



THE AI COMPUTING COMPANY

Romuald Josien



NVIDIA — A LEARNING MACHINE

NVIDIA has continuously reinvented itself over more than two decades.

Our invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics, and revolutionized parallel computing. More recently, GPU computing ignited the era of AI.

NVIDIA is a “learning machine” that constantly evolves by adapting to new opportunities that are hard to solve, that only we can tackle, and that matter to the world.

Founded in 1993 | Jensen Huang, Founder & CEO | 13,000 employees | \$11.7B in FY19



NVIDIA

GRAPHICS

HPC

AI



GAMING



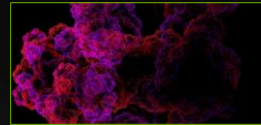
DESIGN



RENDERING



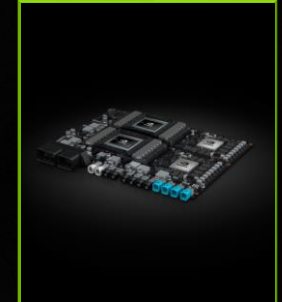
SUPERCOMPUTING



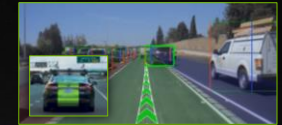
AI TRAINING



AI INFERENCE



ROBOTICS

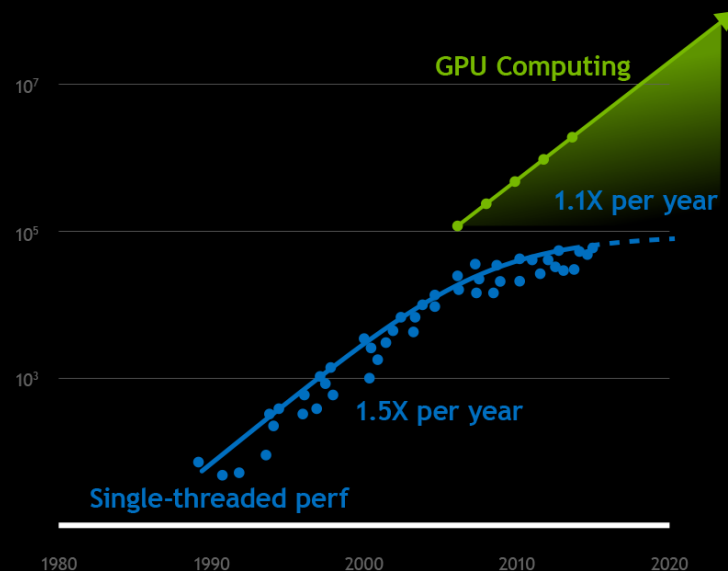


TWO FORCES SHAPING COMPUTING

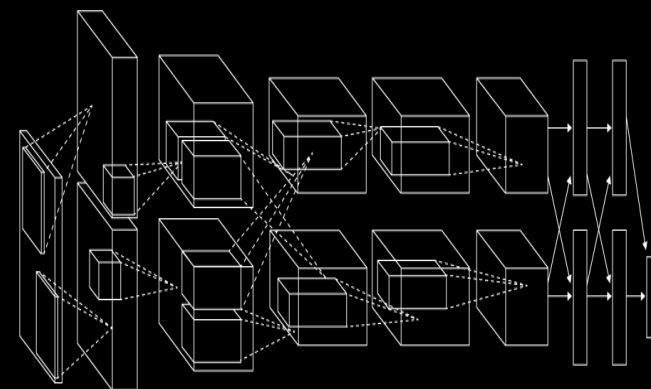
For 30 years, the dynamics of Moore's law held true. But now CPU scaling is slowing while the demand for computing power surges ahead.

With AI, machines can learn. AI can solve grand challenges that have been beyond human reach. But it must be fueled by massive compute power.

Accelerated computing is the path forward beyond Moore's law, delivering 1,000X computing performance every 10 years.



40 YEARS OF CPU TREND DATA

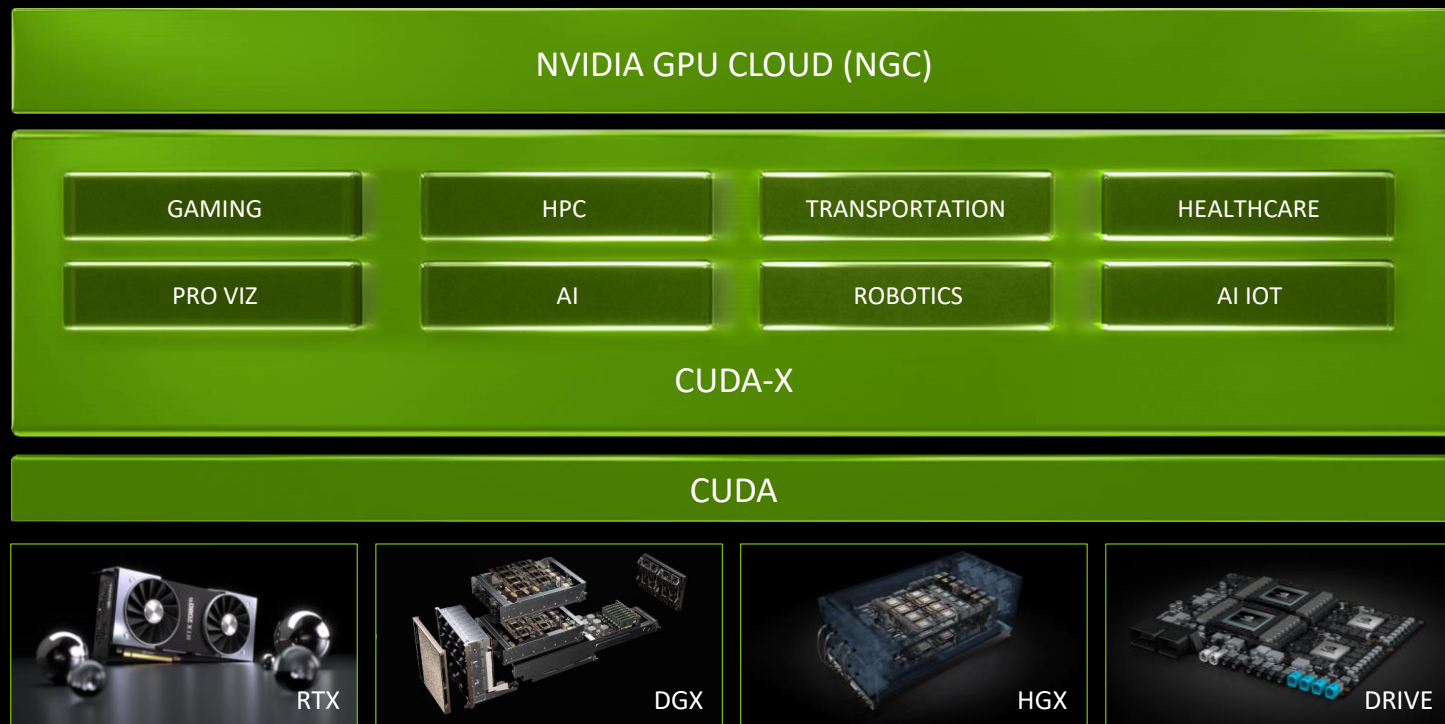


ALEXNET: THE SPARK OF THE MODERN AI ERA

ONE ARCHITECTURE

NVIDIA is an accelerated computing company. It starts with a highly specialized parallel processor called the GPU and continues through system design, system software, algorithms, and optimized applications.

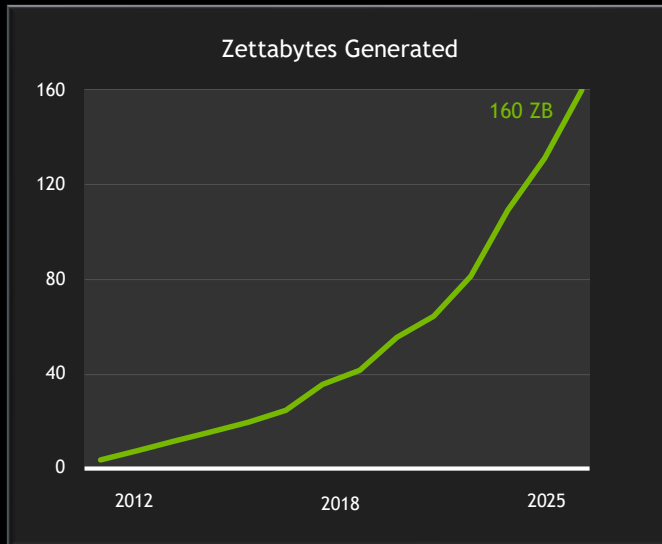
CUDA-X® is a suite of software libraries that accelerate applications for our growth markets — from gaming to transportation to healthcare — all based on a common CUDA architecture supported by more than 1.2 million developers today.



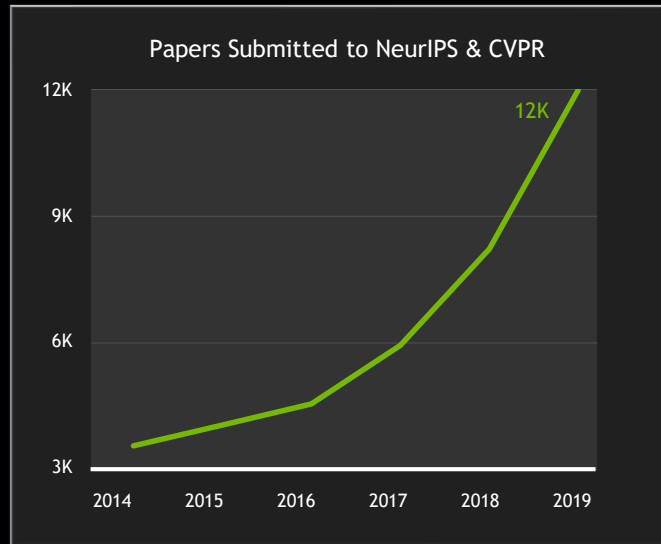
A RACE FOR PERFORMANCE

EXPONENTIAL GROWTH IN COMPUTING DEMAND

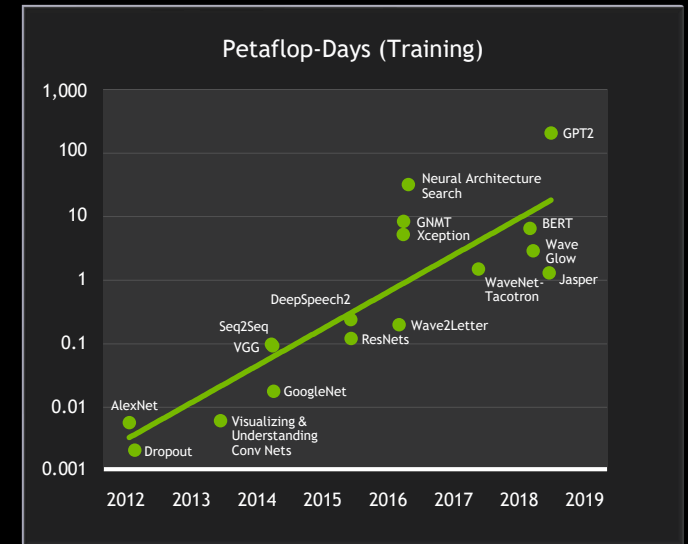
DATA SIZE GROWING



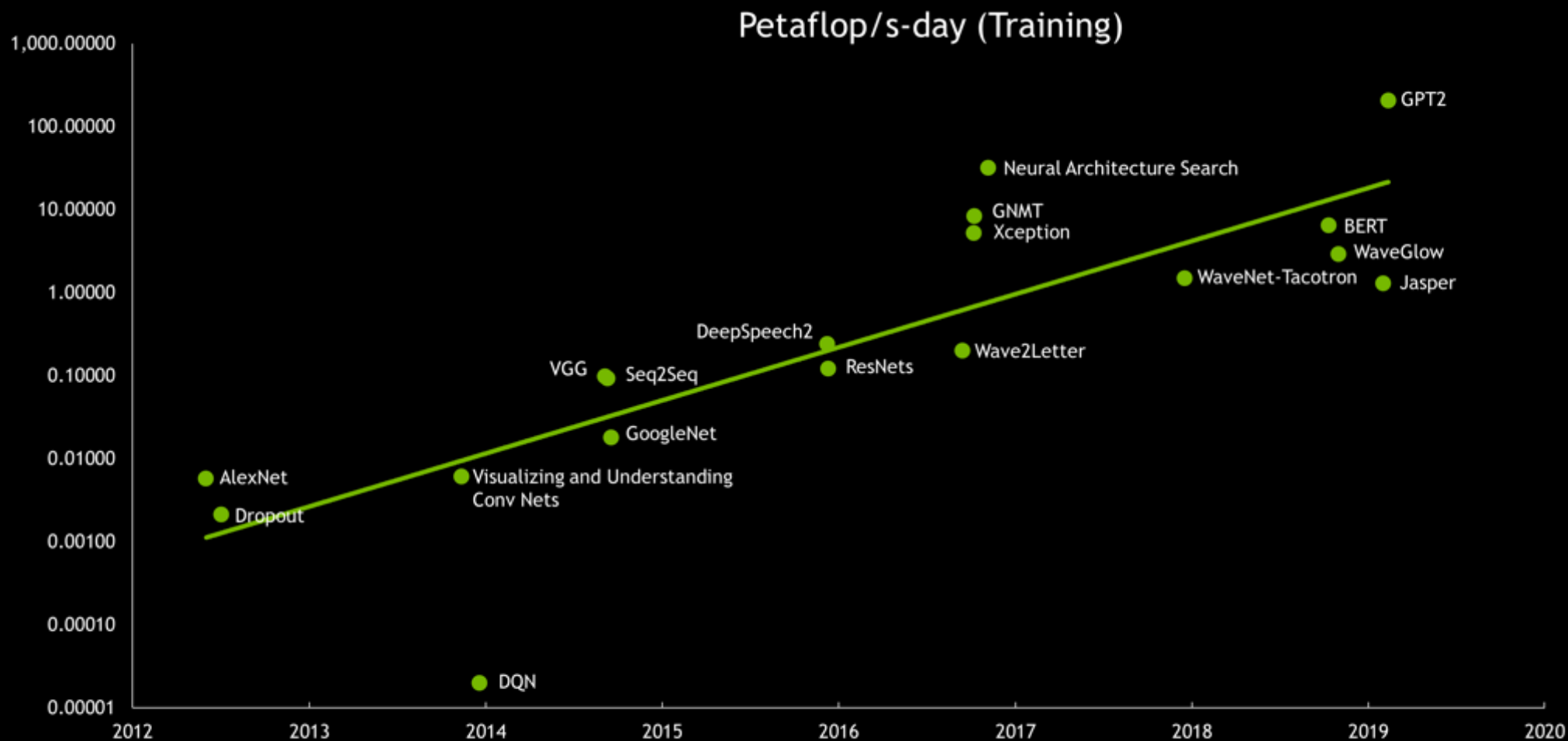
AI RESEARCH GROWING



AI MODEL COMPLEXITY GROWING

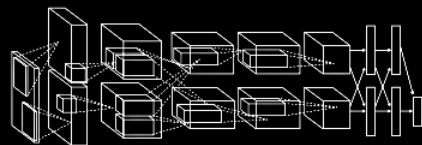


NETWORK COMPLEXITY IS EXPLODING



A CAMBRIAN EXPLOSION OF DL MODELS

CONVOLUTIONAL NETWORKS



Encoder/Decoder



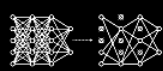
ReLu



BatchNorm



Concat

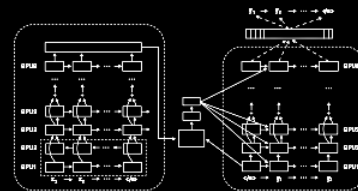


Dropout



Pooling

RECURRENT NETWORKS



LSTM



GRU



Beam Search



WaveNet



CTC



Attention

GENERATIVE ADVERSARIAL NETWORKS



3D-GAN



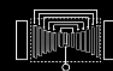
MedGAN



Conditional GAN

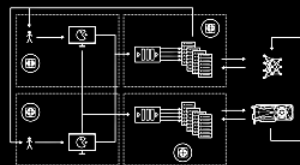


Coupled GAN

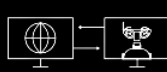


Speech Enhancement GAN

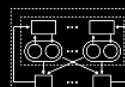
REINFORCEMENT LEARNING



DQN

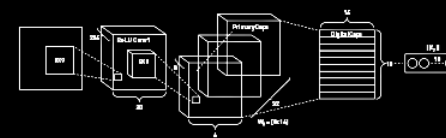


Simulation



DDPG

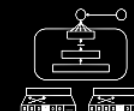
NEW SPECIES



Capsule Nets



Mixture of Experts



Neural
Collaborative
Filtering

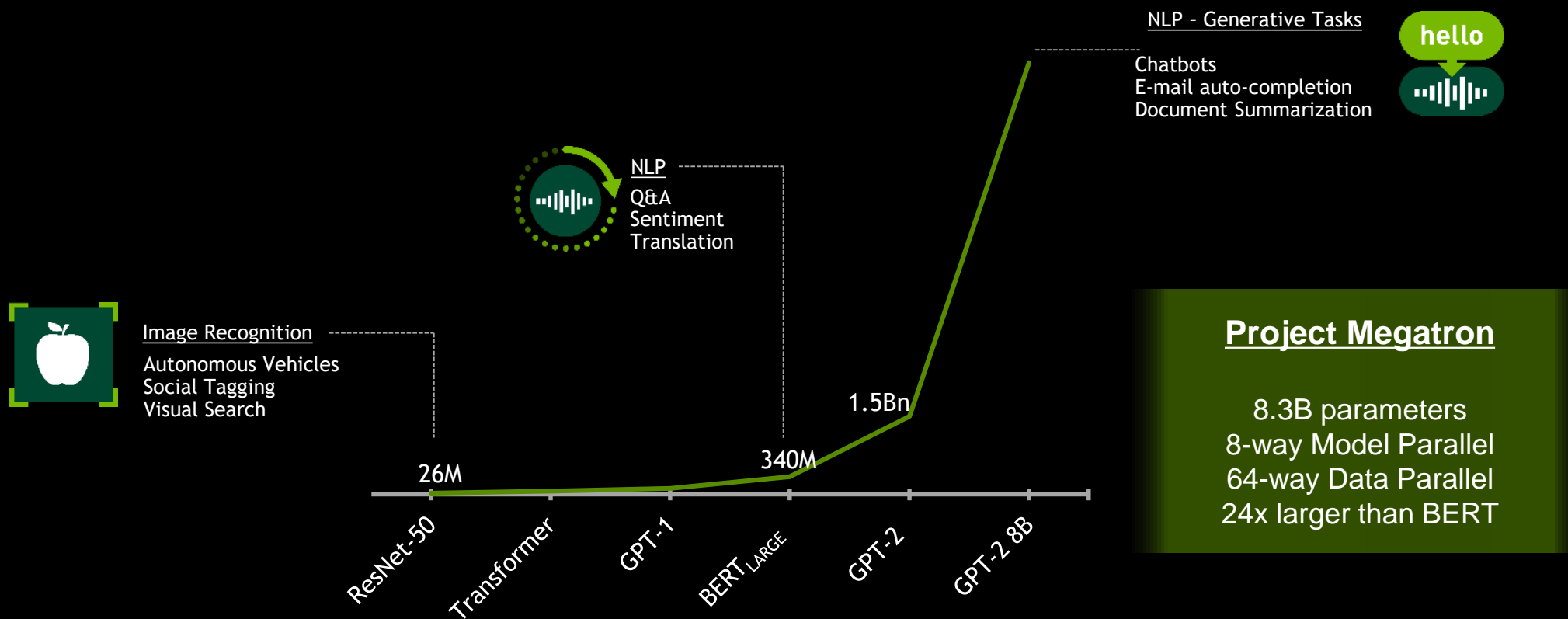


Block Sparse
LSTM

AI INNOVATION IS SHIFTING, AND GROWING

Next-Level Use-Cases Require Gigantic Models

Number of Parameters by Network



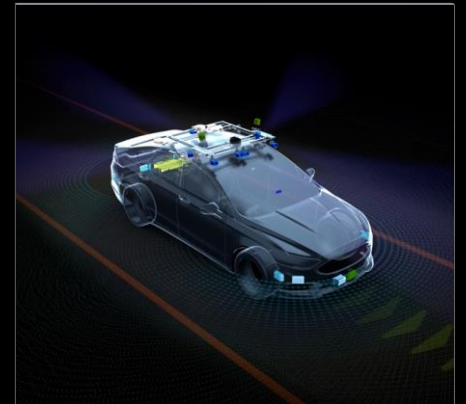
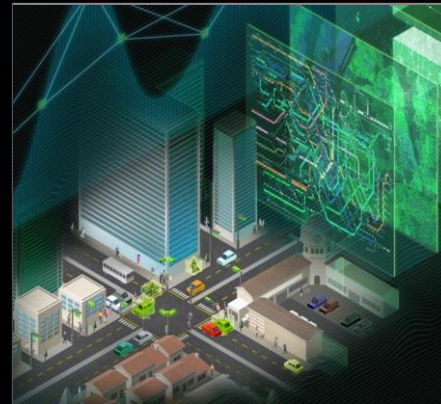
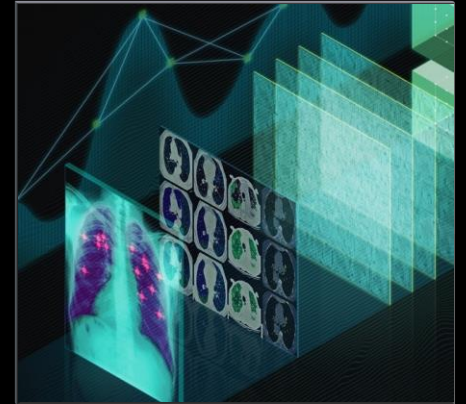
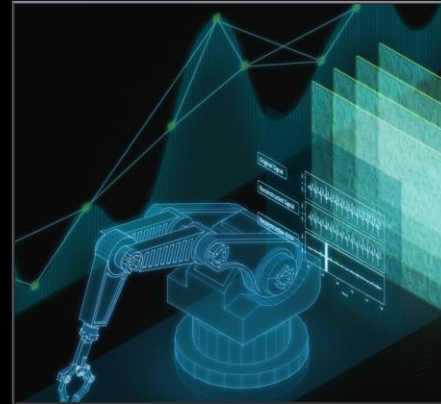
AI LEADERSHIP STARTS WITH AI COMPUTING LEADERSHIP

Researchers racing to advance AI for the world's largest industries - auto, healthcare, manufacturing

Increasingly complex AI models and larger data size demand powerful computers

Iteration speed and time-to-train fuels innovation

NVIDIA created DGX SuperPOD to serve as the essential instrument of AI research

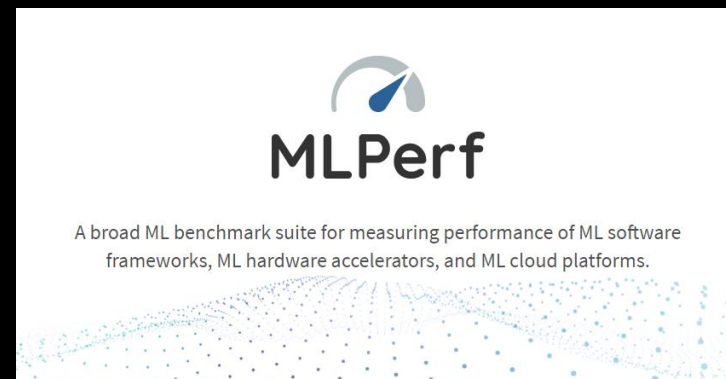


NVIDIA BREAKS RECORDS IN AI PERFORMANCE

Both On At Scale And Per Accelerator

Record Type	Benchmark	Record
Max Scale (Minutes To Train)	Object Detection (Heavy Weight) Mask R-CNN	18.47 Mins
	Translation (Recurrent) GNMT	1.8 Mins
	Reinforcement Learning (MiniGo)	13.57 Mins
Per Accelerator (Hours To Train)	Object Detection (Heavy Weight) Mask R-CNN	25.39 Hrs
	Object Detection (Light Weight) SSD	3.04 Hrs
	Translation (Recurrent) GNMT	2.63 Hrs
	Translation (Non-recurrent)Transformer	2.61 Hrs
	Reinforcement Learning (MiniGo)	3.65 Hrs

INDUSTRY WIDE BENCHMARK SUITE FOR AI PERFORMANCE



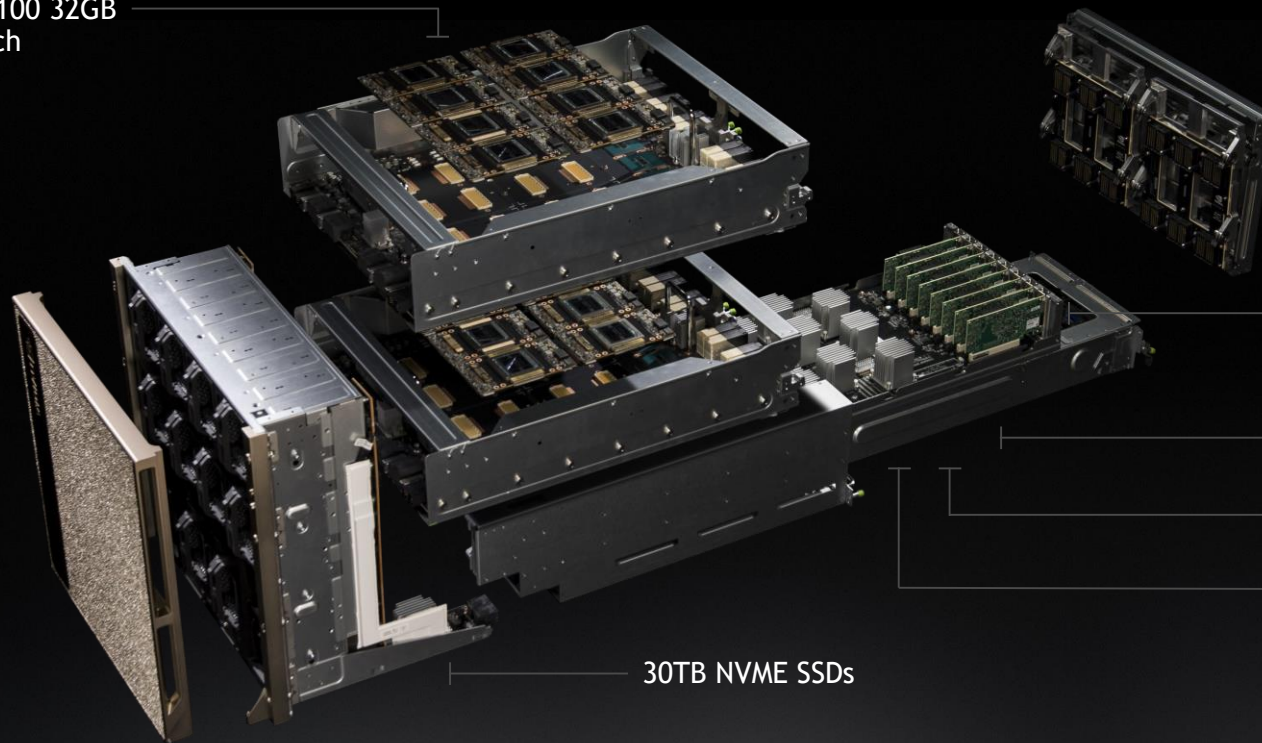
<https://mlperf.org/>

Per Accelerator comparison using reported performance for MLPerf 0.6 NVIDIA DGX-2H (16 V100s) compared to other submissions at same scale except for MiniGo where NVIDIA DGX-1 (8 V100s) submission was used | MLPerf ID Max Scale: Mask R-CNN: 0.6-23, GNMT: 0.6-26, MiniGo: 0.6-11 | MLPerf ID Per Accelerator: Mask R-CNN, SSD, GNMT, Transformer: all use 0.6-20, MiniGo: 0.6-10

NVIDIA DGX-2

The World's Most Powerful AI Computer

16x Tesla V100 32GB
12x NVSwitch



NVLink Plane Card

10x EDR IB/100 GigE

2x Xeon Platinum

1.5TB System Memory

PCIe Switch Complex

30TB NVME SSDs

2 PFLOPS | 512GB HBM2 | 10kW | 350 lbs

TIME MACHINE FOR AI

Smashing Time to Train From 8 Hours to 80 Seconds On V100

2015

K80 | CUDA®

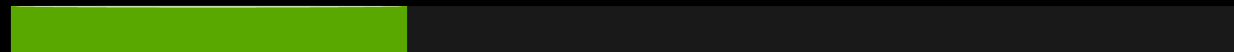
36,000 Mins (25 Days)



2017

NVIDIA® DGX-1™ | Volta | Tensor Cores

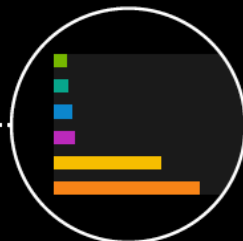
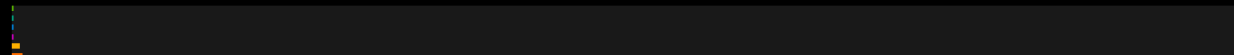
480 Mins (8 Hours)



2019

NVIDIA DGX SuperPOD™ | NVIDIA NVSwitch™ | Mellanox InfiniBand

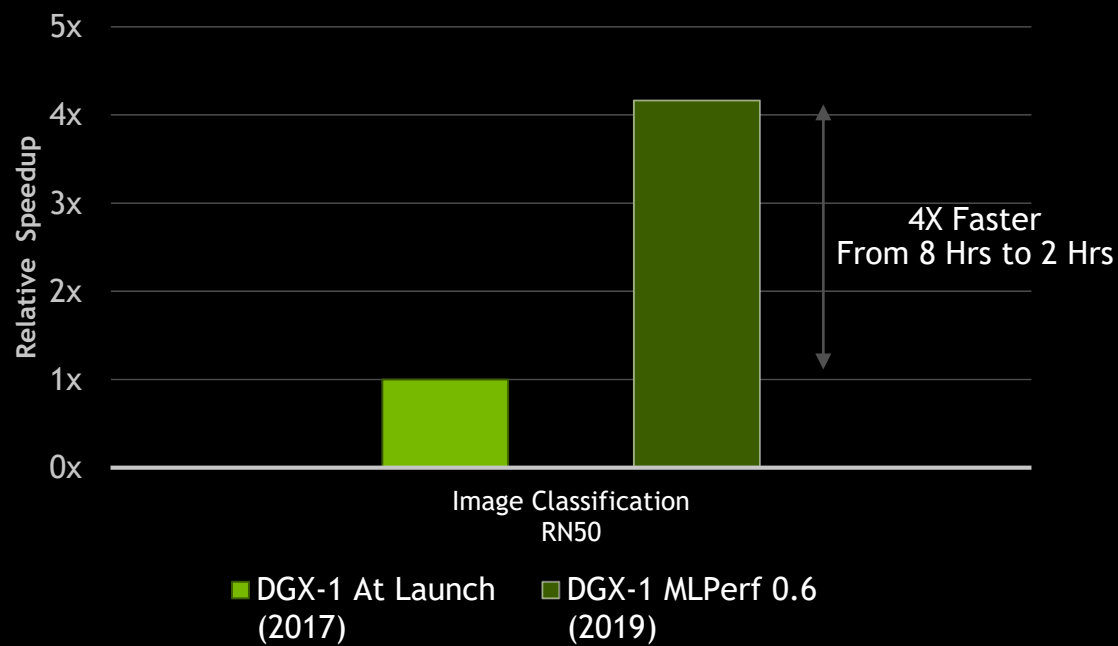
8 HRS TO
80 SECS



- **80 Secs (1.33 Mins)** (ResNet-50, Image Classification)
- **1.59 Mins** (Transformer, Non-Recurrent Translation)
- **1.99 Mins** (GNMT, Recurrent Translation)
- **2.23 Mins** (SSD, Lightweight Object Detection)
- **13.57 Mins** (Reinforcement Learning, Mini-Go)
- **18.47 Mins** (Mask R-CNN, Heavyweight Object Detection)

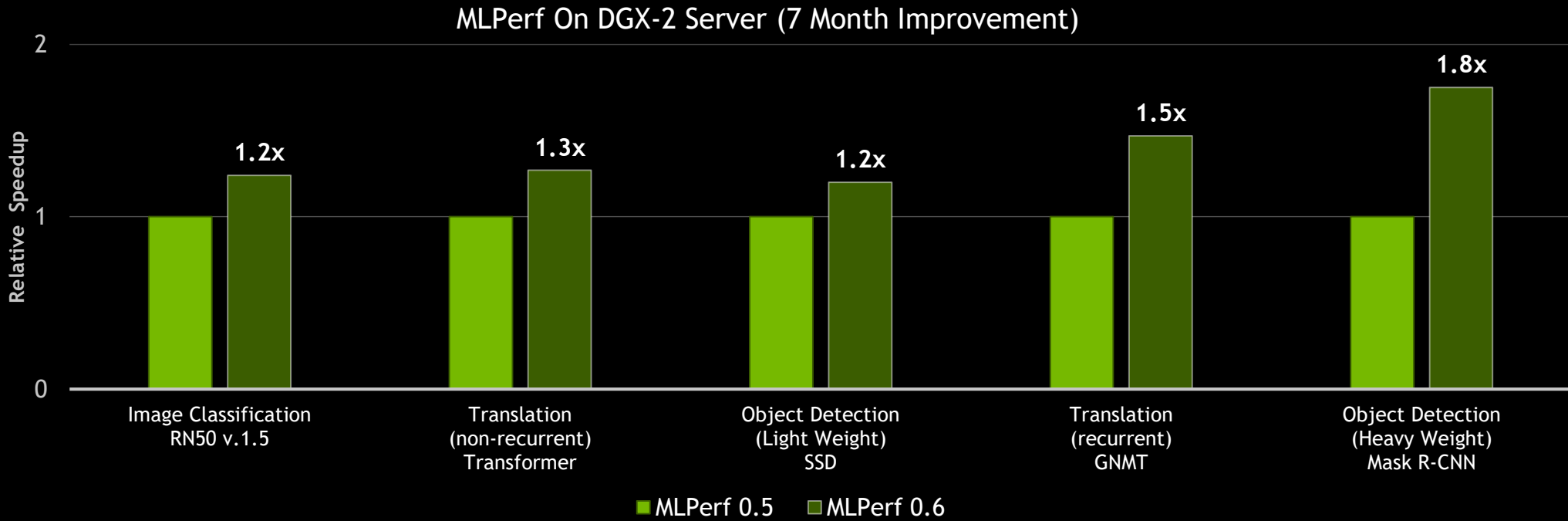
4X MORE PERFORMANCE, SAME SERVER

Rapid Software Innovation Delivers Continuous Improvements



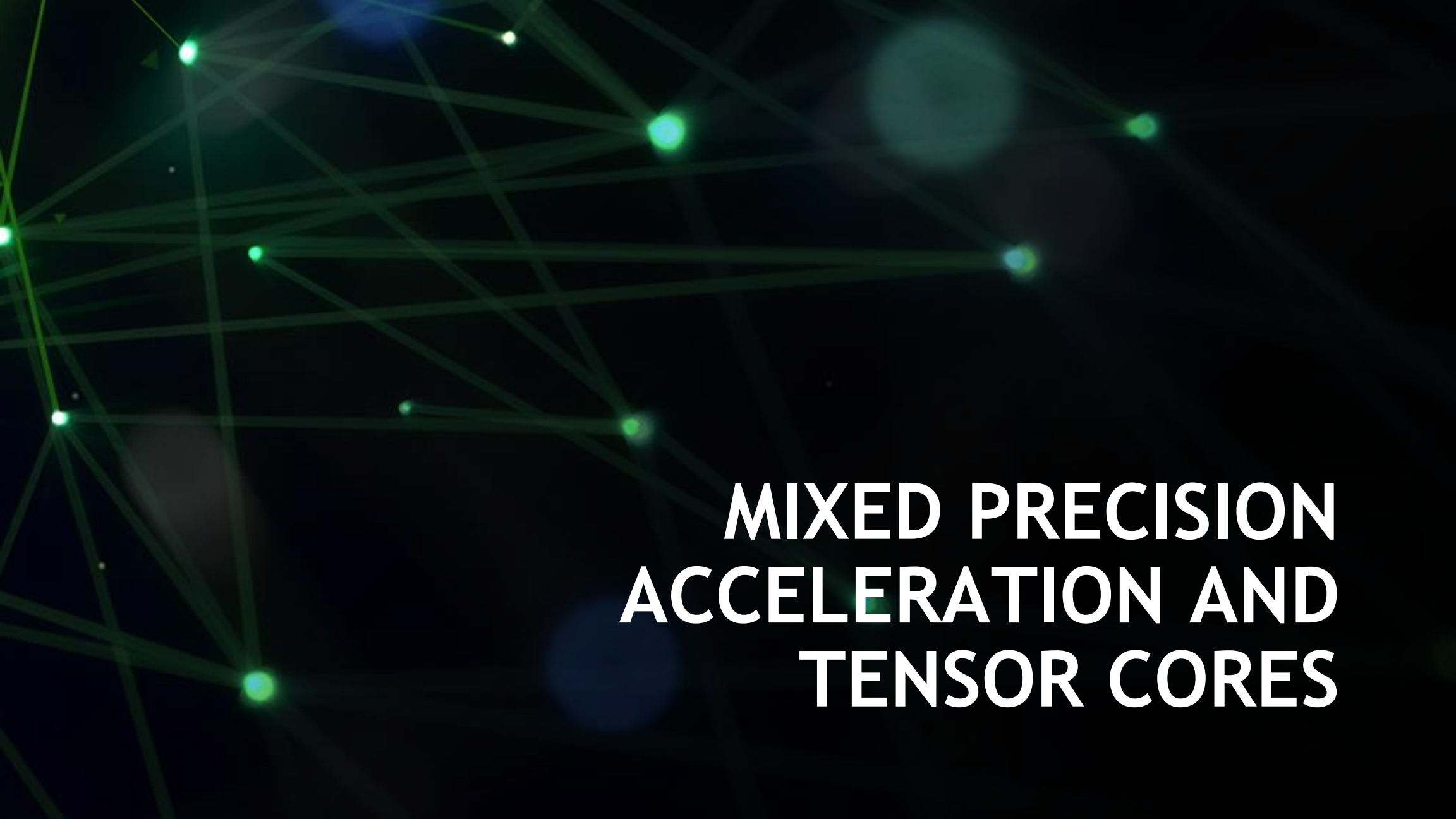
UP TO 80% MORE PERFORMANCE ON SAME SERVER

Software Innovation Delivers Continuous MLPerf Improvements



Comparing the throughput of a single DGX-2H server on a single epoch (Single pass of the dataset through the neural network) | MLPerf ID 0.5/0.6 comparison: ResNet50 v1.5: 0.5-20/0.6-30 | Transformer: 0.5-21/0.6-20 | SSD: 0.5-21/0.6-20 | GNMT: 0.5-19/0.6-20 | Mask R-CNN: 0.5-21/0.6-20

HOW TO REACH THIS PERFORMANCE



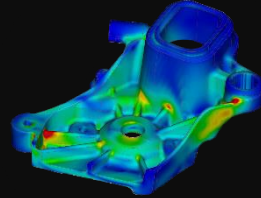
MIXED PRECISION ACCELERATION AND TENSOR CORES

TENSOR CORE GPU FUSES HPC & AI COMPUTING



VOLTA TENSOR CORE GPU

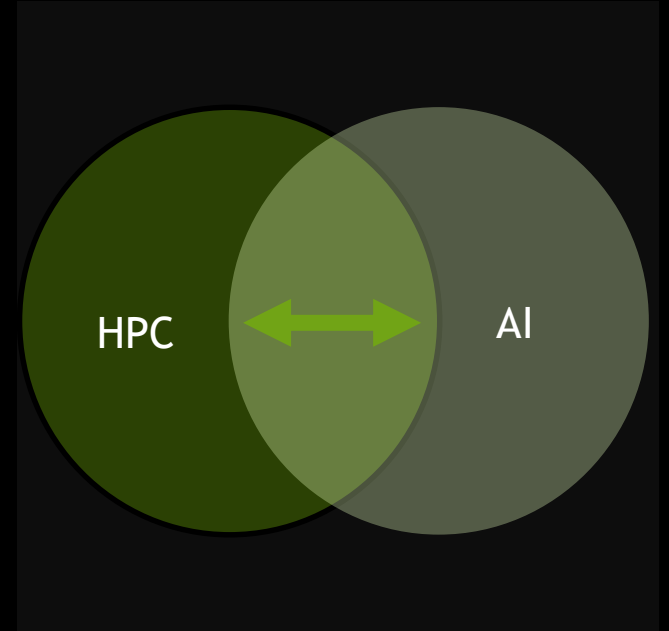
HPC (Simulation) - FP64, FP32



AI (Deep Learning) - FP16, INT8



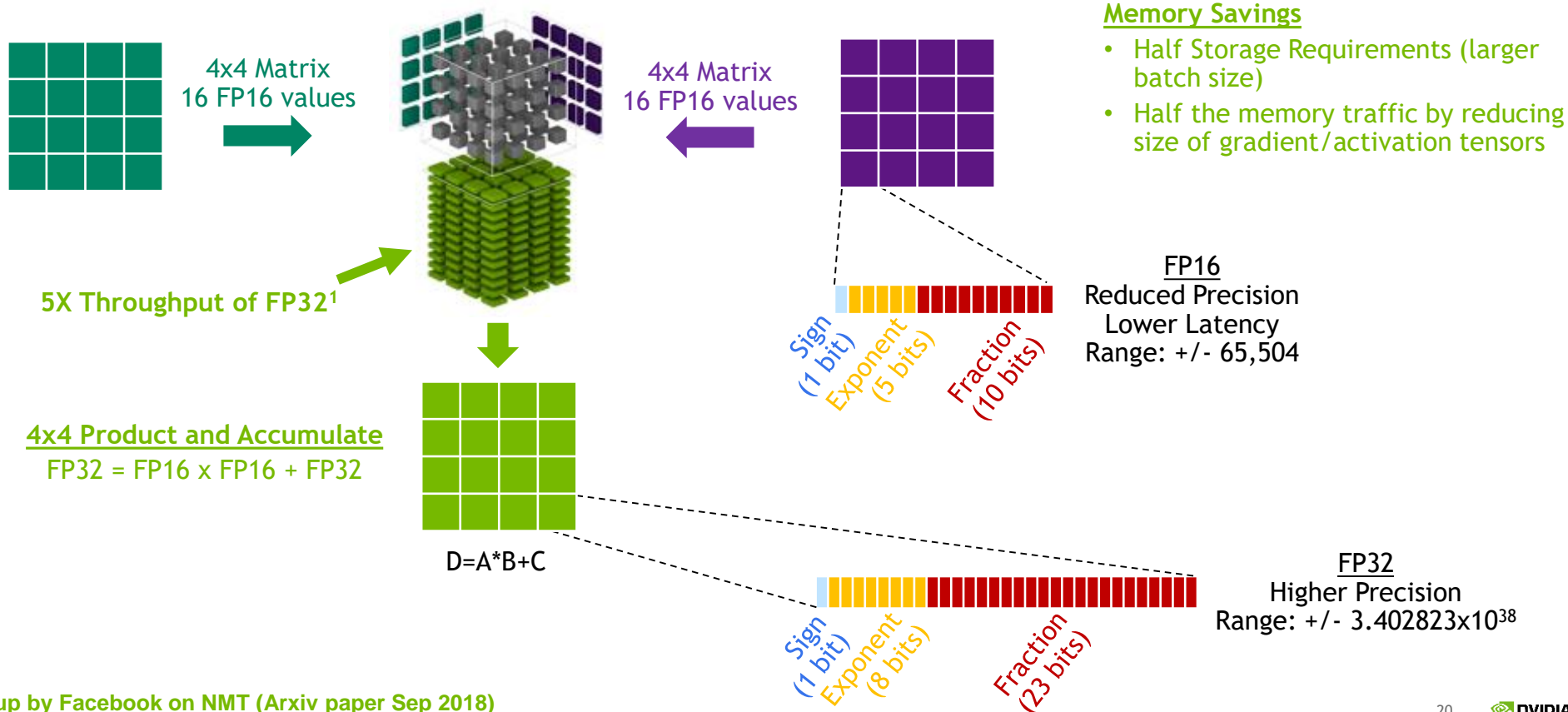
MULTI-PRECISION
COMPUTING



FUSION OF HPC & AI

TENSOR CORES BUILT FOR AI AND HPC

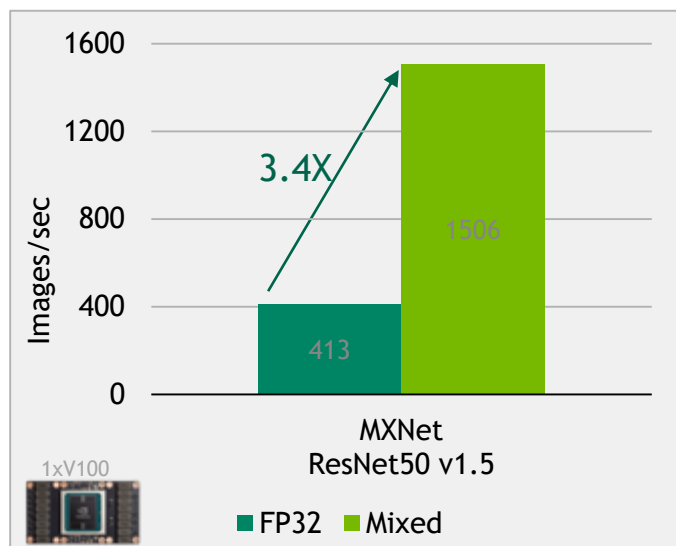
Mixed Precision Accelerator - Delivering Up To 5X Throughput of FP32¹



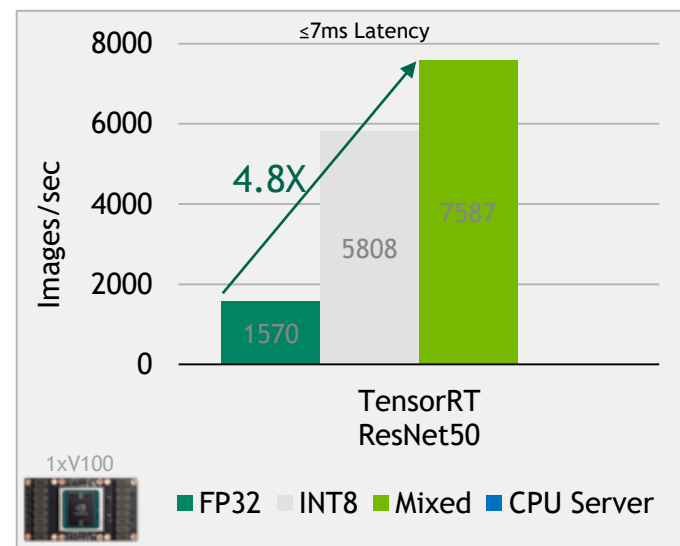
¹Fastest Tensor Core Speedup by Facebook on NMT (Arxiv paper Sep 2018)
<https://arxiv.org/pdf/1806.00187.pdf>

TENSOR CORE AUTOMATIC MIXED PRECISION

3x Speedup With Just One Line of Code



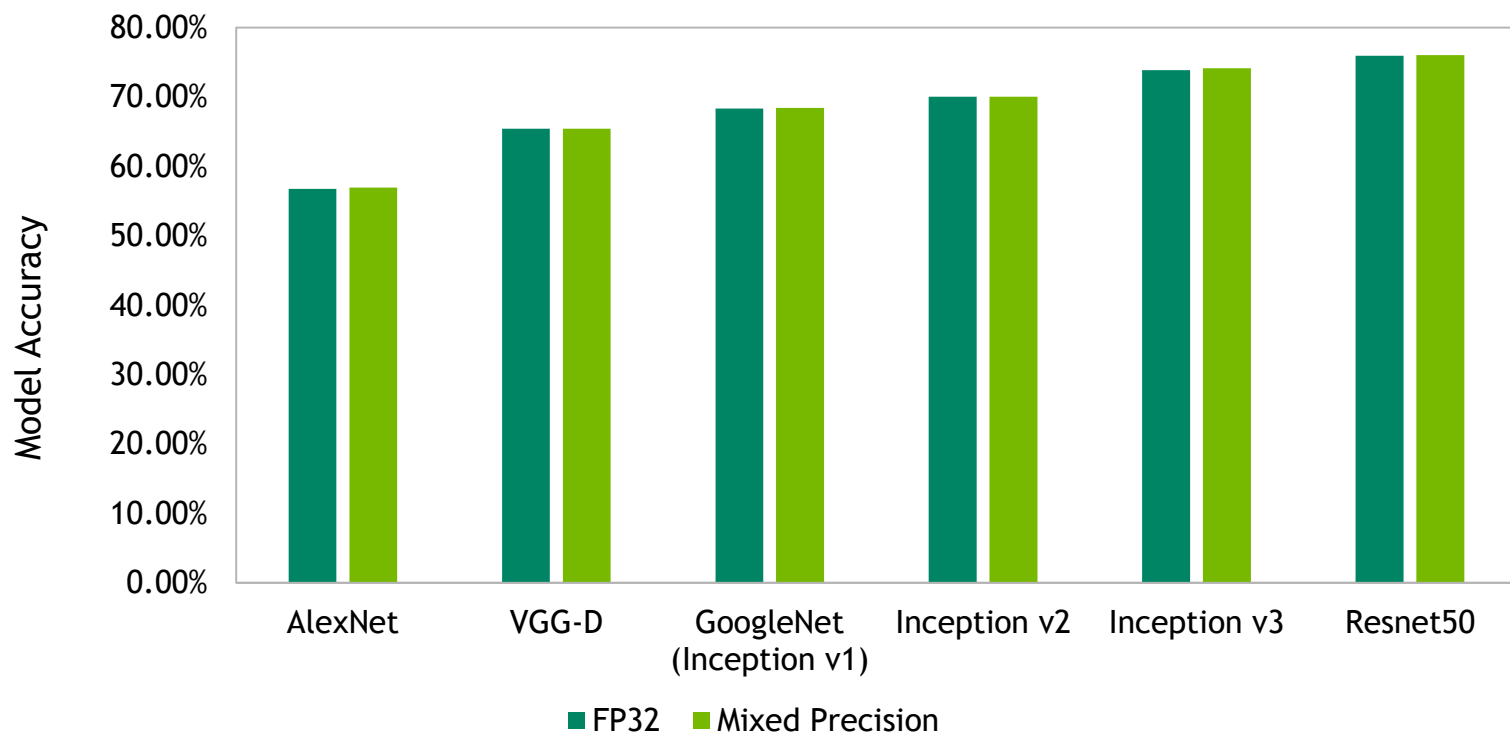
TRAINING SPEEDUP OVER 3X



INFERENCE SPEEDUP OVER 4X

MIXED PRECISION MAINTAINS ACCURACY

Benefit From Higher Throughput Without Compromise



ACTIVATING MIXED PRECISION WITH EASE

Use NVIDIA Optimized Models...



Computer Vision



Speech



Recommender



GANs

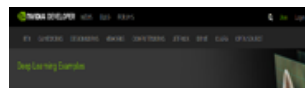
...Available From:



NVIDIA GPU Cloud
Docker Pull Containers



Tensor Core
Journey Page



NVIDIA Deep Learning
Examples



Github
Repository

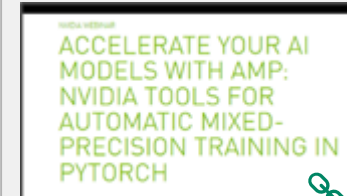


PYTORCH



Or Accelerate Your Own Models

Webinar

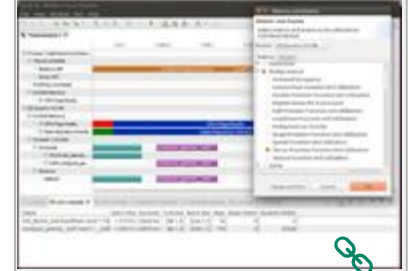


Github



Automatic Mixed Precision
Just Add 2 Lines of Code

Nvprof Nsight Compute



Profiler Tools
Individual Kernel
Optimization



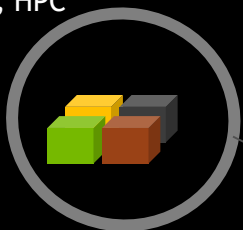
NGC

NGC: GPU-OPTIMIZED SOFTWARE HUB

Ready-to-run GPU Optimized Software, Anywhere

50+ Containers

DL, ML, HPC



15+ Model Training Scripts

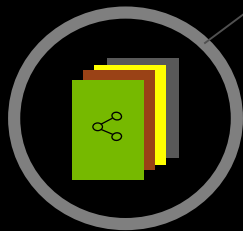
NLP, Image Classification, Object Detection & more



NGC

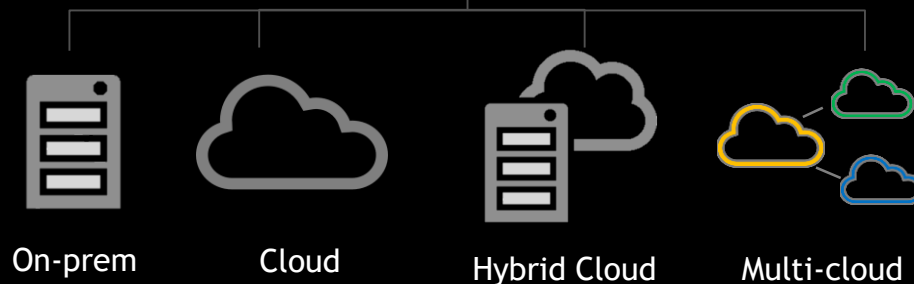
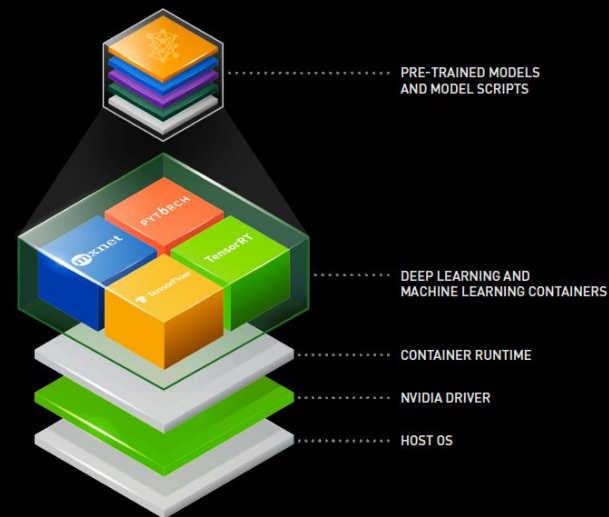
60 Pre-trained Models

NLP, Image Classification, Object Detection & more



Industry Workflows

Medical Imaging, Intelligent Video Analytics



SIMPLIFYING APPLICATION DEPLOYMENTS

Driving Productivity and Faster Discoveries



Data Scientists &
Developers

Superior Performance - Continuous optimizations

Pre-trained Models & Scripts - Speed up AI workflows

On-demand Software - Higher productivity

Scalable - on multi-GPU, multi-node systems

Run Anywhere - On-Prem, Cloud, Hybrid

Designed for Enterprise & HPC - Docker & Singularity



Sysadmins &
DevOps

NGC CONTAINERS: ACCELERATING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments

- Eliminates complex, time-consuming builds and installs

Get started in minutes

- Simply Pull & Run the app

Portable

- Deploy across various environments, from test to production with minimal changes

WHY NGC CONTAINERS

Optimized for Performance

- Monthly DL container releases offer latest features and superior performance on NVIDIA GPUs

Scalable Performance

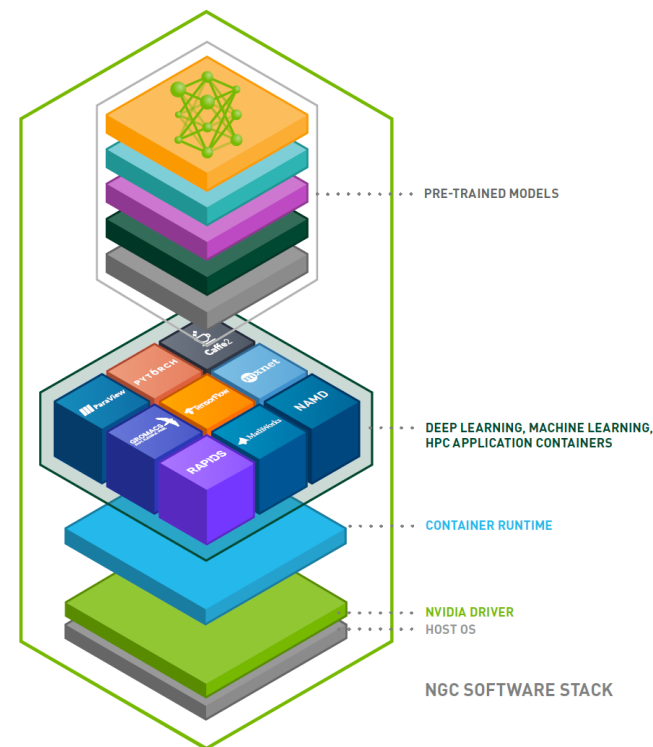
- Supports multi-GPU & multi-node systems for scale-up & scale-out environments

Designed for Enterprise & HPC environments

- Supports Docker & Singularity runtimes

Run Anywhere

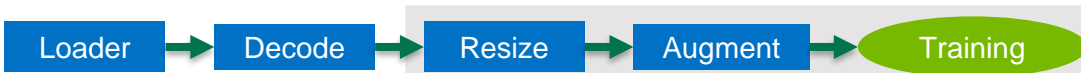
- Pascal/Volta/Turing-powered NVIDIA DGX, PC workstations, and servers
- From Core to the Edge
- On-Prem to Hybrid to Cloud



DALI

Eliminating CPU Bottleneck for DL Workflows

CPU Bottleneck Waste GPU Cycles



DALI Shifts Workloads to GPUs

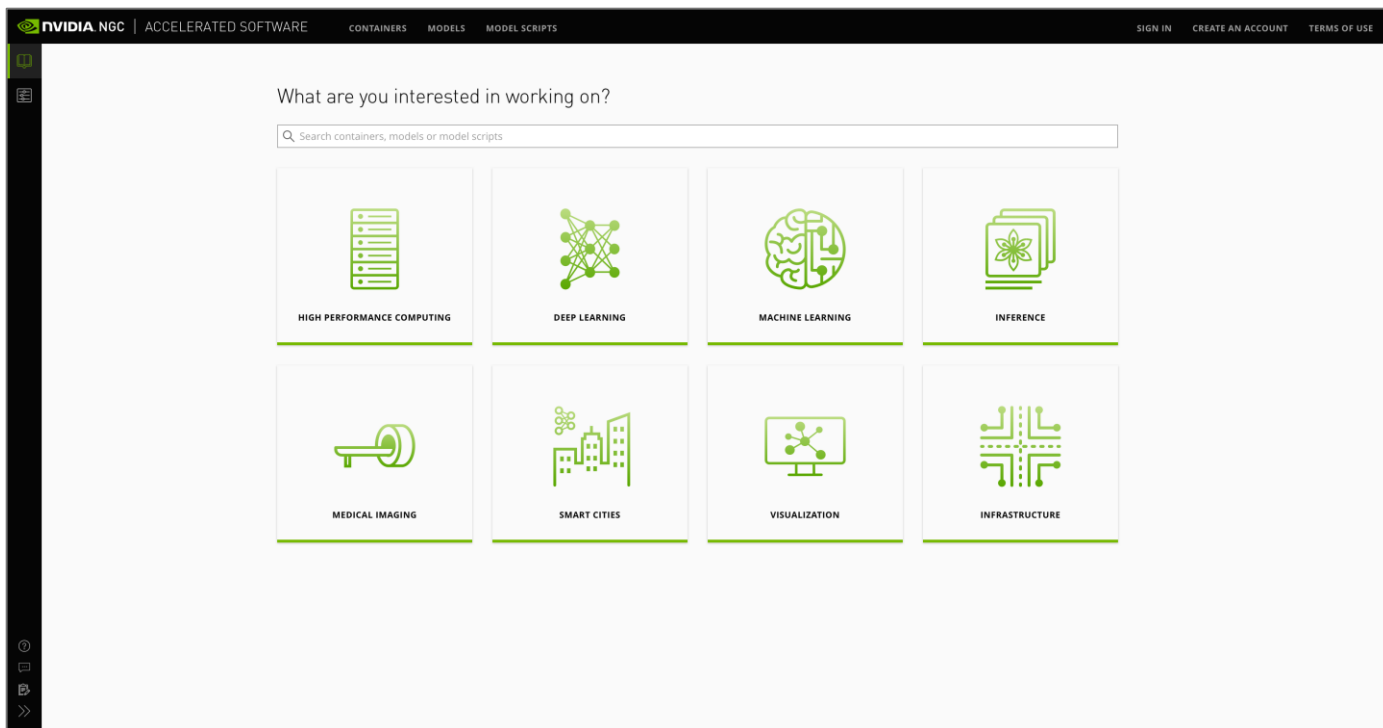


- Complex I/O pipelines
- Multi-pipeline frameworks
- Decreasing CPU:GPU ratio

- Full input pipeline acceleration including data loading and augmentation
- Integrated in PyTorch, TF, MxNET
- Supports Resnet50 & SSD

GET STARTED WITH NGC

Explore the NGC Registry for DL, ML & HPC



Deploy containers:
ngc.nvidia.com

Learn more about NGC offering:
nvidia.com/ngc

Technical information:
developer.nvidia.com

An abstract graphic featuring a complex network of thin, glowing green lines that crisscross the frame. These lines connect various points, some of which are highlighted as bright green nodes. The background is a deep, dark blue or black, which makes the green elements stand out. The overall effect is one of a dynamic, interconnected system, possibly representing a data network or a computational architecture.

GPU ACCELERATED SERVER PLATFORMS

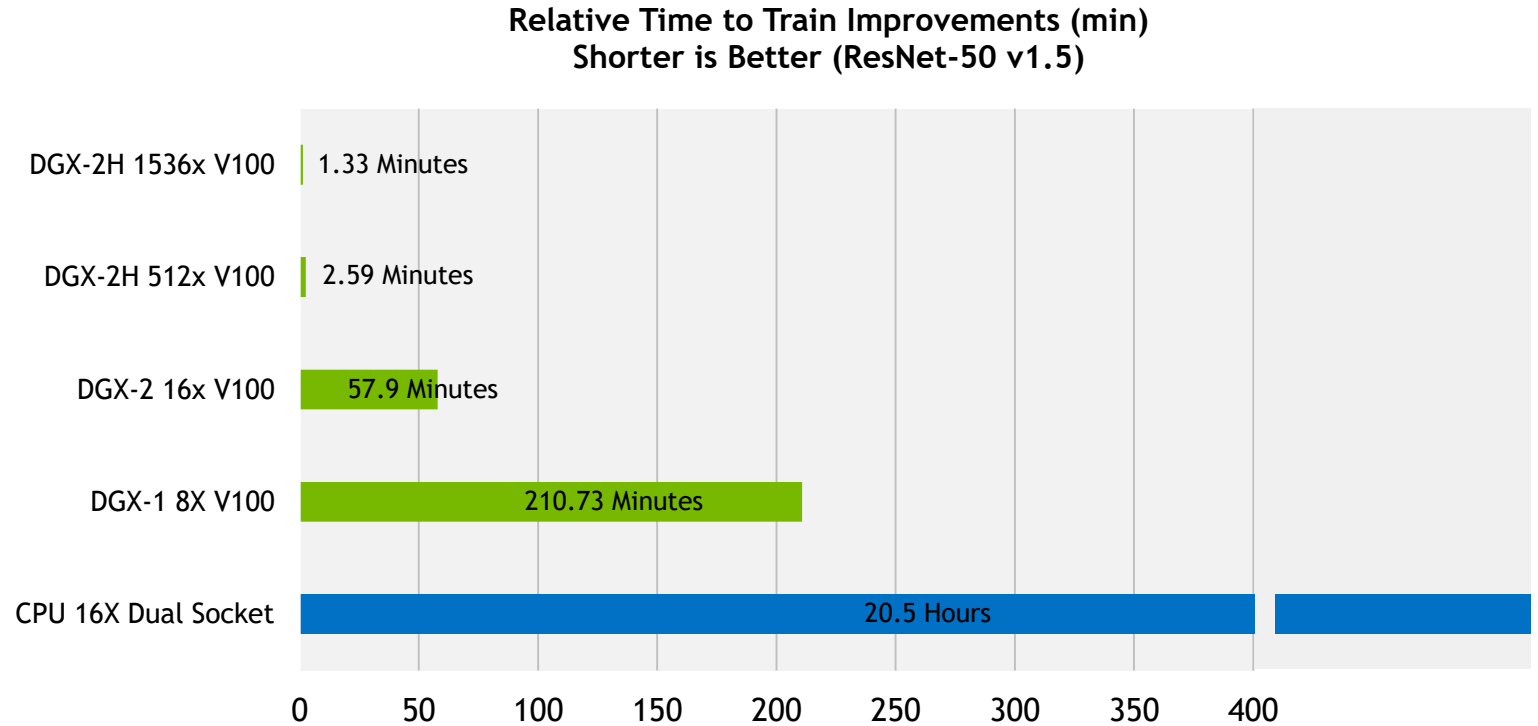
TESLA V100 TENSOR CORE GPU

World's Most Powerful
Data Center GPU

5,120 CUDA cores
640 NEW Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS
| 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32 GB HBM2 @ 900GB/s |
300GB/s NVLink



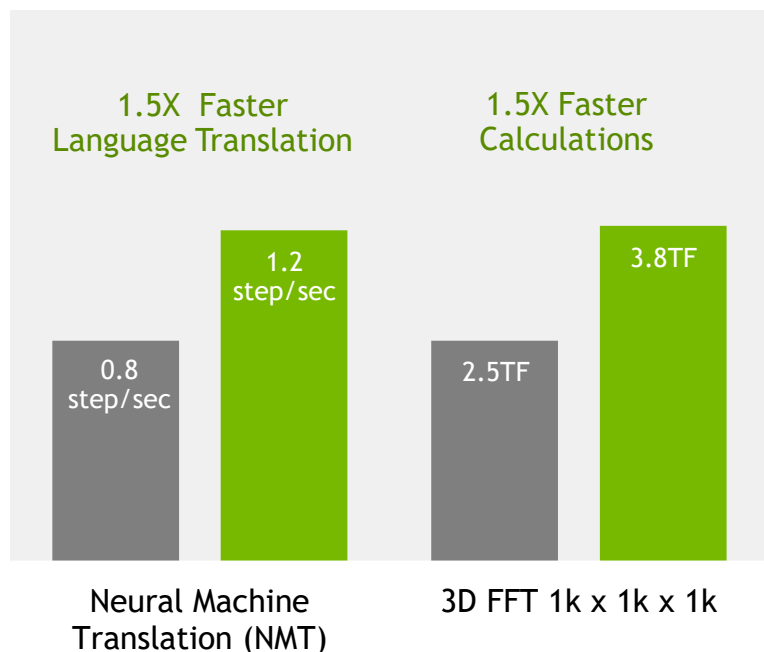
TESLA PLATFORM ENABLES DRAMATIC REDUCTION IN TIME TO TRAIN



UP TO 50% PERFORMANCE IMPROVEMENT

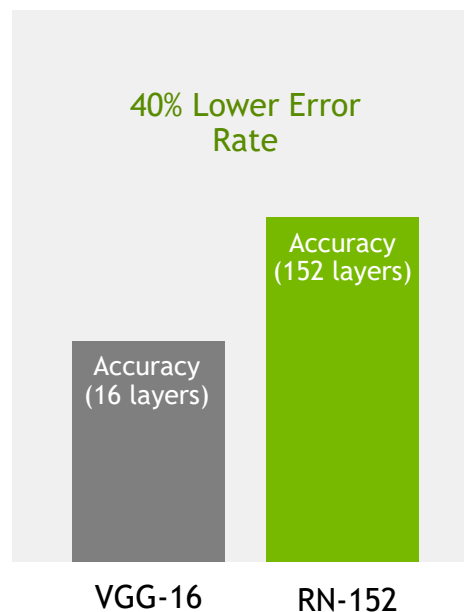
32GB Benefits for AI and HPC

FASTER RESULTS



■ V100 16GB

HIGHER ACCURACY



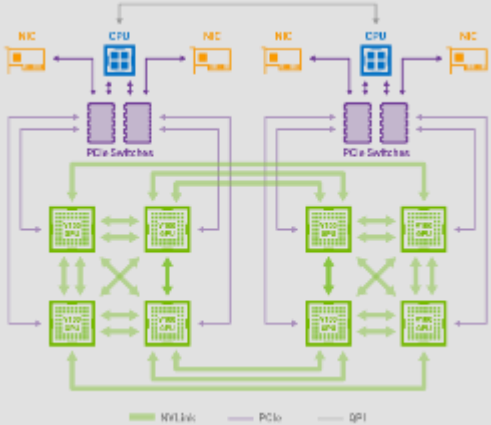
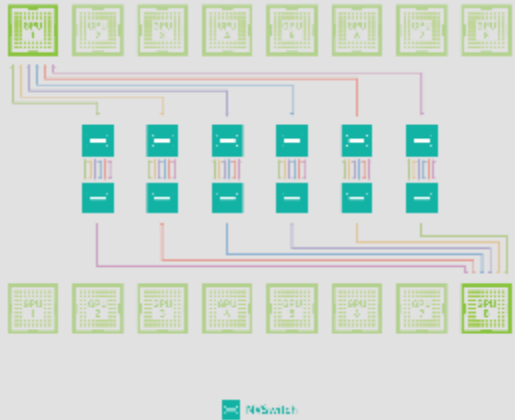
■ V100 32GB

“ The additional memory improved our ability to handle higher definition images on a larger ResNet-152 model, reducing error rates by 40 percent on average. This results in accurate, timely and auditable services at scale. ”

Michael Kemelmakher
Vice President
SAP Innovation Center, Israel



HGX PLATFORMS

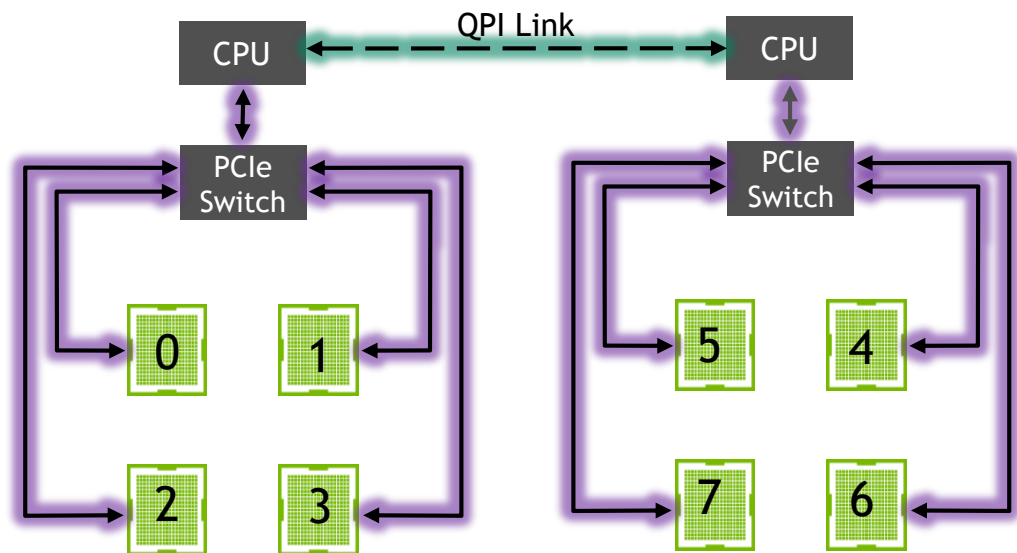
	HGX-1	HGX-2
Topology		
Performance	1 petaFLOP tensor operations 125 teraFLOPS single-precision 62 teraFLOPS double-precision	2 petaFLOPS tensor operations 250 teraFLOPS single-precision 125 teraFLOPS double-precision
GPUs	8x NVIDIA Tesla V100	16x NVIDIA Tesla V100
GPU Memory	256GB total	512GB total
Communication Channel	Hybrid cube mesh powered by NVLink 300GB/s bisection bandwidth	NVSwitch powered by NVLink 2.4TB/s bisection bandwidth

NVLINK AND MULTI-GPU SCALING

For Data Parallel Training



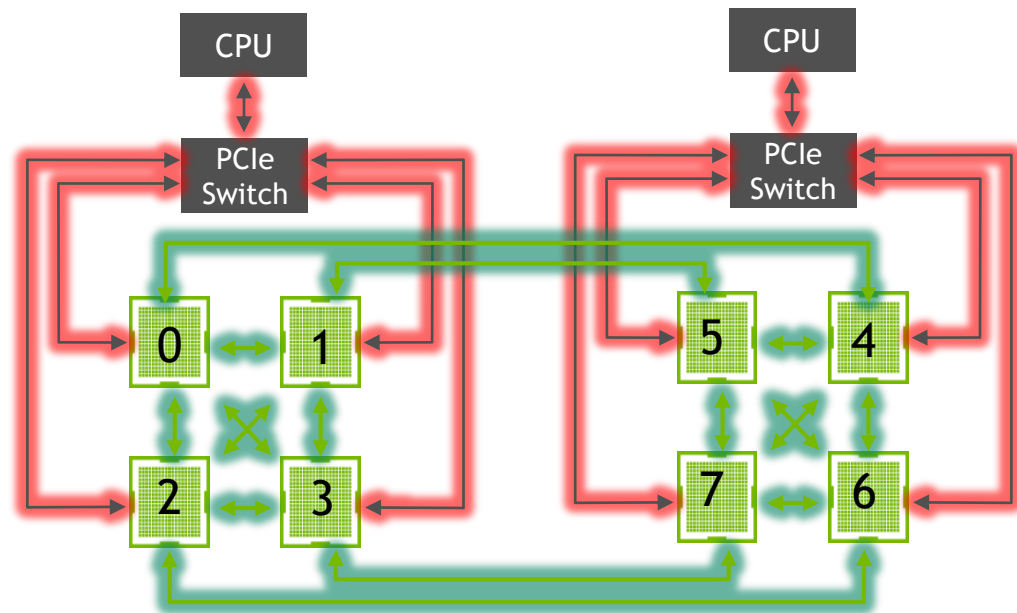
PCIe based system



- Data loading over PCIe
- Gradient averaging over PCIe and QPI
- Data loading and gradient averaging share communication resources: Congestion



NVLINK based system

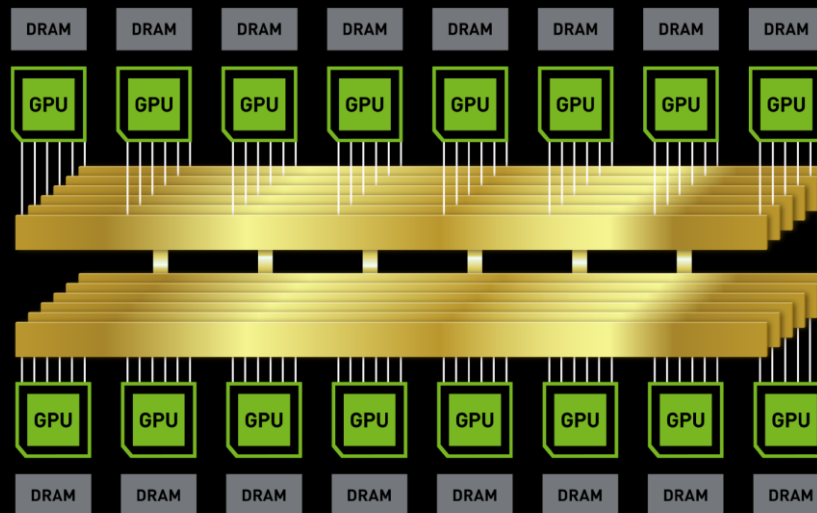


- Data loading over PCIe (red)
- Gradient averaging over NVLink (green)
- No sharing of communication resources: No congestion

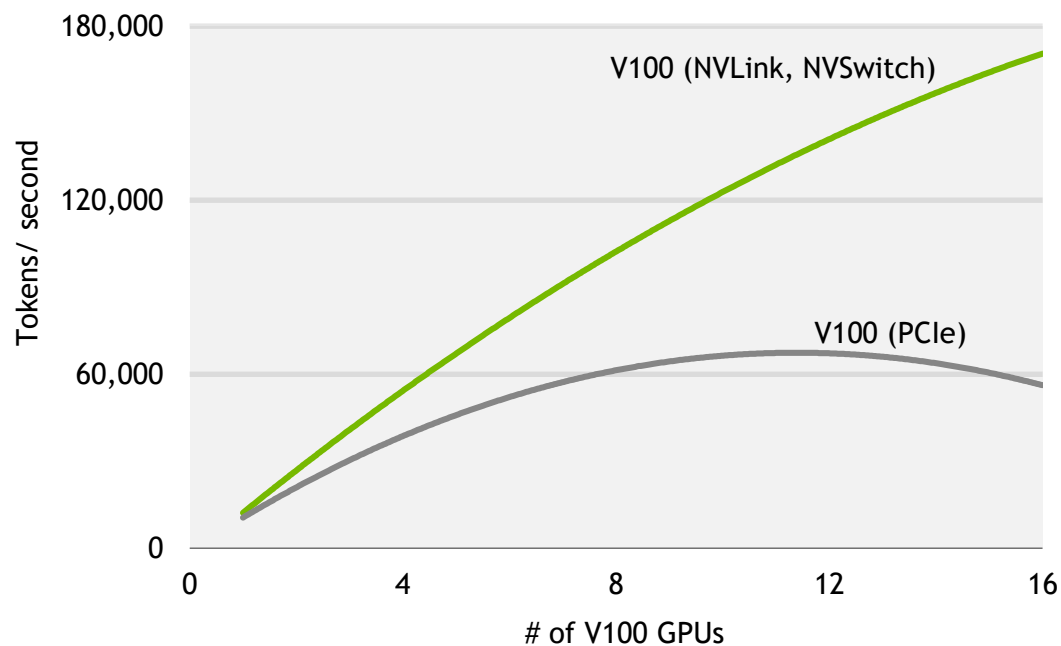
NVSWITCH

World's Highest Bandwidth On-node Switch

7.2 Terabits/sec or 900 GB/sec
18 NVLINK ports | 50GB/s per
port bi-directional
Fully-connected crossbar
2 billion transistors |
47.5mm x 47.5mm package



SCALING-UP PERFORMANCE WITH NVSWITCH



The background is a dark blue field with a complex network of thin, glowing green lines. These lines connect various points, some of which are highlighted as bright green dots. The overall effect is a sense of a dynamic, interconnected system or network.

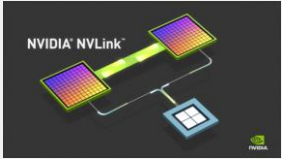
TESLA BRAND PROMISE

THE TESLA BRAND PROMISE

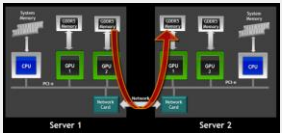
Backed by NVIDIA Products, People, & Processes

Shorten Time-to-Insight

NVLink, NCCL,
NVSwitch



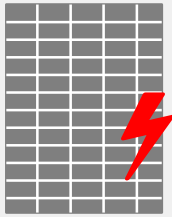
GPUDirect RDMA



Improved throughput
Lower TCO

Run Large Jobs With Confidence

ECC, DPR, Testing &
Qualification



Sys Monitoring and
Management



Lower TCO
Increased productivity

Complex Stacks Just Work

Direct Technical
Support for Developers



GPU Optimized Apps



App Testing
Infrastructure



Increased throughput
Deploy with confidence

Enterprise Grade Reliability

Access to NVIDIA Experts
Defined Escalation Paths

Reliable Software for
Mission Critical
Applications



NVIDIA NVONLINE
Proactive Software Bug
Fix Process

Maximum Uptime
Lower TCO

Available Everywhere

Every OEM and Cloud
Provider



240+ Resellers
Worldwide

Choice of supplier
Lower acquisition costs

Certainty of Supply

Purchase unlimited
quantities

3-year warranty

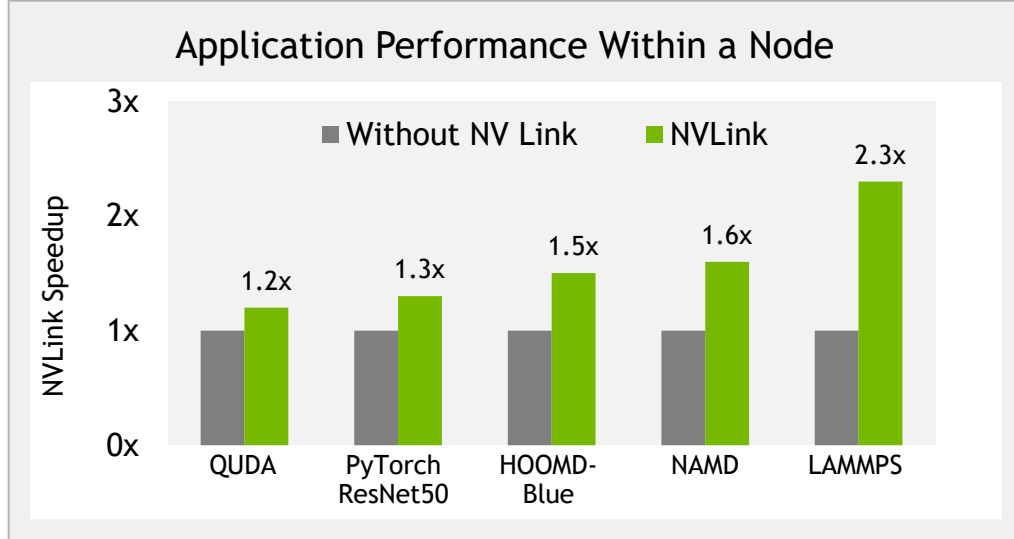
10-month advanced
EOL notification

3-year SKU life

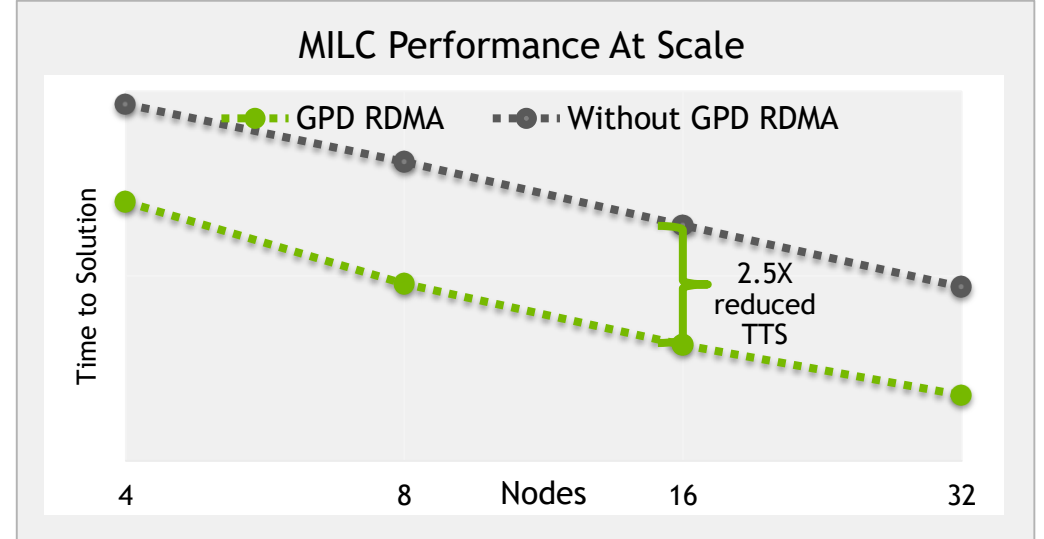
Increased productivity
Avoid supply shortages

SHORTEN TIME TO INSIGHT

Up to 2.5X Faster with NVLink and GPUDirect RDMA



NVLink 8xV100



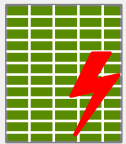
GPUDirect RDMA 8xV100 nodes

RUNNING LARGE JOBS WITH CONFIDENCE

Enterprise Reliability, Management and Live Migration

ERROR CORRECTION CODE (ECC)

GPU MEMORY

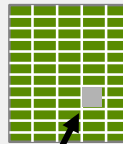


Uncorrectable Data Error
causes application to crash

- Single-bit error correction, double-bit error detection

DYNAMIC PAGE RETIREMENT (DPR)

Tesla GPU with DPR



Weak memory page is retired

- Double-bit error mitigation
- Removes suspect memory locations with simple reset
- No physical work required for IT
- <0.01% of memory is retired

SYSTEM QUALIFICATION AND TESTING



- Long burn-in testing
- Zero error tolerance at aggressive clocks
- Large guard-band for guaranteed quality
- 5% of GPUs are screened out

SYSTEM MONITORING AND MANAGEMENT



- Active health monitoring
- Diagnostics and system validation
- Policy and group config management
- Power and clock management

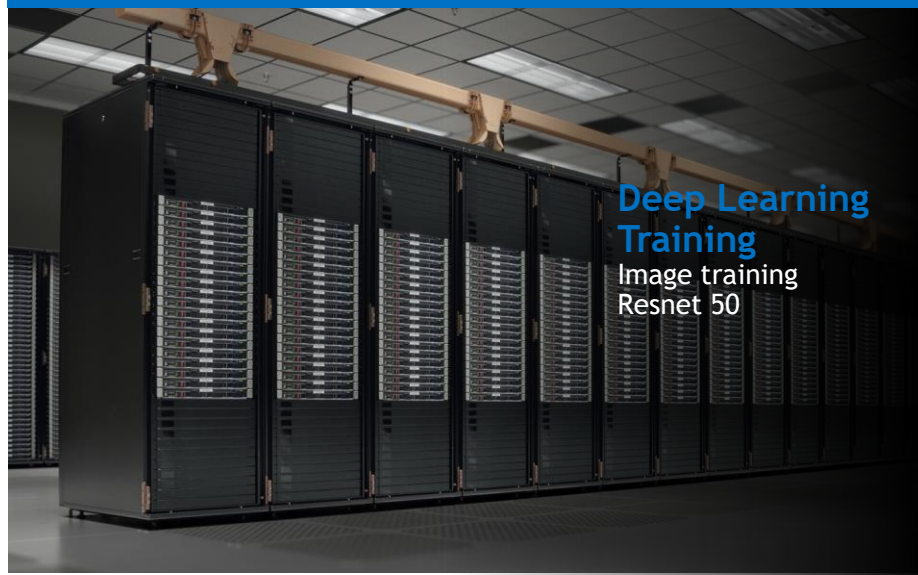
LIVE MIGRATION FOR NVIDIA VGPUS



- Keep large jobs running during patches and upgrades
- Maximize infrastructure investment

DRAMATICALLY MORE FOR YOUR MONEY

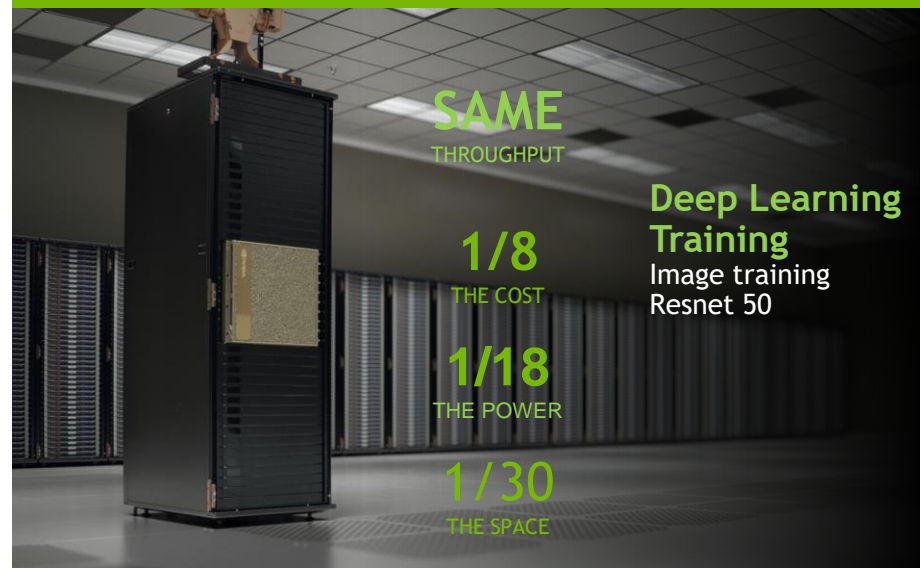
CPU-Only Cluster



300 Self-hosted Broadwell CPU Servers
180 KWatts

=

GPU-Accelerated



1 DGX-2
10 KWatts

DEEP LEARNING INSTITUTE

University Ambassador Program

Preparing today's students and researchers
for tomorrow's AI computing challenges

Want to bring DLI to your campus?

DLI can award qualified academics as certified DLI Ambassadors, enabling them to bring ready-made, free DLI content exclusively to university students and staff

DLI University Ambassadorship is an additional status on top of DLI Instructor Certification with additional benefits

Candidates should have relevant teaching and research experience, and can apply [here](#) for an invitation to an on-site instructor certification event



TEACHING YOU
TO SOLVE PROBLEMS
WITH DEEP LEARNING

New to deep learning or accelerated computing?

Fundamentals training is the place to start. Content is designed for a technical audience of developers, researchers, and data scientists.

