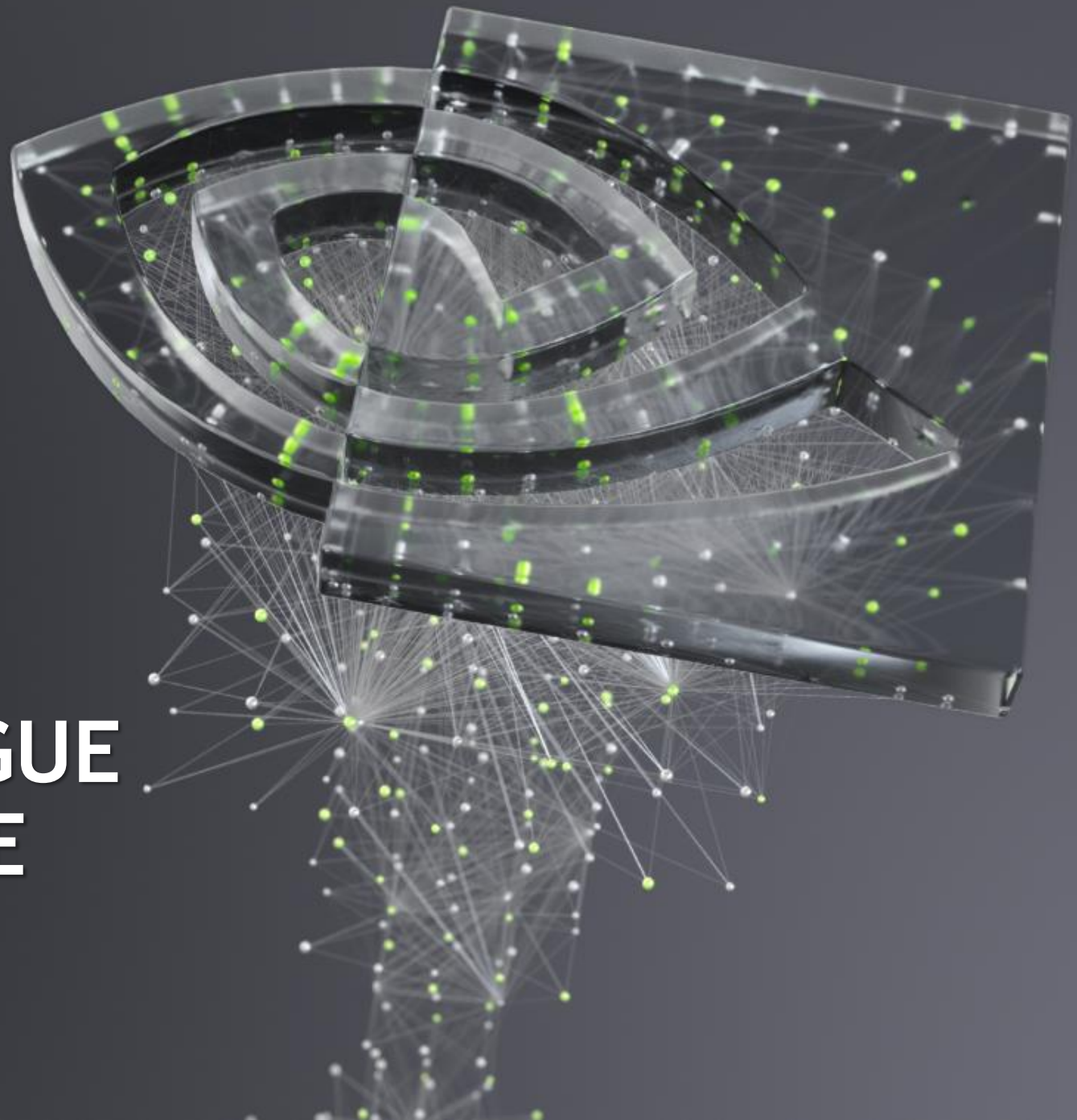


NVIDIA

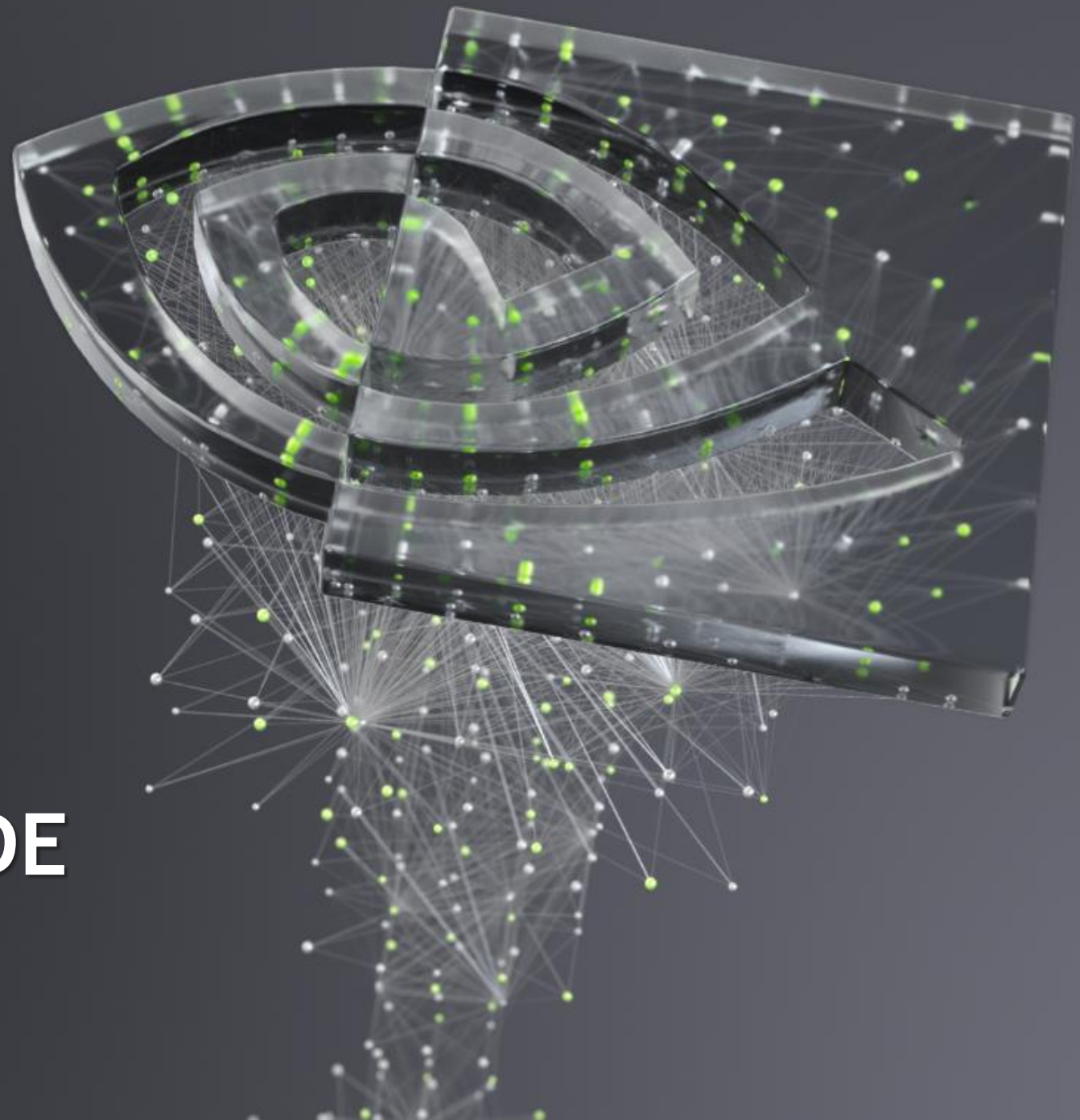
# AI DAYS 2020, PRAGUE
# A100 PERFORMANCE

# CONTACT DETAILS


**Ralph Hinsche**

*Business Development Manager*
*Higher Education & Research*

NVIDIA GmbH
Flößergasse 2
Haus 1 West, 3. OG
81369 München
T+49 (0)173 533 3514
**M** +49 (0)173 533 3514
rhinsche@nvidia.com
www.nvidia.eu
www.facebook.com/NVIDIADeutschland



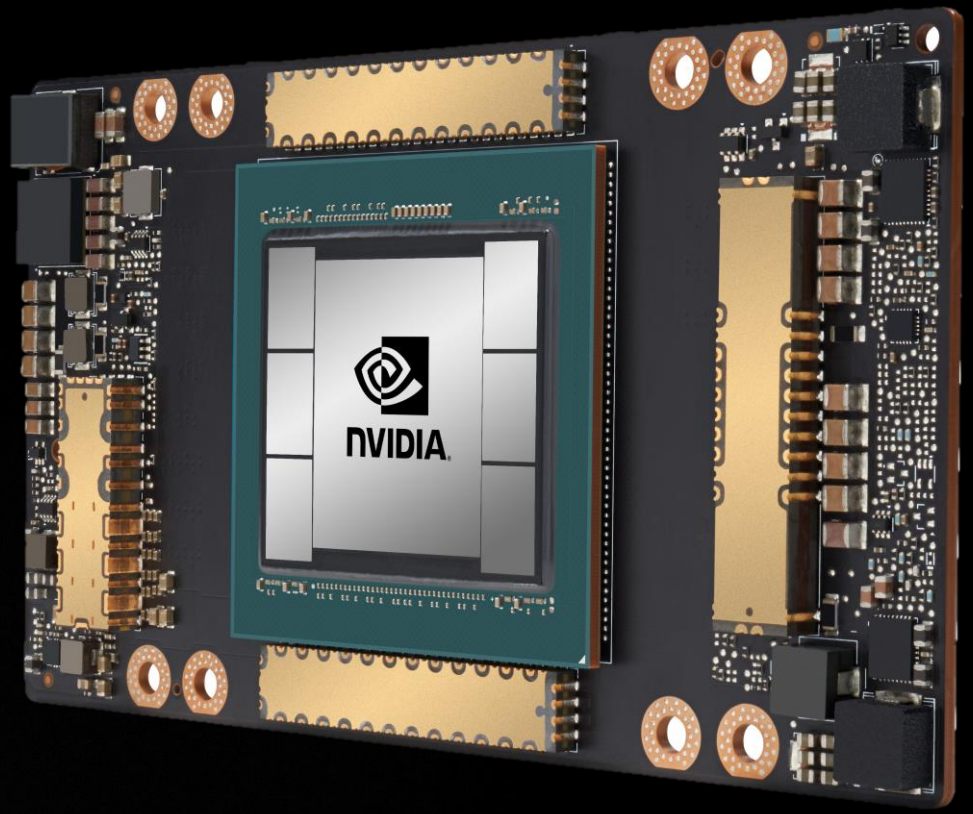
# DEEP LEARNING PERFORMANCE GUIDE

As of October ‘20 using NGC 20.09 containers

# 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 |

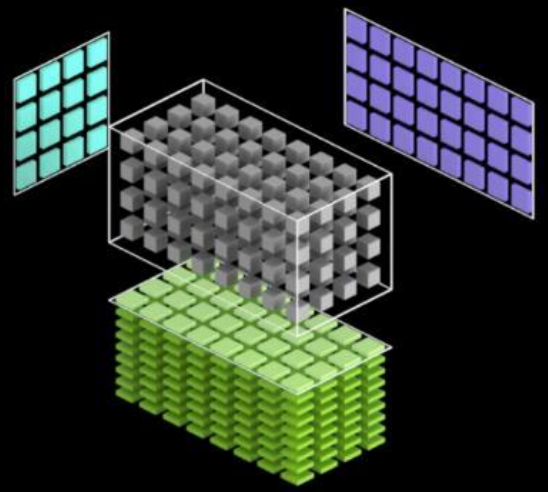


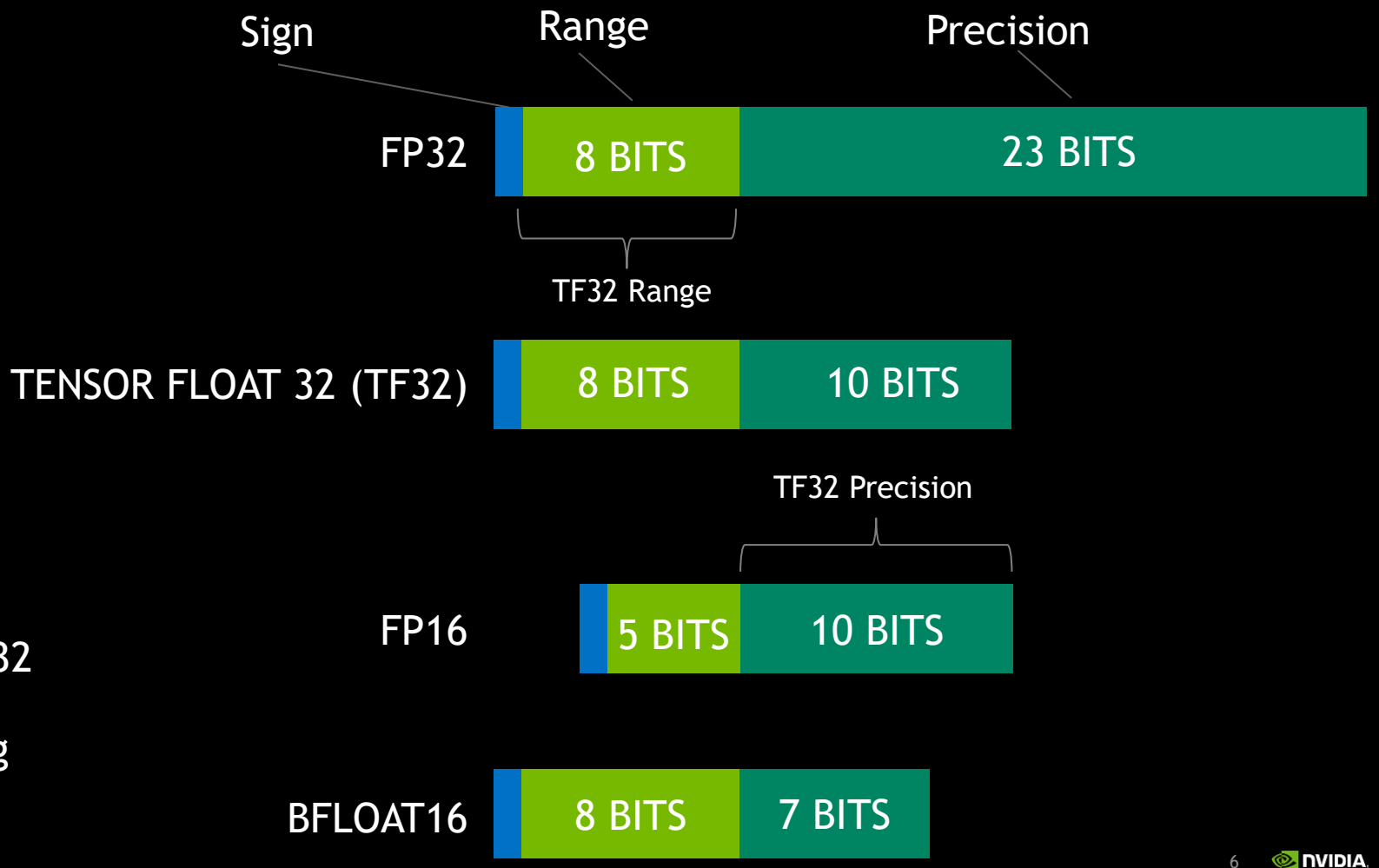

54B XTOR | 826mm2 | TSMC 7N | 40GB Samsung HBM2 | 600 GB/s NVLink

# 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<br>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) |

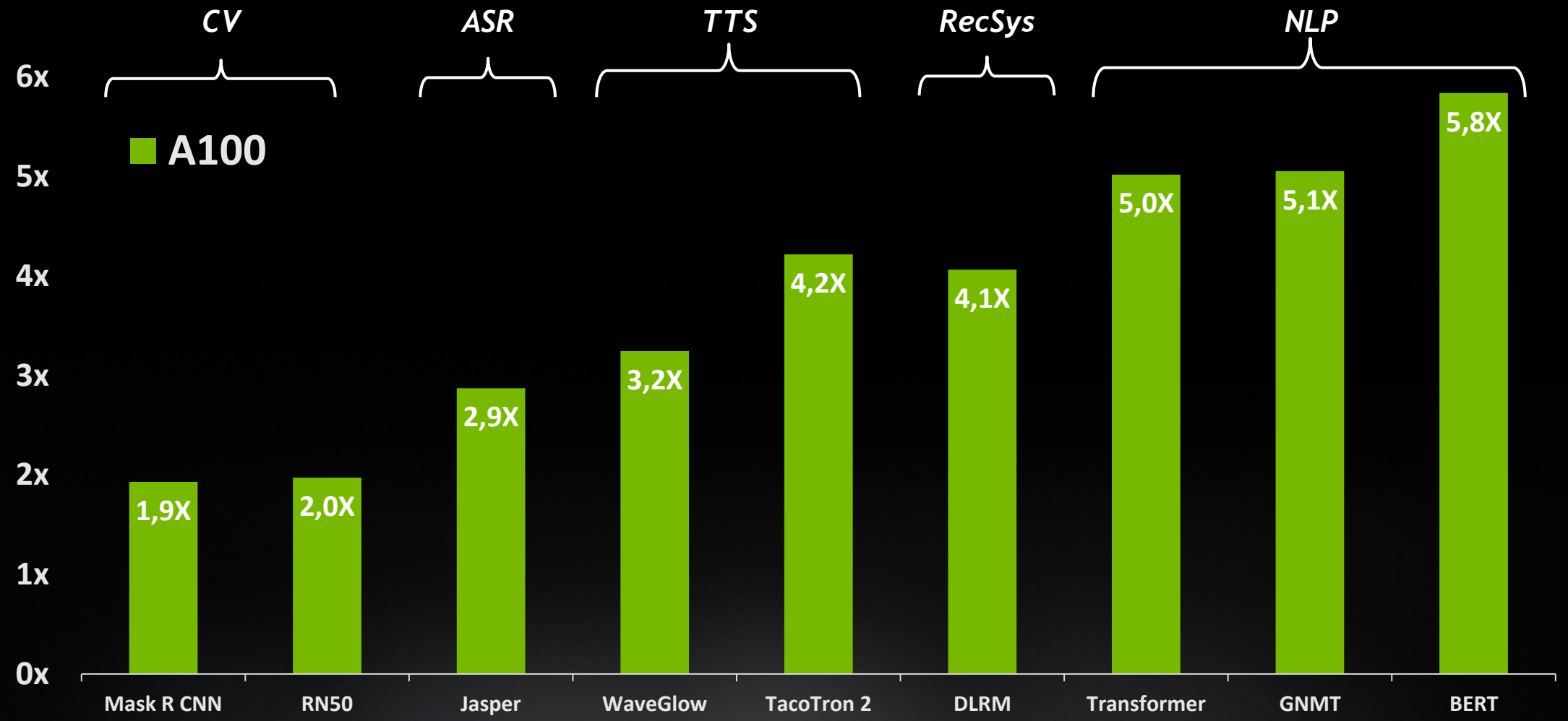
# NEW TF32 TENSOR CORES






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

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

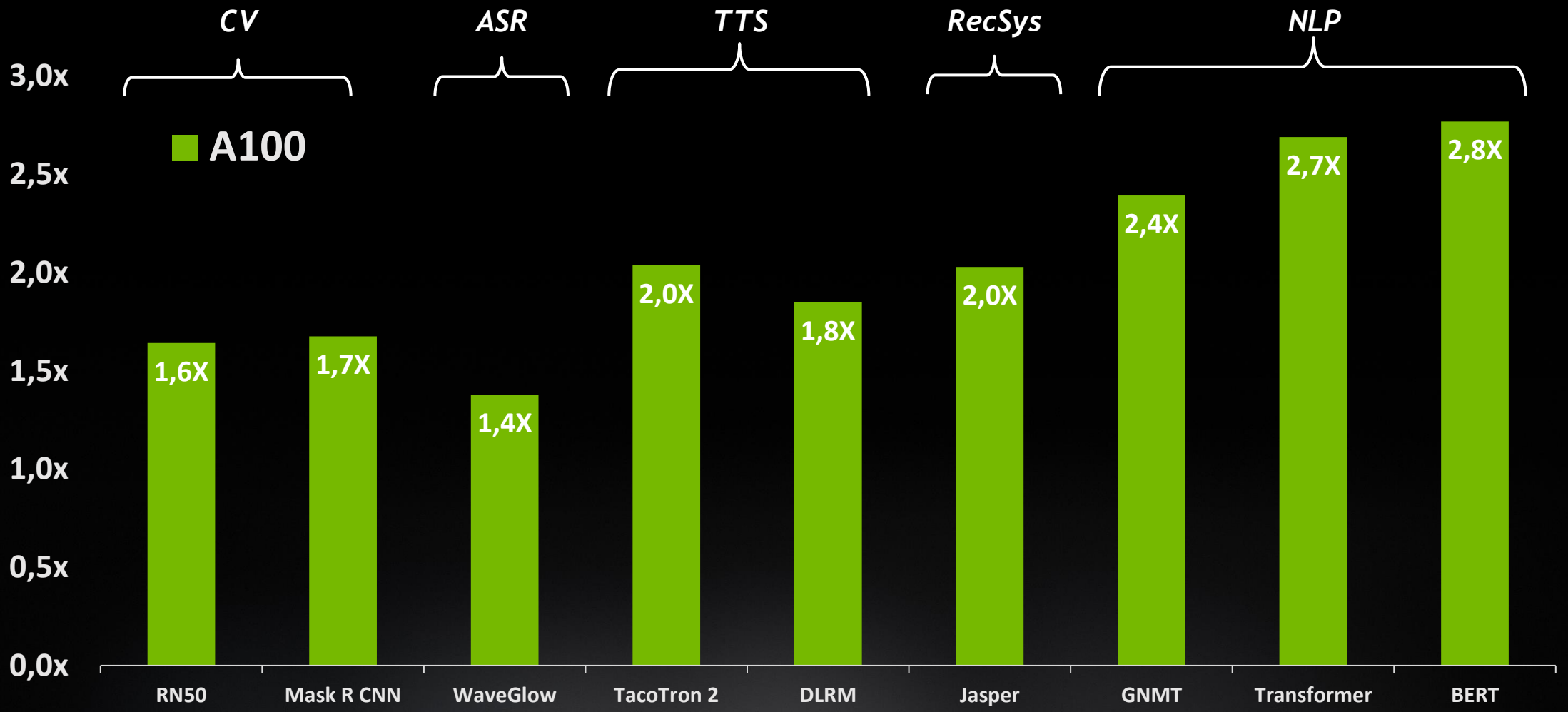



*All results are measured*

*V100 used is DGX-1 (8xV100 16GB). A100 used is s DGX A100 (8xA100 SXM4), except DLRM which uses 1xV100 and 1xA100; V100 uses FP32 and A100 uses TF32*

*RN50 uses MXNET Batch size = 96, Mask R CNN uses PyTorch BS = 4 (V100) and BS=8 (A100), DLRM uses PyTorch and BS=32768, Jasper uses PyTorch and BS=16,, WaveGlow uses PyTorch and BS=4 (V100) and 10 (A100), TacoTron2 uses PyTorch and BS=48 (V100) and 128 (A100), Transformer uses PyTorch and BS=2560 (V100) and 6656 (A100 and GNMT uses PyTorch and BS=128 (V100) and 512 (A100); 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*

# UP TO 3X SPEEDUP WITH FP16 & AMP FOR AI TRAINING



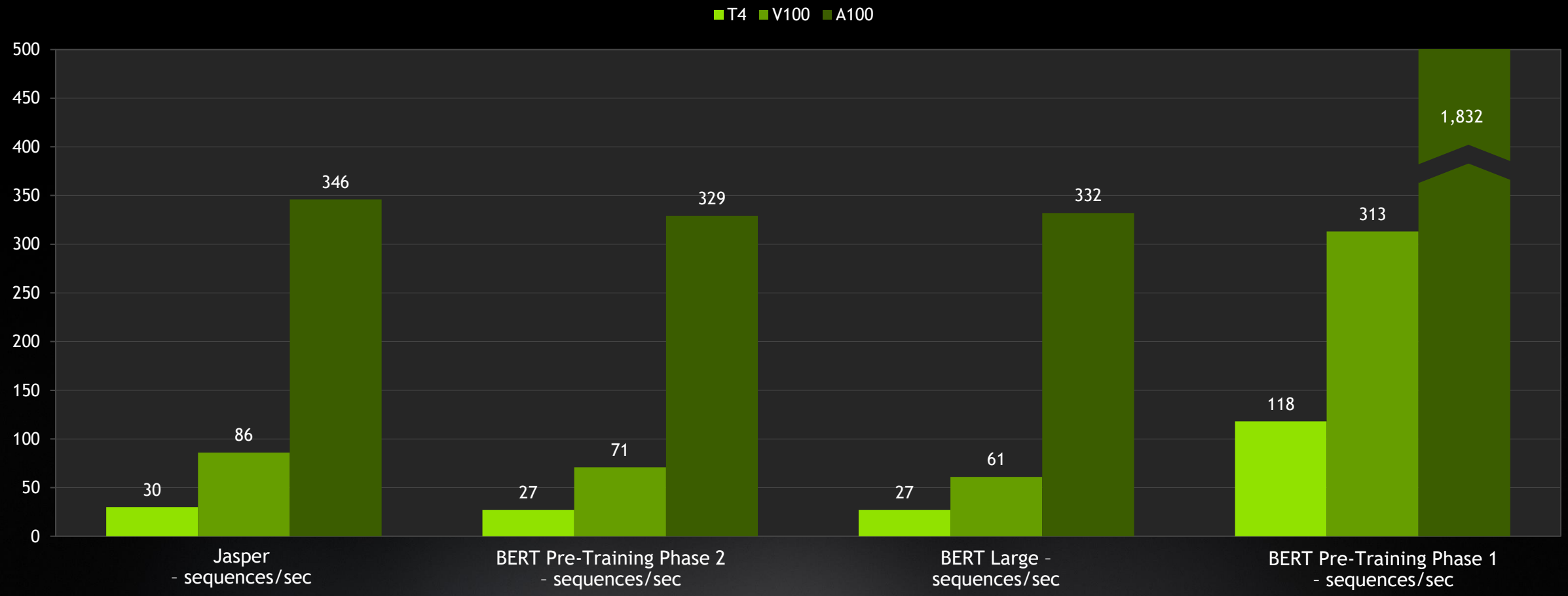

*All results are measured*

*V100 used is DGX-1 (8xV100 16GB). A100 used is s DGX A100 (8xA100 SXM4), except DLRM which uses 1xV100 and 1xA100; all use FP16*

*RN50 uses MXNET Batch size =192, Mask R CNN uses PyTorch BS = 4 (V100) and BS=16 (A100), DLRM uses PyTorch and BS=32768, Jasper uses PyTorch and BS=32 (V100) and 96 (A10), WaveGlow uses PyTorch and BS=10, TacoTron2 uses PyTorch and BS=104 (V100) and 100 (A100), Transformer uses PyTorch and BS=5120 (V100) and 13312 (A100 and GNMT uses PyTorch and BS=128 (V100) and 256 (A100); 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*

# UP TO 6X SPEEDUP WITH TF32 PRECISION FOR AI TRAINING

## Deep Learning Training Performance With A100 On PyTorch

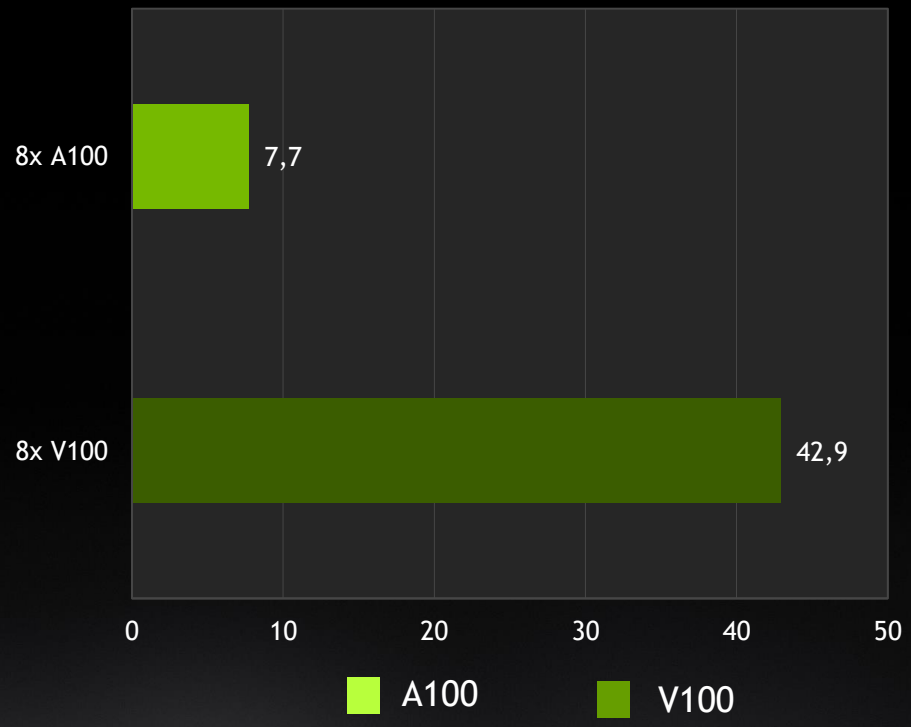



*GPU Server: Dual-Socket EPYC 7742@2.25GHz w/ 8x NVIDIA A100 SXM4, Dual-Socket Xeon E5-2698v4@2.2GHz w/ 8x NVIDIA V100 SXM2 (16GB), and Dual-Socket Xeon Gold 6240@2.6GHz w/ 8x T4*
*Frameworks: PyTorch v1.7.0a0+8deb4fe; Precision: TF32 for A100 and FP32 for v100 and T4; CUDA 11.0.221; NCCL 2.7.8; cuDNN 8.0.4; cuBLAS 11.2.0.252; DALI 0.25.1;*
*NVIDIA Driver: 450.51.06; Dataset: LibriSpeech for Jasper, SQuaD v1.1 for BERT Large, and Wikipedia+BookCorpus for BERT Pre-Training;*
*Batch sizes for Jasper: A100 = 32, V100 and T4 = 16; Batch sizes for BERT Large: A100 = 16, V100 and T4 = 4; Batch sizes for BERT Pre-Training Phase1: A100 = 54, V100 and T4 = 8;*
*Batch sizes for BERT Pre-Training Phase2: A100 = 8, V100 and T4 = 2. Sequence Length: BERT-Large = 384, BERT Pre-training Phase1 = 128, BERT Pre-training Phase2 = 512*

# DEEP LEARNING TRAINING TIME TO SOLUTION

PyTorch: DLRM Time to Solution on FP32 Precision

| GPU | Time to Solution |
|---|---|
| 8x A100 | 7.7 Minutes |
| 8x V100 | 42.9 Minutes |




Time to Train in *Minutes* – LOWER is Better

**Time to Solution Matters**

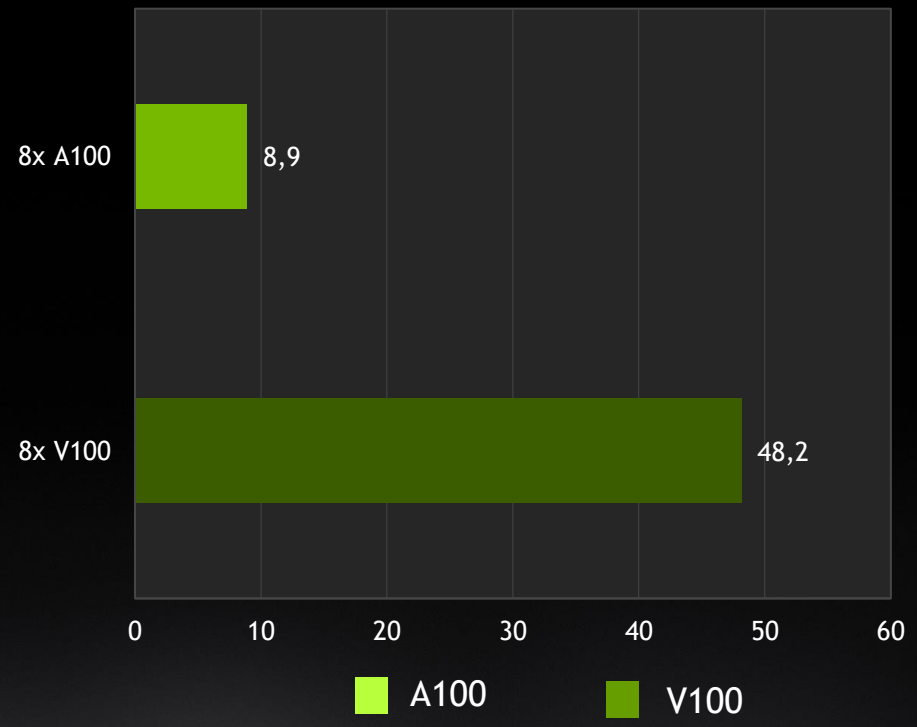
When training a neural network speed is important, but the network needs to converge to a required accuracy to be deployed for inferencing. If the network won't converge, then throughput rate alone isn't useful.

*GPU Server: Dual-Socket EPYC 7742@2.25GHz w/ 4x or 8x NVIDIA A100 SXM4, Dual-Socket Xeon E5-2698 v4@2.2GHz w/ 8x or 16x NVIDIA V100 SXM2 (32GB)*
*Frameworks: PyTorch v1.7.0a0+8deb4fe; Precision: A100 = TF32, V100 = FP32; CUDA 11.0.221; NCCL 2.7.8; cuDNN 8.0.2; cuBLAS 11.2.0.252; DALI 0.24; NVIDIA Driver: 450.51.06;*
*Dataset: Criteo Terabyte Dataset; Batch sizes: 65536 for A100 and V100; DLRM is showing 20.08 container data*

# DEEP LEARNING TRAINING TIME TO SOLUTION

PyTorch: BERT-Large Fine Tuning Time to Solution on FP32 Precision

| GPU | Time to Solution |
|---|---|
| 4x A100 | 17.7 Minutes |
| 8x A100 | 8.9 Minutes |
| 4x V100 | 94.1 Minutes |
| 8x V100 | 48.2 Minutes |




Time to Train in *Minutes* – LOWER is Better

**Time to Solution Matters**

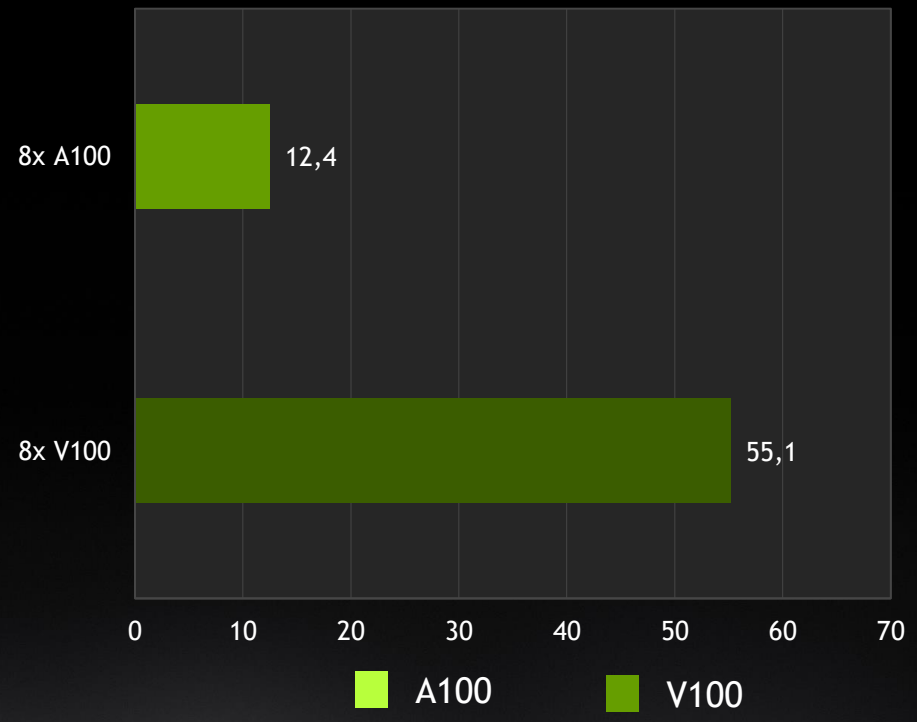
When training a neural network speed is important, but the network needs to converge to a required accuracy to be deployed for inferencing. If the network won't converge, then throughput rate alone isn't useful.

*GPU Server: Dual-Socket EPYC 7742@2.25GHz w/ 4x or 8x NVIDIA A100 SXM4, Dual-Socket Xeon E5-2698 v4@2.2GHz w/ 8x or 16x NVIDIA V100 SXM2 (16GB)*
*Frameworks: PyTorch v1.7.0a0+8deb4fe; Precision: A100 = TF32, V100 = FP32; CUDA 11.0.221; NCCL 2.7.8; cuDNN 8.0.4; cuBLAS 11.2.0.252; DALI 0.25.1; NVIDIA Driver: 450.51.06;*
*Dataset: SQuaD v1.1; Batch sizes: 32 for A100, 10 for V100 and T4; Sequence Length = 384*

# DEEP LEARNING TRAINING TIME TO SOLUTION

TensorFlow: BERT-Large Fine Tuning Time to Solution on FP32 Precision

| GPU | Time to Solution |
|---|---|
| 4x A100 | 18.8 Minutes |
| 8x A100 | 12.4 Minutes |
| 4x V100 | 101.4 Minutes |
| 8x V100 | 55.1 Minutes |




Time to Train in *Minutes* – LOWER is Better

**Time to Solution Matters**

When training a neural network speed is important, but the network needs to converge to a required accuracy to be deployed for inferencing. If the network won't converge, then throughput rate alone isn't useful.

*GPU Server: Dual-Socket EPYC 7742@2.25GHz w/ 4x or 8x NVIDIA A100 SXM4, Dual-Socket Xeon E5-2698 v4@2.2GHz w/ 8x or 16x NVIDIA V100 SXM2 (16GB)*
*Frameworks: TensorFlow v1.15.3; Precision: A100 = TF32, V100 = FP32; CUDA 11.0.221; NCCL 2.7.8; cuDNN 8.0.4; cuBLAS 11.2.0.252; DALI 0.25.1; NVIDIA Driver: 450.51.06;*
*Dataset: SQuaD v1.1; Batch sizes: 16 for A100 and 2 for V100; Sequence Length = 384. A100 is showing 20.08 container data and configuration*

# Deep Learning GPU Acceleration on PyTorch

## Training on BERT Large (Natural Language Processing)


*Server: Dual-Socket EPYC 7742@2.25GHz with A100 and Dual Xeon E5-2698v4@2.2GHz with V100. Framework: PyTorch v1.7.0a0+8deb4fe; Mixed Precision; CUDA 11.0.221; NCCL 2.7.8; cuDNN 8.0.4; cuBLAS 11.2.0.252; DALI 0.25.1; NVIDIA Driver: 450.51.06; Batch size: 32 for A100 and 10 for V100; Sequence Length = 384*

PyTorch

## PyTorch
### Deep Learning Training

PyTorch is a deep learning framework that puts Python first.

**VERSION**
1.7.0a0+8deb4fe

**ACCELERATED FEATURES**
Full framework accelerated

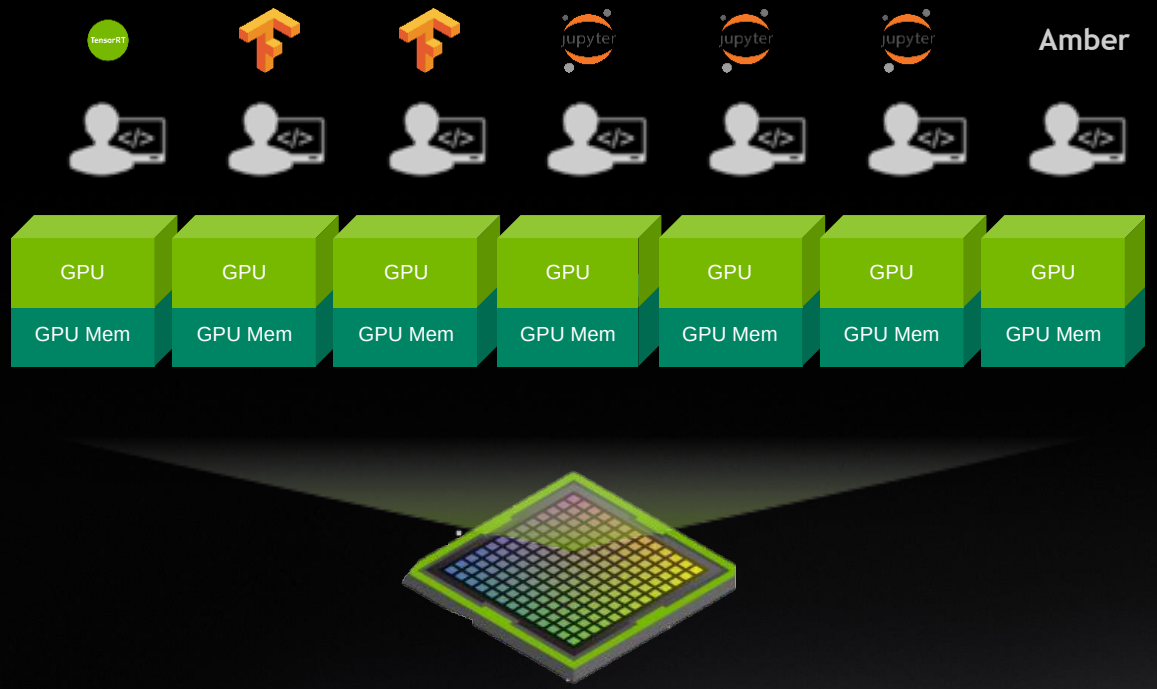
**SCALABILITY**
Multi-GPU, multi-node

**More Information**
www.pytorch.org
PyTorch on NGC


Deep Learning
Performance
INFERENCE

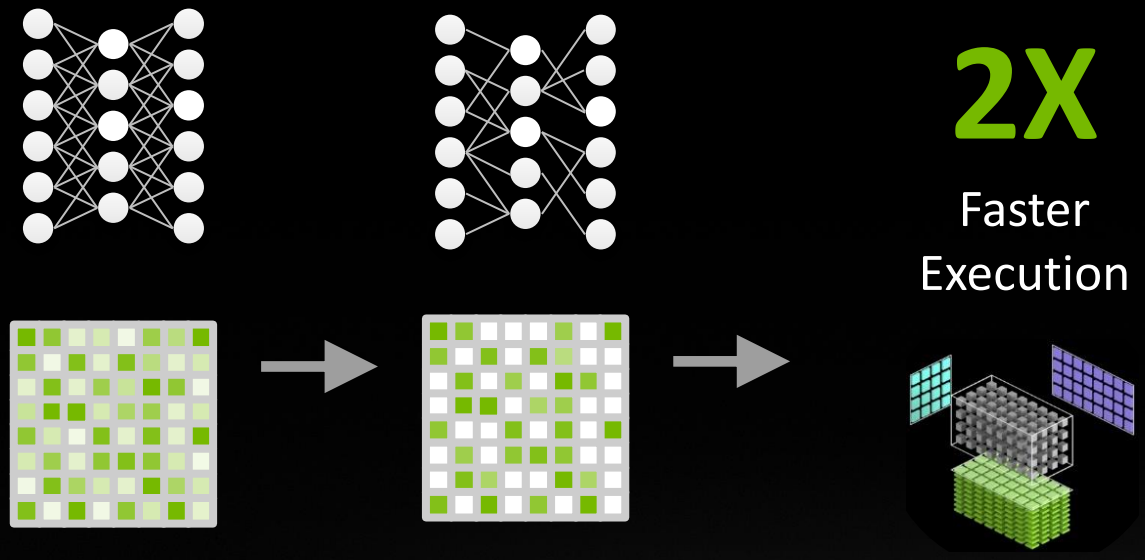
# NEW MULTI-INSTANCE GPU (MIG)

## 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.

# STRUCTURAL SPARSITY BRINGS ADDITIONAL SPEEDUPS



Dense Matrix

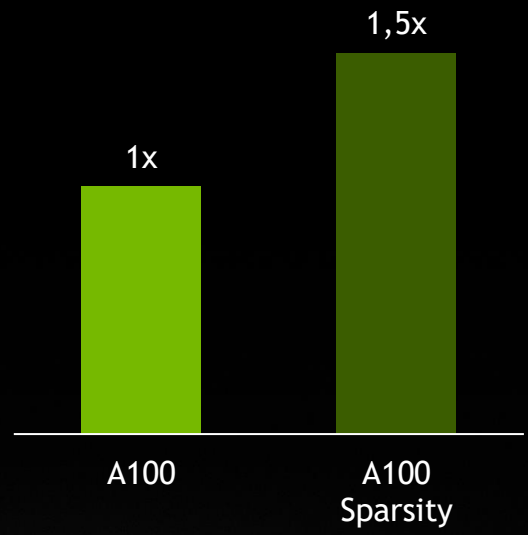
Sparse Matrix

**2X**

Faster Execution

A100 Tensor Core

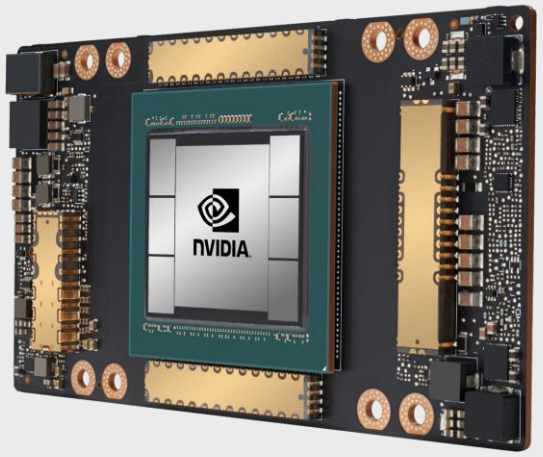
BERT Large Inference



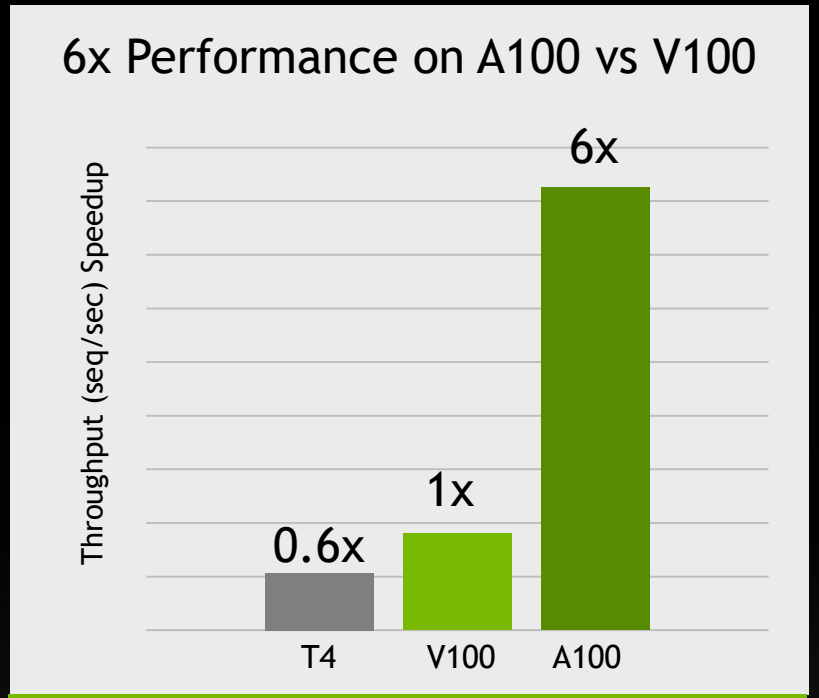

- Structured sparsity: Half the values are zero
- Skip half of the compute and mem fetches
- Compute up to 2x rate vs non-sparse

BERT Large Inference | precision = INT8 with and without sparsity | Batch sizes - no sparsity: bs256, with sparsity: bs49, A100 with 7 MIGS

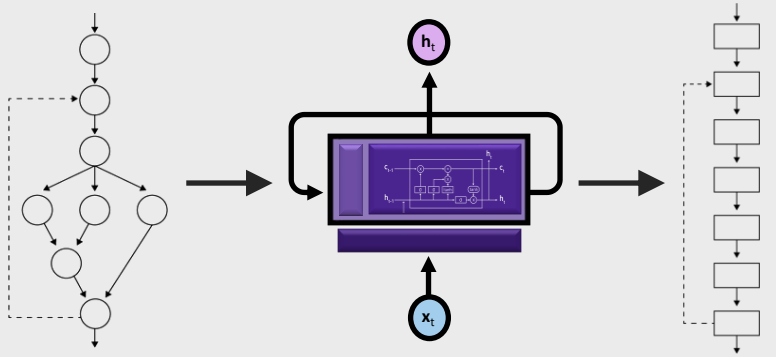
# ANNOUNCING TENSORRT 7.1




Tuned for A100 GPUs

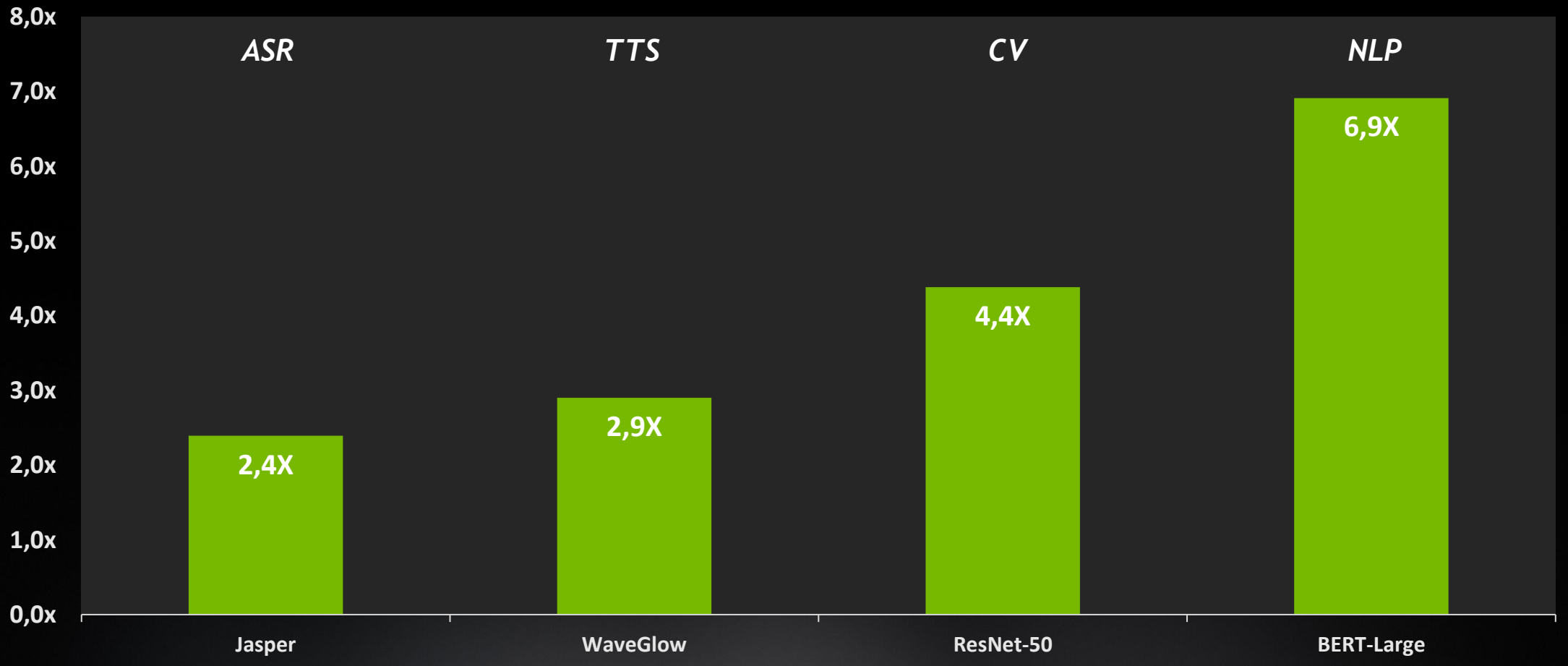



New BERT-Large
INT8 optimizations



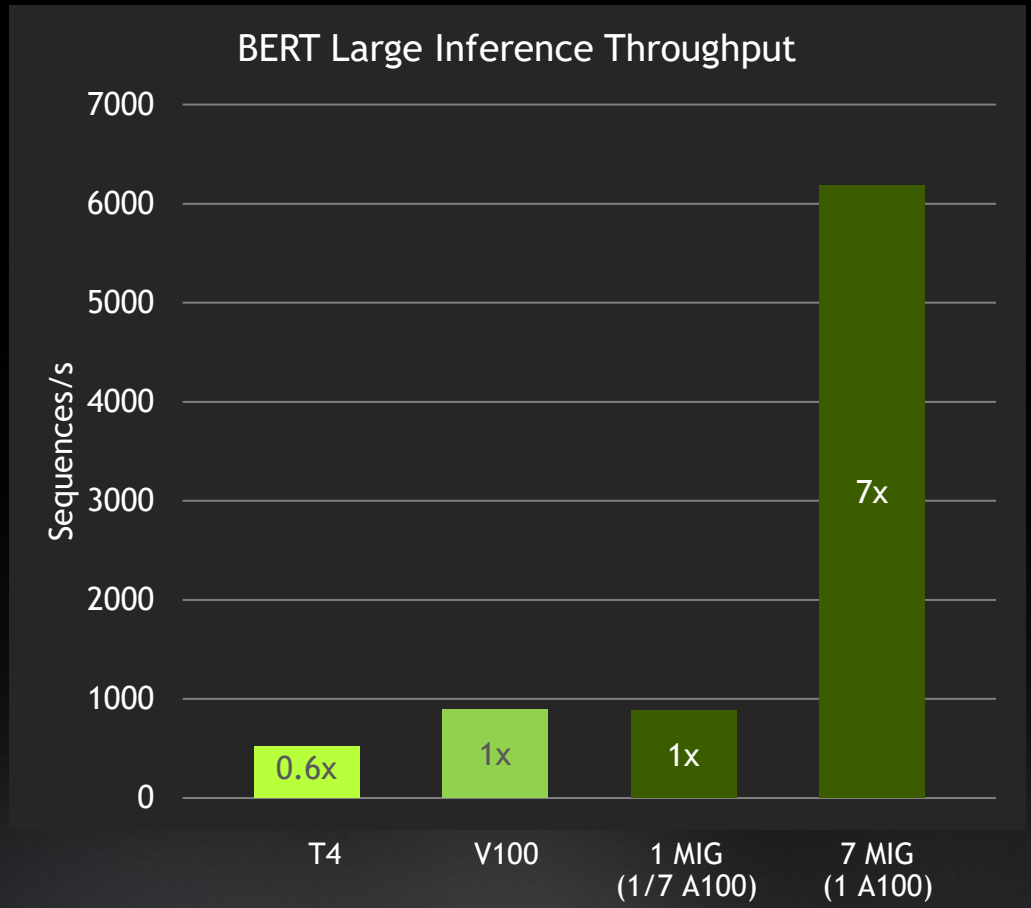
Compiler Supports RNNs,
Transformers and CNNs

# A100 DELIVERS UP TO 7X MORE INFERENCE PERFORMANCE




*Jasper: batch size 1, sequence length 5.12s, precision FP16, BERT-Large: batch size 1, sequence length 128, V100: FP16 precision, A100: INT8 precision*
*WaveGlow: batch size 1, ResNet-50: batch size 128, precision FP16. System configuration: Dual Xeon Platinum 8174 CPUs, 1.5 TB system memory*

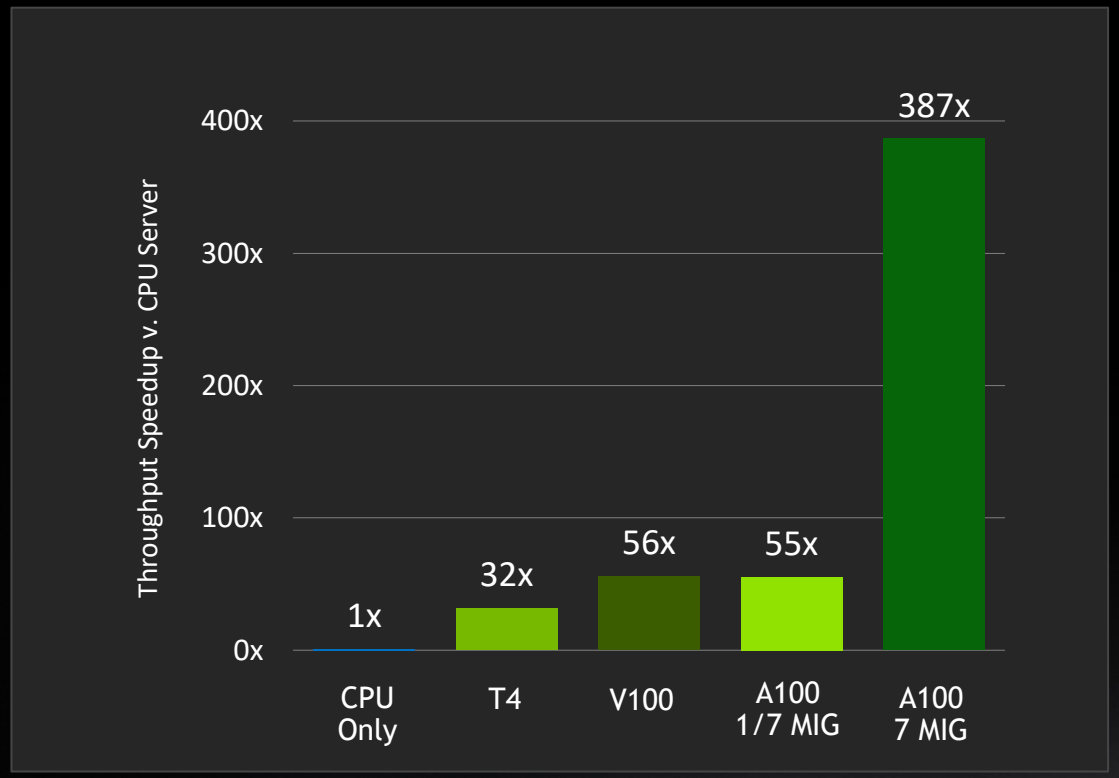
# A100 BRINGS 7X HIGHER INFERENCE THROUGHPUT WITH MIG



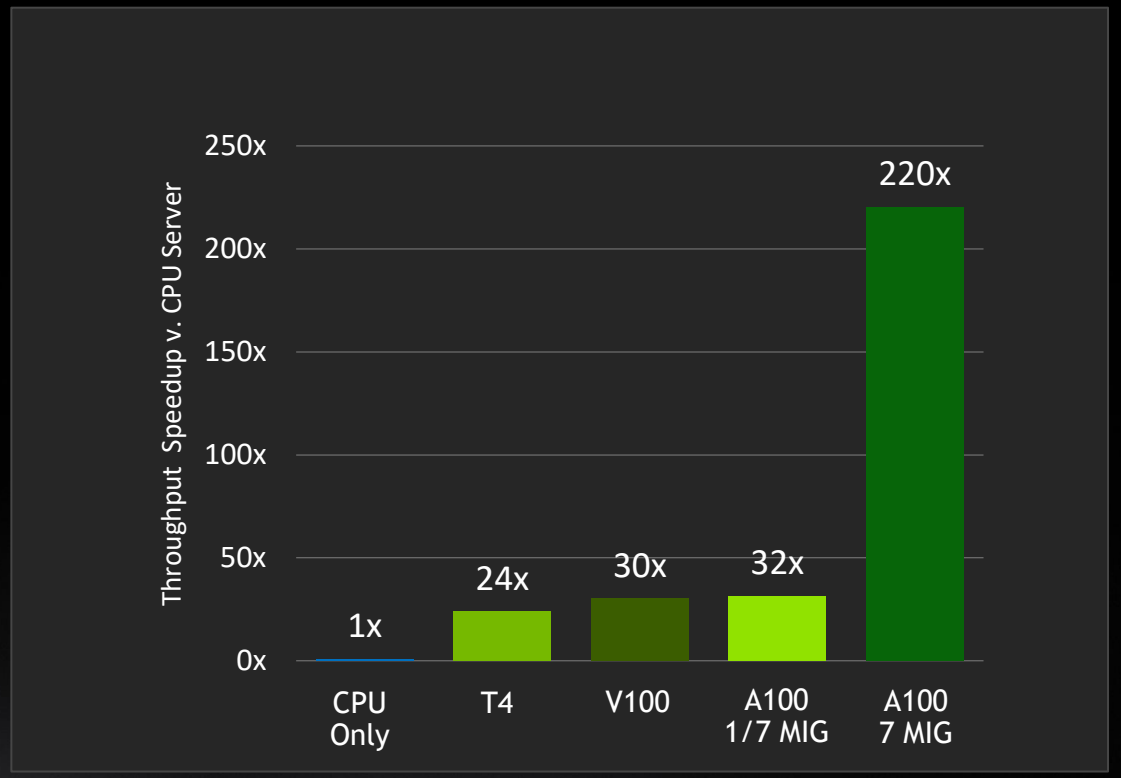

*BERT Large Inference | T4: TRT 7.1, Precision = INT8, Batch Size =256, V100: TRT 7.1, Precision = FP16, Batch Size =256 | A100 with 7 MIG instances of 1g.5gb : Pre-production TRT, Batch Size =94, Precision = INT8 with Sparsity*

# UP TO 387X INFERENCE THROUGHPUT SPEEDUP WITH MIG

## Natural Language Processing Massive Throughput




LANGUAGE PERFORMANCE
Max Throughput
BERT-Large (sequences/sec)



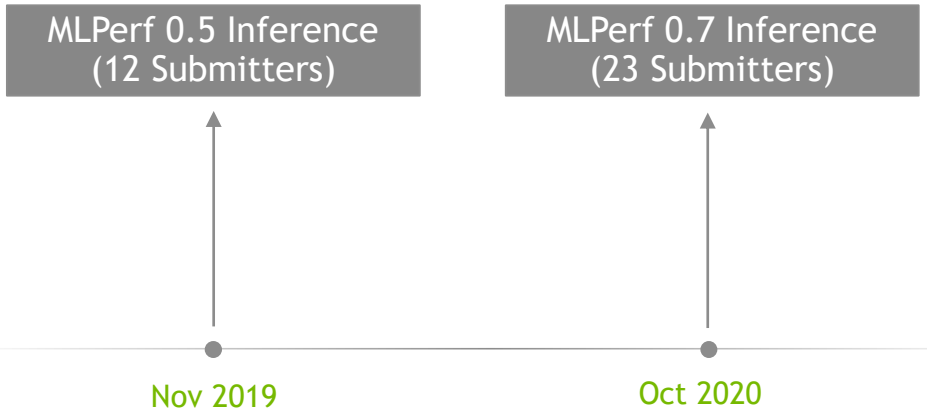

LANGUAGE PERFORMANCE
Throughput at Batch Size 1
BERT-Large (sequences/sec)

*CPU Server: Dual-Socket Xeon Platinum 8280@2.70GHz using OpenVINO Toolkit;*
*GPU Server: EPYC 7742@2.25GHz with a single A100, Xeon Gold 6240@2.60GHz with a single T4 or Dual-socket Xeon Platinum 8168@2.70GHz with a single V100;*
*TensorRT 7.1; Pre-Release Container; Batch-size: 256 for max throughput and 128 for CPU Only;*
*Sequence length =128; CPU data is measured in FP32 which is the best performing precision for a standalone CPU. Precision: T4 = INT8, V100 = FP16, A100 = INT8 with Sparsity*

# Deep Learning Performance

## MLPERF 0.7 - INFERENCE

# INDUSTRY-WIDE BENCHMARK SUITE FOR AI PERFORMANCE




MLPerf

Fair and useful benchmarks for measuring training and inference performance of ML hardware, software, and services.

## COMPANIES

AI Labs.tw 台灣人工智慧實驗室 · Alibaba Group 阿里巴巴集团 · AMD · ANDES TECHNOLOGY · AONdevices · arm · Automaton AI · Baidu 百度 · 北京智源人工智能研究院 BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE · cadence · CALYPSO · CenTaur technology · cerebras · CEVA · CIRRUS · cisco · CODE REEF · CRAY · criteo · Tuning Foundation · DELL EMC · dv/dt · DDN STORAGE · Edgify · Enflame · Esperanto TECHNOLOGIES · facebook · FURIOSA · Google · groq · habana · Hewlett Packard Enterprise · Hop · 地平线 Horizon Robotics · 天数智芯 Iluvatar CoreX · inspur · intel AI · IQT IN-Q-TEL · Lanner · Lenovo · MEDIATEK · Mentor Graphics A Siemens Business · Microsoft · myrtle.ai · MYTHIC · NetApp · NVIDIA · One Convergence · oppo · PATHPARTNER · PURESTORAGE · Qualcomm · rpa2ai Automation & AI Advisors · SambaNova SYSTEMS · SAMSUNG Exynos · SIGOPT · SiMa.ai · skymizer · SUPERMICRO · SYNOPSYS · Tencent 腾讯 · TENSYR driving deployment · TERADYNE · Transpire Ventures · trustworthy ai · verifAI · VMind · vmware · volley.com · WAVE COMPUTING · wiwynn · WEKA.IO World's Fastest File System · XILINX

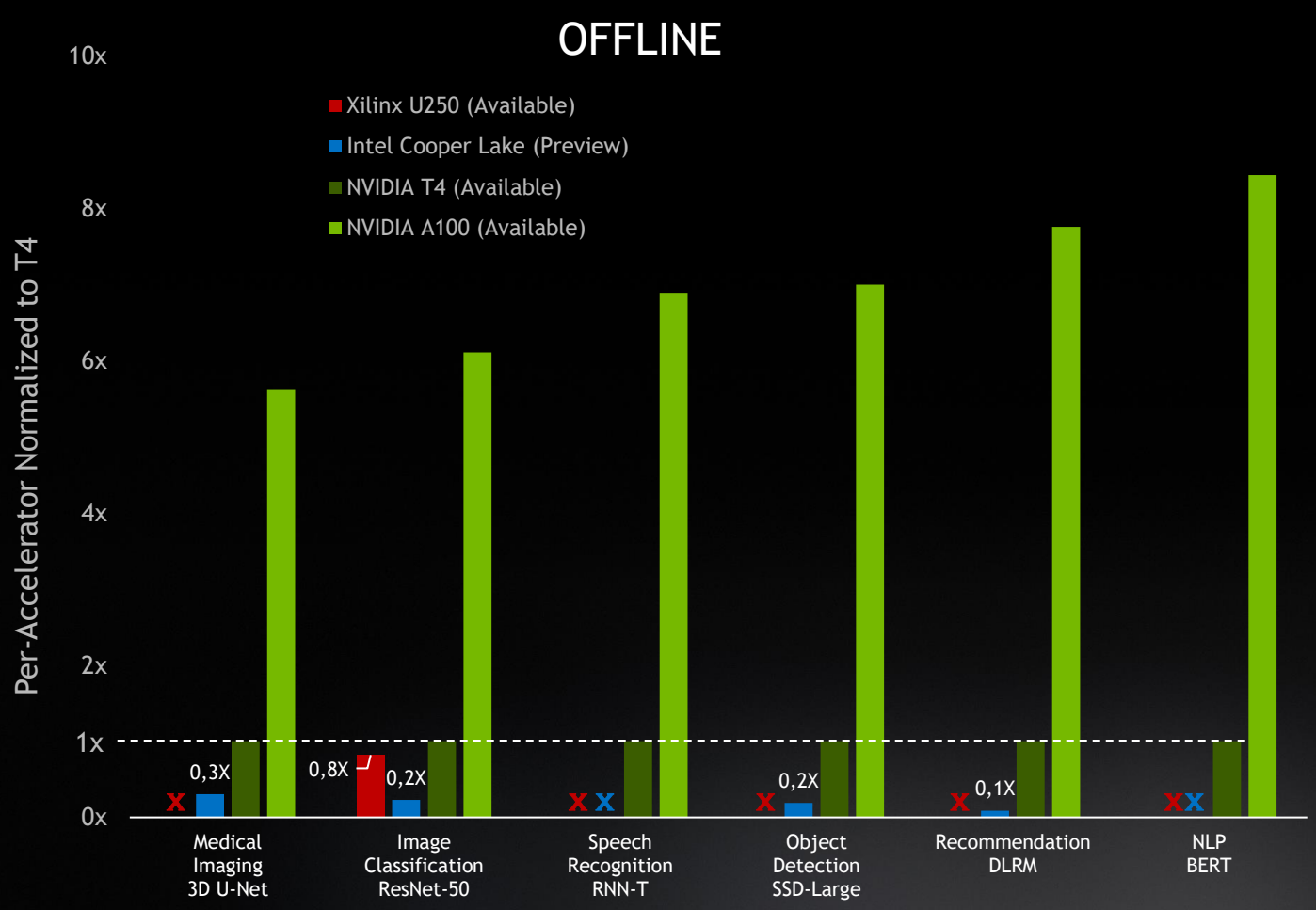
## RESEARCHERS

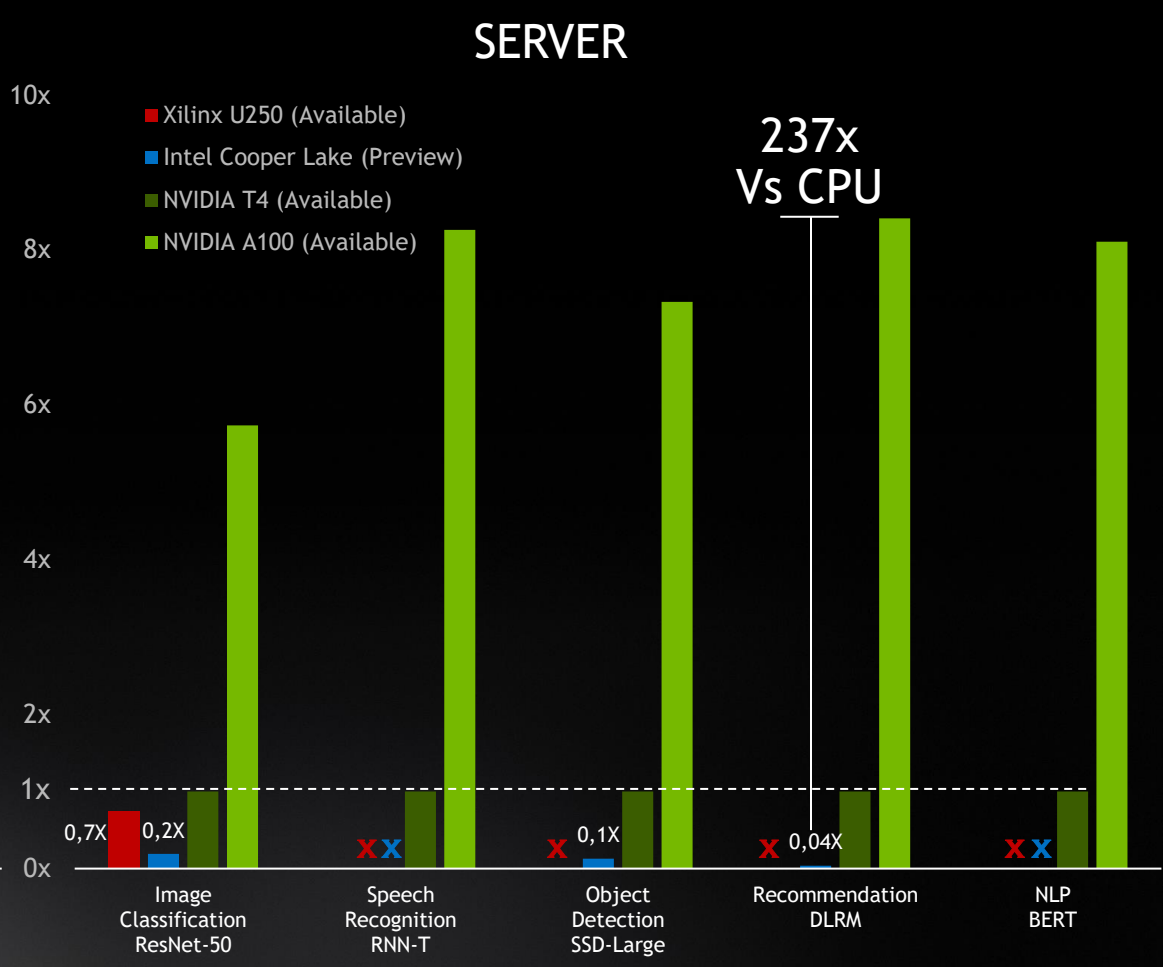
HARVARD School of Engineering and Applied Sciences · Stanford ENGINEERING · "El saber de mis hijos hará mi grandeza" · UA LITTLE ROCK · Berkeley UNIVERSITY OF CALIFORNIA · UC SANTA CRUZ · ILLINOIS · UNIVERSITY OF MINNESOTA · The University of Texas at Austin Cockrell School of Engineering · UNIVERSITY OF TORONTO

https://mlperf.org/

# NVIDIA TOPS MLPERF DATA CENTER BENCHMARKS

## A100 Is Up To 237x Faster Than The CPU

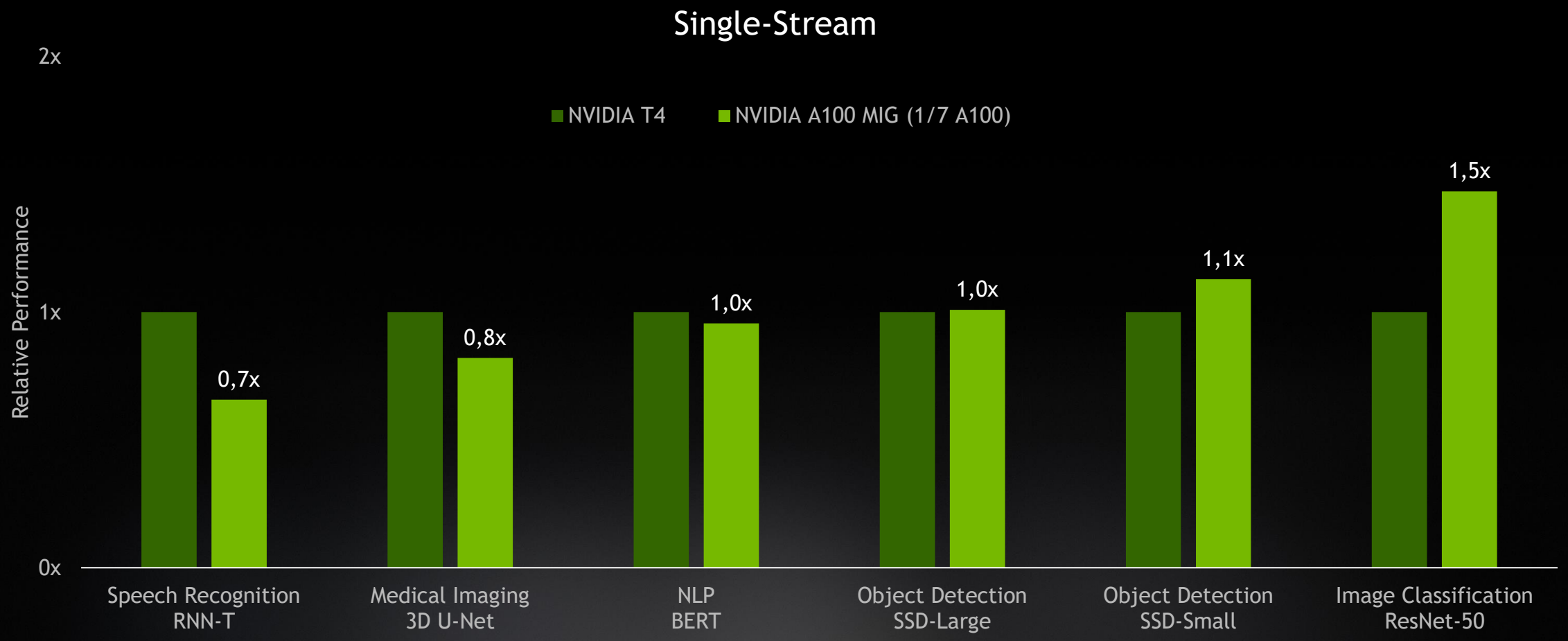


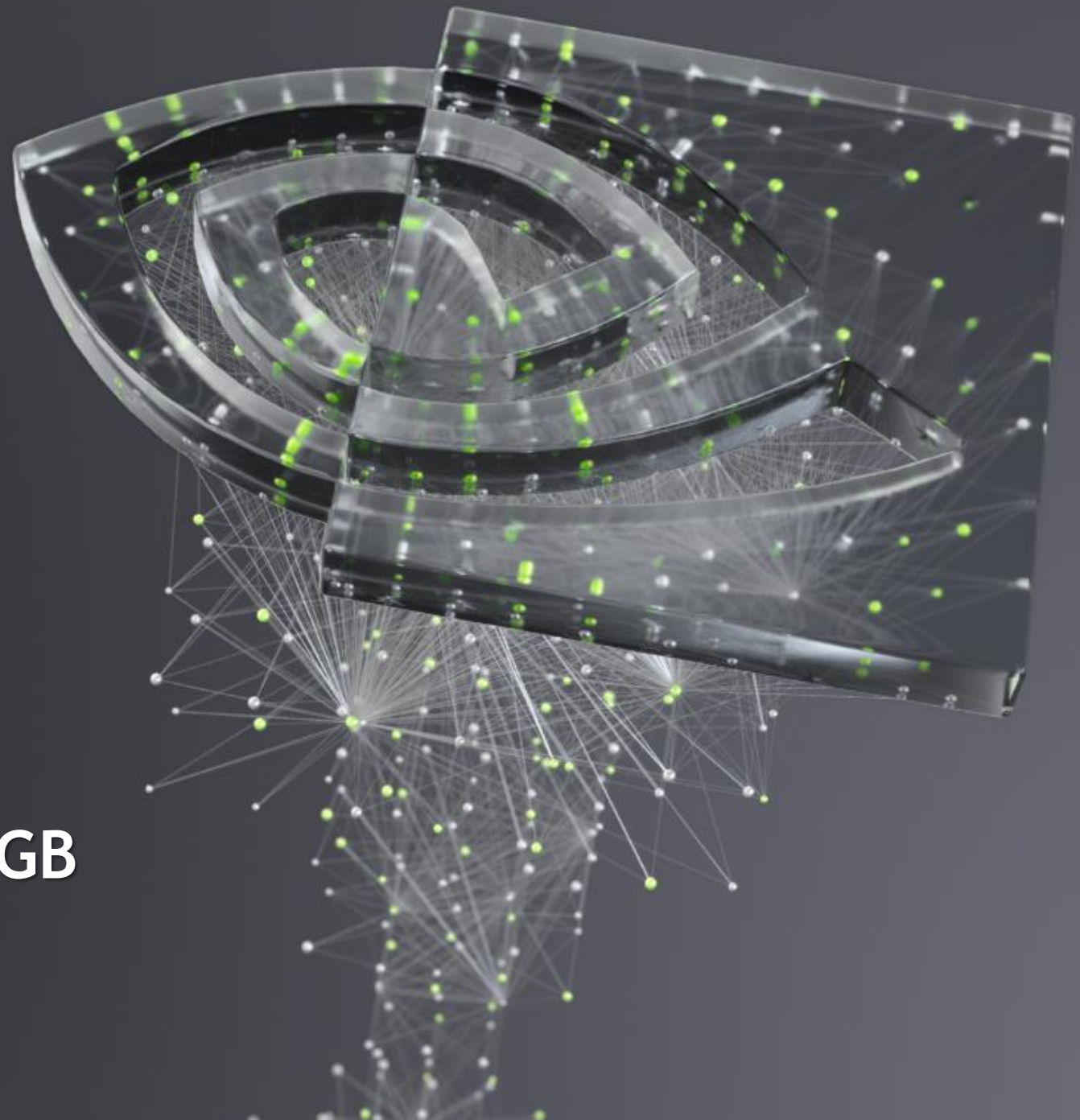

*MLPerf v0.7 Inference Closed; Per-accelerator performance derived from the best MLPerf results for respective submissions using reported accelerator count in Data Center Offline and Server. 3D U-Net 99.9%,: 0.7-125, 0.7-113, 0.7-111, ResNet-50: 0.7-119, 0.7-124, 0.7-113, 0.7-111, SSD-Large: 0.7-123, 0.7-113, 0.7-111 DLRM 99.9%: 0.7-126, 0.7-113, 0.7-111, RNN-T, BERT 99.9%: 0.7-111, 0.7-113 MLPerf name and logo are trademarks. See www.mlperf.org for more information.*

**X** = No result submitted

# MIG ENABLES 7 T4s IN AN A100




*MLPerf v0.7 Inference Closed; Edge Single-Stream results derived from reported latencies. RNN-T, 3D U-Net 99.9%, BERT 99%, SSD-Large, SSD-Small, ResNet-50: 0.7-153, 0.7-150.*
*MLPerf name and logo are trademarks. See www.mlperf.org for more information.*

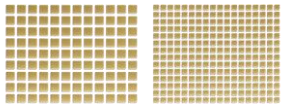


# UPDATE TESLA A100 / 80GB

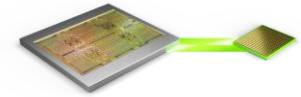
Ralph Hinsche

# ANNOUNCING NVIDIA A100 80GB

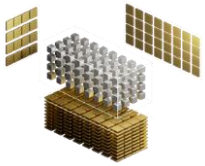
## Supercharging The World's Highest Performing AI Supercomputing GPU



80GB HBM2e
For largest datasets and models



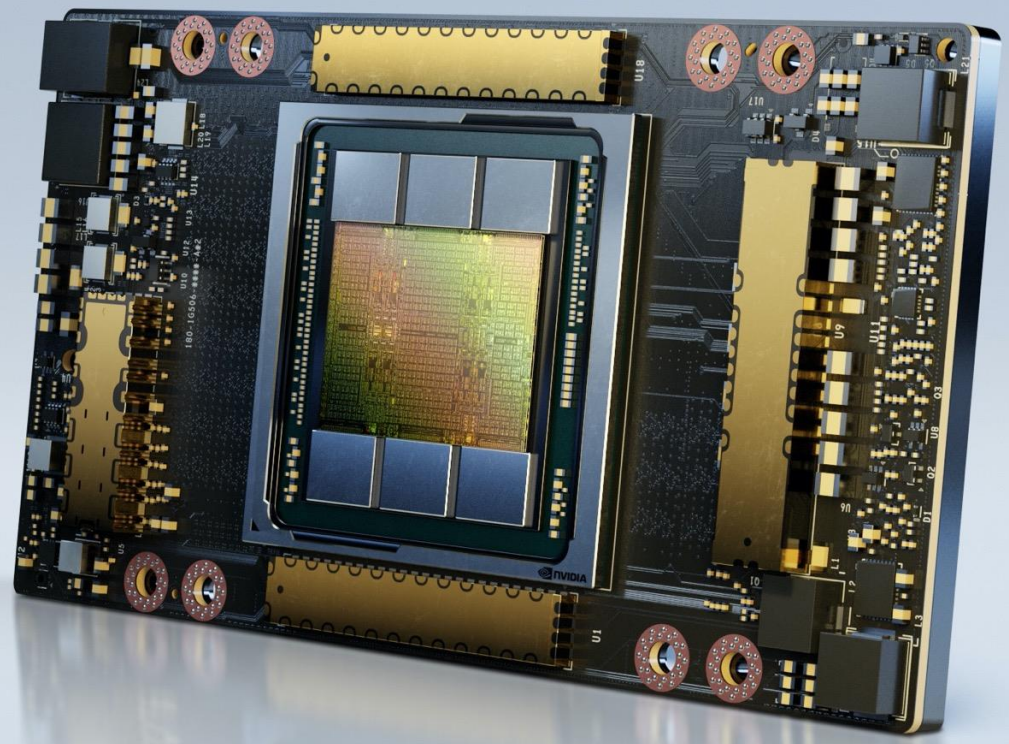
2TB/s +
World's highest memory bandwidth to feed the world's fastest GPU



3rd Gen Tensor Core

Multi-Instance GPU

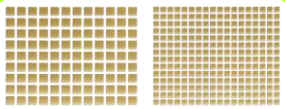
3rd Gen NVLink



# ANNOUNCING NVIDIA A100 80GB

## Supercharging The World's Highest Performing AI Supercomputing GPU

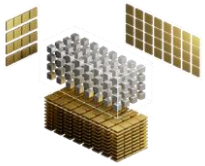
+100%



80GB HBM2e
For largest datasets and models

+25%

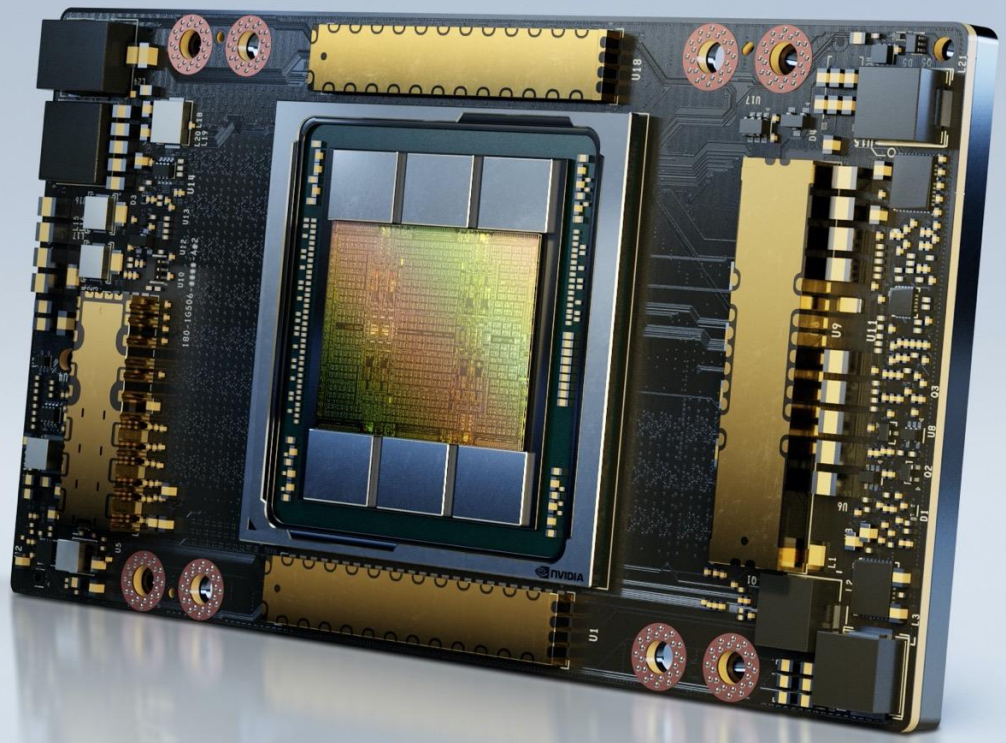
2TB/s +
World's highest memory bandwidth to feed the world's fastest GPU



3rd Gen Tensor Core

Multi-Instance GPU

3rd Gen NVLink



# SUPERCHARGED AI SUPERCOMPUTING WITH A100 80GB

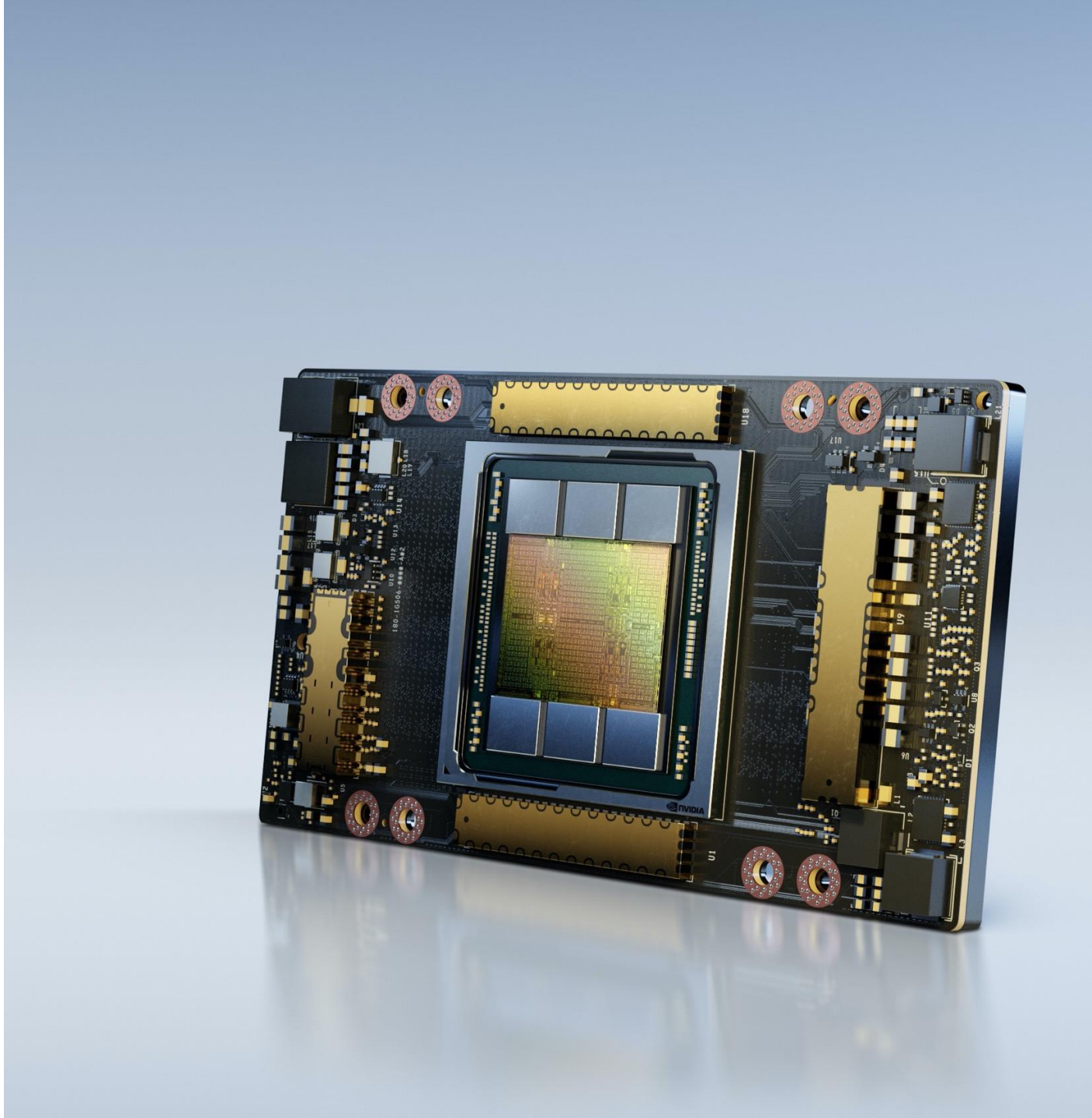
## World's Fastest GPU with World's Fastest Memory

A100 80GB Throughput vs A100 40GB

| 2X | 2X | 3X |
|---|---|---|
| **Simulation** | **Big Data Analytics** | **AI Training** |
| Quantum Espresso | 10 TB Retail Benchmark | DLRM Recommender |

| 1.25X | 1.25X |
|---|---|
| **MIG Inference** | **Energy Efficiency** |
| RNN-T Speech Recognition | Shatters 25 GF/W |

*Speedups Normalized to Number of GPUs | Comparisons A100 80GB to A100 40GB | Measurements performed DGX A100 servers | AI Training running DLRM using Huge CTR framework on a 450 GB Criteo dataset. Normalized speedup ~2.6X | Data Analytics: big data benchmark with 10TB dataset, 30 analytical retail queries, ETL, ML, NLP. Normalized speedup ~1.9X | HPC: Quantum Espresso with CNT10POR8 dataset on a 1.6TB dataset. Normalized speedup ~1.8X | AI Inference: RNN-T Single stream latency on A100 80GB on 1MIG@10GB when configured for 7MIGs. Normalized speedup ~1.25X*



# WHAT IS A PETAFLOPS ?

## “PETA”

1 Peta = $10^{15}$ = 1 Quadrillion = 1.000 Trillion

1 Trillion (AE) = $10^{12}$ = 1.000 Billion (AE)

## WORLD POPULATION

Current World population (Status 2020) :

ca. 7,8 Billion

## 1 PETAFLOPS

- Every person in the world gets his own hand-held calculator
- Enters 125.000 calculations on this pocket calculator, EVERY SECOND

## TESLA A100 (1X GPU)

625 TF = 0,625 PetaFLOP (FP16 Tensor/Sparse)
→ 75.000 per Second

312 TF = 0,312 PetaFLOP (TF32 Tensor/Sparse)
→ 39.000 per Second

# MEMORY BANDWIDTH

## PCI-E GEN 4.0

Theoretical Bandwidth (x16) : 32 GB/s
(Burstrate incl. Protocol-Overhead)

In practice (sustained) : 20 GB/s

## TESLA A100 (80 GB)

2,0 TB/s = 2.000 GB/s (HBM2e Bandwidth)

➞ 100x faster than PCI-E

(4K Film/2hours = 100GB, 20 Films/s)

## MEMORY-FILL

To fully fill 40GB of a Tesla A100/40GB you would require (via PCI-E) :

2 sec

## TIME LOST …

In these 2 sec you could have done :

1.250.000.000.000.000 (2 * 625 TF)

Calculations instead of waiting

# NEW DGX A100 640GB SYSTEM

## For the Largest AI Workloads

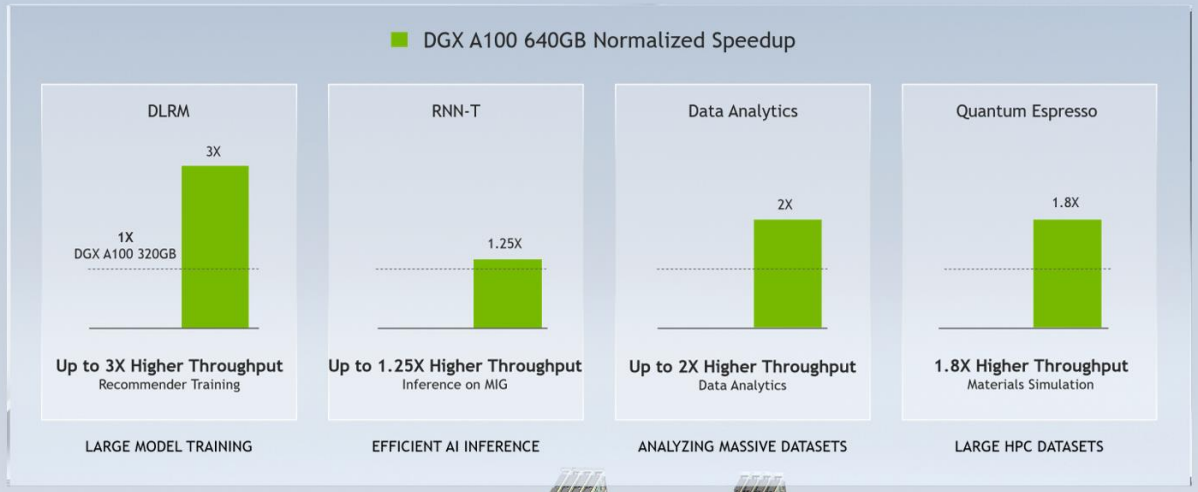
640 GB of GPU memory per system to increase model accuracy and reduce-time-to-solution

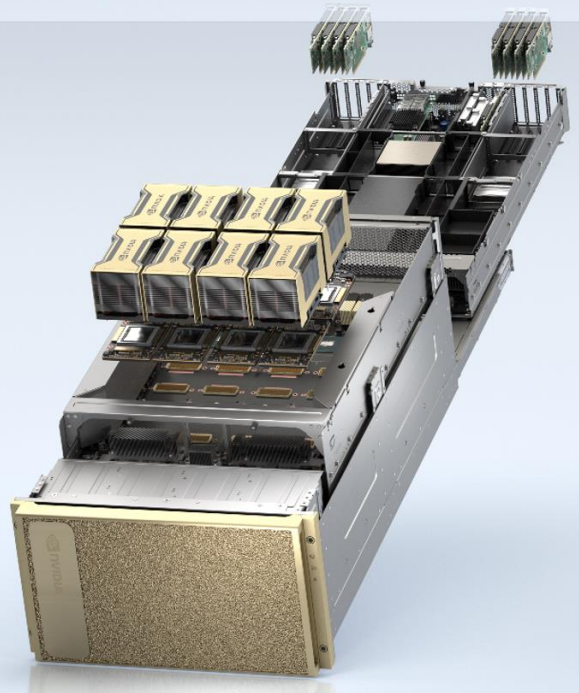
Up to 3X higher throughput for large-scale workloads

Double the GPU memory for MIG for more flexible AI development, analytics, and inference

Available individually, or part of DGX SuperPOD Solution for Enterprise

**Upgrade option for current DGX A100 customers**

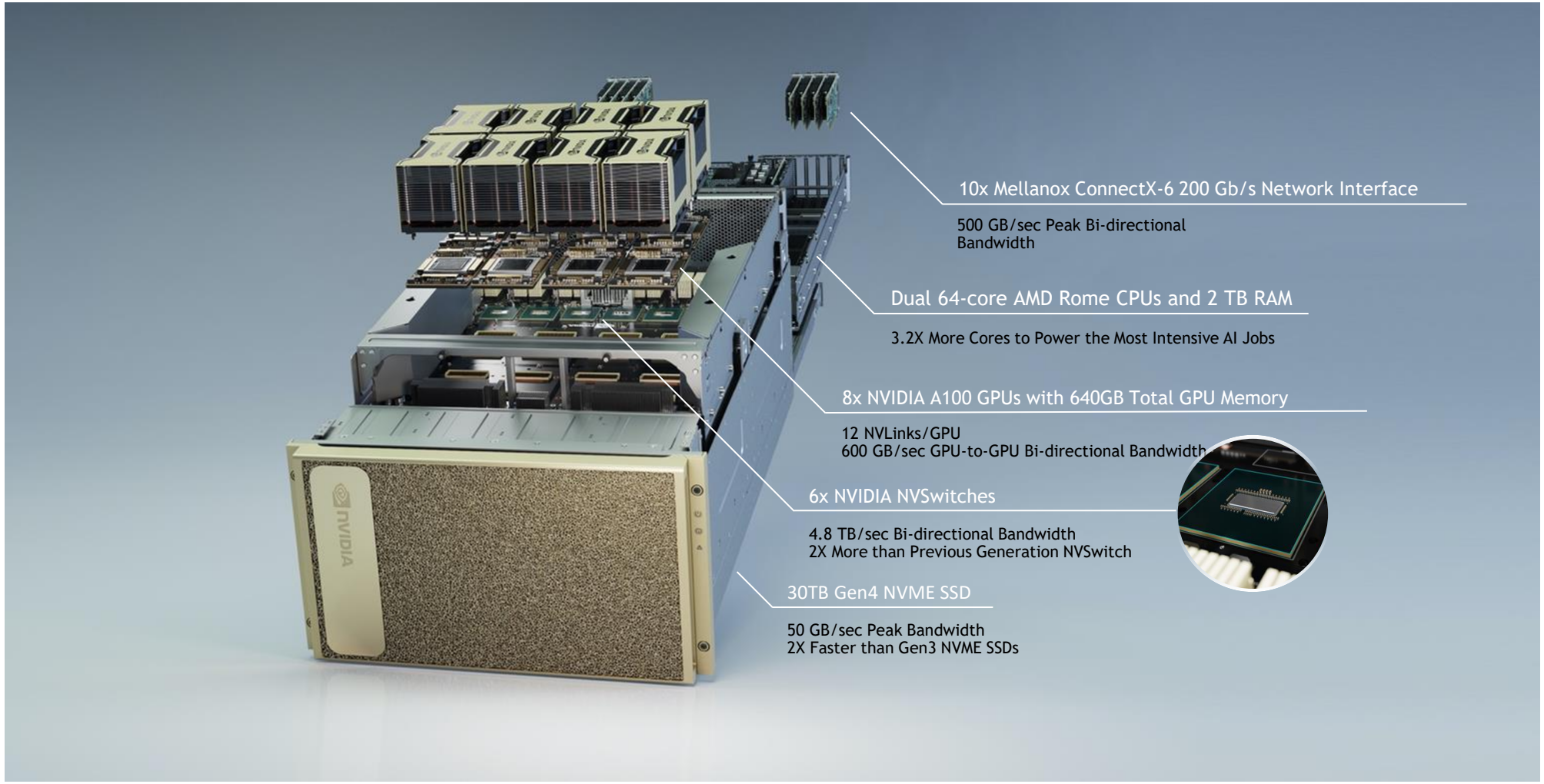


*Speedups Normalized to Number of GPUs | Comparisons to A100 40GB | Measurements performed DGX A100 servers . AI Training: DLRM (Huge CTR) | DGX A100: 16x A100 40GB vs 8x A100 80GB | speedup = 1.4X. Speedup normalized to number of GPUs = 2.8X. AI Inference: RNN-T (MLPerf 0.7 Single stream latency) | DGX A100: A100 40GB vs A100 80GB on 1MIG@10GB when configured for 7MIGs | Data Analytics: big data benchmark with RAPIDS(0.16), BlazingSQL(0.16), DASK(2.2.0) | 30 analytical retail queries, ETL, ML, NLP | 96x A100 40GB vs 48x A100 80GB | HPC: Quantum Espresso - CNT10POR8 40x A100 40GB vs 24x A100 80GB | Speedup normalized to number of GPUs = 1.8X*

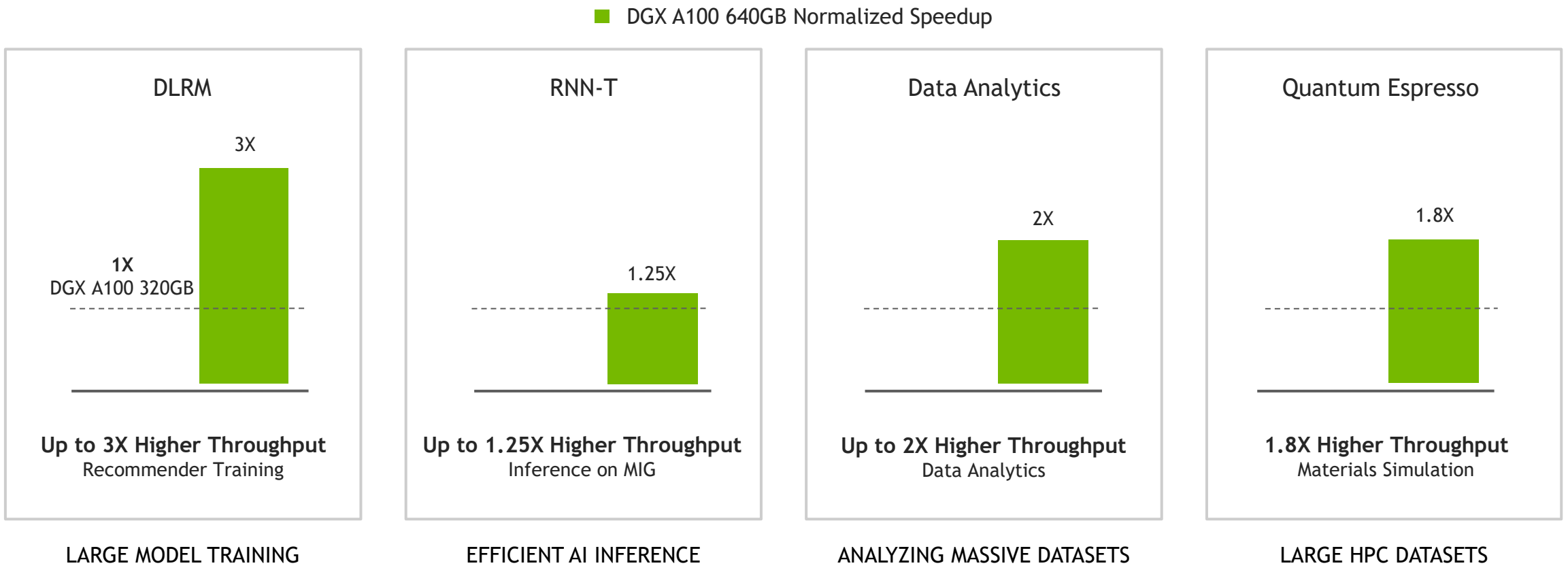
# GAME-CHANGING PERFORMANCE FOR INNOVATORS

## NVIDIA DGX A100 640GB System



# DGX A100 640GB ACCELERATES THE LARGEST WORKLOADS

## Up to 3X Faster on the Largest Models and Datasets




*Speedups Normalized to Number of GPUs | Comparisons to A100 40GB | Measurements performed DGX A100 servers. AI Training: DLRM (Huge CTR) | DGX A100: 16x A100 40GB vs 8x A100 80GB | speedup = 1.4X. | Speedup normalized to number of GPUs = 2.8X. AI Inference: RNN-T (MLPerf 0.7 Single stream latency) | DGX A100: A100 40GB vs A100 80GB on 1MIG@10GB when configured for 7MIGs | Data Analytics: big data benchmark with RAPIDS(0.16), BlazingSQL(0.16), DASK(2.2.0) | 30 analytical retail queries, ETL, ML, NLP | 96x A100 40GB vs 48x A100 80GB | HPC: Quantum Espresso - CNT10POR8 | 40x A100 40GB vs 24x A100 80GB | Speedup normalized to number of GPUs = 1.8X*


NVIDIA SELENE
Now Featuring NVIDIA DGX A100 640GB
4,480 A100 GPUs
560 DGX A100 system
850 Mellanox 200G HDR switches
14 PB of high-performance storage
2.8 EFLOPS of AI peak performance
63 PFLOPS HPL @ 24GF/W

# DGX STATION A100

## Workgroup Appliance for the Age of AI

AI Supercomputing for Data Science Teams

Data center performance without the data center

An AI appliance you can place anywhere

Bigger models, faster answers

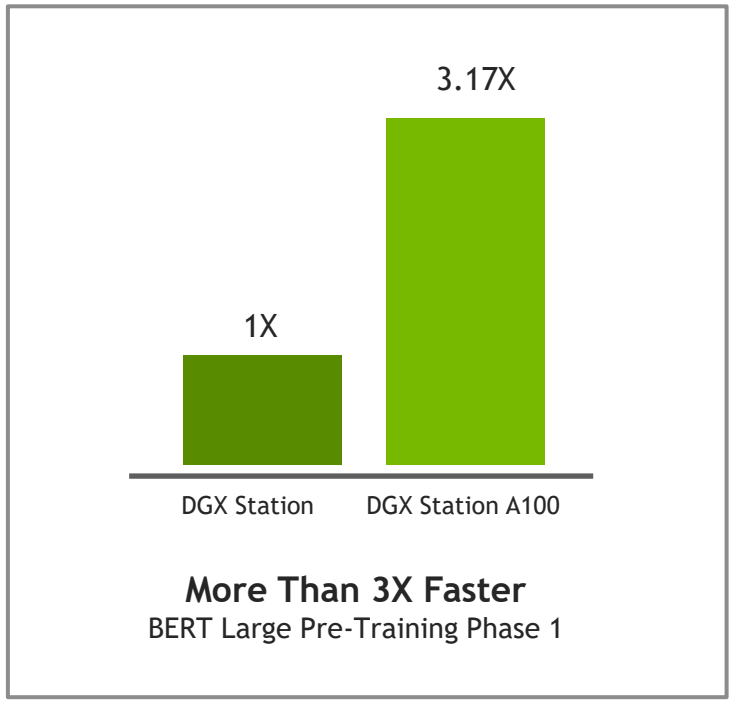

2.5 PFLOPS AI

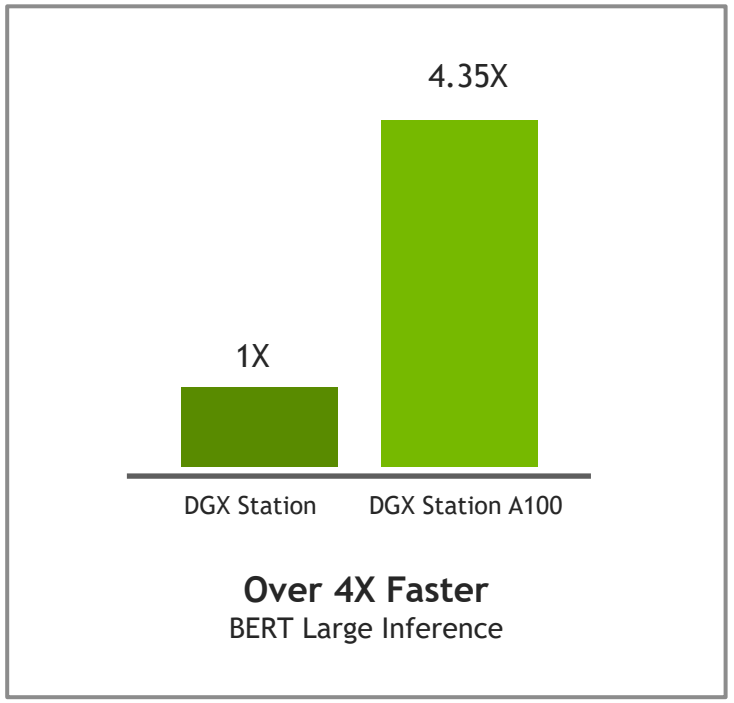
320 GB GPU MEMORY

Only workstation with 4-way NVLink and Multi-instance GPU (MIG)

# A DATA CENTER-IN-A-BOX

## DGX Station A100 is More Than 4X Faster

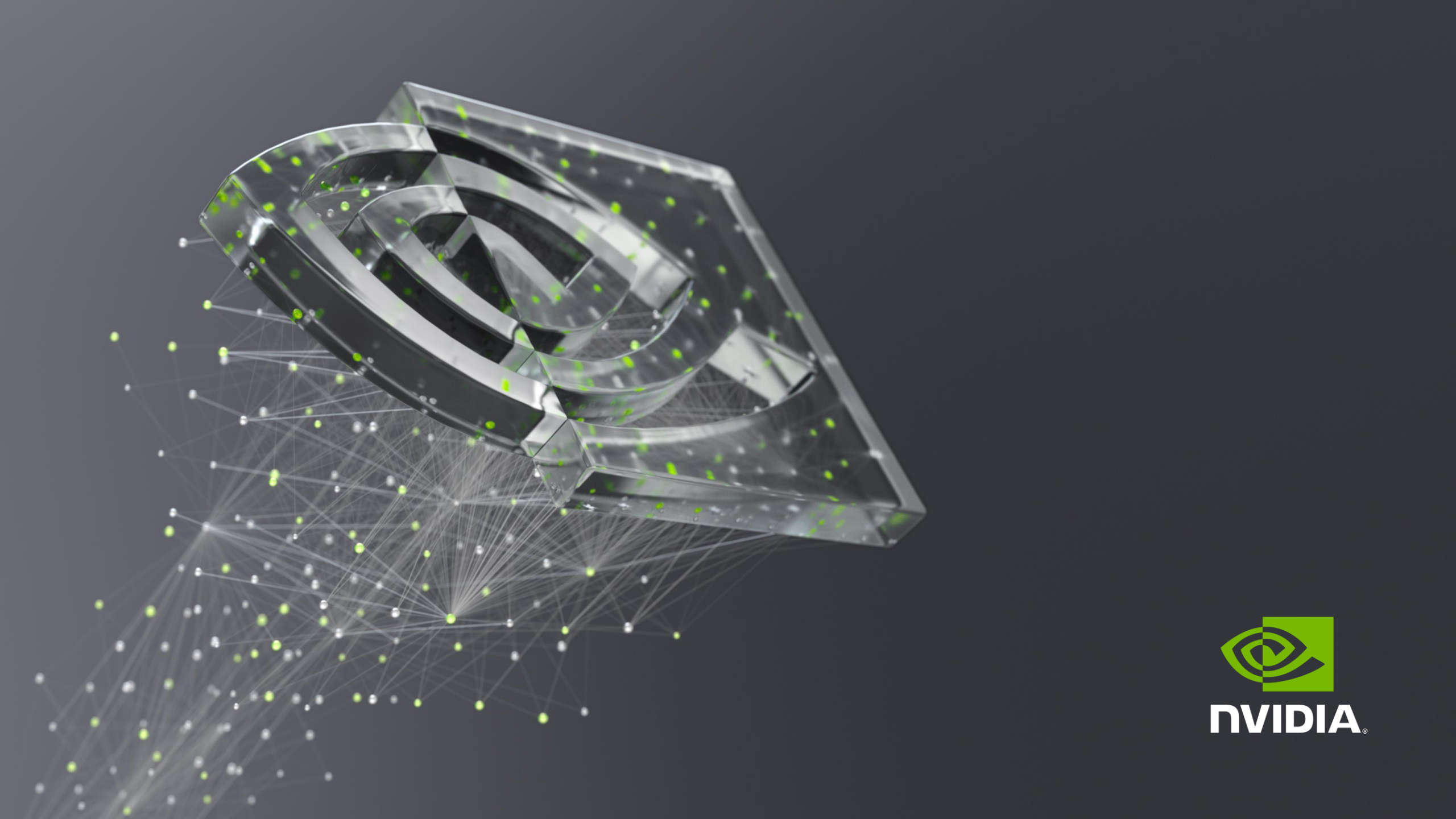




Training | Inference | HPC

*Training: Batch Size=64; Mixed Precision; With AMP; Real Data; Sequence Length=128*
*Inference: Batch Size=256; INT8 Precision; Synthetic Data; Sequence Length=128, cuDNN 8.0.4*
*HPC: FP32 Precision; Dataset/Input=Cellulose (h-bond) | Best value listed. Average is 1.5X across all these inputs: ADH Dodec (h-bond); Cellulose (h-bond); STMV (h-bond)*


NVIDIA