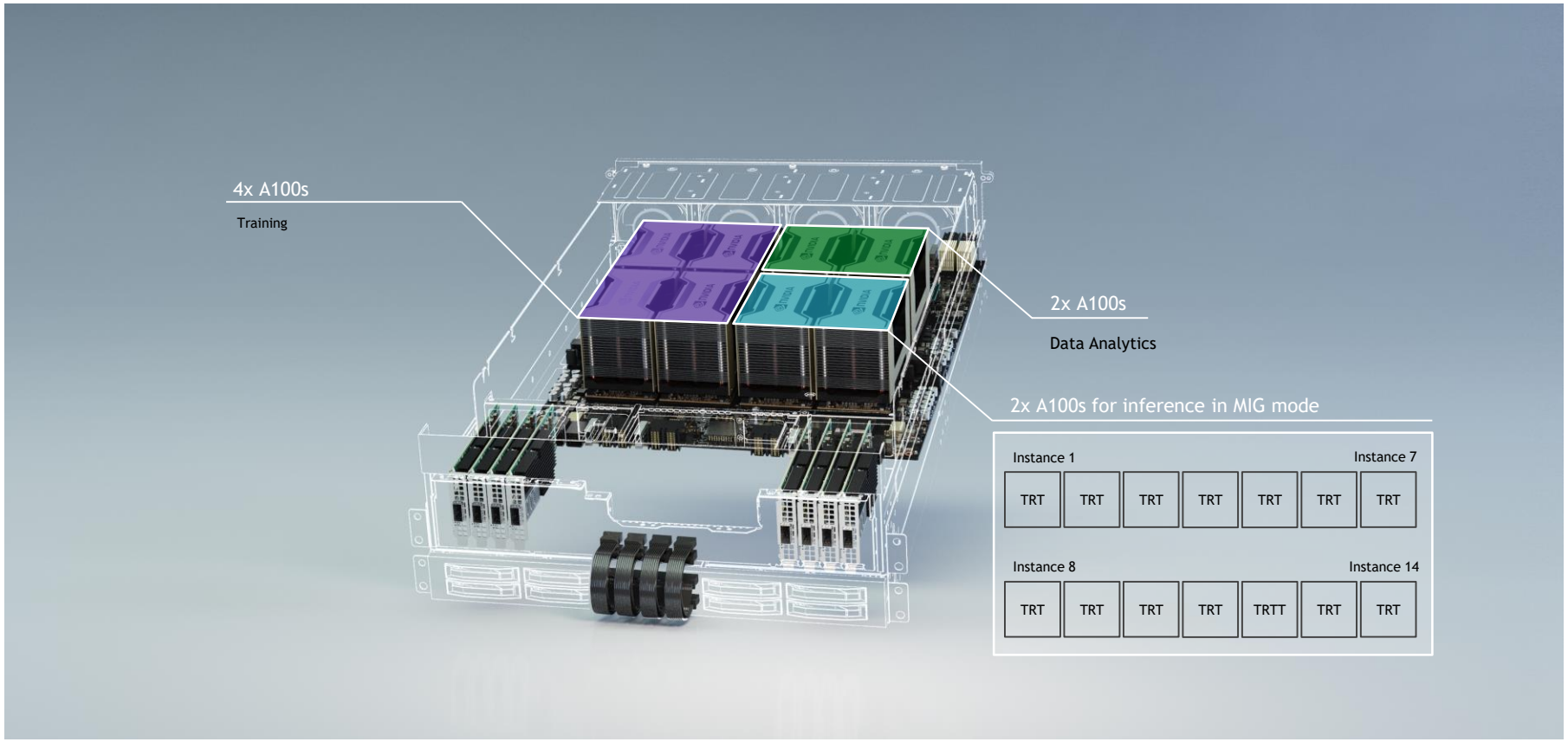
- Over 180 teraFLOPS for training
  = (2) reserved cloud V100 instances

or

- Over 357 teraOPS for inference
  = (6) dedicated 28-core dual CPU servers

CONSOLIDATING DIFFERENT WORKLOADS ON DGX A100

One Platform for Training, Inference and Data Analytics



DGX A100: NEW A100 GPUS AND 2X FASTER NVSWITCH

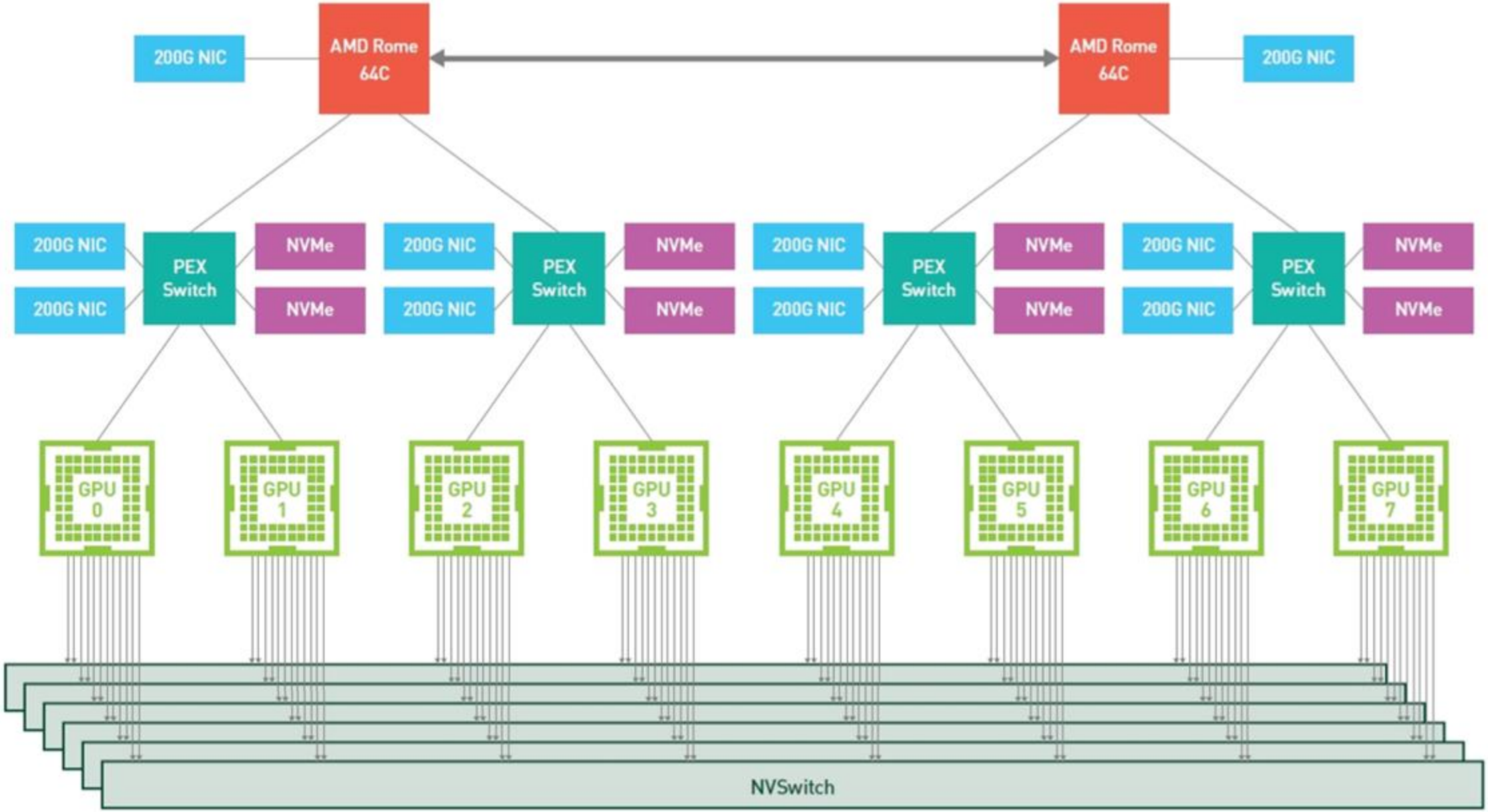
5 PetaFLOPS AI Performance

Eight new A100 Tensor Core GPUs/320GB total HBM2

- Twelve NVLinks per GPU, 2x more than V100
- 600GB/s bi-directional bandwidth between any GPU pair
- ~10X PCIe Gen4 bandwidth with next-gen NVLink

All GPUs fully connected with six next-gen NVSwitch

- 4.8TB/s bi-directional bandwidth
- In one second we could transfer 426 hours of HD video


200G NIC
AMD Rome 64C
AMD Rome 64C
200G NIC
200G NIC
200G NIC
PEX Switch
NVMe
NVMe
200G NIC
200G NIC
PEX Switch
NVMe
NVMe
200G NIC
200G NIC
PEX Switch
NVMe
NVMe
200G NIC
200G NIC
PEX Switch
NVMe
NVMe
GPU 0
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
NVSwitch

NGC – GPU-OPTIMIZED SOFTWARE

Build AI Faster, Deploy Anywhere

NGC

CONTAINERS ▶ TRAINED MODELS ▶ HELM CHARTS



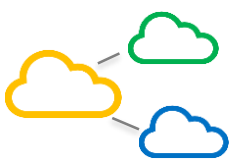
100+ | 100+ | ML, Inference

AI TOOLKITS & SDKs

Healthcare | Smart Cities | Conversational AI | Robotics | more

ENCRYPTED

x86 | ARM | POWER


MULTI-CLOUD

ON-PREM

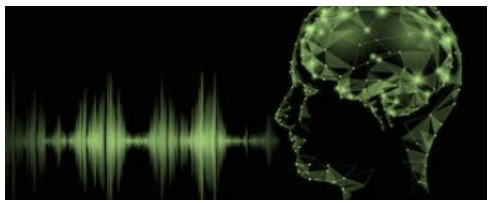

HYBRID CLOUD


EDGE

NEW NGC FEATURES

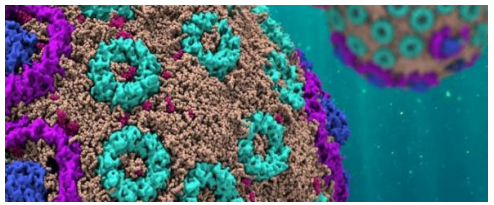
SDKs & CONTAINERS FOR A100

Q2



Industry SDKs - Jarvis, Aerial…

DL - TF, PyT, MxNet, Triton…



HPC - NAMD, Chroma, LAMMPS…

NGC PRIVATE REGISTRY

Now

Easily grant and manage content access

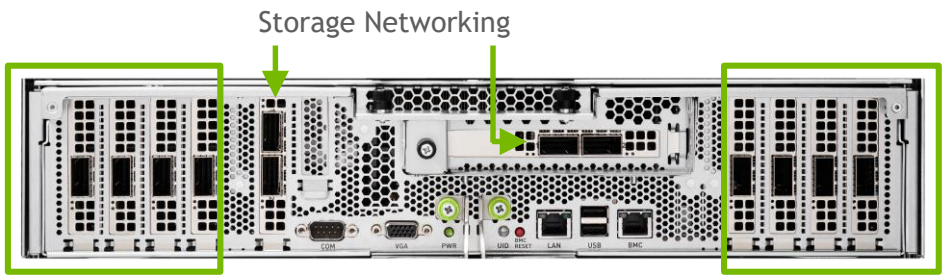
Container scanning and signing. Model versioning and encryption

Securely share and collaborate

Multi-arch support - x86, Arm, POWER

UNMATCHED SCALABILITY WITH MELLANOX NETWORKING

Highest Network Throughput for Data and Clustering




Single-port CX-6 NIC

For clustering networking:

- Eight Mellanox single-port ConnectX-6
- Supporting HDR/HDR100/EDR InfiniBand default or 200GigE

For data/storage networking:

- One Mellanox dual-port ConnectX-6
  - Supporting: 200/100/50/40/25/10Gb Ethernet default or HDR/HDR100/EDR InfiniBand
- One optional Dual-Port CX-6 available as add-on

450GB/sec peak bi-directional bandwidth

All I/O now PCIe Gen4, 2x performance increase over Gen3

Scale up multiple DGX A100 nodes with Mellanox Quantum Switch, the world's smartest network switch

RACK-SCALE INFRASTRUCTURE

Building an AI Center of Excellence with DGX POD Built on DGX A100

4-node
DGX POD

8-node
DGX POD

- DGX POD more attainable than ever with DGX A100
- Experience a faster start with building flexible AI infrastructure
- Proven architectures, with leading storage partners
- Up to 40 PFLOPS computing power in just 2 racks

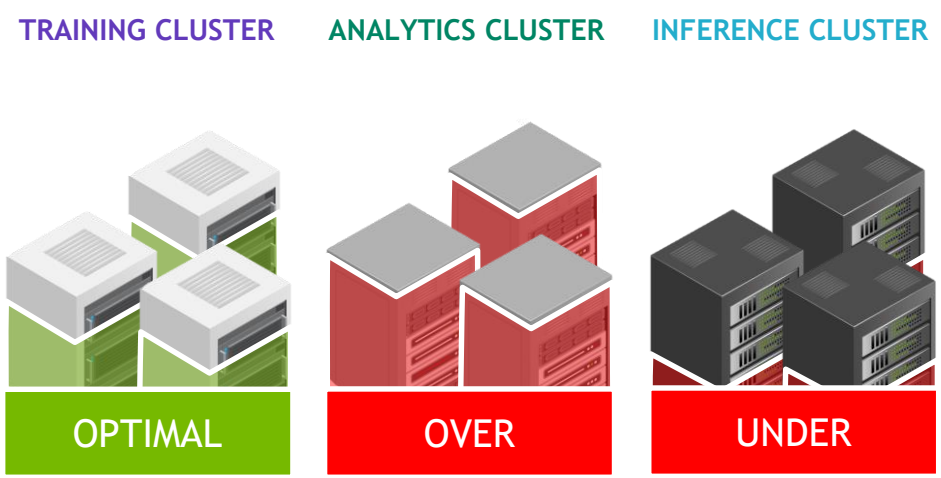
Complete AI infrastructure solutions:
DGX, storage, networking, services, software

ELASTIC AI INFRASTRUCTURE WITH DGX A100

DGX A100 with MIG Delivers New Agility for Today's Enterprise Data Center

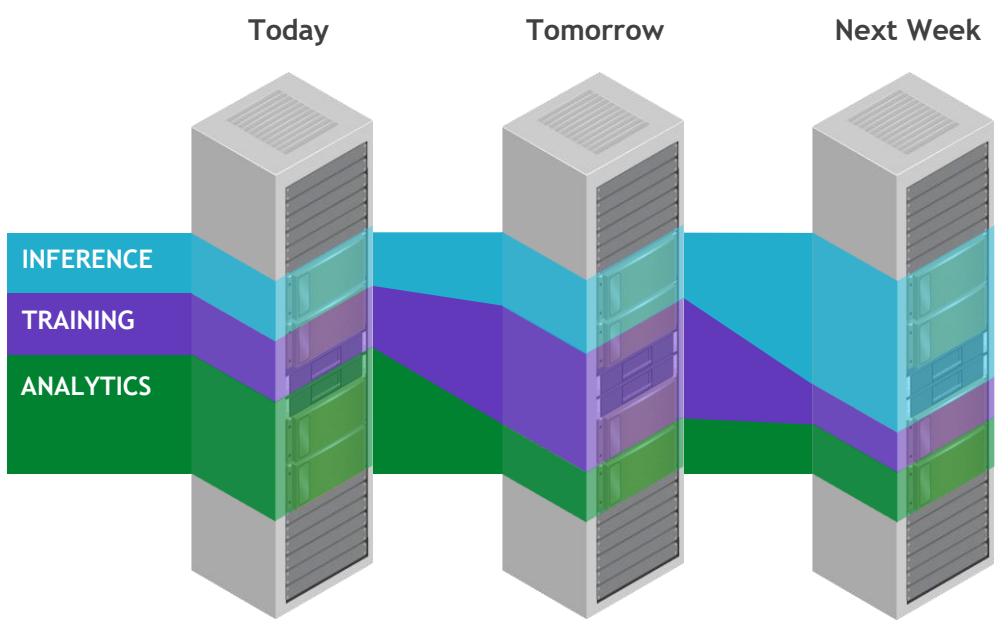
Traditional Infrastructure is Constrained

Infrastructure silos starve AI workloads or waste capacity



DGX A100 Infrastructure is Agile

DGX A100 infrastructure uses MIG to allocate GPU resources to workloads



NVIDIA DGX SUPERPOD WITH DGX A100

Unmatched Data Center Scalability — Deployed in Under 3 Weeks

Leadership-class AI infrastructure

- The blueprint for AI power and scale using DGX A100
- Infused with the expertise of NVIDIA's AI practitioners
- Designed to solve the previously unsolvable
- Configurations start at 20 systems

NVIDIA DGX SuperPOD deployed in SATURNV

- 1,120 A100 GPUs
- 140 DGX A100 systems
- 170 Mellanox 200G HDR switches
- 4 PB of high-performance storage
- 700 PFLOPS of power to train the previously impossible

ARGONNE NATIONAL LABORATORY

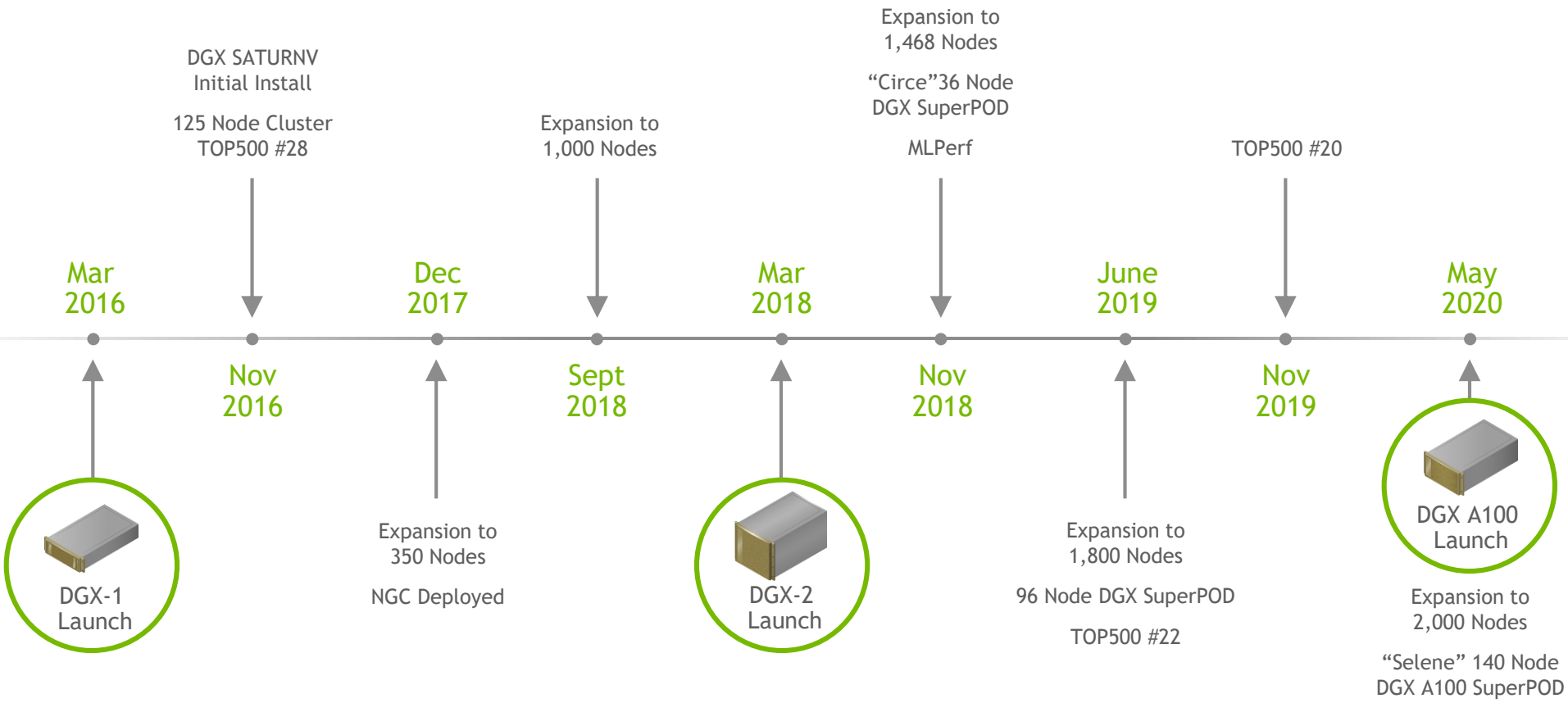
World's First DGX A100 Supercomputer Fighting COVID-19

- 24-node Cluster of DGX A100 Systems
- 192 A100 GPUs
- Mellanox High-Speed Low-Latency Network Fabric
- 120 PetaFLOPS of AI Computing Power for Scientific Research

Argonne NATIONAL LABORATORY

NVIDIA DGX SATURNV EVOLUTION

World's Largest Infrastructure Purpose-Built for AI Research and Development



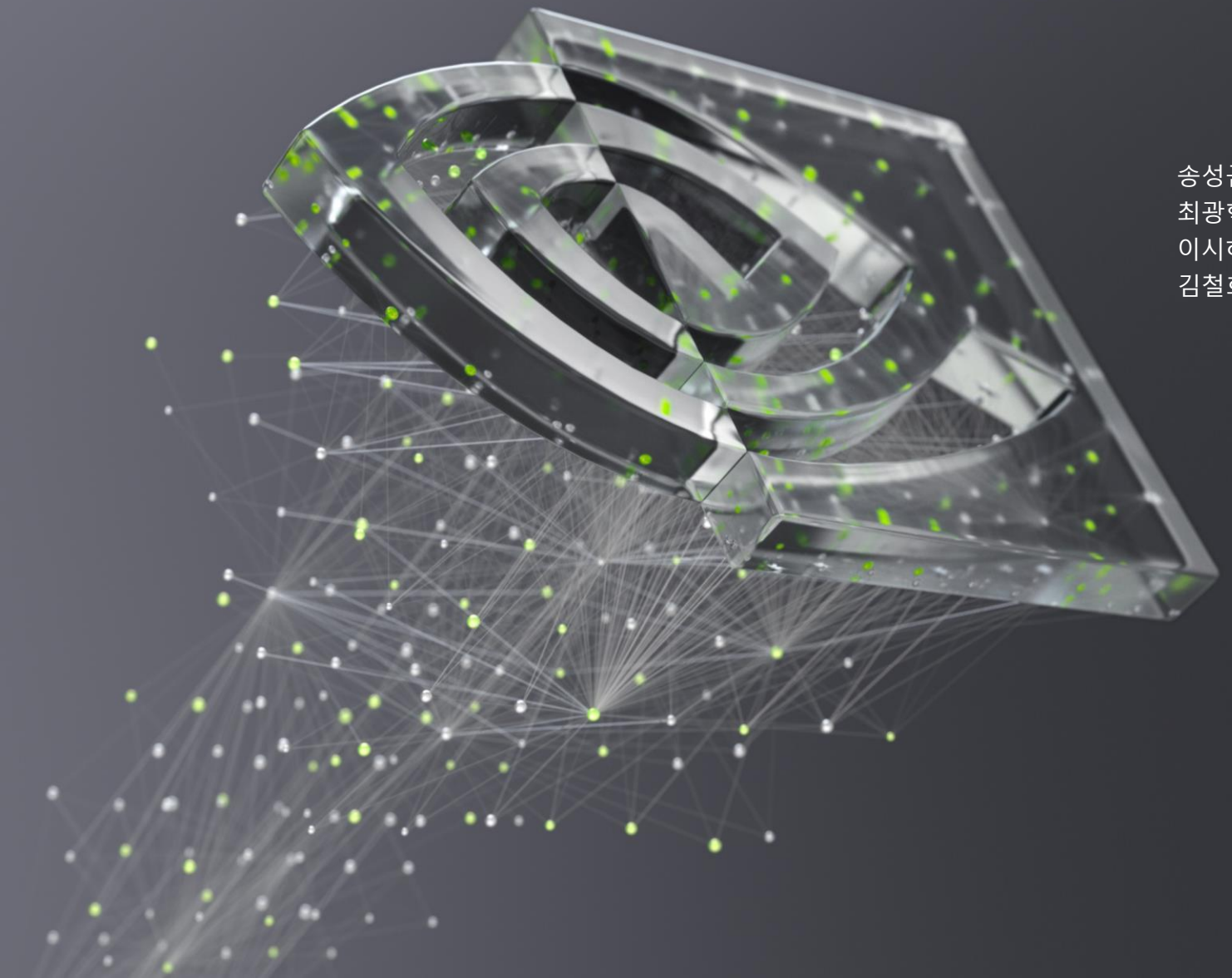
INTRODUCING: NVIDIA DGXpert

With Every DGX System - Your Trusted Navigator in AI Transformation

14,000+
AI-Fluent Experts

DESIGN | PLAN | BUILD | TEST | DEPLOY | OPERATE | MONITOR

With you every step of the way - Included with every DGX system



문의처

㈜에즈웰플러스 (www.azwell.co.kr)

송성근 전무	010-9255-8570	sk.song@azwell.co.kr
최광혁 상무	010-4321-0001	khchoi@azwell.co.kr
이시하 과장	010-2078-9268	siha@azwell.co.kr
김철희 대리	010-5002-8152	chulhee@azwell.co.kr

NVIDIA