TECH TAKE

NVIDIA INSIGHTS: GPU COMPUTING LANDSCAPE FROM SOFTWARE TO HARDWARE.

Michael Lang

Solutions Architecture Manager – APAC South

June 29 2021

NVIDIA IS A FULL STACK COMPUTING PLATFORM

Amazing Innovation and Expansion of NVIDIA Ecosystem

BUSINESS APPLICATIONS

- Customer Engagement
- Patient Diagnostics
- Fraud Detection
- Quality Assurance
- Industrial Automation
- Precision Marketing
- Molecular Simulations
- ++

NGC

SOFTWARE HUB

HELM

Certified Containers

Pre-trained Models

SDKs

NVIDIA CERTIFIED

VALIDATED SOLUTIONS

APPLICATION FRAMEWORKS

- SMART CITY — *Metropolis*
- CONVERSATIONAL AI — *Jarvis*
- AUTONOMOUS VEHICLES — *Drive*
- RECOMMENDATION SYSTEMS — *Merlin*
- HEALTHCARE — *Clara*
- ++ ...

DEVELOPER TOOLKITS

- ML & DATA ANALYTICS — RAPIDS, dmlc XGBoost
- AI TRAINING & INFERENCE — TensorFlow, TensorRT, PYTORCH, mxnet
- HIGH PERFORMANCE COMPUTING — *NVIDIA HPC SDK*
- RENDERING & VISUALIZATION — *IndeX*, *OptiX*, *CloudXR*, *MDL*

ACCELERATION LIBRARIES

- COMPUTE — *CUDA-X*
- NETWORKING, STORAGE & SECURITY — *DOCA*, *MAGNUM IO*

SERVERS & CLOUD

- DGX HGX — Purpose Built
- EGX — Mainstream & Edge
- CSP Instances

HARDWARE TECHNOLOGIES

- GPU
- NVSwitch
- BF DPU
- SMART NIC
- NVIDIA Switch

OPERATIONS

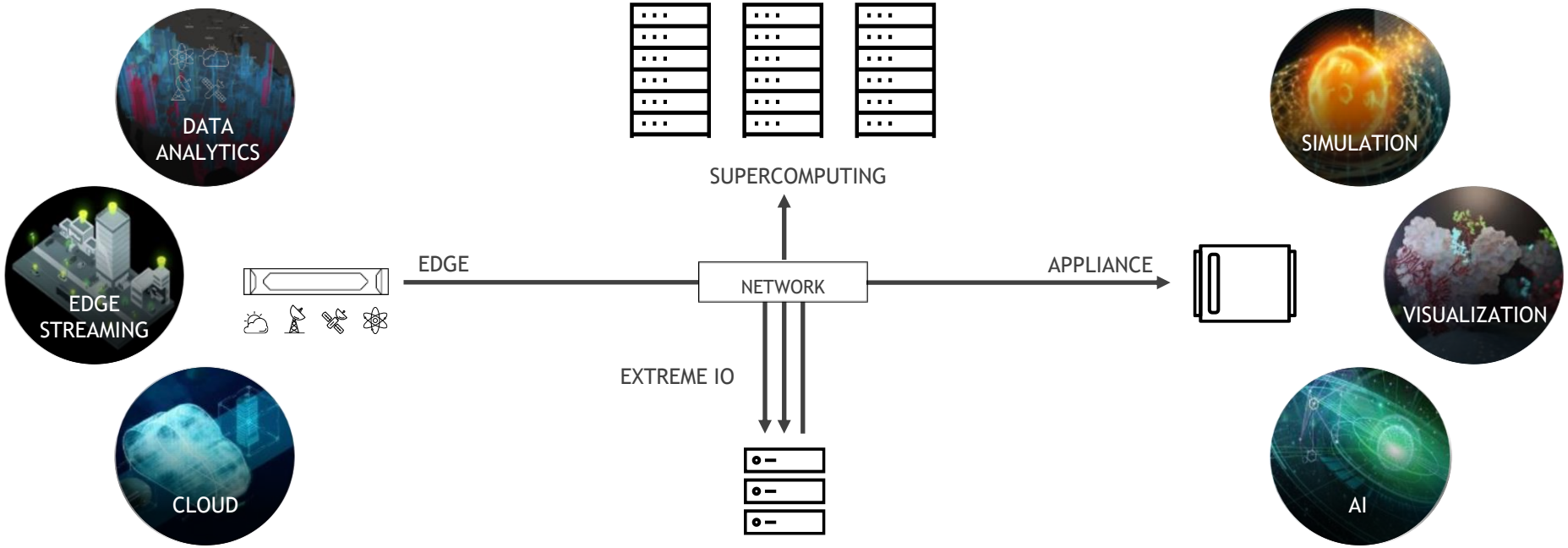
- VIRTUAL GPU SW
- TRITON INFERENCE SERVER
- FLEET COMMAND
- NVIDIA GPU Operator
- Red Hat
- vmware

MONITORING

- *DCGM*
- Grafana
- Prometheus
- *UFM*

NVIDIA

EXPANDING UNIVERSE OF SCIENTIFIC COMPUTING



NVIDIA POWER WORLD'S FASTEST AND MOST EFFICIENT, AI SUPERCOMPUTERS

TOP 500 The List.

8
of Top 10

68%
Overall

20%
More InfiniBand Systems v. ISC20

The GREEN 500

9
of Top 10

3.5X
Higher Energy Efficiency v. non-GPU Green500 Systems

29.5GF/W
Greenest NVIDIA System

HPL-AI

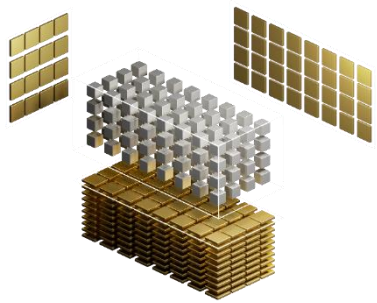
8
of Top 10

1.1EF
Mixed Precision AI Performance on SUMMIT

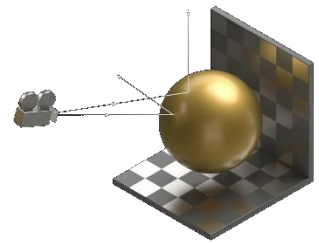
2X
Submission Over Previous List

NVIDIA

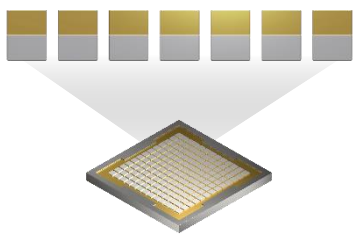
AMPERE ARCHITECTURE



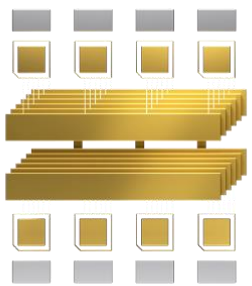
3rd Gen Tensor Cores
Faster, Flexible, Easier to use
20x AI Perf with TF32
2.5x HPC Perf



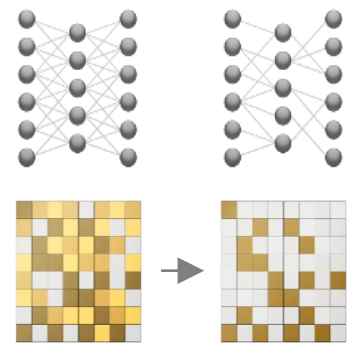
2nd Gen RT Cores
Up to 2X throughput of previous generation



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



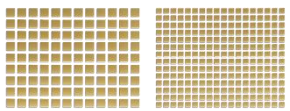
3rd Gen NVLINK and NVSWITCH
Efficient Scaling to Enable Super GPU
2X More Bandwidth



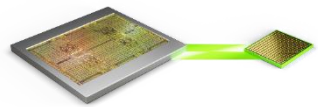
Sparsity
Harness Sparsity in AI Models
2x AI Performance

ANNOUNCING NVIDIA A100 80GB

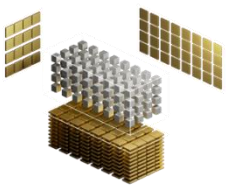
Supercharging The World's Highest Performing AI Supercomputing GPU



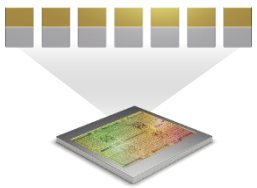
80GB HBM2e
For largest datasets and models



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

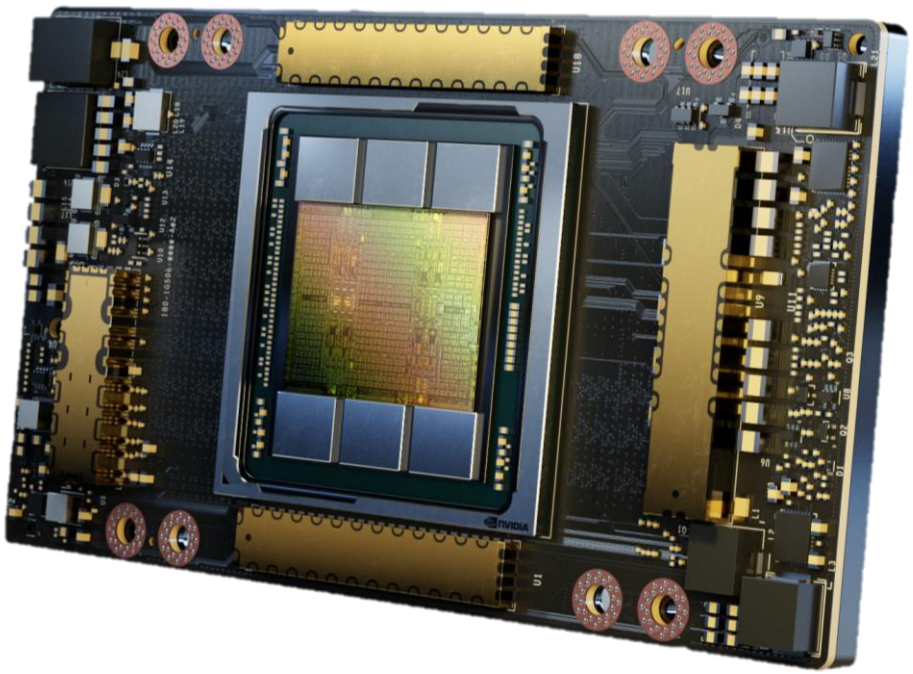


3rd Gen Tensor Core

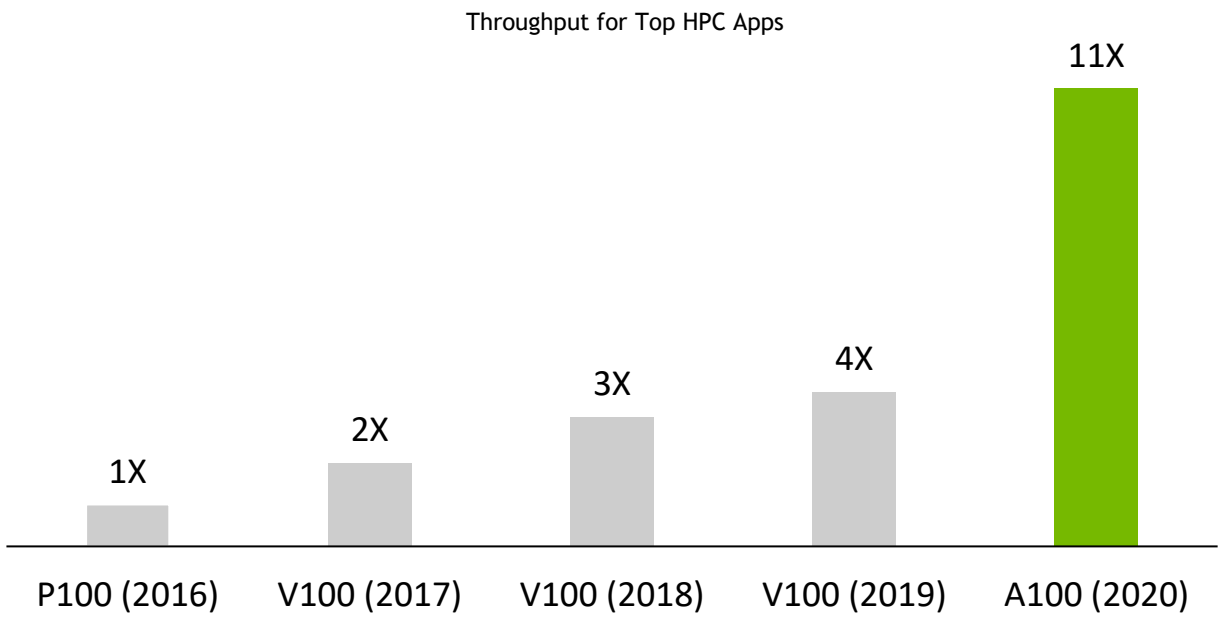


Multi-Instance GPU

3rd Gen NVLink



11X MORE HPC PERFORMANCE IN FOUR YEARS

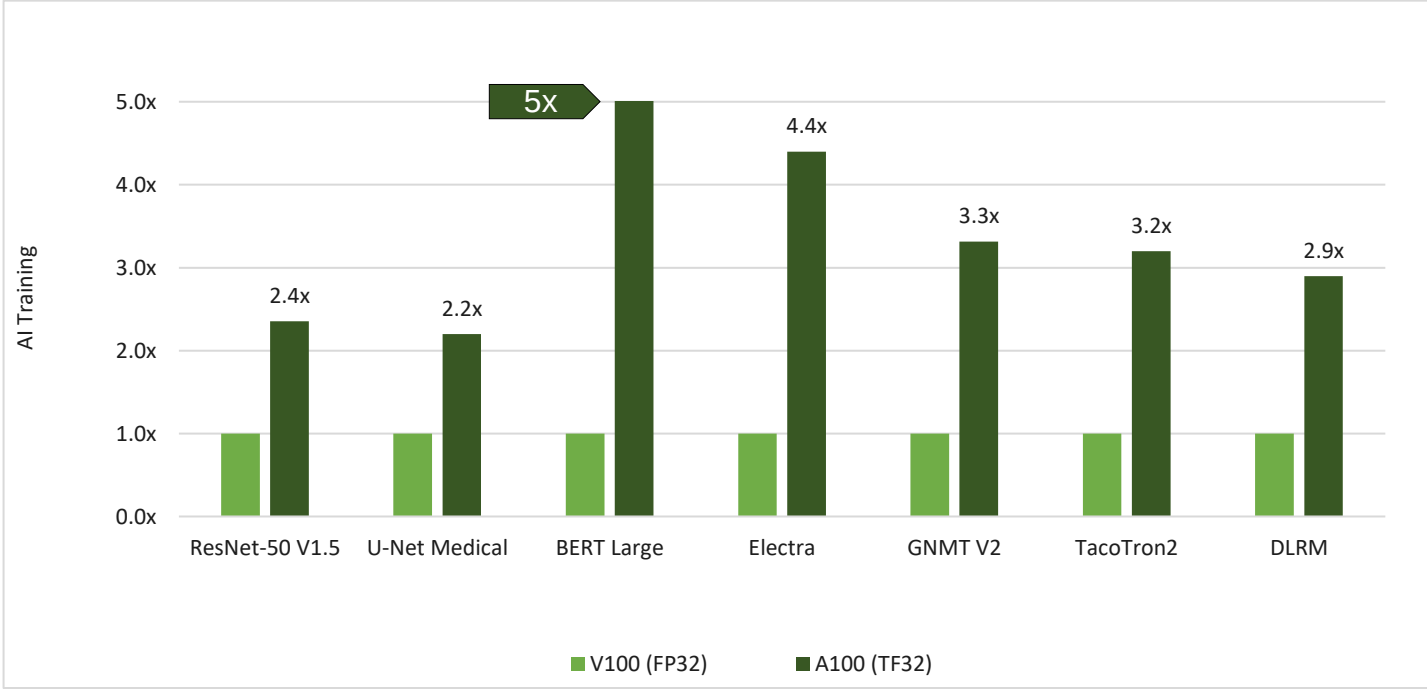


Geometric mean of application speedups vs. P100: Benchmark application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT-Large Fine Tuner], Quantum Espresso [AUSURF112-jR]; Random Forest FP32 [make_blobs (160000 x 64 : 10)], TensorFlow [ResNet-50], VASP 6 [Si Huge] | GPU node with dual-socket CPUs with 4x NVIDIA P100, V100, or A100 GPUs.

3RD GENERATION MULTI-PRECISION TENSOR CORES IN A100

Greatest Generational Leap – 20X Volta

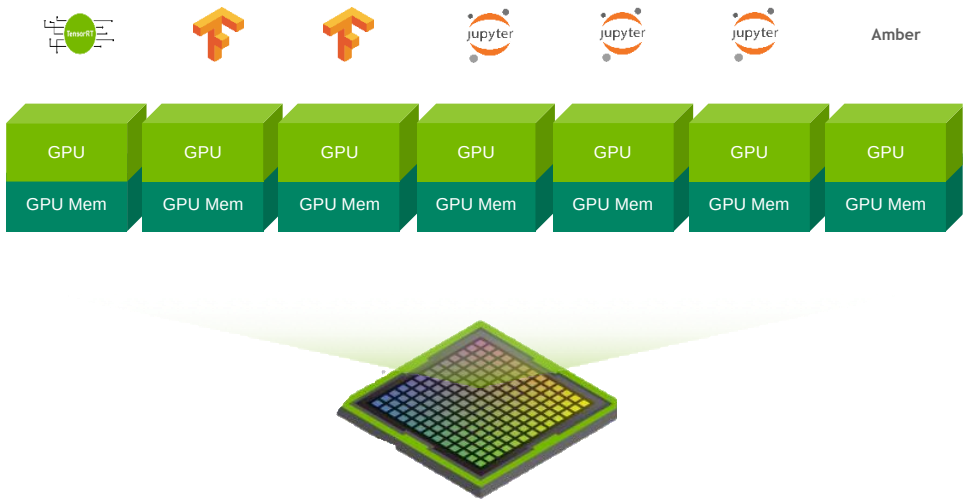
Precision	Peak	Vs Volta
FP64 HPC	19.5 TFLOPS	2.5x
FP32 TRAINING	312 TFLOPS	20x
INT8 INFERENCE	1,248 TOPS	20x

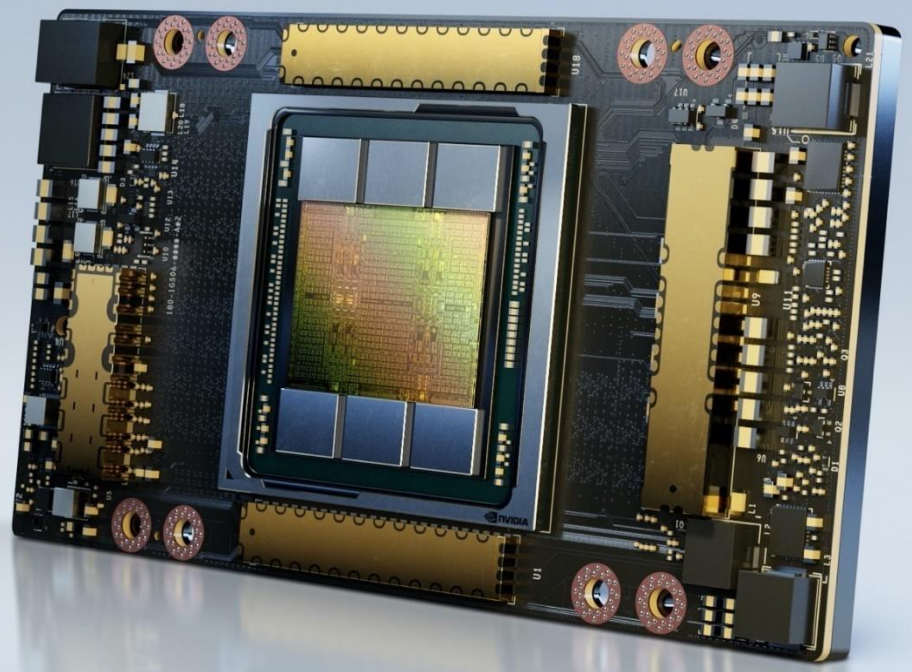


TF32 = Default Precision in TensorFlow, PyTorch, MXNet

NVIDIA A100 MULTI-INSTANCE GPU

Greatest Generational Leap – 20X Volta





Up To 7 hardware isolated GPU Instances per A100

Simultaneous Workload Execution With Guaranteed Quality Of Service

Right-sized GPU

A100 AVAILABLE VIA NVIDIA HGX A100 AND A100 PCIE

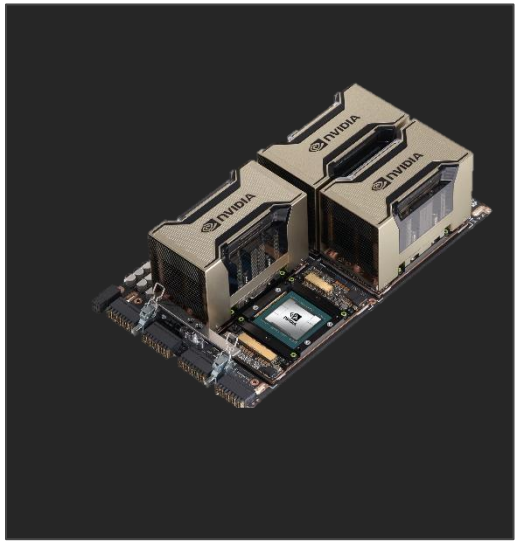
A100 PCIe

For Mainstream Servers

1-8 GPUs per server, optional NVLink Bridge between 2 GPUs

40GB option only

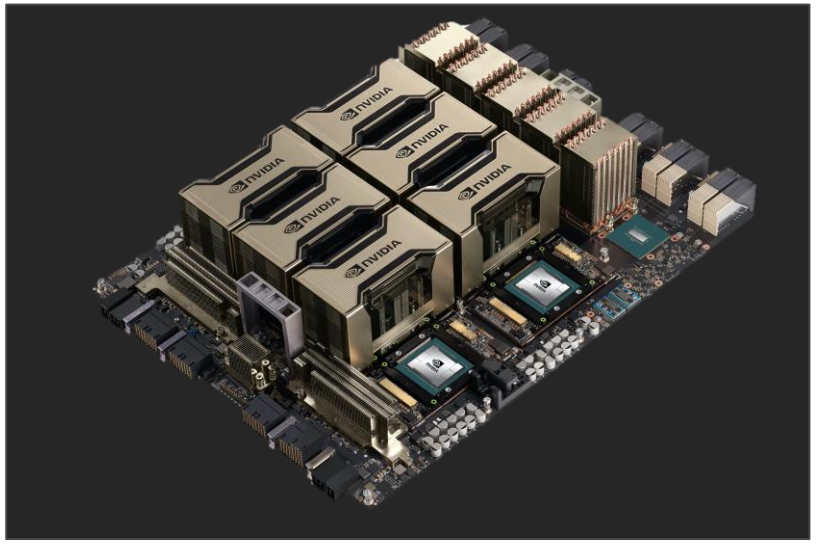
HGX A100 4-GPU



Scale-Up – Mixed AI & HPC

4 A100s, Fully Connected w/ shared NVLinks

HGX A100 8-GPU



Scale-up - Fastest Time-to-solution for AI

8 GPUs, Full NVLink B/W between all GPUs with NVSwitch

GPUs FOR EVERY VIRTUAL WORKLOAD

Expanding Workloads Drive The Need for Specialized Accelerators

A100	A30	T4	A10	A40	A16
DL Training Scientific Research Data Analytics	Language Processing Conversational AI Recommender Systems	Edge AI Edge Video Mobile Cloud Games	Virtual Workstation Video Conferencing 4K Cloud Games	Cloud Rendering Cloud XR Omniverse	Virtual Desktop Transcoding
Highest Perf Compute AI, HPC, Data Processing	**AI Inference & Mainstream Compute**	**Small Footprint Datacenter and Edge Inference**	**Mainstream Graphics & Video with AI**	**Highest Perf Graphics Visual Computing**	**Highest Density Virtual Desktop**
Fastest Compute, FP64 Up to 7 MIG instances	Versatile Mainstream Compute FP64, Up to 4 MIG instances	High density Video & Graphics Compact & Versatile	4K Cloud Gaming Graphics, Video with AI	Fastest RT Graphics Largest render models	4K Resolution Max # of encode/decode streams
250W & 300W | 40G & 80G 2-slot FHFL | NVLINK	165W | 24GB 2-slot FHFL | NVLINK	70W | 16GB 1-slot Low Profile	150W | 24GB 1-slot FHFL	300W | 48GB 2-slot FHFL | NVLINK	250W | 4 x 16GB 2-slot FHFL

Compute | *Graphics*

NVIDIA A30

Versatile Compute Acceleration for Mainstream Enterprise Servers

Purpose built for Inference and Flexible Enterprise Compute
20X T4 AI perf (A30 TF32 FLOPS vs T4 FP32)

Multi-Instance GPU
Up to 4 concurrent instances per GPU (QoS)

Compute
3rd Gen Tensor cores, Fast FP64

High Bandwidth Memory
Ultra-low latency

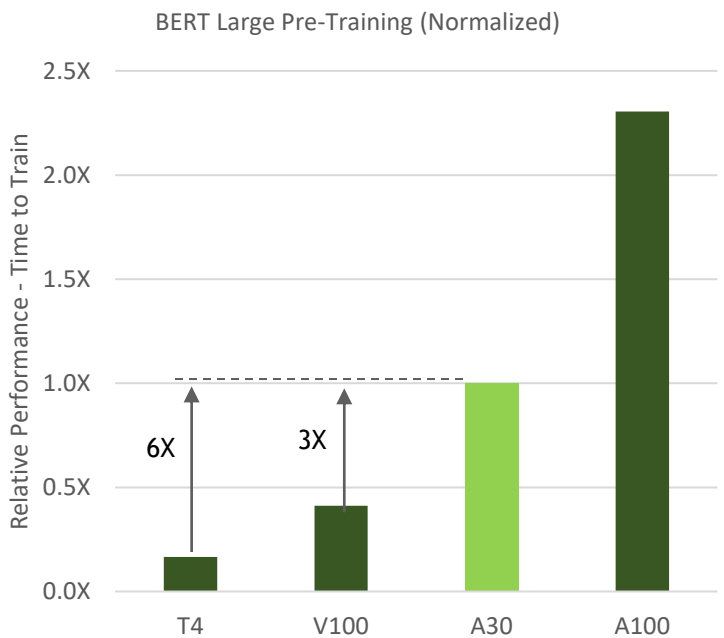
Power Efficient
Excellent Perf/W

Sparsity Acceleration
Further 2x speed up

	A30
GPU Architecture	NVIDIA Ampere
Multi-Instance GPU	4 instances @ 6GB each 2 instances @ 12GB each
GPU Memory	24GB HBM2
Memory Bandwidth	933 GB/s
Interconnect	PCIe Gen 4 (x16) 1x NVLINK Bridge
Form Factor	2 Slot FHFL
Max Power	165W
Schedule	Production

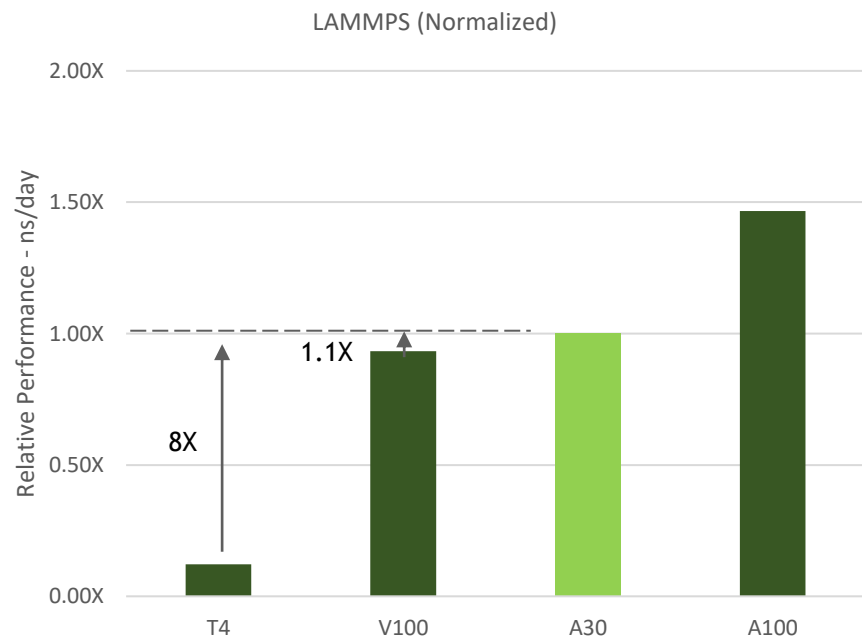
A30 DELIVERED APPLICATION PERFORMANCE – TRAINING AND HPC

AI TRAINING—UP TO 3X HIGHER THROUGHPUT THAN V100 AND 6X HIGHER THAN T4



BERT-Large Pre-Training (9/10 epochs) Phase 1 and (1/10 epochs) Phase 2, Sequence Length for Phase 1 = 128 and Phase 2 = 512, dataset = real, NGC™ container = 21.03, 8x GPU: T4 (FP32, BS=8, 2) | V100 PCIE 16GB (FP32, BS=8, 2) | A30 (TF32, BS=8, 2) | A100 PCIE 40GB (TF32, BS=54, 8) | batch sizes indicated are for Phase 1 and Phase 2 respectively

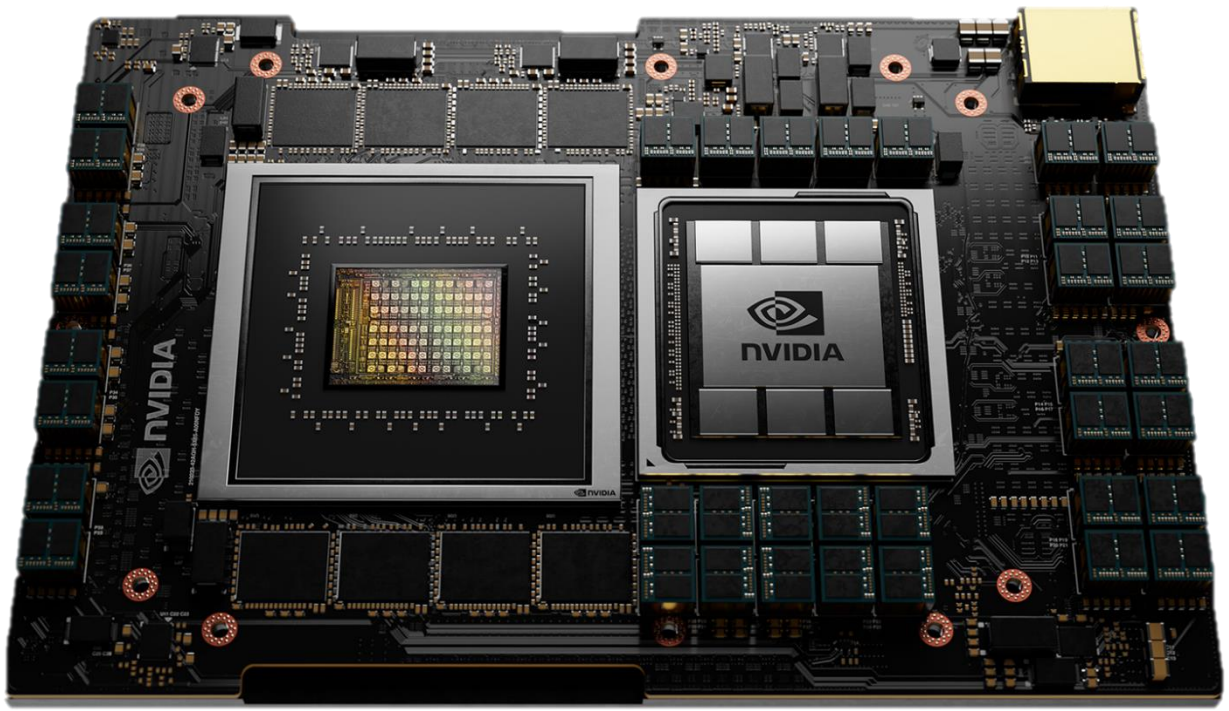
HPC— 1.1X HIGHER THROUGHPUT THAN V100 AND 8X HIGHER THAN T4



Dataset: ReaxFF/C, FP64 | 4x GPU: T4, V100 PCIE 16GB, A30

ANNOUNCING NVIDIA GRACE

Breakthrough CPU Designed for Giant-Scale AI and HPC Applications



FASTEST INTERCONNECTS

>900 GB/s Cache Coherent NVLink CPU To GPU (14x)
>600GB/s CPU To CPU (2x)

HIGHEST MEMORY BANDWIDTH

>500GB/s LPDDR5x w/ ECC
>2x Higher B/W
10x Higher Energy Efficiency

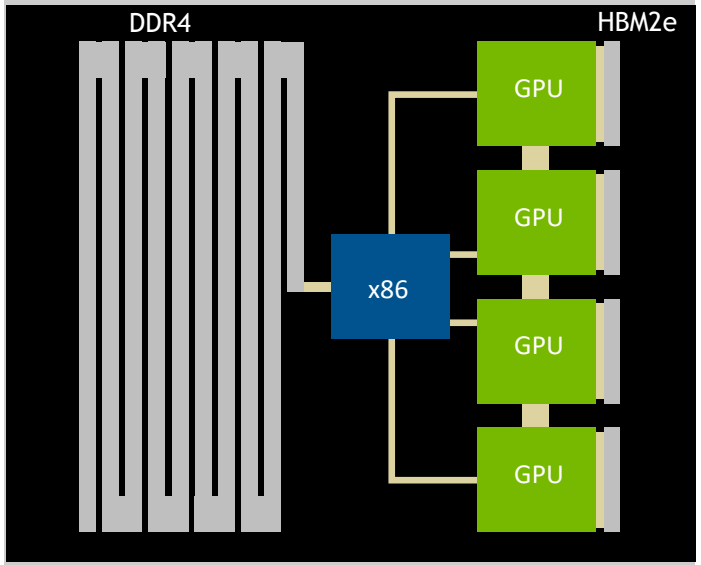
NEXT GENERATION ARM NEOVERSE CORES

>300 SPECrate2017_int_base est.
Availability 2023

TURBOCHARGED TERABYTE SCALE ACCELERATED COMPUTING

Evolving Architecture For New Workloads

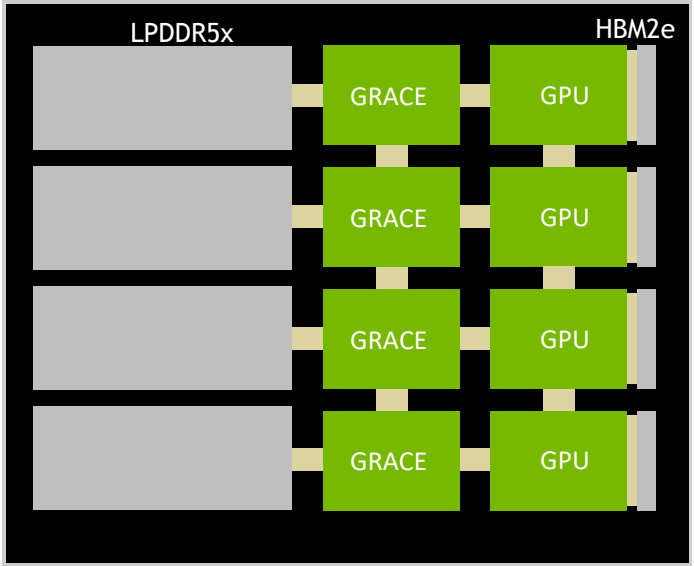
CURRENT x86 ARCHITECTURE



GPU	8,000	GB/sec
CPU	200	GB/sec
PCIE Gen4 (Effective Per GPU)	16	GB/sec
Mem-to-GPU	**64**	**GB/sec**

Transfer 2TB in 30 secs

INTEGRATED CPU-GPU ARCHITECTURE



GPU	8,000	GB/sec
CPU	500	GB/sec
NVLink	500	GB/sec
Mem-to-GPU	**2000**	**GB/sec**

Transfer 2TB in 1 secs

3 DAYS FROM 1 MONTH
Fine-Tune Training of 1T Model

REAL-TIME INFERENCE ON 0.5T MODEL
Interactive Single Node NLP Inference

Bandwidth claims rounded to nearest hundred for illustration.
Performance results based on projections on these configurations Grace : 8xGrace and 8xA100 with 4th Gen NVIDIA NVLink Connection between CPU and GPU and x86: DGX A100.
Training: 1 Month of training is Fine-Tuning a 1T parameter model on a large custom data set on 64xGrace+64xA100 compared to 8xDGXA100 (16xX86+64xA100)
Inference: 530B Parameter model on 8xGrace+8xA100 compared to DGXA100.

ANNOUNCING

THE WORLD'S FASTEST SUPERCOMPUTER FOR AI

20 Exaflops of AI

Accelerated w/ NVIDIA Grace CPU and NVIDIA GPU

HPC and AI For Scientific and Commercial Apps

Advance Weather, Climate, and Material Science

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Hewlett Packard Enterprise

NVIDIA

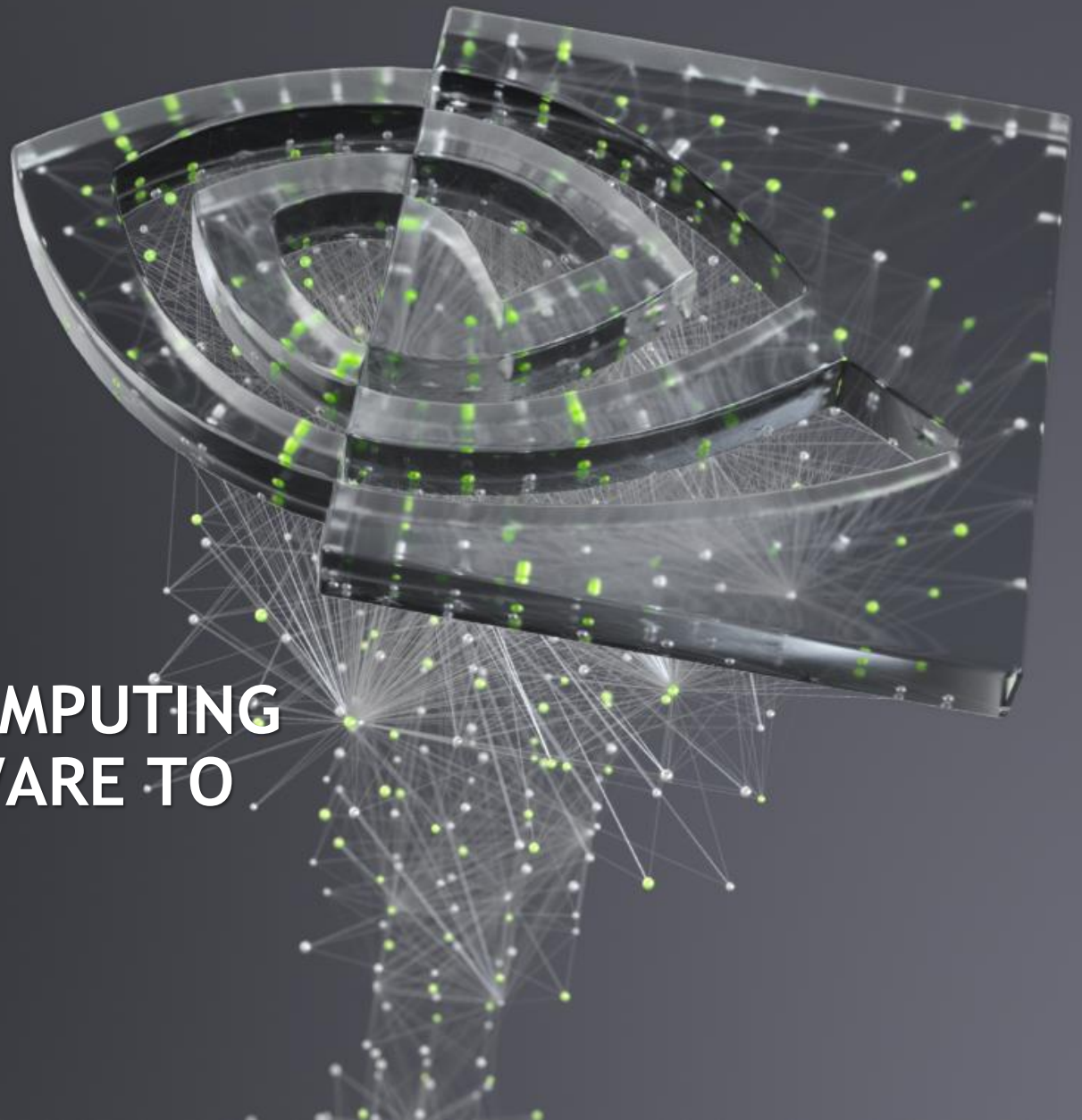
TECH TAKE

NVIDIA INSIGHTS: GPU COMPUTING LANDSCAPE FROM SOFTWARE TO HARDWARE.

Gabriel Noaje, PhD

Senior Solutions Architect, NVIDIA APAC South

June 29, 2021



NVIDIA IS A FULL STACK COMPUTING PLATFORM

Amazing Innovation and Expansion of NVIDIA Ecosystem

BUSINESS APPLICATIONS

- Customer Engagement
- Patient Diagnostics
- Fraud Detection
- Quality Assurance
- Industrial Automation
- Precision Marketing
- Molecular Simulations
- ++

SOFTWARE HUB

- HELM
- *Certified Containers*
- *Pre-trained Models*
- *SDKs*

APPLICATION FRAMEWORKS

SMART CITY	CONVERSATIONAL AI	AUTONOMOUS VEHICLES	RECOMMENDATION SYSTEMS	HEALTHCARE	++
Metropolis	*Jarvis*	*Drive*	*Merlin*	*Clara*	...

DEVELOPER TOOLKITS

- **ML & DATA ANALYTICS**: RAPIDS, dmlc XGBoost
- **AI TRAINING & INFERENCE**: TensorFlow, TensorRT, PYTORCH, mxnet
- **HIGH PERFORMANCE COMPUTING**: *NVIDIA HPC SDK*
- **RENDERING & VISUALIZATION**: *IndeX*, *OptiX*, *CloudXR*, *MDL*

ACCELERATION LIBRARIES

- **COMPUTE**: *CUDA-X*
- **NETWORKING, STORAGE & SECURITY**: *DOCA*, *MAGNUM IO*

VALIDATED SOLUTIONS

SERVERS & CLOUD

- DGX, HGX — Purpose Built
- EGX — Mainstream & Edge
- CSP Instances

HARDWARE TECHNOLOGIES

- GPU
- NVSwitch
- BF DPU
- SMART NIC
- NVIDIA Switch

OPERATIONS

- VIRTUAL GPU SW
- TRITON INFERENCE SERVER
- FLEET COMMAND
- NVIDIA GPU Operator
- Red Hat
- vmware

MONITORING

- *DCGM*
- Grafana
- Prometheus
- *UFM*

NVIDIA IS A FULL STACK COMPUTING PLATFORM

Amazing Innovation and Expansion of NVIDIA Ecosystem

150
SDKs

RTX | HPC | RAPIDS | AI | CLARA | METRO | DRIVE | ISAAC | AERIAL 5G

CUDA-X-AI

CUDA

MAGNUM IO

COMPLETE SOFTWARE STACK

SYSTEMS

CHIPS

FULL STACK INNOVATION

1B
CUDA GPUs

2.5M
Developers

24M
CUDA Downloads

2,000
GPU-Accelerated Applications

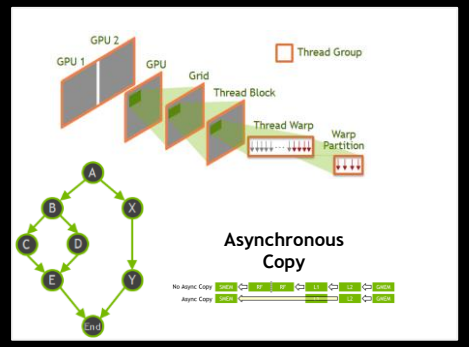
7,500
AI Startups

GROWING ECOSYSTEM

CUDA 11.4

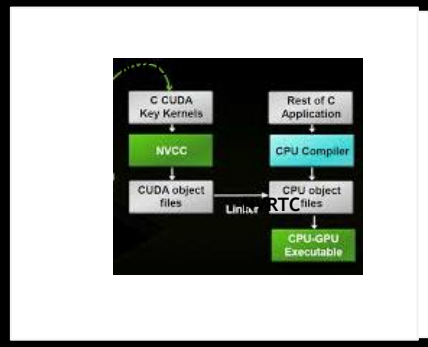
Major Feature Areas

New Platform Capabilities



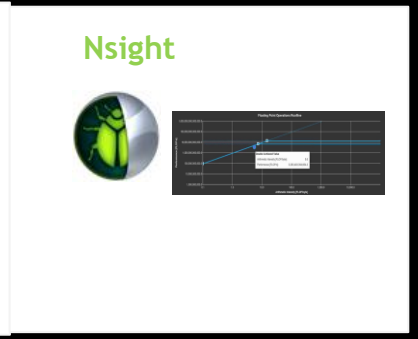
Programming Model

- *CUDA Graphs*
- *cudaMallocAsync*
- *Driver symbols*
- *Language support:*
  - *CUDA Python*



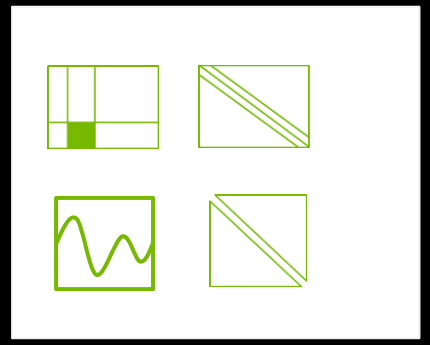
Compiler

- *Standalone demangler tool*
- *alloca() builtin-function*



Developer Tools

- *Nsight VS Code edition*
- *GPU metrics sampling*

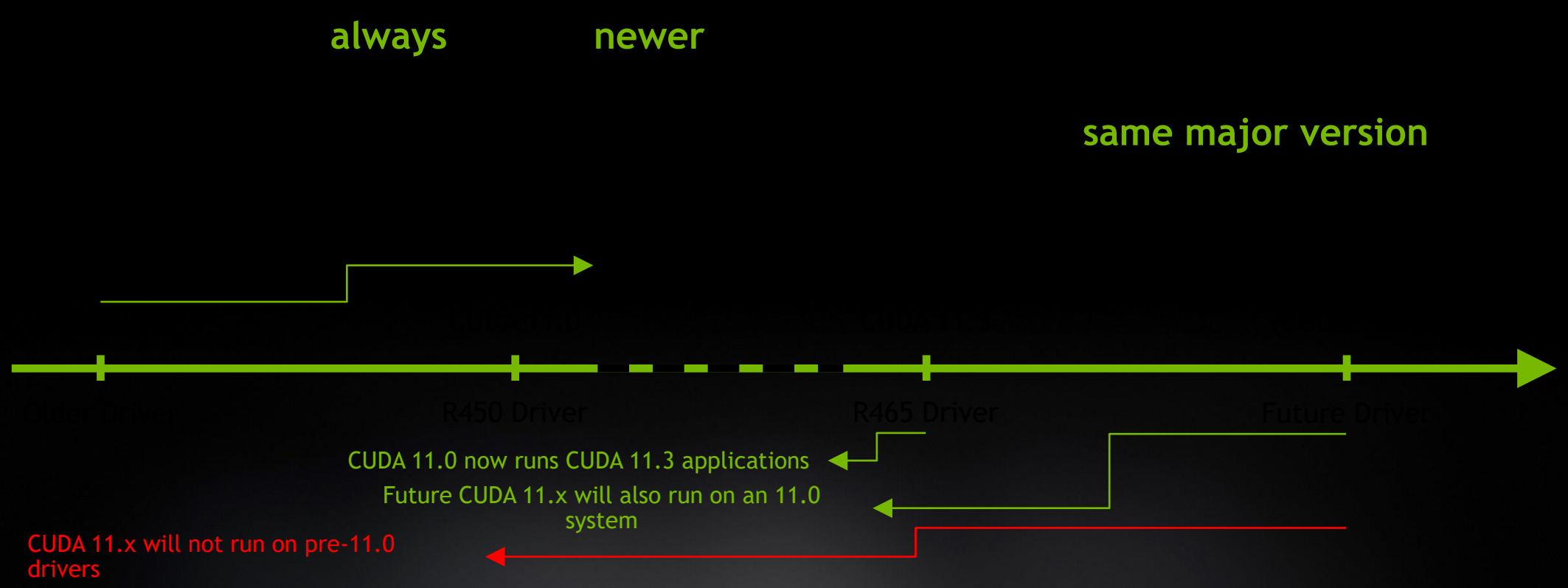


Math Libraries

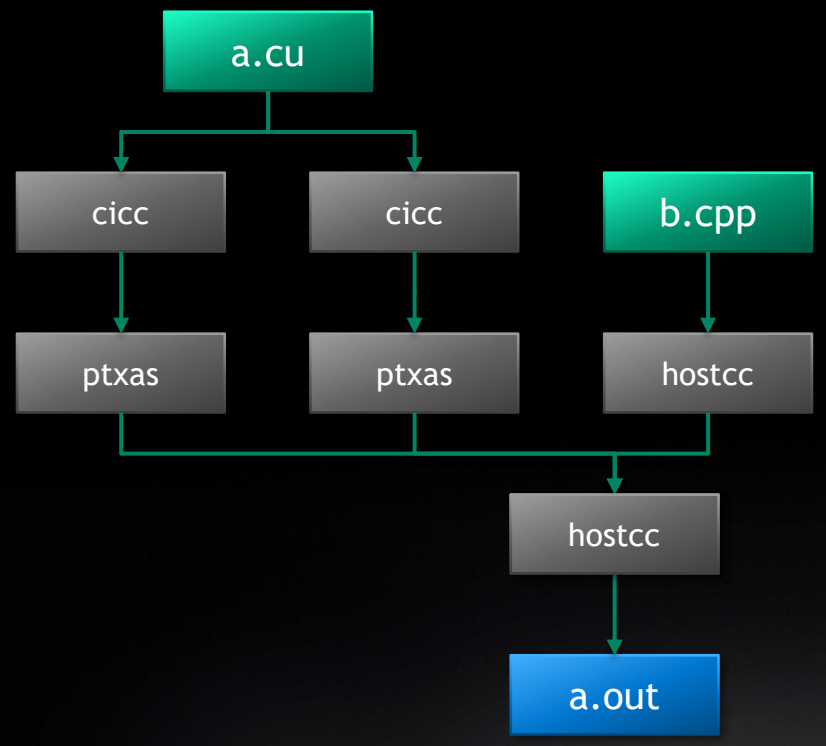
- *CUB*

CUDA 11.4 Release notes: https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html#title-new-features

MINOR VERSION COMPATIBILITY



NVCC MULTI-TARGET PARALLEL COMPILATION




```
# nvcc -t x.cu y.cu z.cu    ← Multi-target, single arch

# nvcc -t a.cu b.cpp -gencode=arch=compute_70,code=sm_70
\
                     -gencode=arch=compute_80,code=sm_80
```

Parallel compilation for multiple architectures

NVIDIA HPC SDK

Now including NVIDIA HPC-X Communications Stack

DEVELOPMENT

Programming Models
- Standard C++ & Fortran
- OpenACC & OpenMP
- CUDA

Compilers
- nvcc
- nvc
- nvc++
- nvfortran

Core Libraries
- libcu++
- Thrust
- CUB

Math Libraries
- cuBLAS
- cuTENSOR
- cuSPARSE
- cuSOLVER
- cuFFT
- cuRAND

Communication Libraries
- HPC-X
  - MPI
  - UCX
  - SHMEM
  - SHARP
  - HCOLL
- NVSHMEM
- NCCL

ANALYSIS

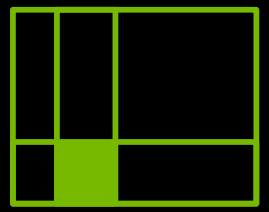
Profilers
- Nsight
  - Systems
  - Compute

Debugger
- cuda-gdb
  - Host
  - Device

Develop for the NVIDIA Platform: GPU, CPU and Interconnect
Libraries | Accelerated C++ and Fortran | Directives | CUDA
7-8 Releases Per Year | Freely Available

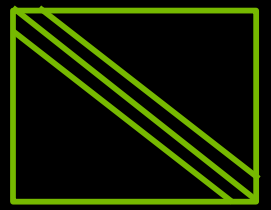
GPU ACCELERATED MATH LIBRARIES IN CUDA 11

OVERVIEW OF NEW NVIDIA A100 FEATURES



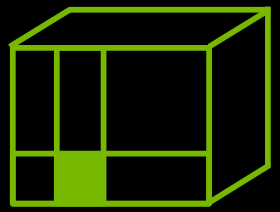
cuBLAS

BF16, TF32 and FP64 Tensor Cores



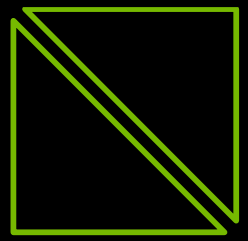
cuSPARSE

BF16, Increased memory BW, Shared Memory and L2



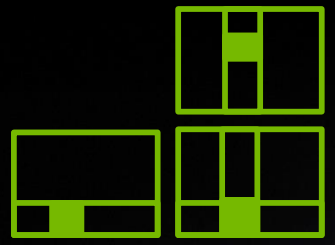
cuTENSOR

BF16, TF32 and FP64 Tensor Cores



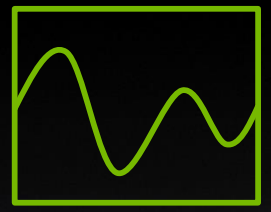
cuSOLVER

BF16, TF32 and FP64 Tensor Cores



CUTLASS

BF16, TF32 and FP64 Tensor Cores

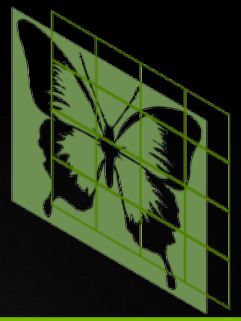


cuFFT

Increased memory BW, Shared Memory and L2

CUDA Math API

BF16 Support



nvJPEG

Hardware Decoder

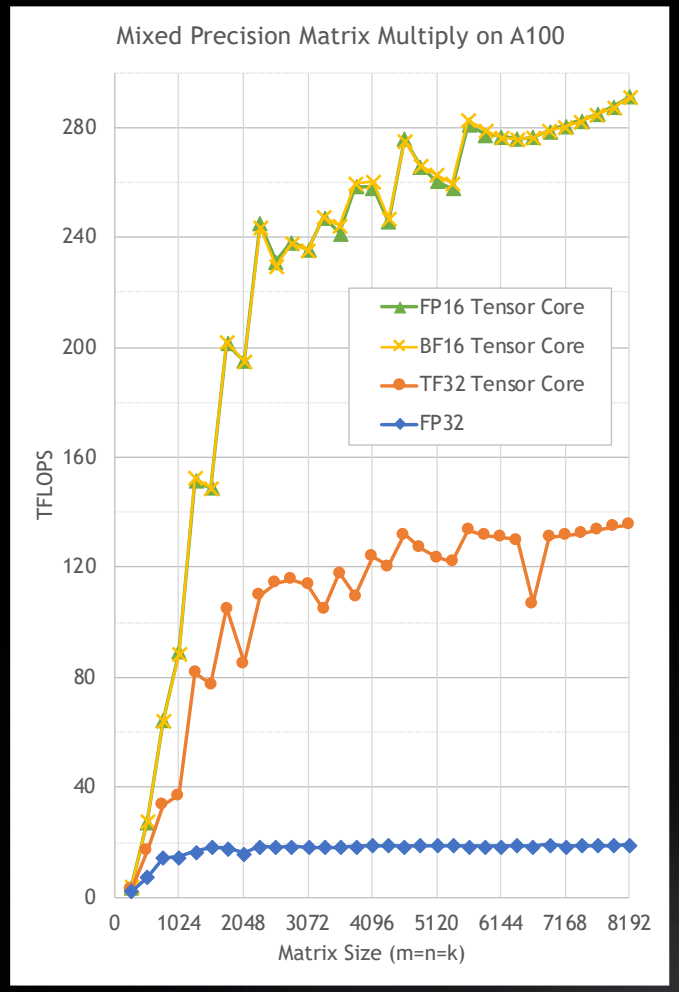
TENSOR CORE SUPPORT IN MATH LIBRARIES

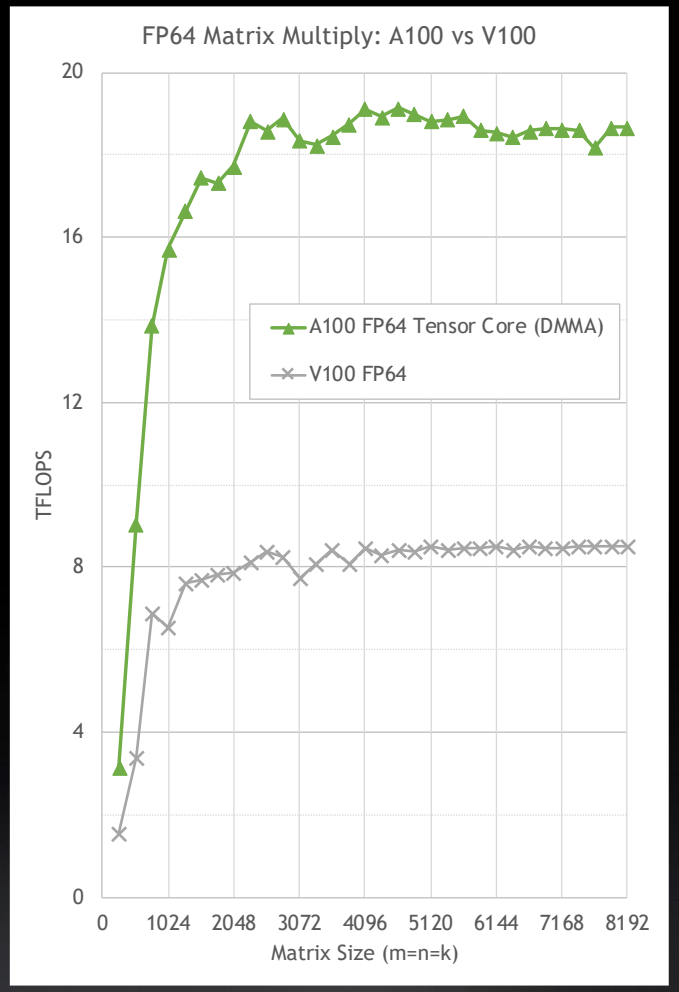
High-level overview of supported functionality by each library

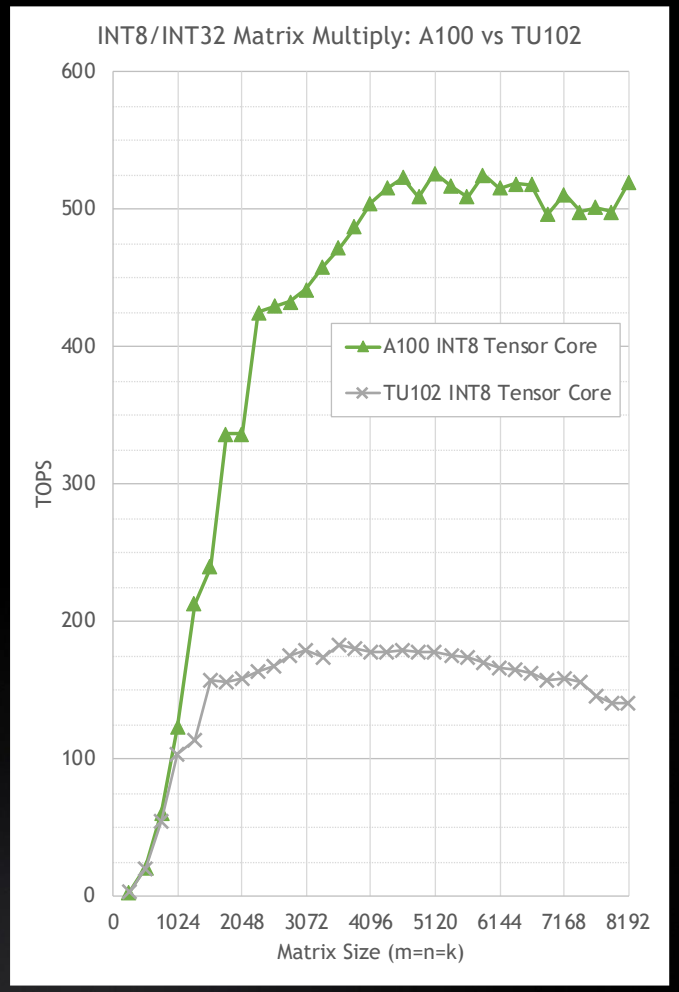
	INT4		INT8		FP16		BF16		TF32		FP64
	Dense	Sparse	Dense	Sparse	Dense	Sparse	Dense	Sparse	Dense	Sparse	Dense
			✓		✓		✓		✓		✓
					✓		✓		✓		✓
					✓		✓		✓		✓
cuSPARSE Block SpMM			✓		✓		✓		✓		✓
cuSPARSELt SpMM				✓		✓		✓		✓	
CUTLASS Dense GEMM and SpMM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CUTLASS Convolutions	✓		✓		✓		✓		✓		

cuBLAS

3rd GENERATION TENSOR CORES ADD SUPPORT FOR FP64 & NEW TYPE BF16 & COMPUTE TYPE TF32

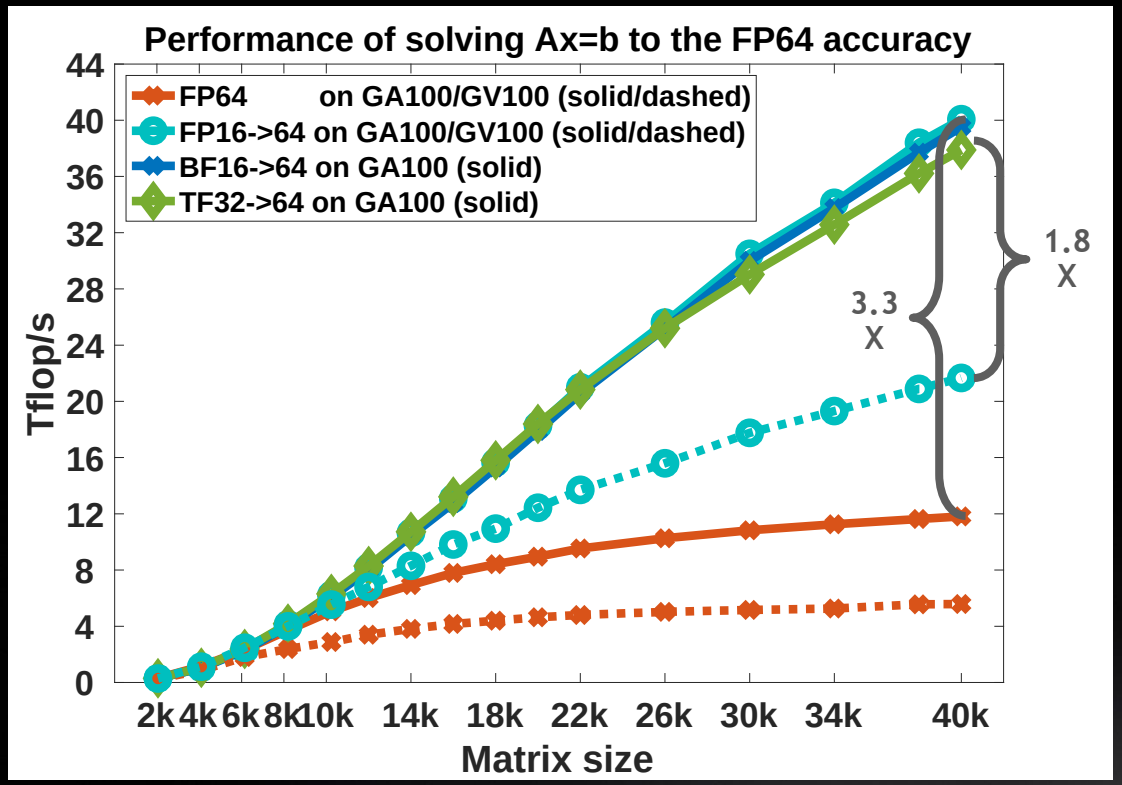


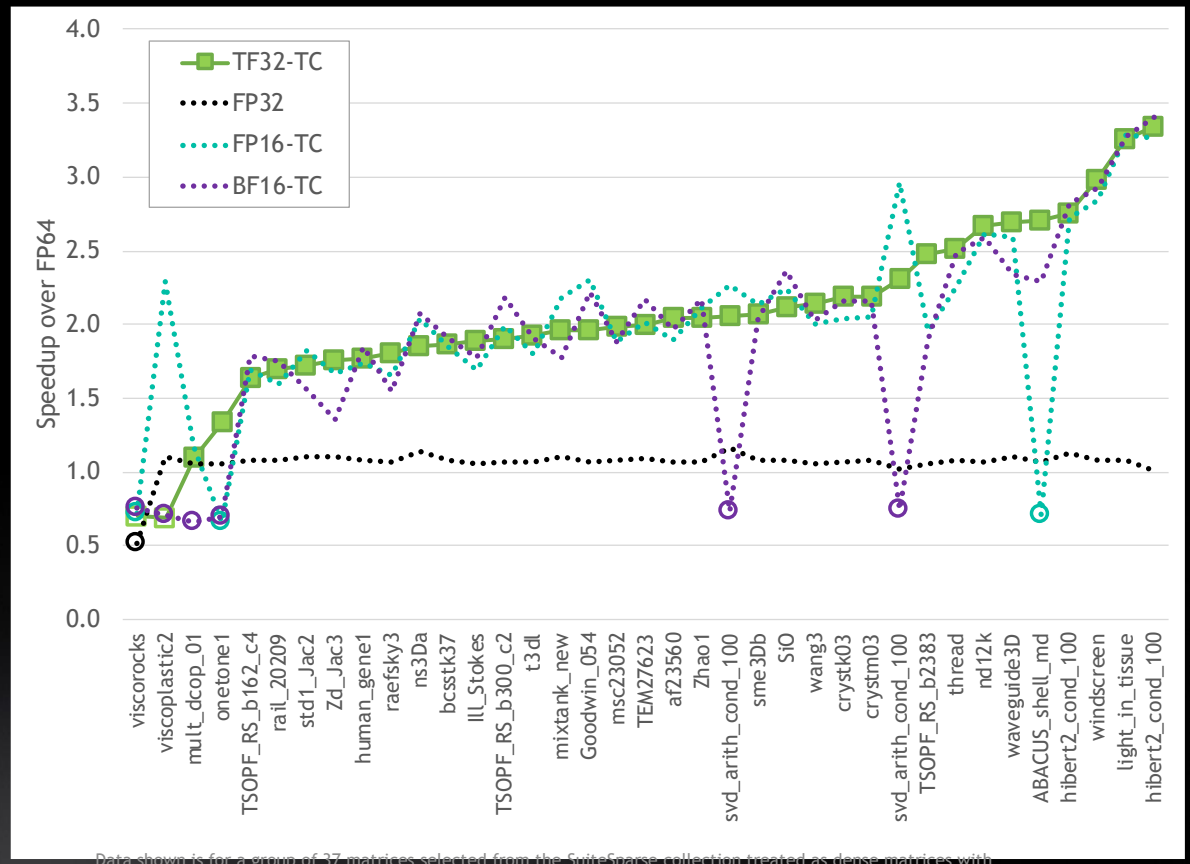




cuSOLVER

TENSOR CORE ACCELERATED ITERATIVE REFINEMENT SOLVER (TCAIRS)





Matrices used in this performance chart are Hilbert matrices. TCAIRS solver is ~3.3X faster than full FP64 solver using DMMA tensor cores and A100 is ~1.8X faster than V100 .

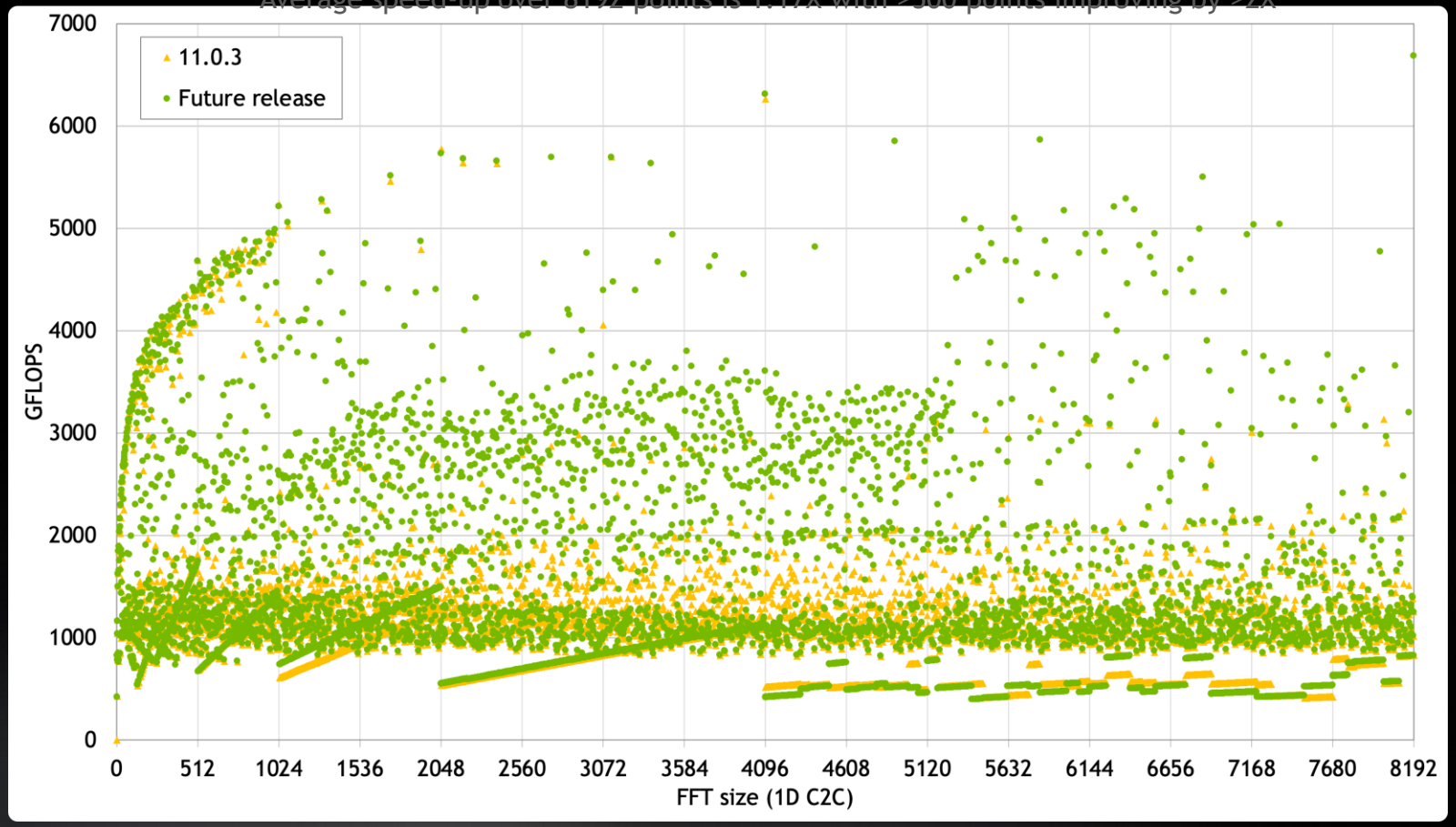
Data shown is for a group of 37 matrices selected from the SuiteSparse collection treated as dense matrices with sizes ranging from 20k to 38k. Performance data is average of all cases and all runs were done on an NVIDIA A100 with cuSOLVER from CUDA Toolkit 11.0. FP16 cases were scaled to ensure matrix entries are within the type range.

IMPROVED PERFORMANCE cuFFT

Updated performance and features since CUDA 11.0

- Optimized FFTs up to size 512 per dimension
- Improved C2C, R2C and C2R transforms
- bfloat16 support since 11.1.1
- Coming soon:
  - Optimized FFT sizes up to size 32768 per dimension
  - Improved decompositions for large FFT sizes (>16384) and 1D/2D/3D

cuFFT performance measured on A100 80GB
Average speed-up over 8192 points is 1.17X with >500 points improving by >2X

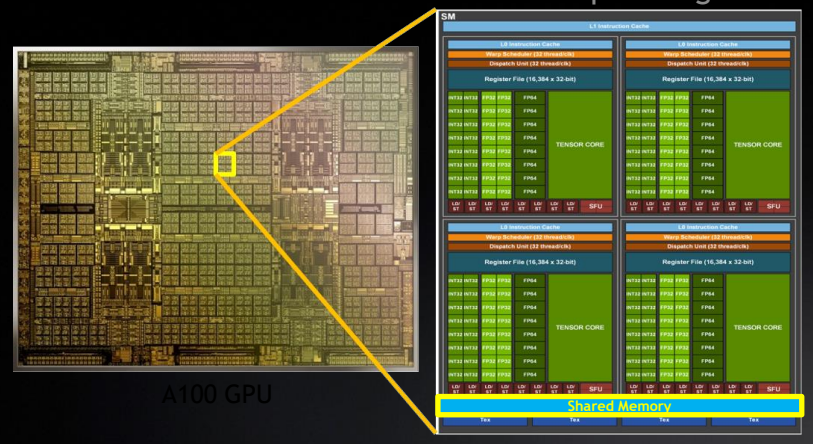


DEVICE EXTENSION LIBRARIES

Enabling kernel fusion of high-performance numerical method implementations

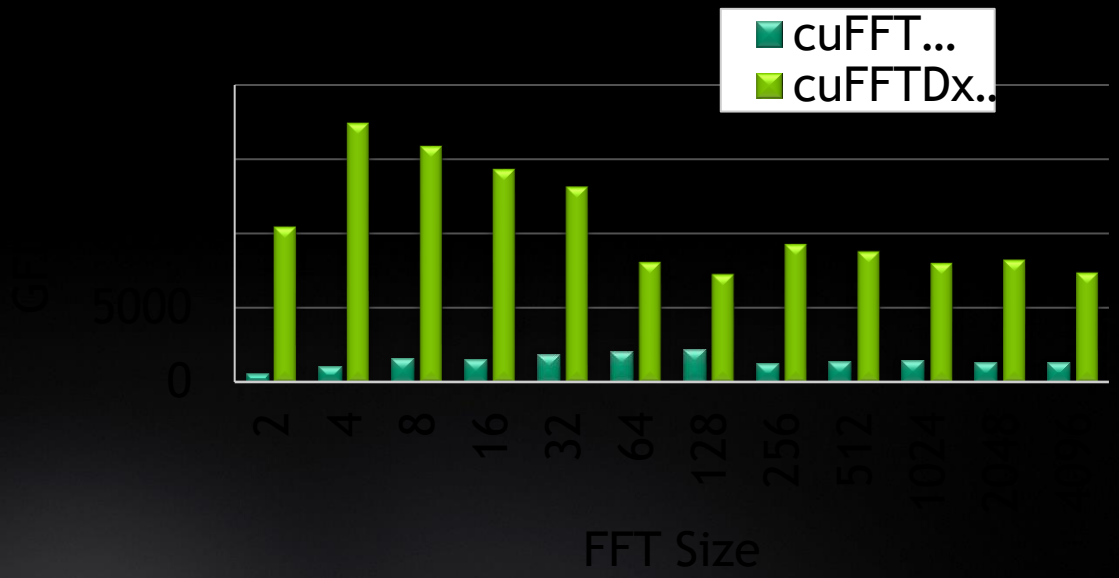
- cuFFTDx: In Math Library EA Program
  - Retain and reuse of on-chip data
  - Inline FFTs in user kernel with sizes up to 32k (on A100)
  - Combine FFT operations
  - Expecting cuFFTDx GA release later this year
- Linear algebra device extensions
  - cuSOLVERDx with non-pivoting LU also in EA
  - cuBLASDx with GEMM and TRSM are upcoming

```
using namespace cufftdx;

using FFT = decltype(Block() + Size<128>() + Type<fft_type::c2c>() +
                     Direction<fft_direction::forward>() + Precision<float>() +
                     ElementsPerThread<8>() + FFTsPerBlock<2>() + SM<700>());
```



A100 GPU

A100 Streaming Multiprocessor

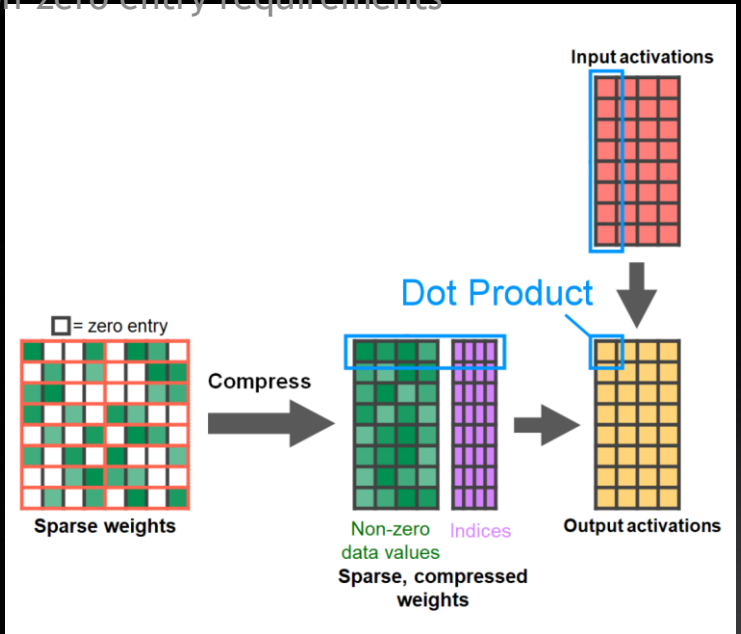


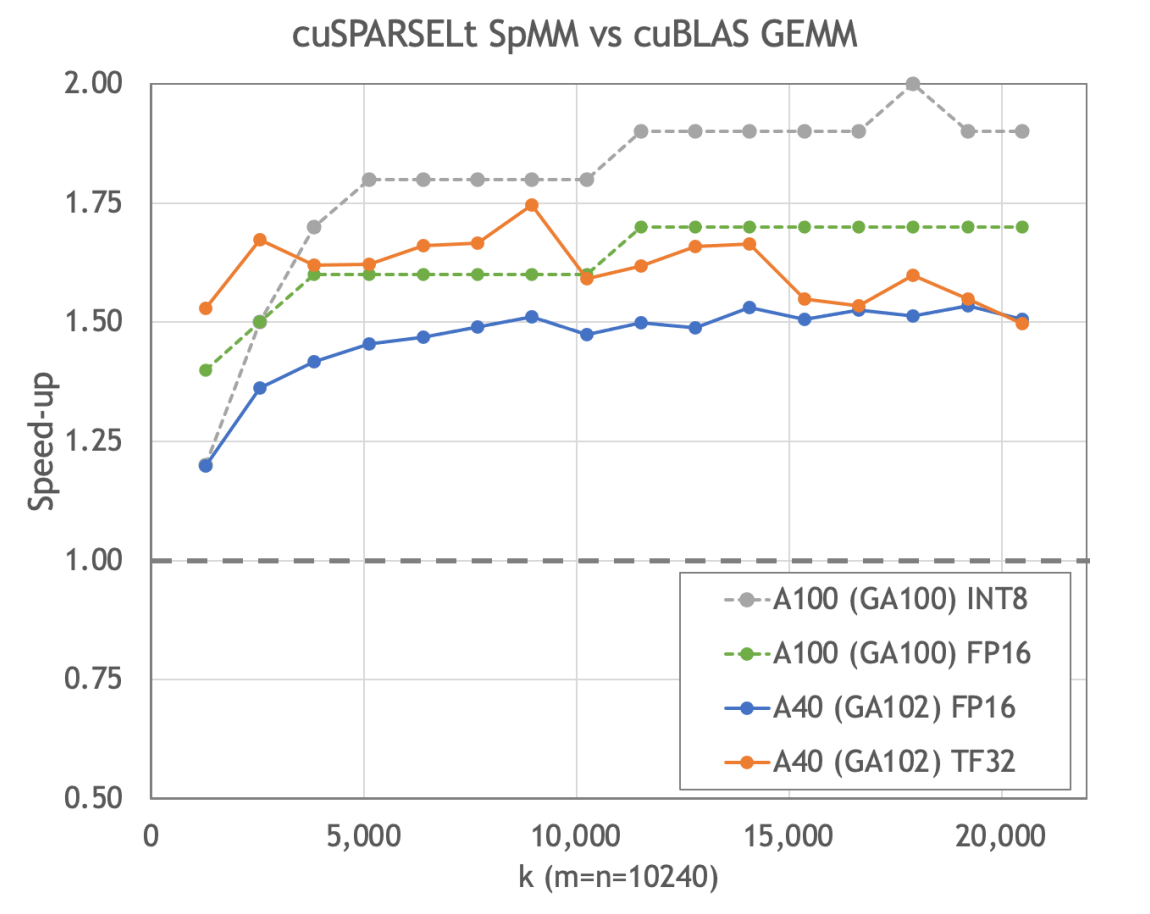
https://developer.nvidia.com/CUDAMathLibraryEA

EXPLOITING STRUCTURED SPARSITY WITH cuSPARSELt ON NVIDIA AMPERE ARCHITECTURE GPUS

Most recent update is available at https://developer.nvidia.com/cusparselt/downloads

- Provides easy access to 3rd generation Tensor Core Sparse MMAs on NVIDIA Ampere architecture GPUs to perform Sparse x Dense Matrix Multiplies
- cuSPARSELt provides essential utilities to easily create 50% sparse matrices that conform to the 2:4 non-zero entry requirements

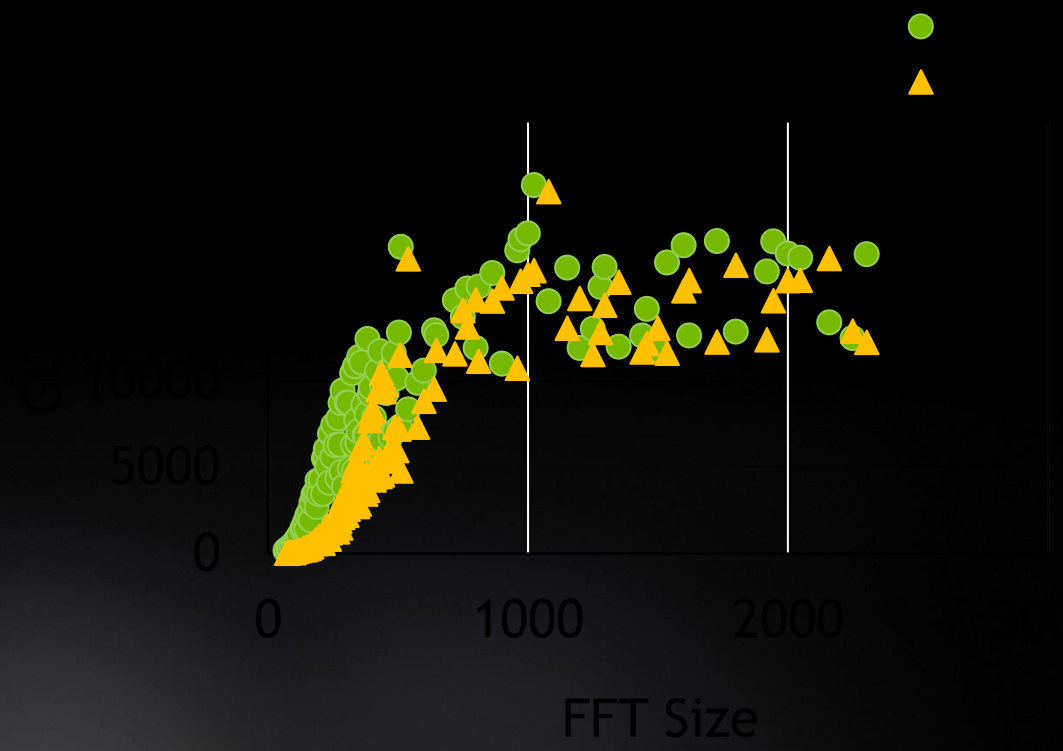




MULTI-GPU cuFFT

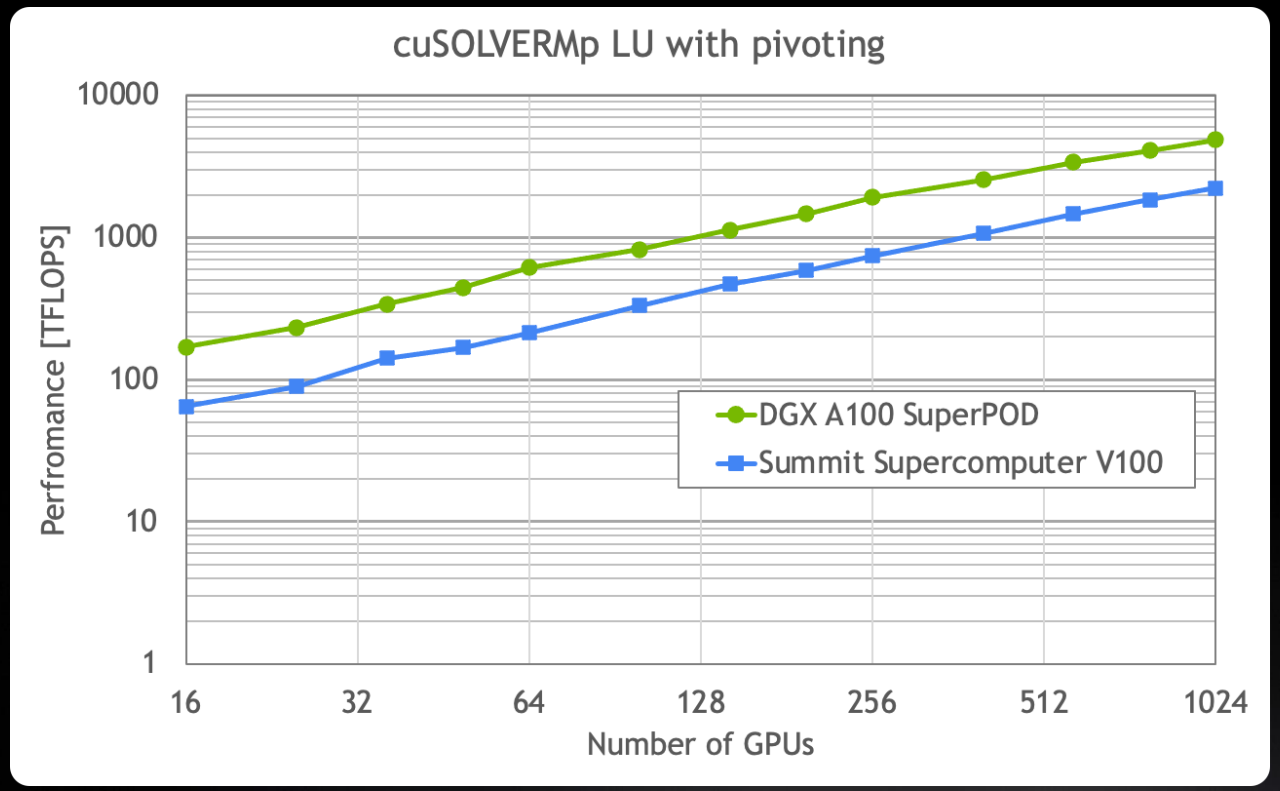
Single-node multi-GPU improvements available with CUDA 11.2

- Improved strong scaling performance for small 2D/3D FFTs with size <512
- Up to 22 TFLOPS SP & 11 TFFLOPS DP of performance on a DGX A100 node for 2D/3D FFT transforms
- Maximum size per node increased to 3072 cubed 3D FFT (SP)
- Stream ordering/graph capturing added to multi-GPU plans
- Upcoming later this year: Multi-process API



cuSOLVER DISTRIBUTED MULTI-GPU LU SOLVER

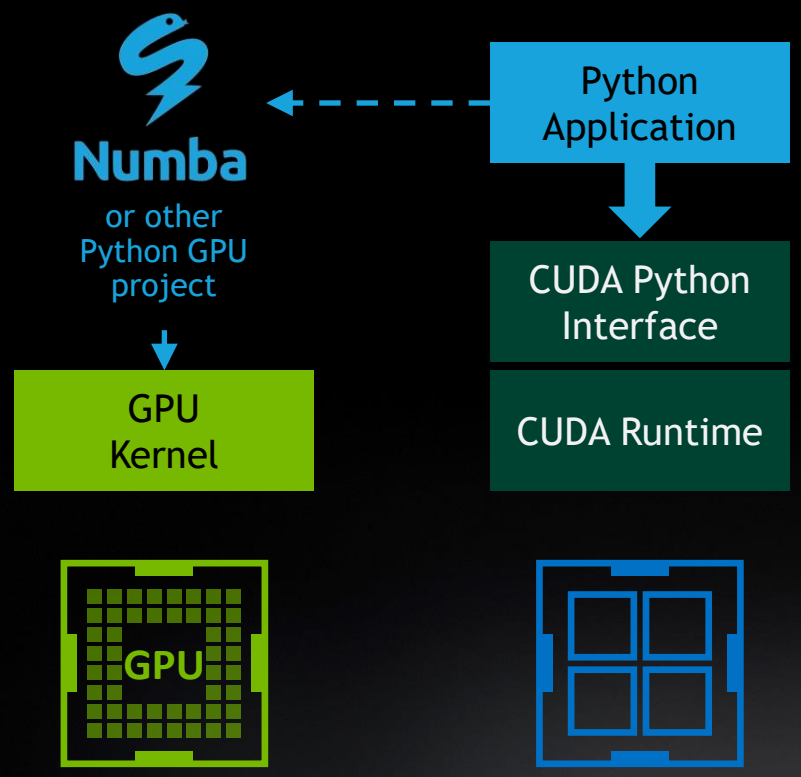
Scalability on Top 5 supercomputers



Data shown is for double complex LU factorization for 27882 rows/cols per device

- Supports standard 2D block cycling distribution of matrix and rhs between processes
- > 2.2X speed-up going from V100 to A100
- First release is planned to be part of HPC SDK in 1st half of 2021
- Early Access available by end of April https://developer.nvidia.com/cudamathlibraryea
- Further performance improves to come after initial release

CUDA PYTHON



Distribution

- Source available on GitHub
- PIP & Conda packages
- Redistributable license

Bindings

Full coverage of and access to the CUDA host APIs from Python

Platforms

- Linux: x86_64, sbsa, ppc64le
- Windows: x86_64

INTRODUCING LEGATE

Accelerated and Distributed

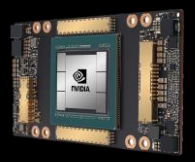
A framework for programming large numbers of GPUs as if they were a single processor

Pass data between Legate libraries without worrying about distribution or synchronization requirements

Legate NumPy and Pandas aim to transparently scale existing Numpy and Pandas workloads

Legate Numpy and Legate Pandas available now and opensource!

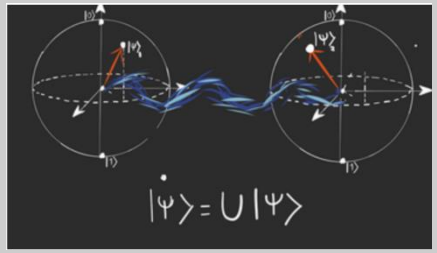
```
import legate.numpy as np

def cg_solve(A, b, tol=1e-10):
    x = np.zeros(A.shape[1])
    r = b - A.dot(x)
    p = r
    rsold = r.dot(r)
    for i in xrange(b.shape[0]):
        Ap = A.dot(p)
        alpha = rsold / (p.dot(Ap))
        x = x + alpha * p
        r = r - alpha * Ap
        rsnew = r.dot(r)
        if np.sqrt(rsnew) < tol:
          break
        beta = rsnew / rsold
        p = r + beta * p
        rsold = rsnew
        return x
```



ANNOUNCING NVIDIA CUQUANTUM

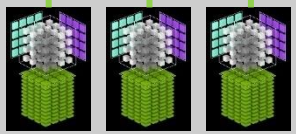
Researching the Computer of Tomorrow on the Most Powerful Computer Today

SDK FOR GPU-ACCELERATED QUANTUM SIMULATIONS

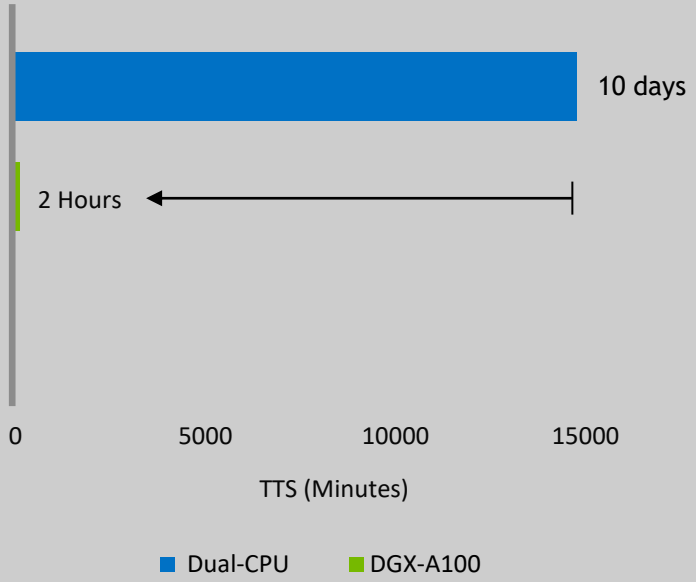


Quantum Circuit Simulation Frameworks

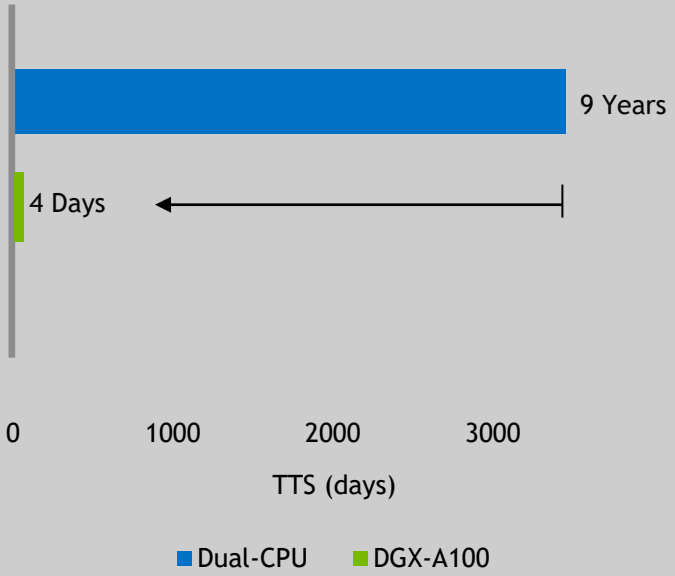
cuQuantum



STATE VECTOR SIMULATION
Scales to 10s of Qubits



TENSOR NETWORK SIMULATION
Scales to 1000's of Qubits



Footnotes: State Vector- 1,000 circuits , 36 qubits depth m=10, complex 64 | CPU: Qiskit (IBM) on Dual AMD EPYC 7742 | GPU: Qgate (NVAITC) on DGX-A100
Tensor Network - 53 qubits, depth m=20 | CPU: Estimated Quimb (Caltech) on Dual AMD EPYC 7742 | GPU: Quimb (Caltech) on DGX-A100

Data Analytics & Training

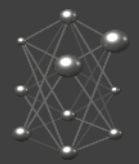
HugeCTR
mxnet
ONNX
PaddlePaddle
PYTORCH
RAPIDS
Apache Spark
TensorFlow

CUDA-X-AI

CUDA

MAGNUM IO

Inference



Conversational AI
Computer Vision
Recommendations
Reinforcement Learning

Pre-Trained Models
SOTA Models

Frameworks → TIS ← Query
TIS → Result

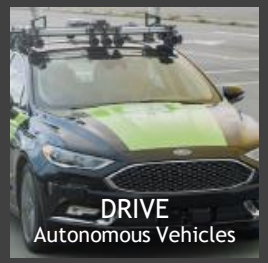
Triton
Inference Serving

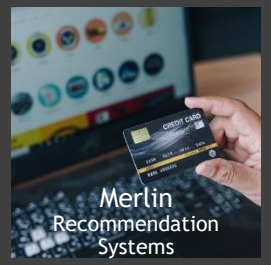
Pre-Trained Model → TLT → Customized Models
Data → TLT

Transfer Learning Toolkit
Zero Coding | Speech, Vision & NLU

TensorRT 7.2

Application Frameworks





Build AI Faster, Deploy Anywhere

HPC | DL | ML

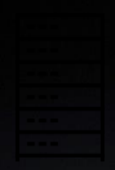
COMPUTER VISION | NLP | DLRM

CLARA | JARVIS | ISAAC

TRITON | GPU OPERATOR

CLARA DISCOVERY | TLT-JARVIS | RECSYS

x86 | ARM | POWER



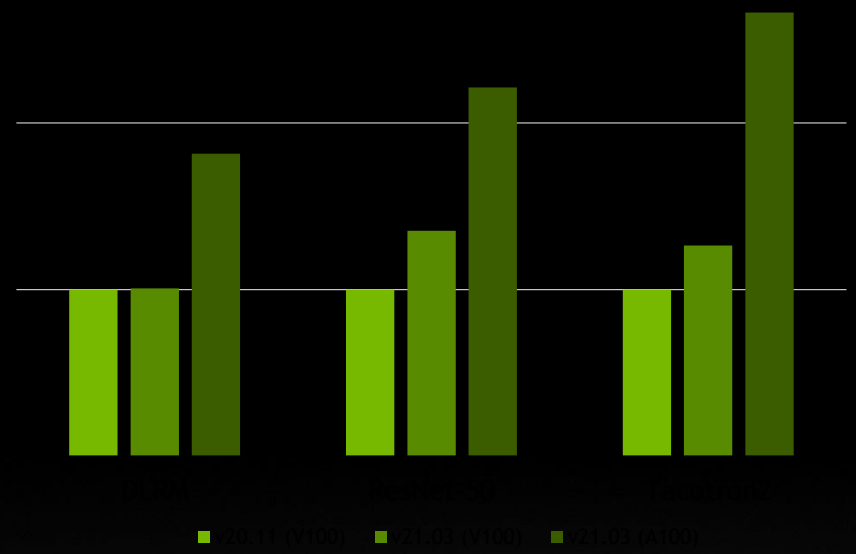
ON-PREM

CLOUD

HYBRID CLOUD

EDGE

DO WHAT YOU DO BEST, FASTER



ENTERPRISE READY SOFTWARE

Scanned for CVEs, malware, crypto

Tested for reliability

Backed by Enterprise support

PERFORMANCE OPTIMIZED

Scalable

Updated Monthly

Better performance on the same system

DEPLOY ANYWHERE

Docker | cri-o | containerd | Singularity

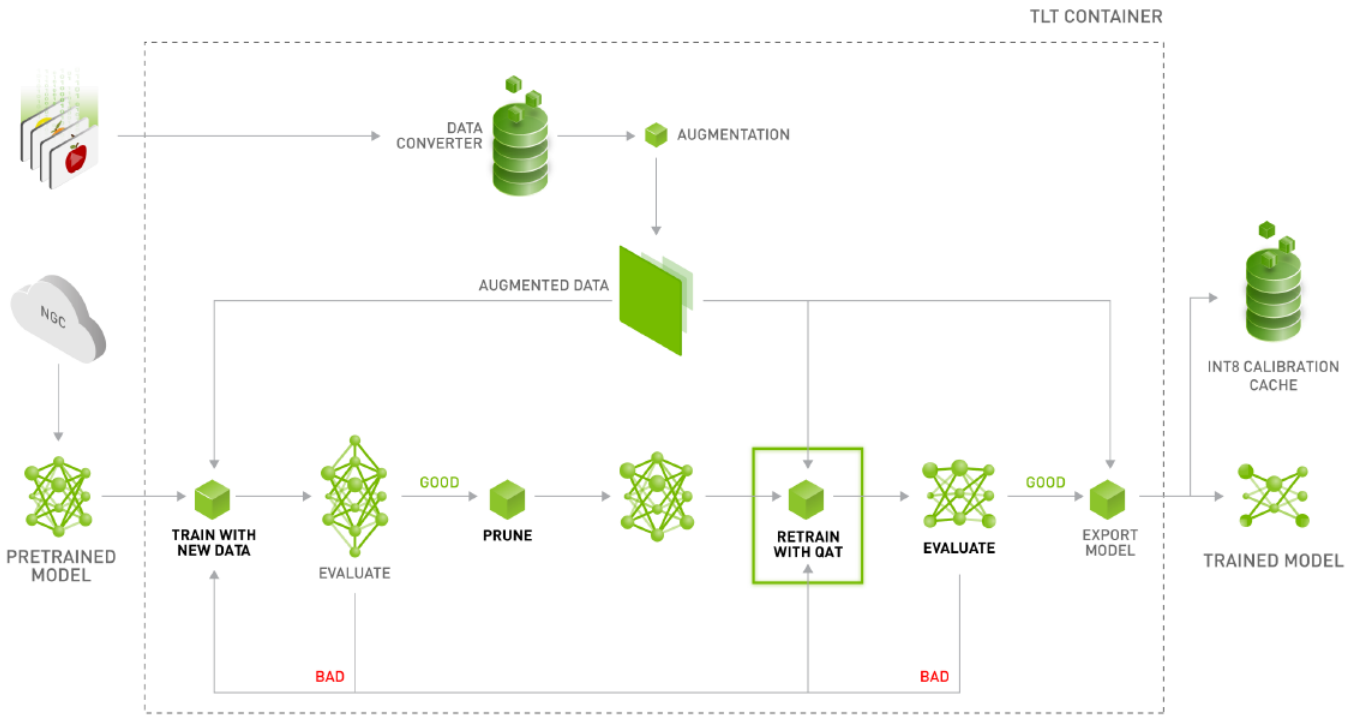
Bare metal, VMs, Kubernetes

Multi-cloud, on-prem, hybrid, edge

on PyTorch container on 8x V100 (32GB) & A100 (40GB). Mixed Precision. Batch size: 104 (V100). 128 (A100). DLRM training with PyTorch on 1x V100 (32GB) & A100 (40GB). Mixed Precision. Batch size: 32,768. ResNet-50 v1.5 training with TensorFlow on 8x V100 (32GB) & A100 (40GB). Mixed Precision.

TLT WORKFLOW

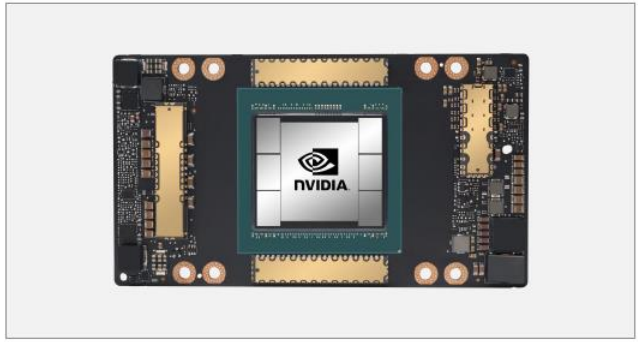
Automatic Mixed Precision | Quantization Aware Training | Pruning



https://developer.nvidia.com/transfer-learning-toolkit

ANNOUNCING TLT 3.0 DEVELOPER PREVIEW

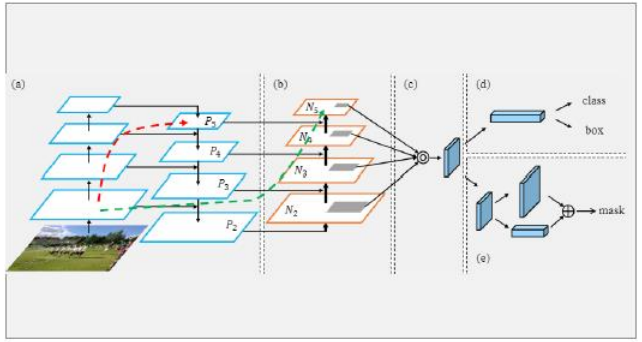
TRAIN CONVERSATIONAL AI MODELS
Simplify training of complex ASR & NLP tasks



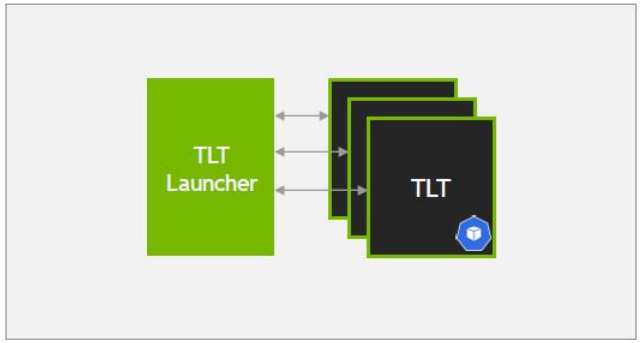
ACCELERATE TRAINING WITH AMPERE
10x speedup in training time

15 DOMAIN-SPECIFIC PRE-TRAINED MODELS
Accelerate time to market

MULTI GPU & MULTI NODE TRAINING
Speedup training time by 8x

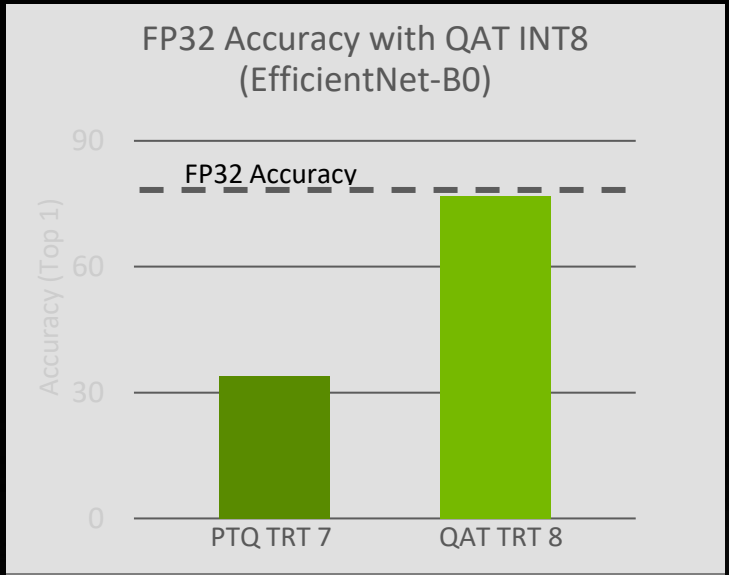


NEW MODEL ARCHITECTURES
EfficientNet, YOLOV4, Facial Landmark Detection, OCR

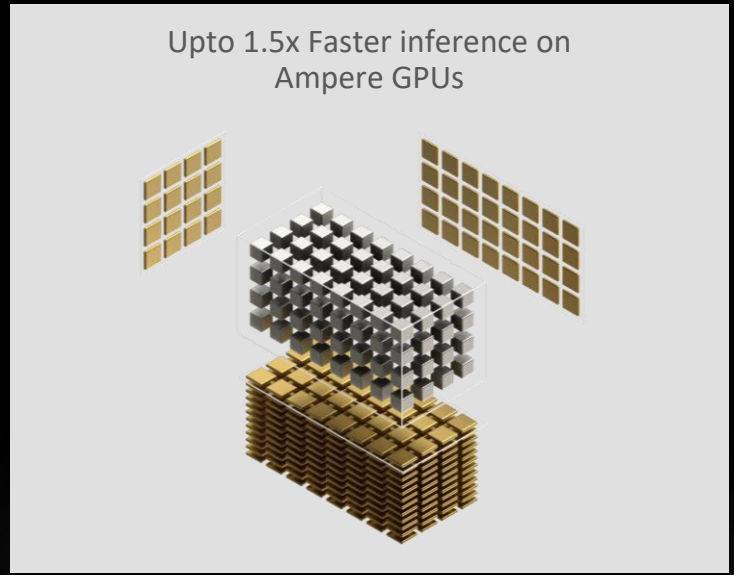


EASE OF USE
A Unified TLT launcher to manage and orchestrate your workflow

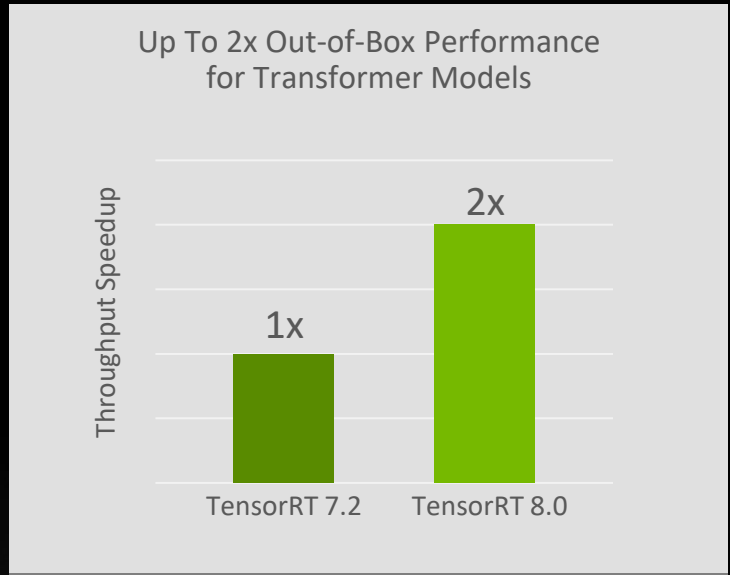
QAT, Sparsity | CNNs, RNNs, MLPs, Transformers



Quantization Aware Training delivers FP32 accuracy with INT8 precision

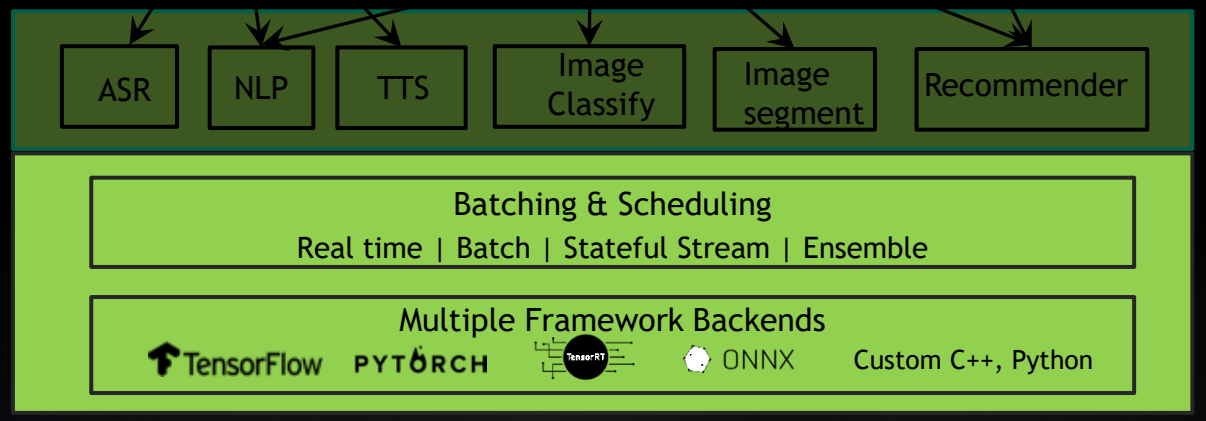


New Sparsity support for improved performance on Tensor Cores with INT8



New optimizations for Transformer based models like BERT

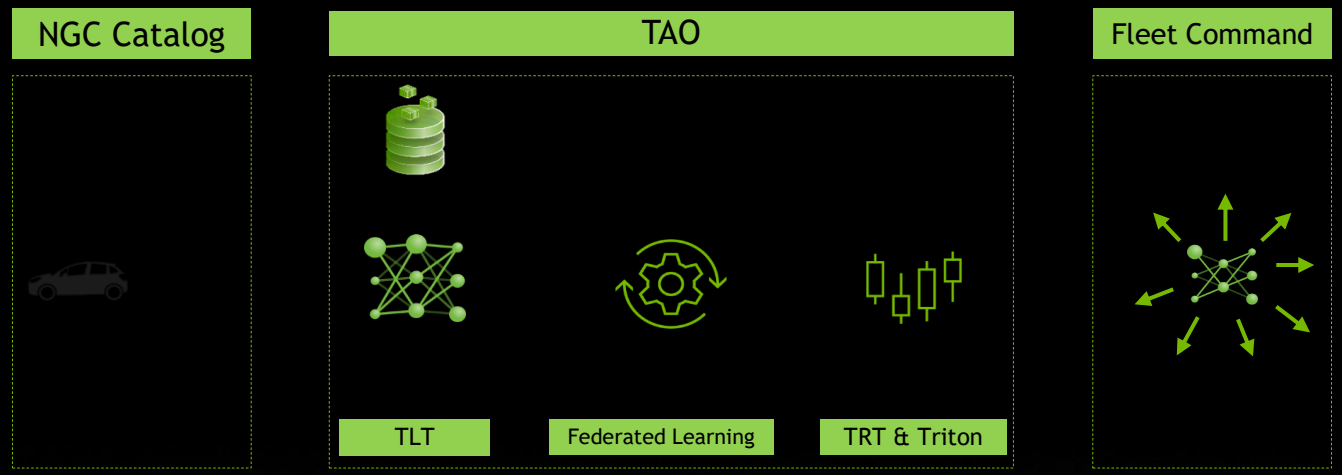
To learn more, signup for the NVIDIA Developer Program from
developer.nvidia.com/tensorrt



X86 CPU T4 GPU A100 GPU A100 MIG V100 GPU ARM CPU

Optimized for All Processors

Train | Adapt | Optimize



TRAIN

AI based framework simplifies AI development

Training specific models in hours v. months

ADAPT & OPTIMIZE

Increase model accuracy with federated learning

Optimize with TensorRT

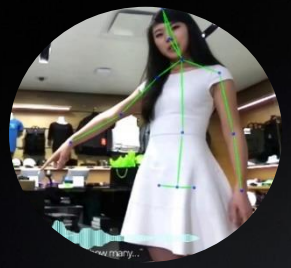
DEPLOY

Deploy from anywhere to everywhere

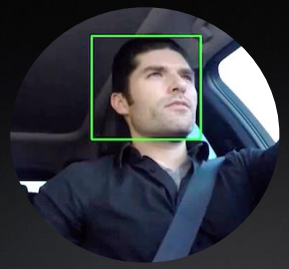
Effortless management and secure updates

NVIDIA TAO availability: 2H, 2021

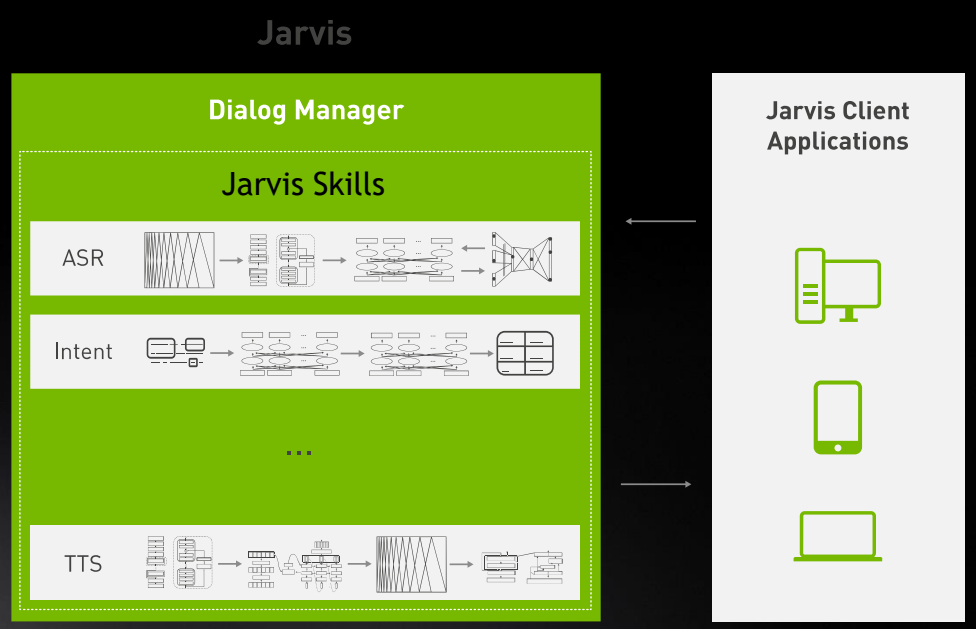
CONVERSATIONAL AI-JARVIS



RETAIL ASSISTANTS
12M Retail Stores



IN-CAR ASSISTANTS
75M New Cars per Year

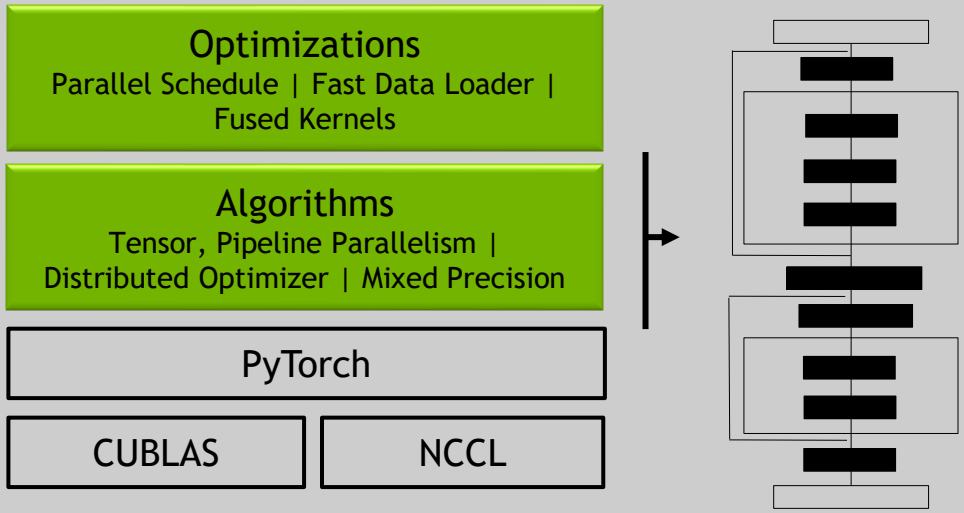


NVIDIA MEGATRON TRAINING FRAMEWORK FOR GIANT TRANSFORMERS

Train Multi-Trillion Parameter Models

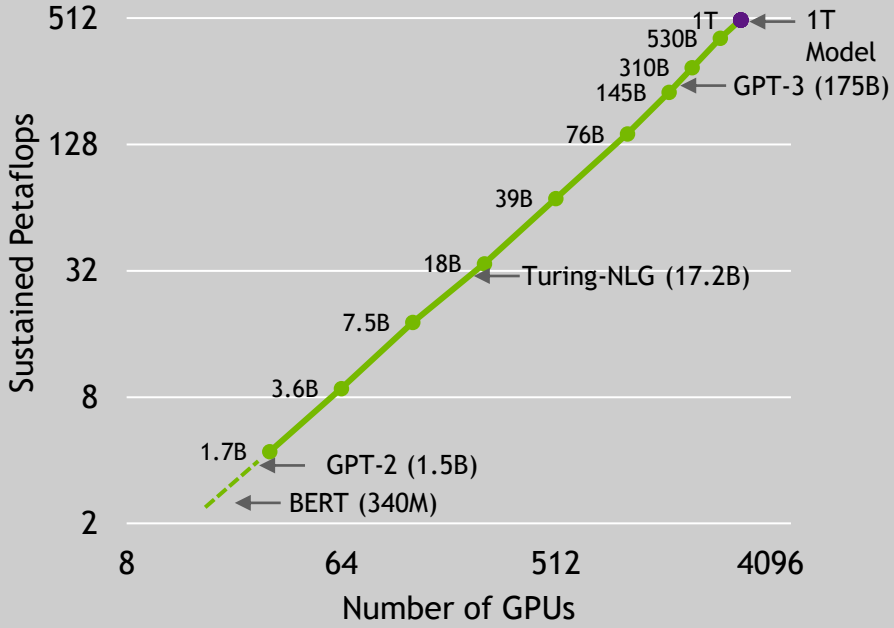
MEGATRON

Optimally Train Giant NLP Models On Large Clusters



LINEAR SCALING 1 TRILLION PARAMETERS

Over 50% Sustained Peak w/ 3000+ GPUS



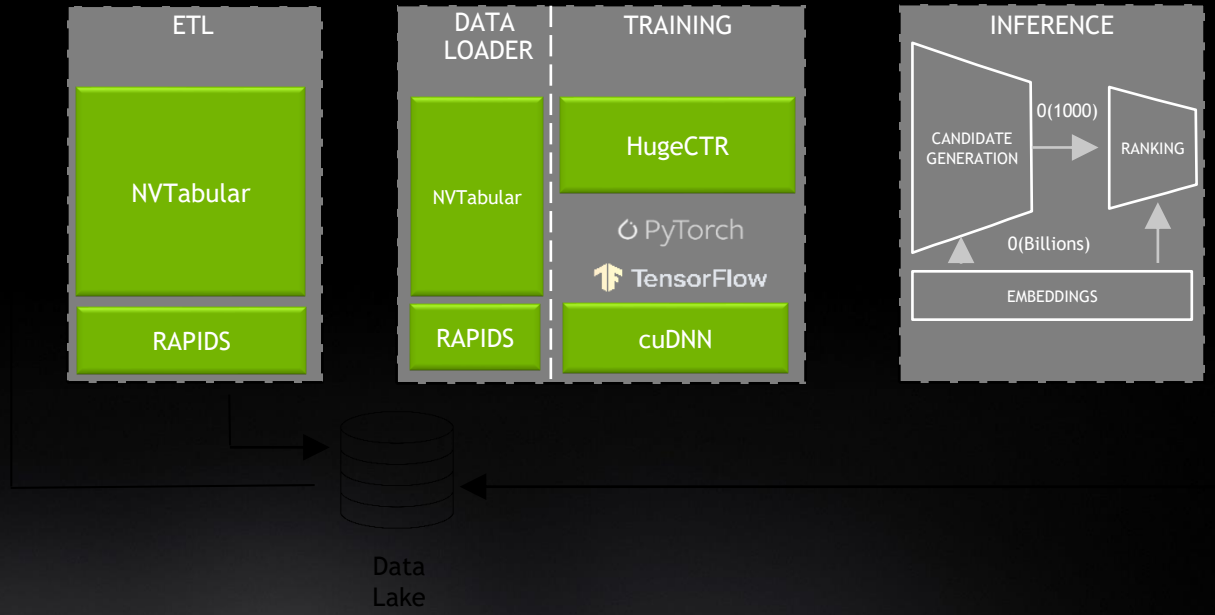
https://developer.nvidia.com/blog/language-modeling-using-megatron-a100-gpu/

Data label: Model size (parameters)
A few hundred iterations of model on # GPUs (x-axis) to estimate petaflops (not trained to convergence)

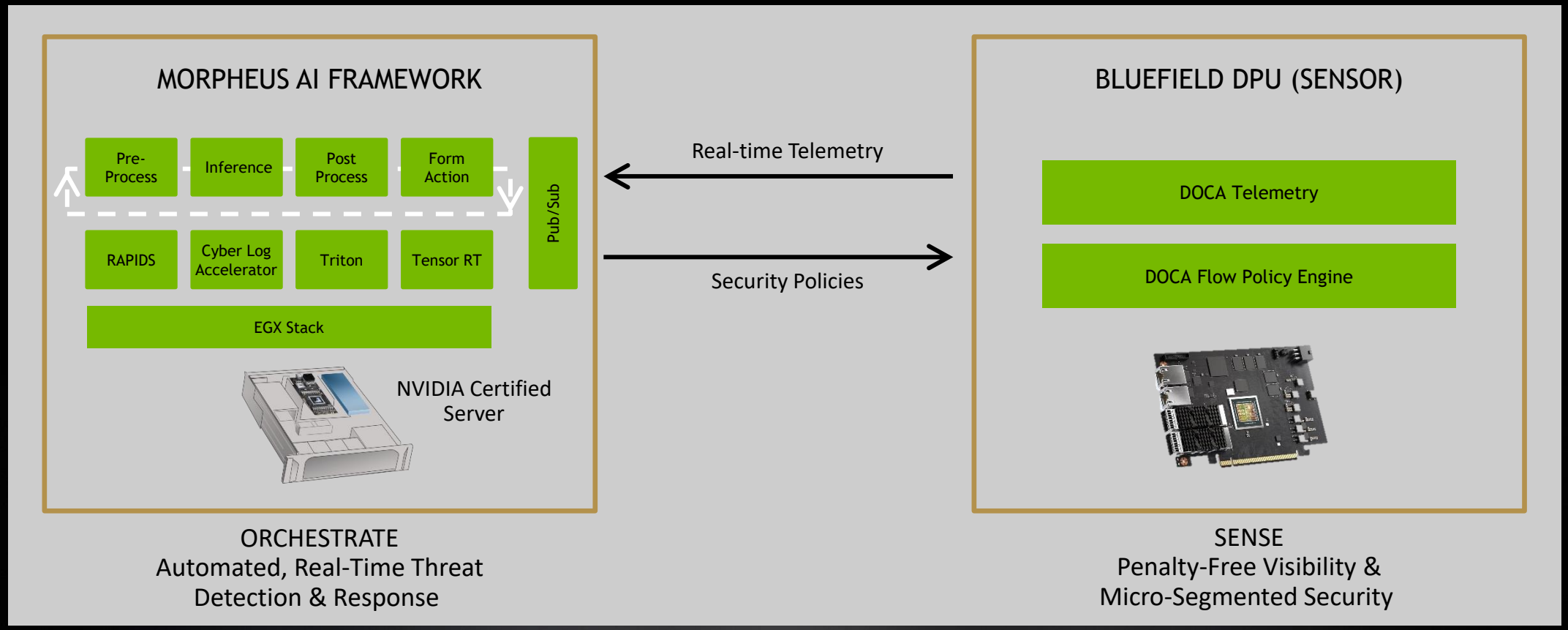
NVIDIA MERLIN END-TO-END ACCELERATED RECOMMENDER SYSTEM

MUSIC
[illegible] Songs Purchased 2020

VIDEO
1B Streaming Subscriptions



Open AI Compute Framework to Accelerate Cyber Ecosystem

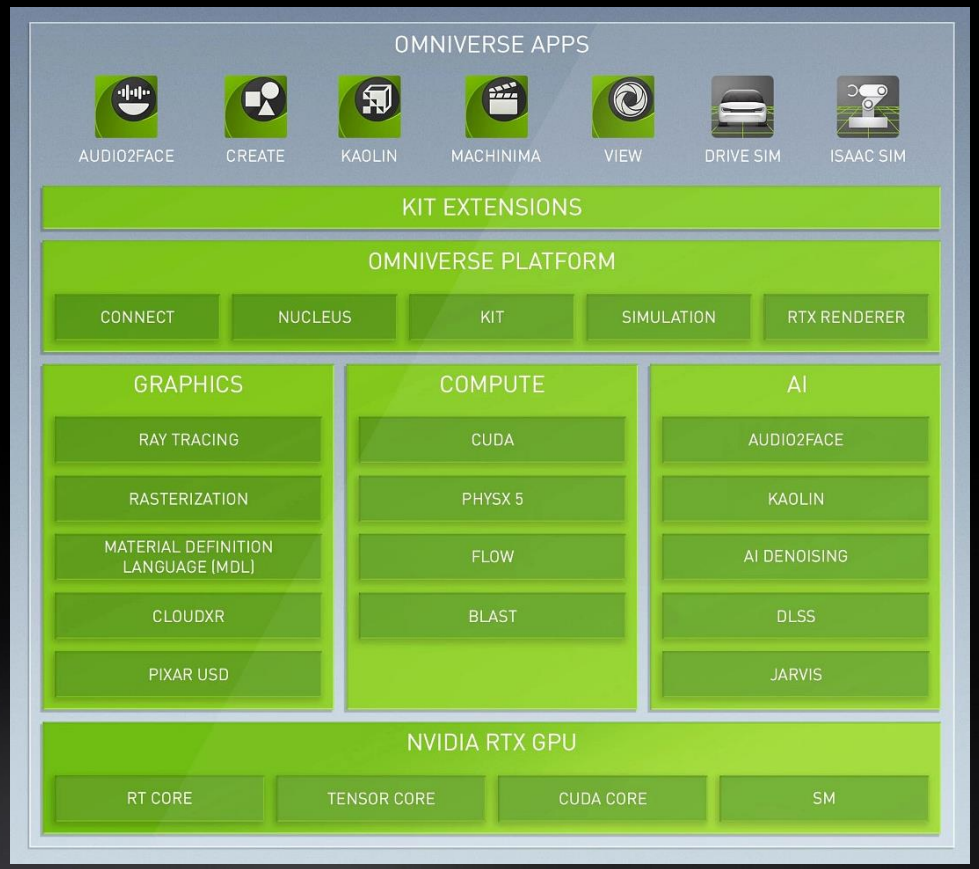


Morpheus: AI Inferencing for Cybersecurity Pipelines [S33291]

A PLATFORM BUILT FOR THE FUTURE

Multi-layer for multiple audiences

- Cloud-native
- Multi-GPU Enabled
- Open standards for cross-team, tool and workflow collaboration, built on Pixar's USD
- Scalable computing to address all workloads
- Works on all NVIDIA RTX™ solutions, from laptops to data centers



INDUSTRY USERS

DEVELOPERS, ISV PARTNERS

NVIDIA OMNIVERSE

ACCELERATE SCIENTIFIC DISCOVERIES WITH NVIDIA OMNIVERSE

ParaView Omniverse Connector

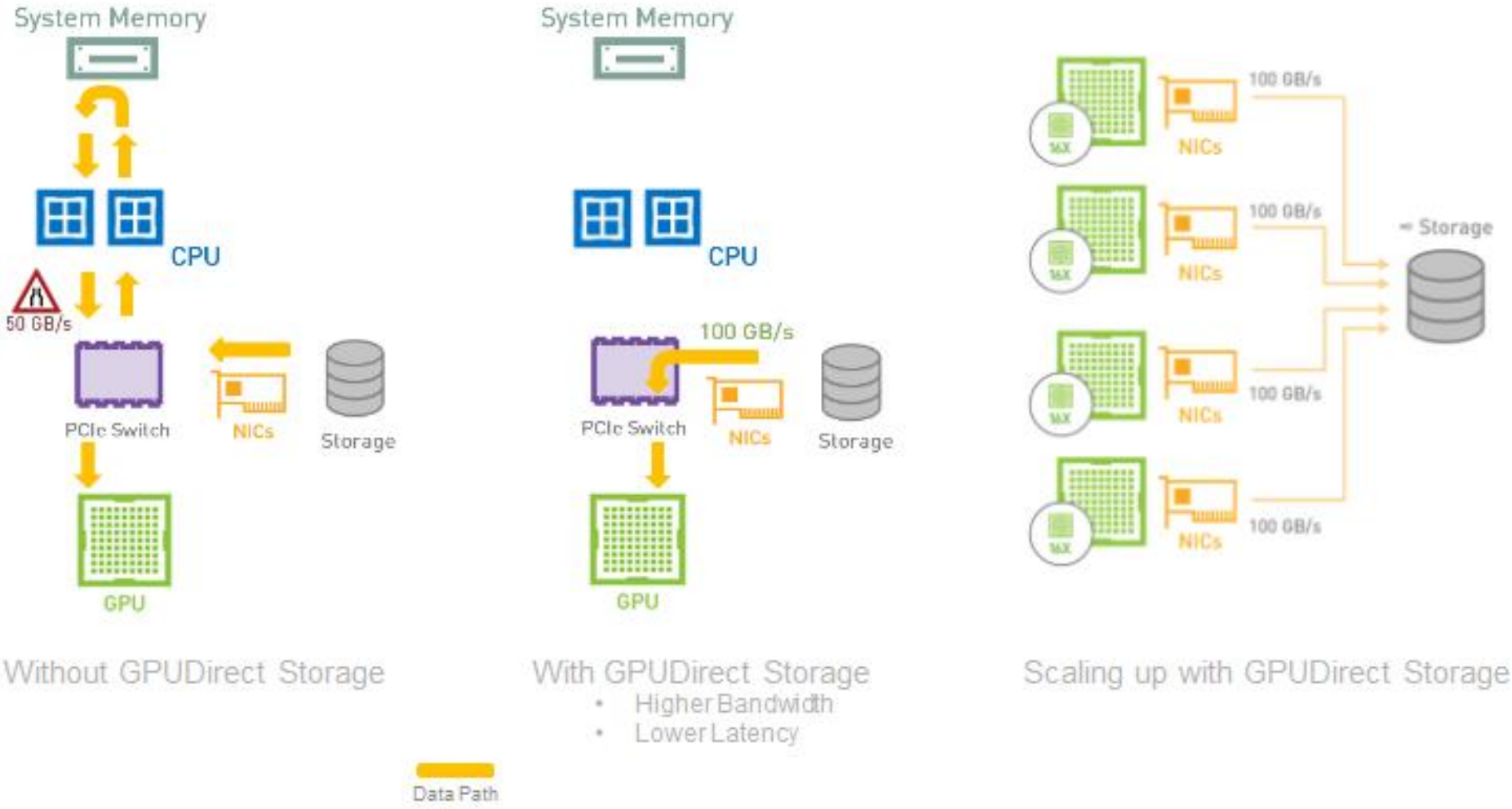
Import Assets from Various Sources

Real-time Collaboration Between Remote Teams

Photorealistic Scientific Visuals

Available Now in Open Beta

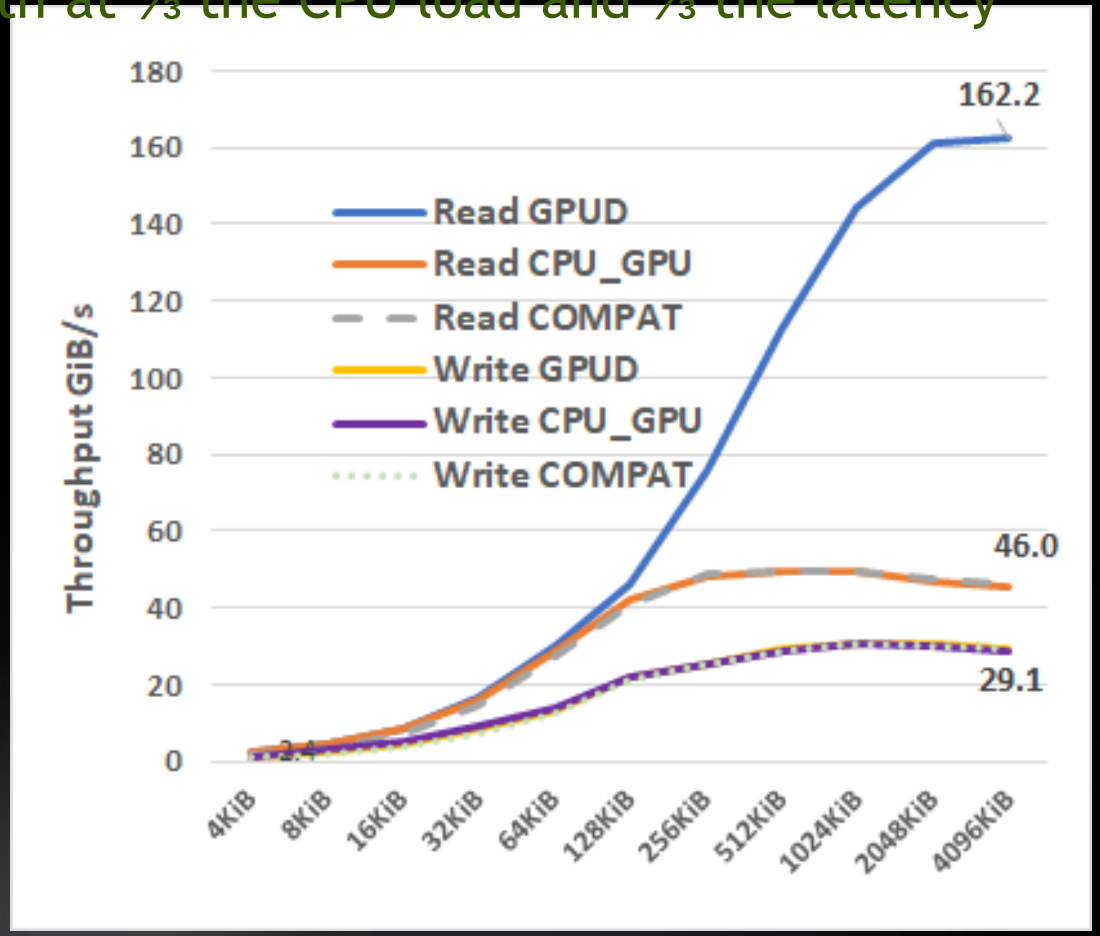
WHAT IS GPUDIRECT STORAGE?



GDS provides 3x the bandwidth at ⅓ the CPU load and ⅓ the latency

-
-
-

GDS control path
GDS avoids CPU's data movement
GDS reads 162 GiB/s @ 14% CPU, 19ms
CPU reads 50 GiB/s @ 49% CPU, 66ms
GDS writes 30 GiB/s @ 3% CPU, 12ms
CPU writes 29 GiB/s @ 4% CPU, 52ms



GA

BETA

Emerging

Early Exploration

NVIDIA DOCA

Enabling Broad DPU Partner Ecosystem

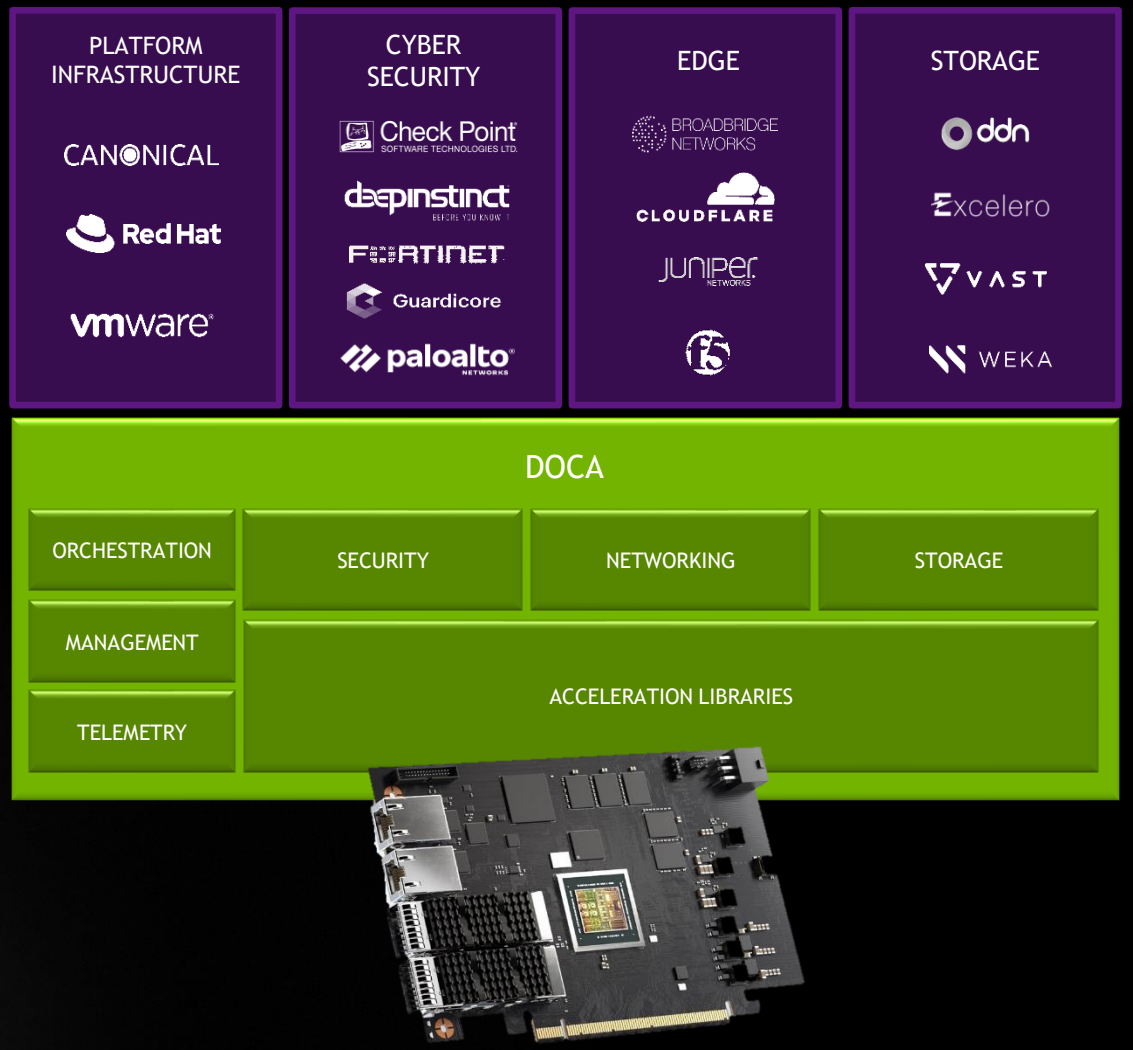
Software application framework for BlueField DPUs

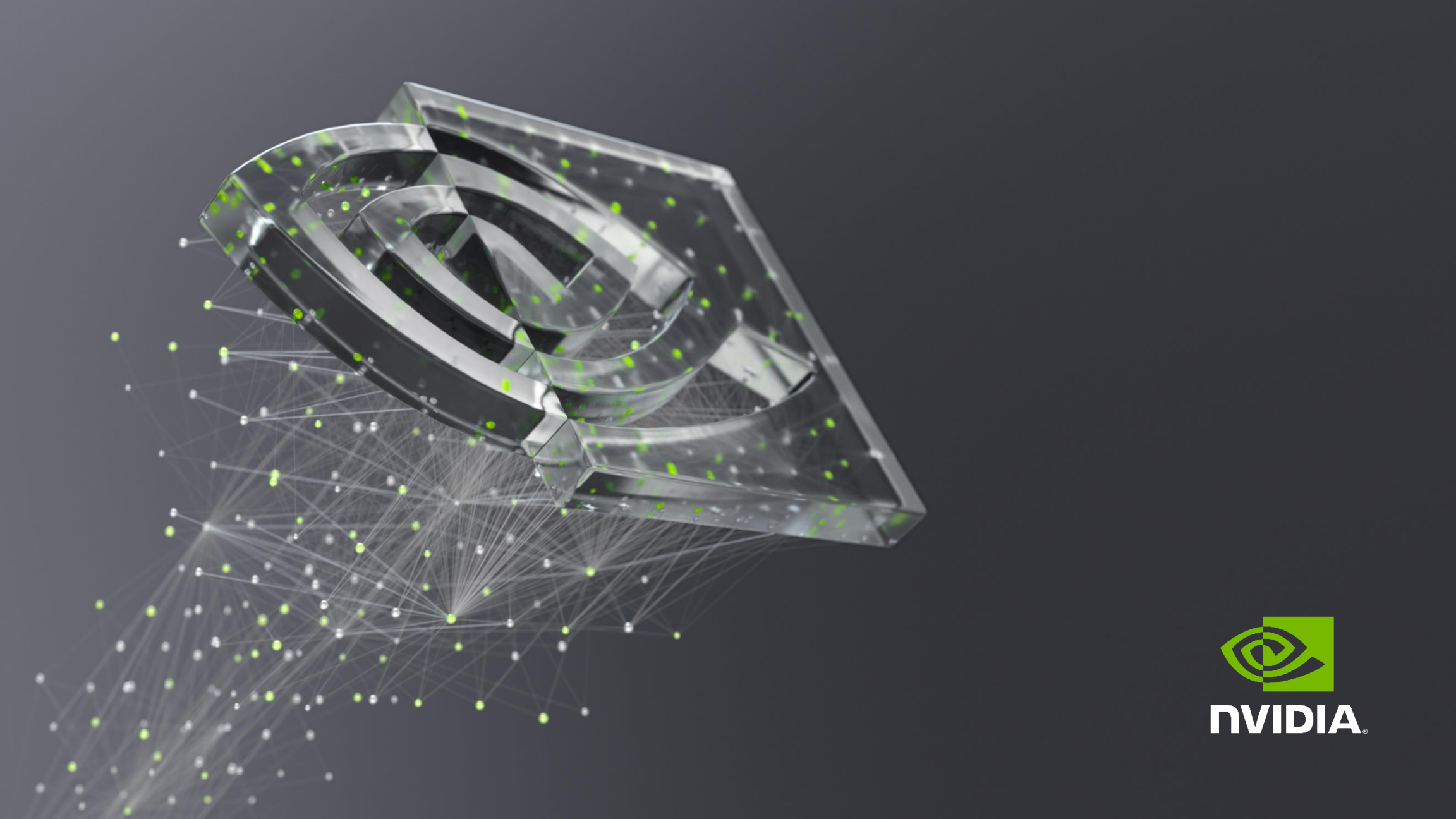
DOCA is for DPUs what CUDA is for GPUs

Protects developer investment for future DPUs

Certified reference applications, APIs & partner solutions

Rich partner ecosystem across industries and workloads




NVIDIA