GPU COMPUTING TRENDS

Romuald Josien - Oct, 2018

Agenda:

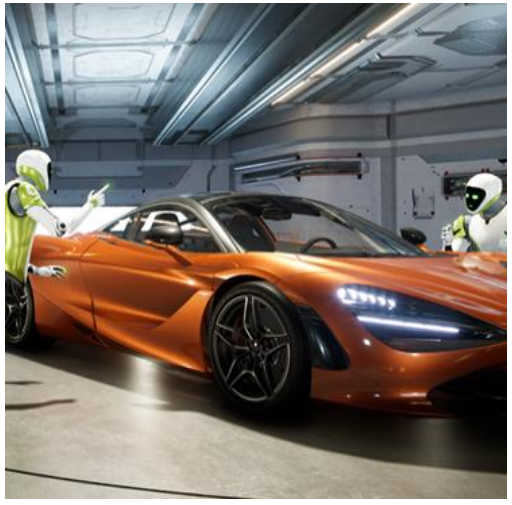
- NVIDIA the company
- NVIDIA contribution to AI/HPC use cases
- What a distance covered these past +5 years!
- Thanks to NVIDIA innovation
- Nvidia TESLA portfolio
- The more you buy, the more you save!
- NVIDIA program for Education

NVIDIA FACTS

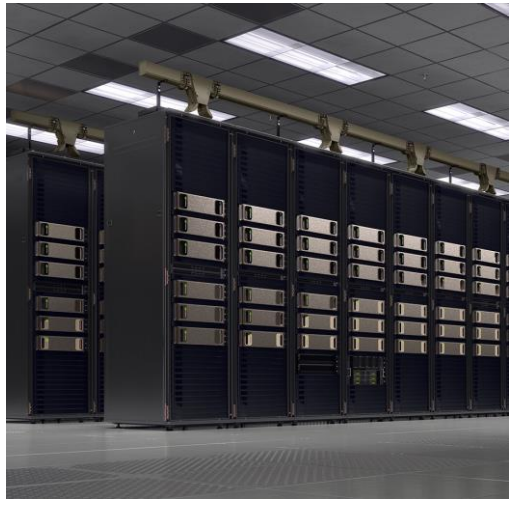
- Founded in 1993
- HQ in Santa Clara (CA – USA)
- > Jensen Huang, Founder & CEO
- > 12,000 employees WW
- $9.7B revenue in FY18 (+41%)
- >1B GPUs shipped to date
- 6,000 patents WW

NVIDIA - GPU COMPUTING

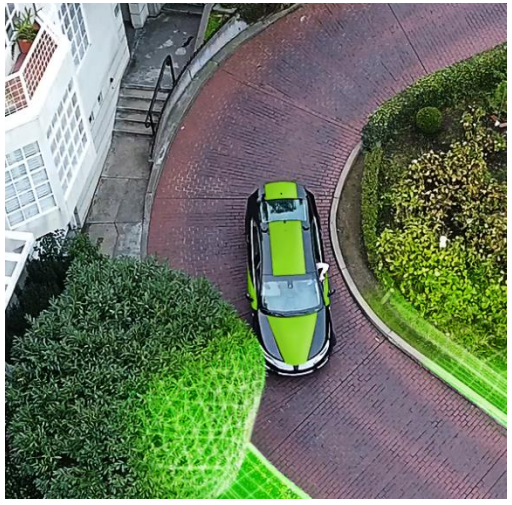
Gaming



VR



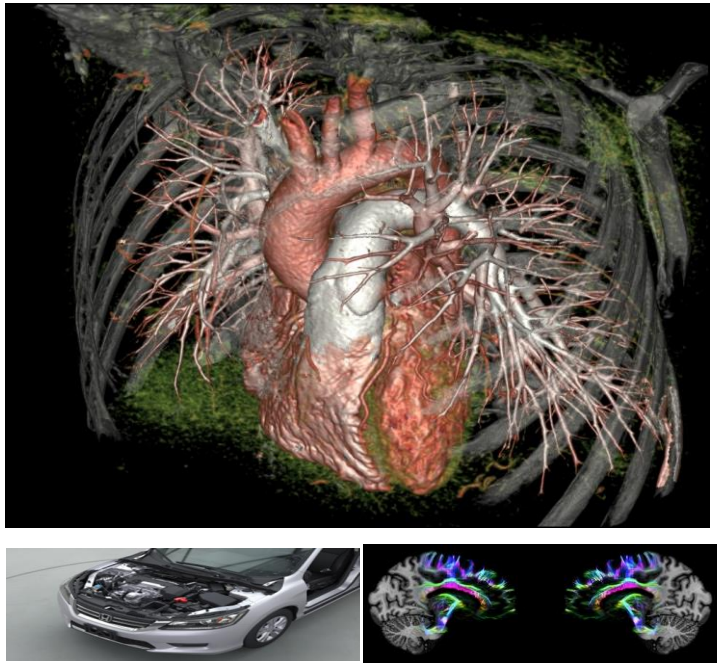
AI & HPC



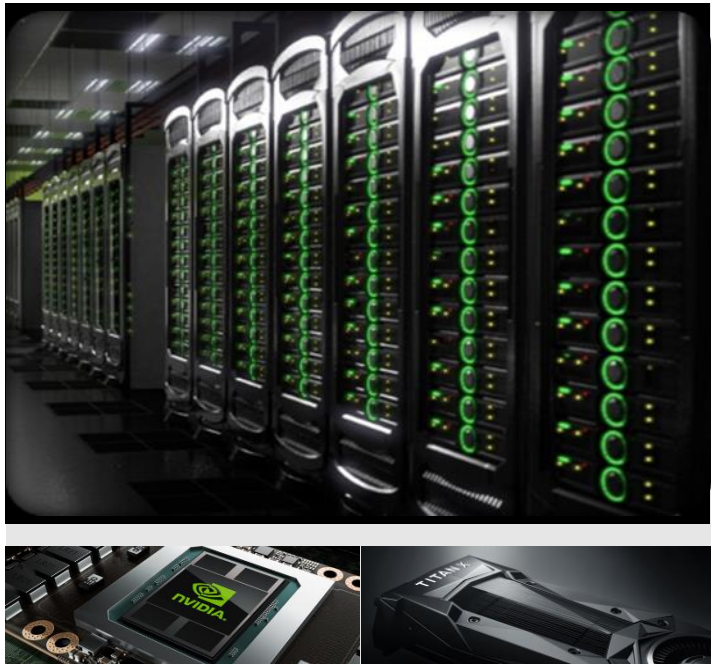
Self-Driving Cars
& Autonomous Machines

ONE ARCHITECTURE — CUDA

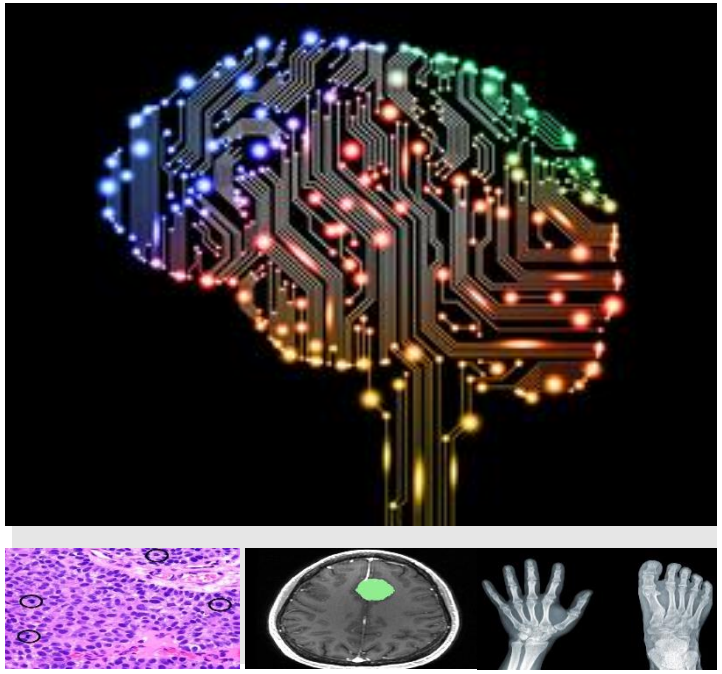
NVIDIA “THE AI COMPUTING COMPANY”



Computer Graphics



GPU Computing

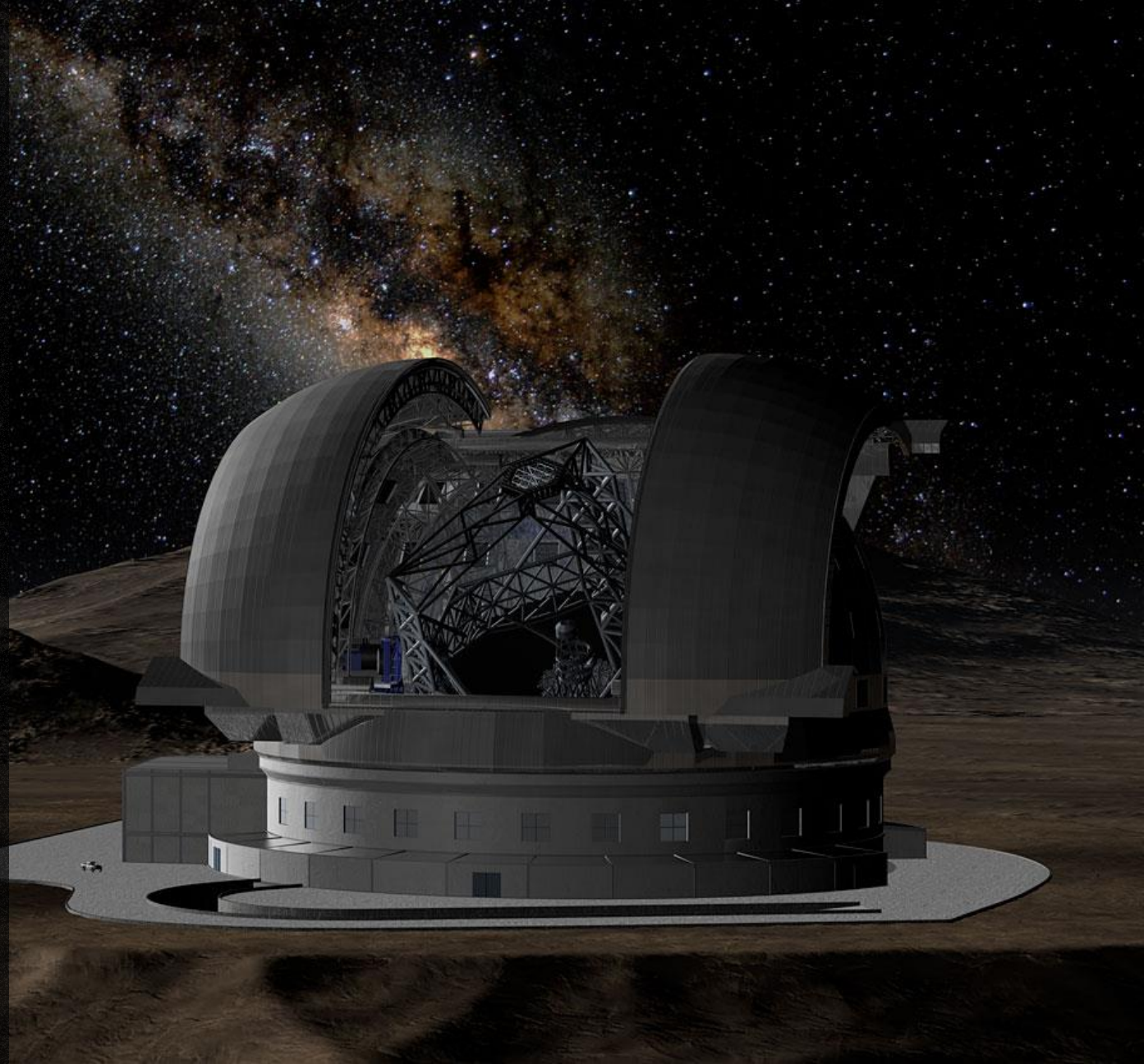


Artificial Intelligence

NVIDIA CONTRIBUTES TO IMPROVE OUR WORLD

POWERING THE WORLD'S BIGGEST EYE ON THE SKY

The European Extremely Large Telescope, which when completed in 2024 will be the largest telescope ever built, will advance astrophysical knowledge by enabling detailed studies of planets, the first galaxies in the universe, super-massive black holes, and other phenomenon with images 16 times sharper than those from the Hubble Space Telescope. Scientists at the Université Paris Diderot and the Observatoire de Paris are using the NVIDIA DGX-1 to revolutionize adaptive optics, a technique used to compensate for atmospheric turbulence which distorts images and impedes discovery. The GPU-powered real-time optical system controller will provide millions of commands/second and significantly stabilize the telescope's image quality to increase its chances of finding Earth-like planets orbiting distant stars.



l'Observatoire de Paris

université PARIS DIDEROT PARIS 7

DEVELOPING THE VEHICLES OF THE FUTURE

Zenuity, a joint venture of Volvo and Veoneer, aims to build autonomous driving software for production vehicles by 2021. They chose to build their deep learning infrastructure with NVIDIA DGX-1 servers and Pure FlashBlade systems to accelerate their AI initiative.

NVIDIA. PURESTORAGE®

ZENUITY
Make it real.

HUNTING "GHOST PARTICLES" WITH DEEP LEARNING

Tiny particles called neutrinos are the most abundant form of matter in the universe and understanding their basic properties is the focus of a world-wide campaign of experiments. Observing these 'ghost particles' in action requires instruments of incredible size and scale. Fermilab's NOvA experiment applies two enormous, cutting-edge technology detectors with a total weight of 30 million pounds spaced 500 miles apart. It is effectively one of the world's largest cameras, snapping two million images per second and analyzing them for neutrino activity.

NOvA's scientists developed deep neural networks trained on NVIDIA GPUs to improve the machine's detection rate by 33% - increasing the discovery potential of NOvA and other large scale experiments probing fundamental questions of the universe.

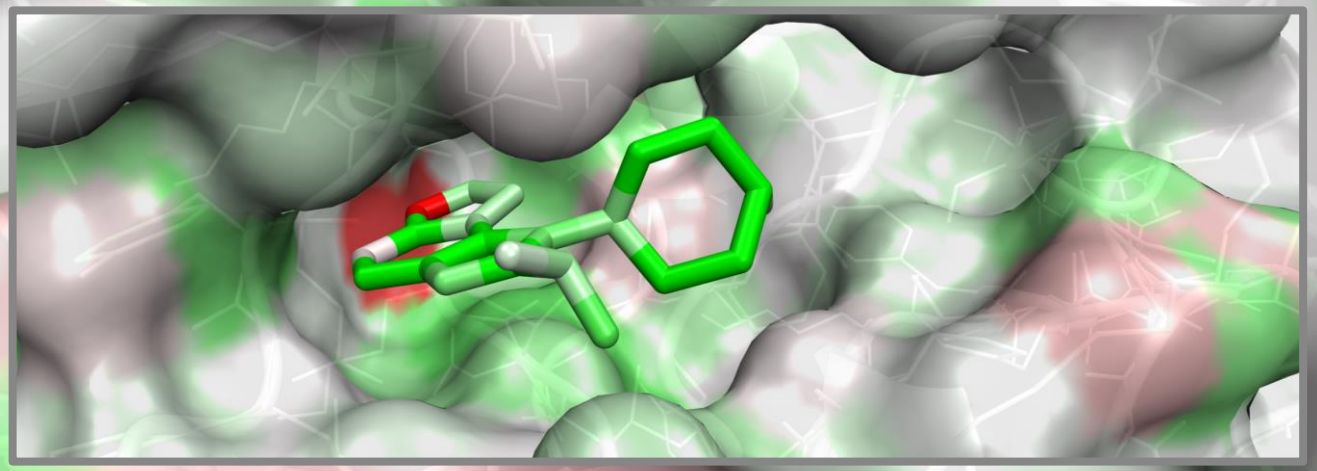
THE BRAINS BEHIND SMART CITIES

Verizon's Smart Communities Group is on a mission to make cities safer, smarter and greener. Using NVIDIA Metropolis, an edge-to-cloud video platform for building smarter, faster AI-powered applications, Verizon is working to collect and analyze multiple streams of video data to improve traffic flow, enhance pedestrian safety, optimize parking and more.

NVIDIA verizon

SPEEDING UP DRUG DISCOVERY

Classic Molecular Dynamics simulations are time-consuming and expensive, but machine learning models can help predict the probability of target molecules to bond With chemical compounds. Researchers at the University of Pittsburgh are improving model Performance and prediction accuracy. Their Convolutional neural network, accelerated by NVIDIA GPU's, improved prediction accuracy from ~52% to 70% which could reduce the time and costs to bring new drugs to market.



“SEEING” GRAVITY FOR THE FIRST TIME

In September 2015, 100 years after Einstein predicted them, gravitational waves were observed for the first time. Astronomers at the Laser Interferometer Gravitational-wave Observatory have since used GPU-powered deep learning to process gravitational wave data 100x faster than previous methods, making real-time analysis possible and putting us one step closer to understanding the universe’s oldest secrets.

Physics Letters B - Deep learning for real-time gravitational wave detection and parameter estimation: Results with advanced LIGO data

Daniel George, E.A. Huerta

AI IS ON TRACK TO SAFEGUARD RAILWAY INTEGRITY

To maintain the integrity of its 3,232 km of tracks, the Swiss Federal Railways (SBB) runs diagnostics trains to photograph and monitor tracks in real-time. But traditional data processing methods produce false positives/negatives. To remedy this, SBB and CSEM (Swiss Research and Development Center) launched the *Railcheck* project which applies deep learning, powered by the NVIDIA DGX Station, to improve the automatic detection and classification of faults.

NVIDIA. :: csem

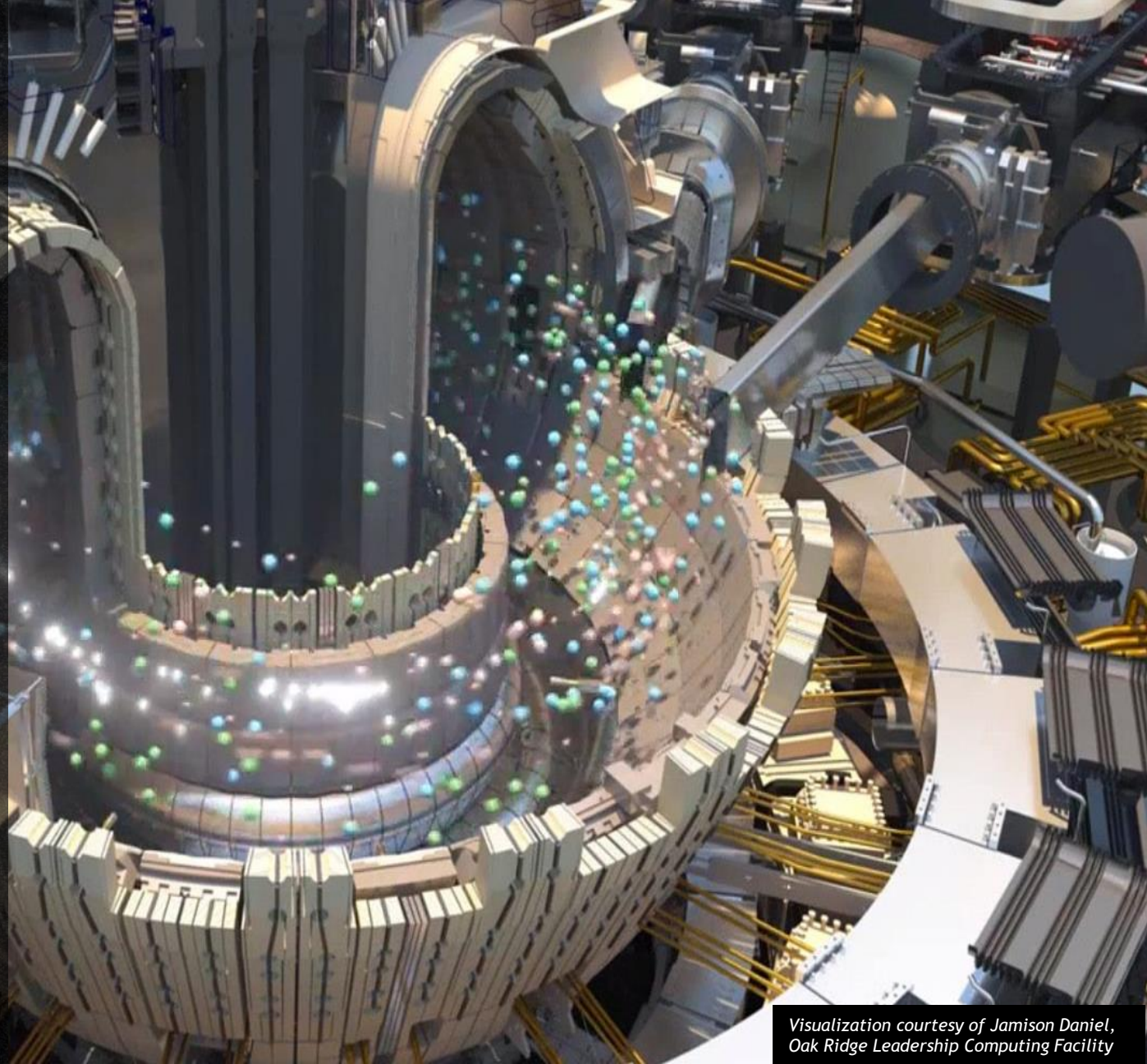
SBB CFF FFS

AI IS SPEEDING THE PATH TO FUSION ENERGY

Fusion is the future of energy on Earth. But it's a highly sensitive process where even small environmental disruptions can stall reactions and damage multi-billion machines. Current models can predict the disruptions with 85% accuracy, but ITER will need something more precise.

Researchers at Princeton University have developed the Fusion Recurrent Neural Network (FRNN) using deep learning and NVIDIA GPUs to predict disruptions and make adjustments to minimize damage and downtime. Even a 1% improvement in the prediction accuracy can be transformative considering the immense scale and cost of fusion science. FRNN has achieved 90% accuracy and is on the path to achieving 95% accuracy for ITER's tests.

PRINCETON UNIVERSITY



Visualization courtesy of Jamison Daniel, Oak Ridge Leadership Computing Facility

AI HELPS DOCTORS DIAGNOSE BREAST CANCER

Every day, pathologists are tasked with providing cancer diagnosis to guide patient treatment. However, sifting through millions of normal cells to identify a few malignant cells is extremely laborious using conventional methods. PathAI combines GPU deep learning with traditional pathology to improve accuracy, speed diagnosis, and reduce error rates by 85%.

NVIDIA PathAI

AI HELPS PERSONALIZE IMMUNOTHERAPY

Immunotherapy has a success rate of only 40% and a risk that it may attack healthy cells. Max Kelsen is using sophisticated AI approaches with NVIDIA V100 GPUs to integrate genomic, transcriptomic and patient information to identify a classifier and develop a test that can predict treatment response.

NVIDIA Max Kelsen

DGX 3 NODE CLUSTER TO ADVANCE GENOMIC RESEARCH

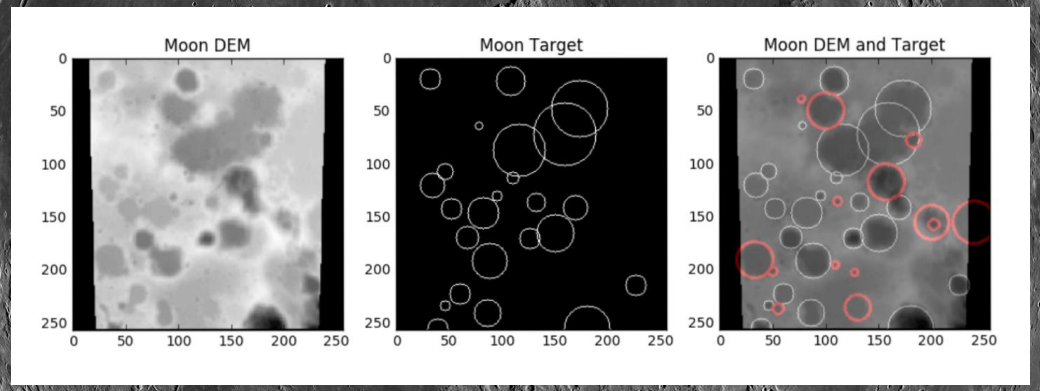
In 2003 the Human Genome Project successfully decoded the human genome and unlocked the door to new genetic discoveries. With 3 billion nucleotide pairs in the human DNA, genome analysis is computationally intensive. The Tohoku Medical Megabank Organization (ToMMo) is using the power of its DGX-1 AI supercomputer cluster to accelerate understanding the complicated correlations between human genotype and phenotype.

And, to further deep learning based genomics research, ToMMo will open its DGX-1 supercomputer cluster to external contracted researchers.

TOHOKU MEDICAL MEGABANK ORGANIZATION

MAPPING MOON CRATERS

Studying moon craters provides insight into the history of our solar system. Researchers at U of T and Penn State developed a CNN, powered by NVIDIA Tesla P100 GPUs on the SciNet P8 supercomputer, that automatically detects and classifies characteristics of craters from lunar digital elevation map images. Upon implementation the system identified 6,000 new craters in just a few hours, making it orders of magnitude faster than human counting.

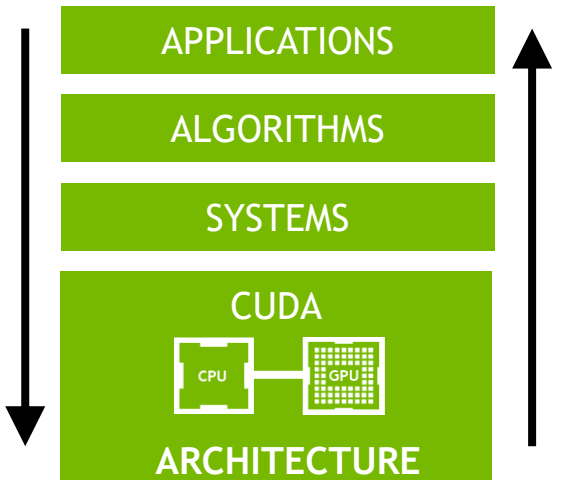


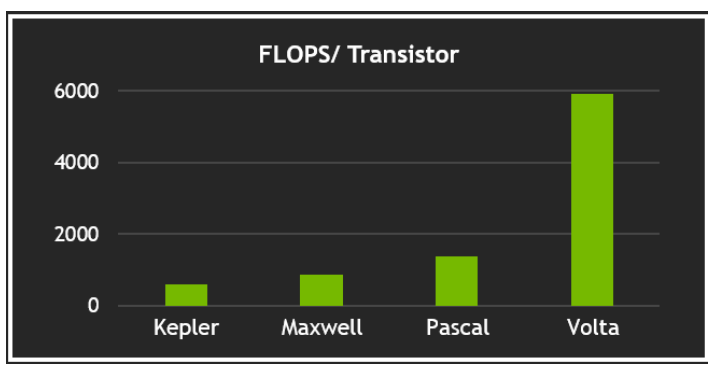
Inset: Sample Moon image (left) and target (center) from the dataset, with the two overlaid (right). Red circles show craters detected by the AI system that are absent from previous datasets.

Lunar image courtesy of NASA

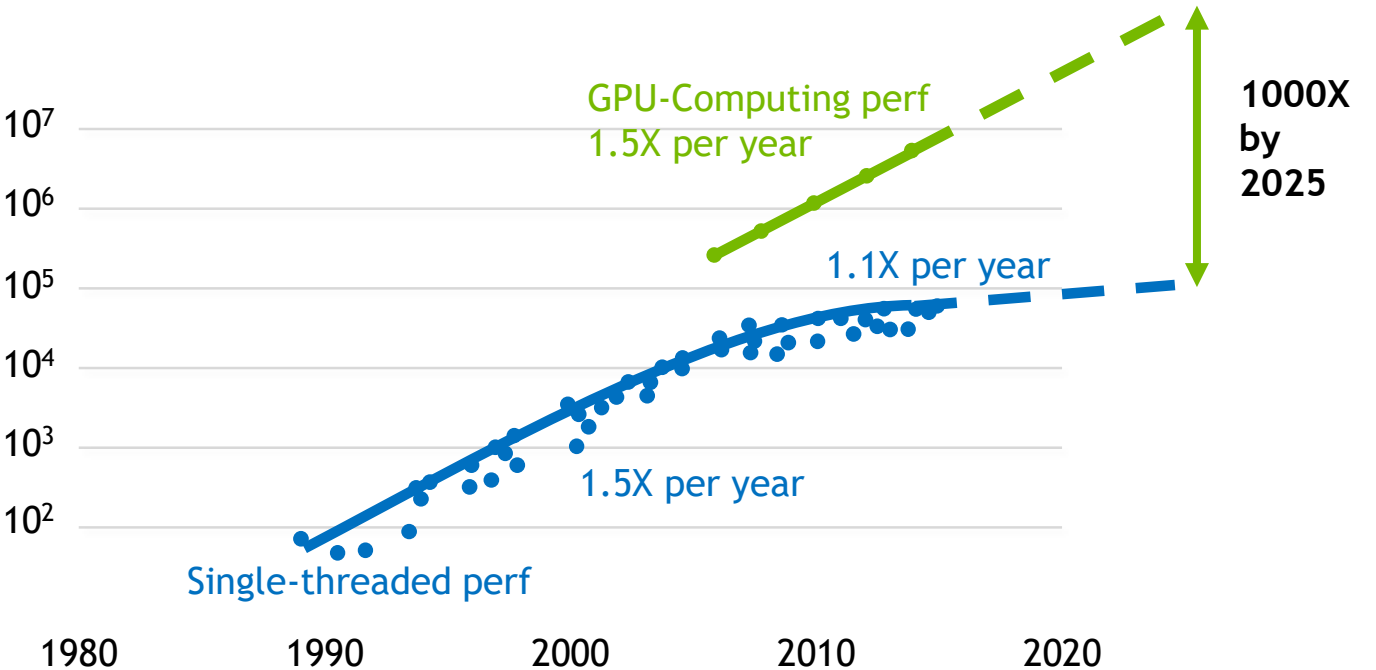
What a distance covered since 2012!

RISE OF GPU COMPUTING





CUDA – Domain Specific Computing Architecture
10X in 5 Years



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

NVIDIA CONFIDENTIAL – DO NOT DISTRIBUTE

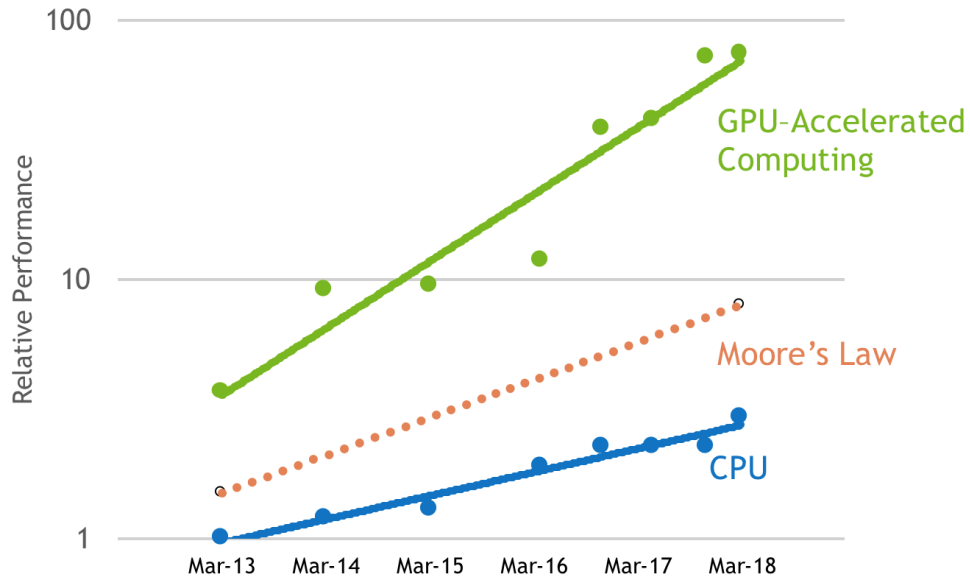
HPC: 20X PERFORMANCE GAIN IN 5 YEARS

Beyond Moore's Law

2013

- cuBLAS: 5.0
- cuFFT: 5.0
- cuRAND: 5.0
- cuSPARSE: 5.0
- NPP: 5.0
- Thrust: 1.5.3
- CUDA: 5.0
- Resource Mgr: r304
- Base OS: CentOS 6.2

Accelerated Server With Fermi



Measured performance of Amber, CHROMA, GTC, LAMMPS, MILC, NAMD, Quantum Espresso, SPECFEM3D

2018

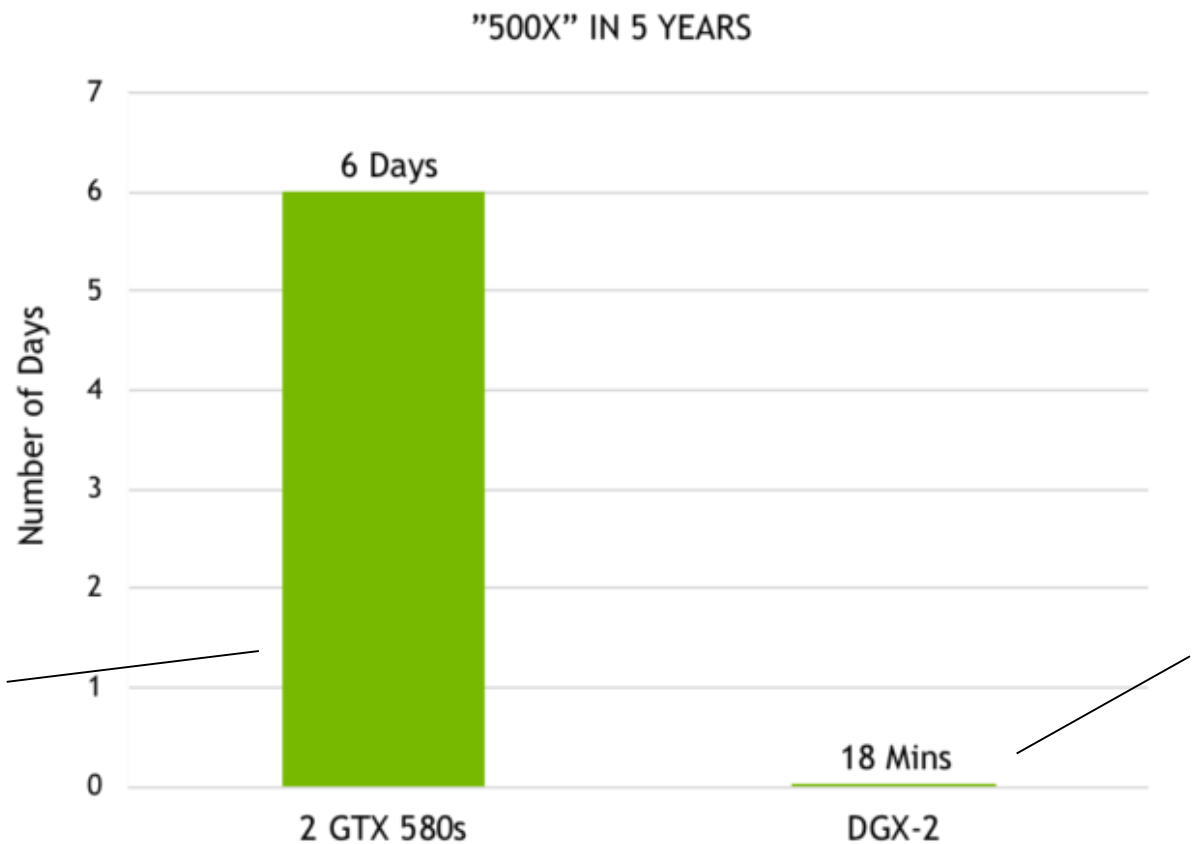
- cuBLAS: 9.0
- cuFFT: 9.0
- cuRAND: 9.0
- cuSOLVER: 9.0
- cuSPARSE: 9.0
- NPP: 9.0
- Thrust: 1.9.0
- CUDA: 9.0
- Resource Mgr: r384
- Base OS: Ubuntu 16.04

Accelerated Server with Volta

DEEP LEARNING: EXPONENTIAL PERFORMANCE IMPROVEMENTS

500X in 5 YEARS!

Alex Krizhevsky won the Imagenet competition in 2012

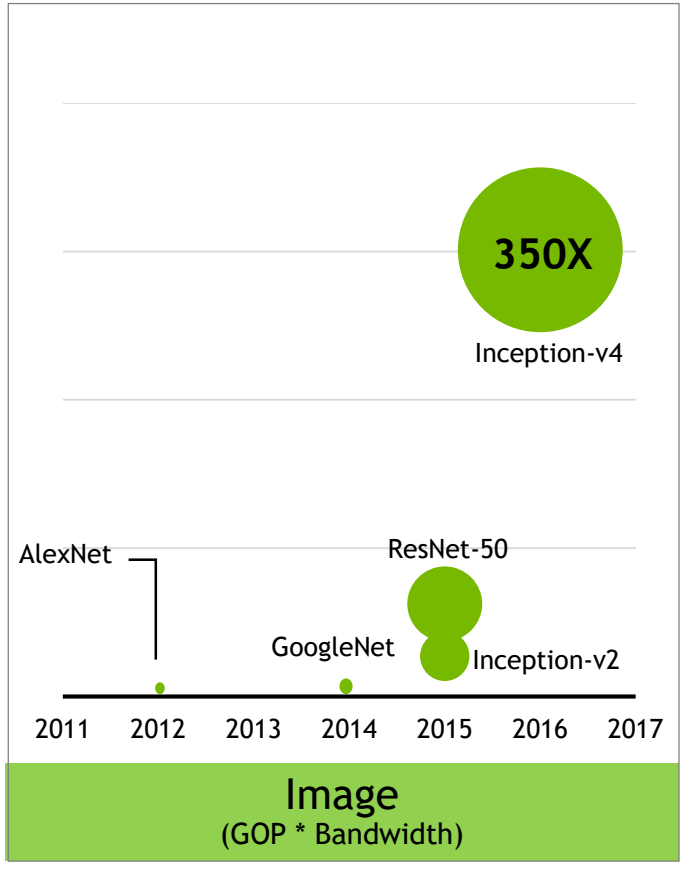


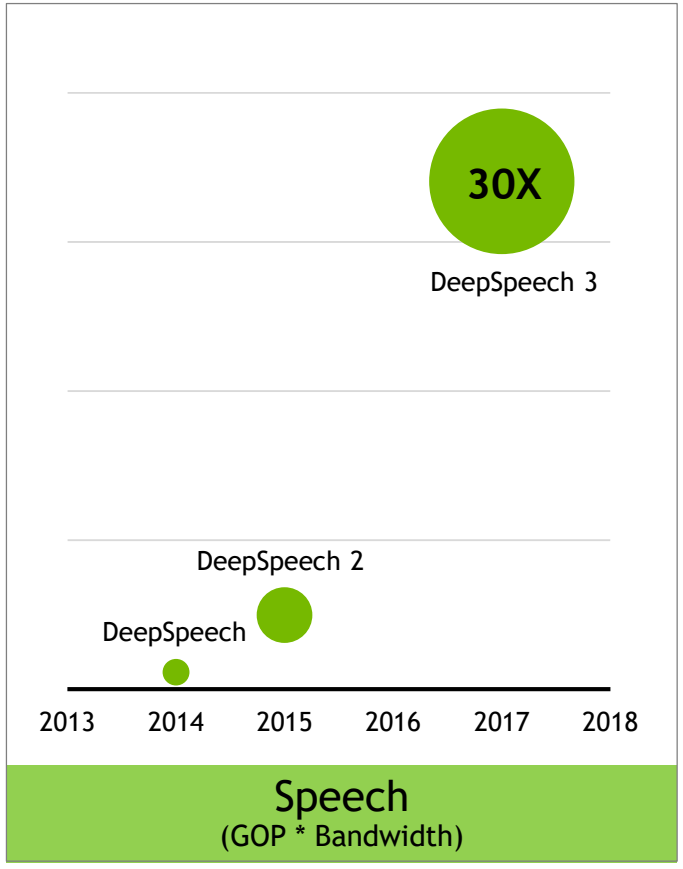
2018

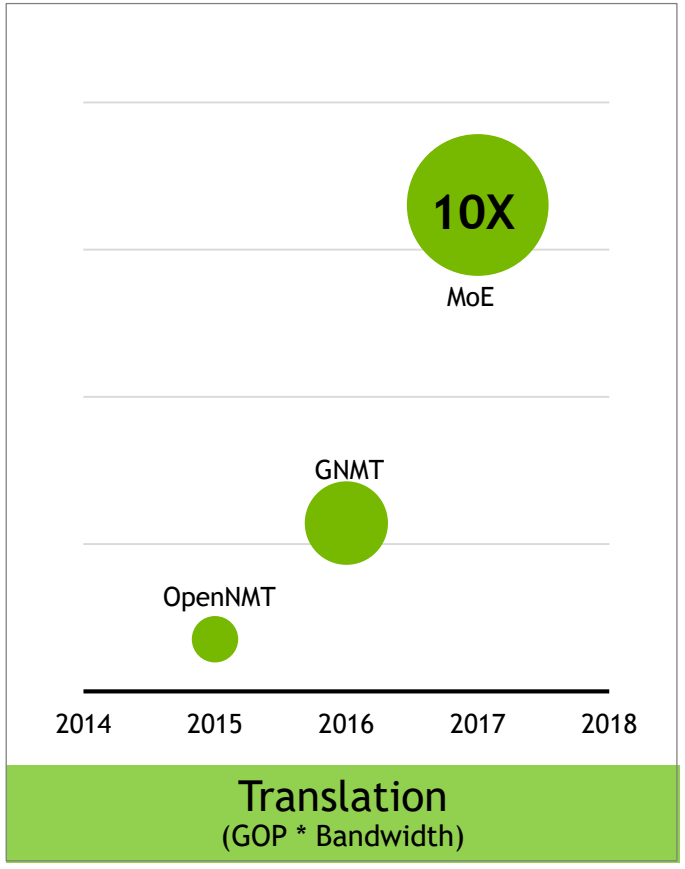
Time to Train AlexNet on the Imagnet Dataset

NEURAL NETWORK COMPLEXITY IS EXPLODING

Bigger and More Compute Intensive

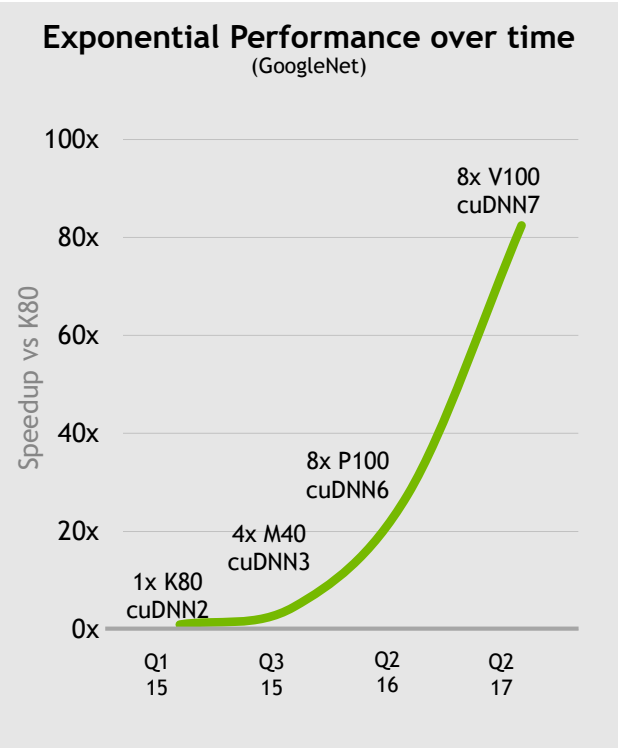


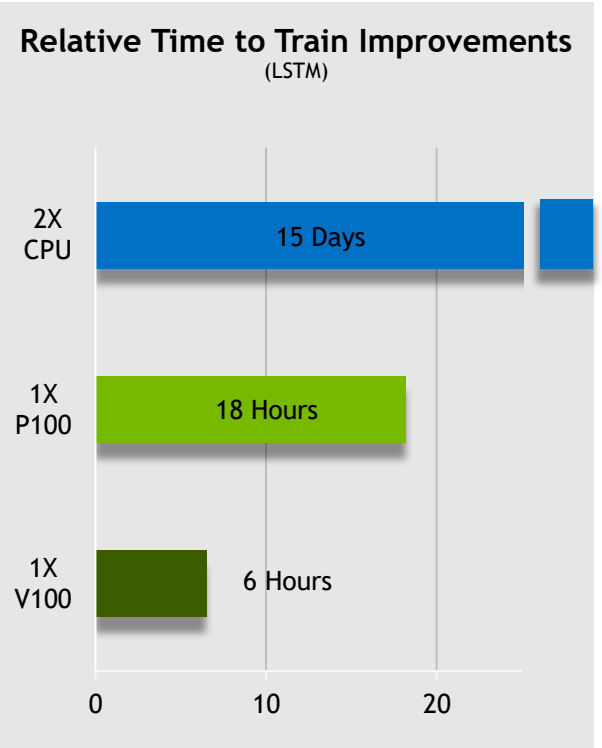




REVOLUTIONARY AI PERFORMANCE

3X Faster DL Training Performance





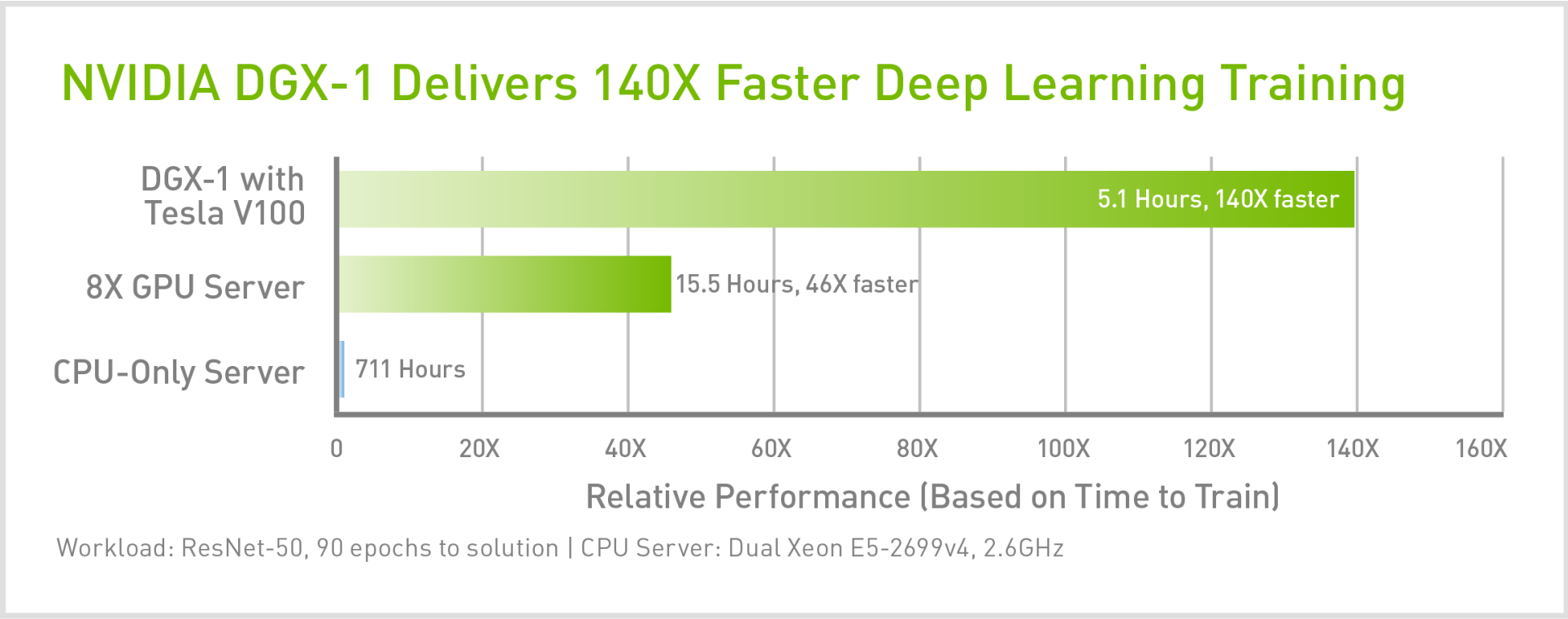
Over 80X DL Training Performance in 3 Years

3X Reduction in Time to Train Over P100

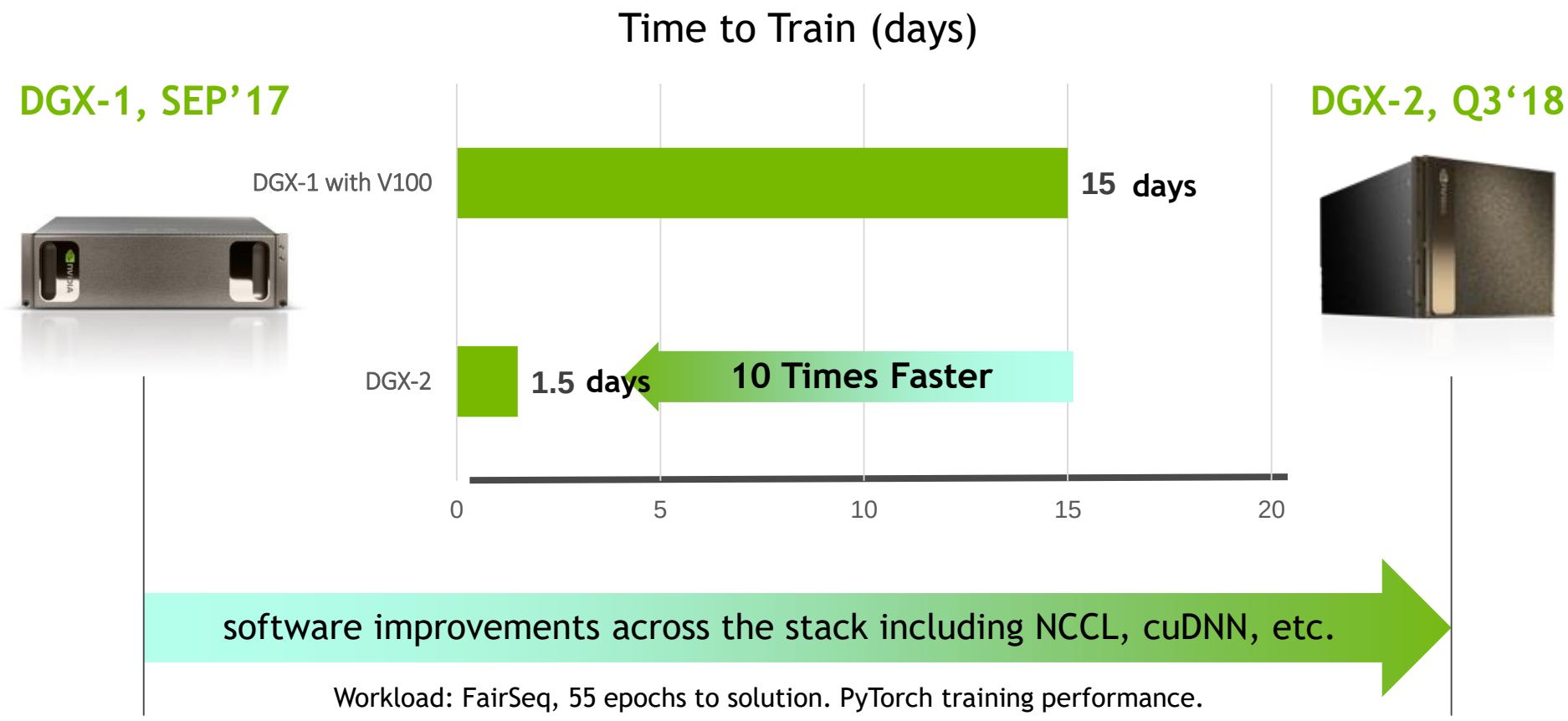
GoogleNet Training Performance on versions of cuDNN
Vs 1x K80 cuDNN2

Neural Machine Translation Training for 13 Epochs |German ->English, WMT15 subset | CPU = 2x Xeon E5 2699 V4

DGX-1: 140X FASTER THAN CPU

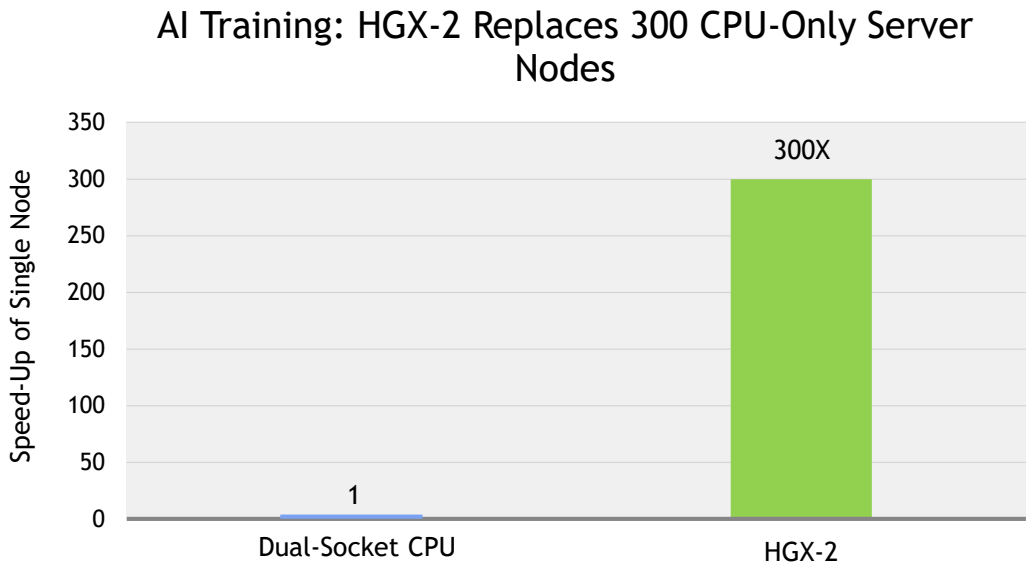


10X PERFORMANCE GAIN IN LESS THAN A YEAR

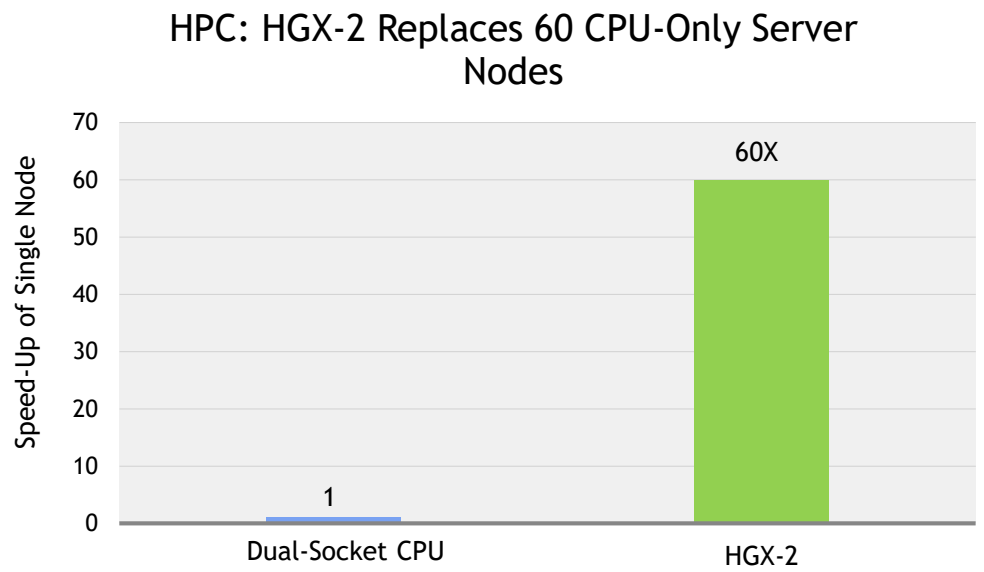


AI AND HPC BENCHMARKS: DGX-2 VS CPU

Replace CPU Nodes - Save Money, Power and Space in the Data Center



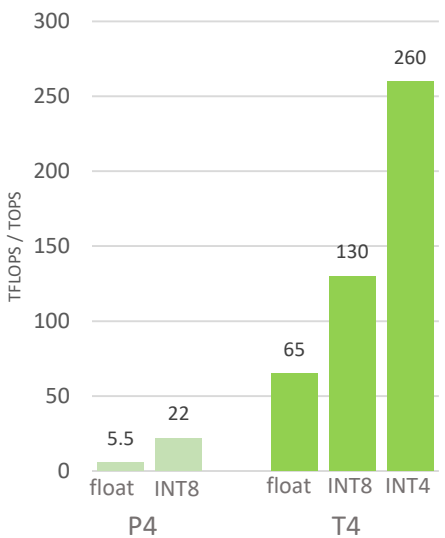
Workload: ResNet50, 90 epochs to solution | CPU Server: Dual-Socket Intel Xeon Gold 6140

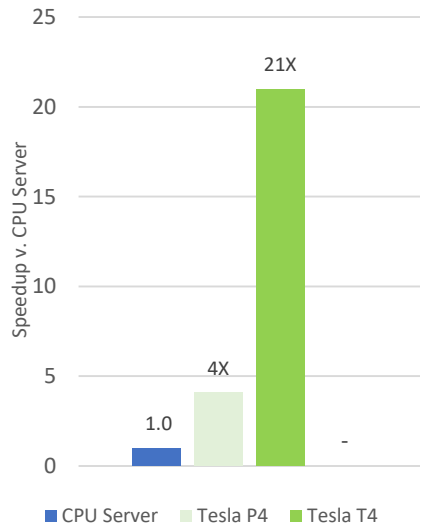


Workload: MILC (particle physics HPC application) | CPU Server: Dual-Socket Intel Xeon Gold 6140

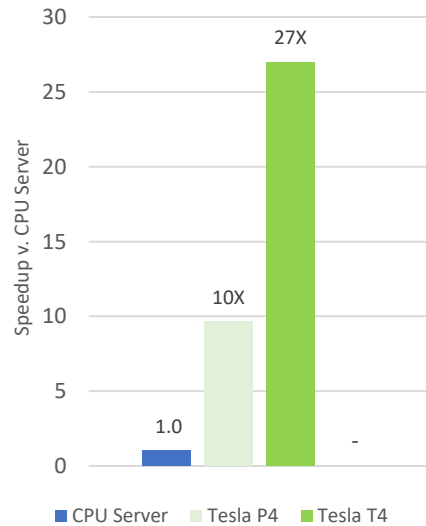
WORLD'S MOST PERFORMANT INFERENCE PLATFORM

Up To 36X Faster Than CPUs | Accelerates All AI Workloads

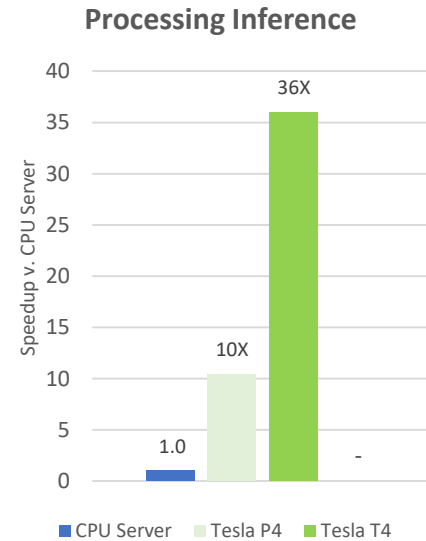




Speedup: 21X faster
DeepSpeech 2

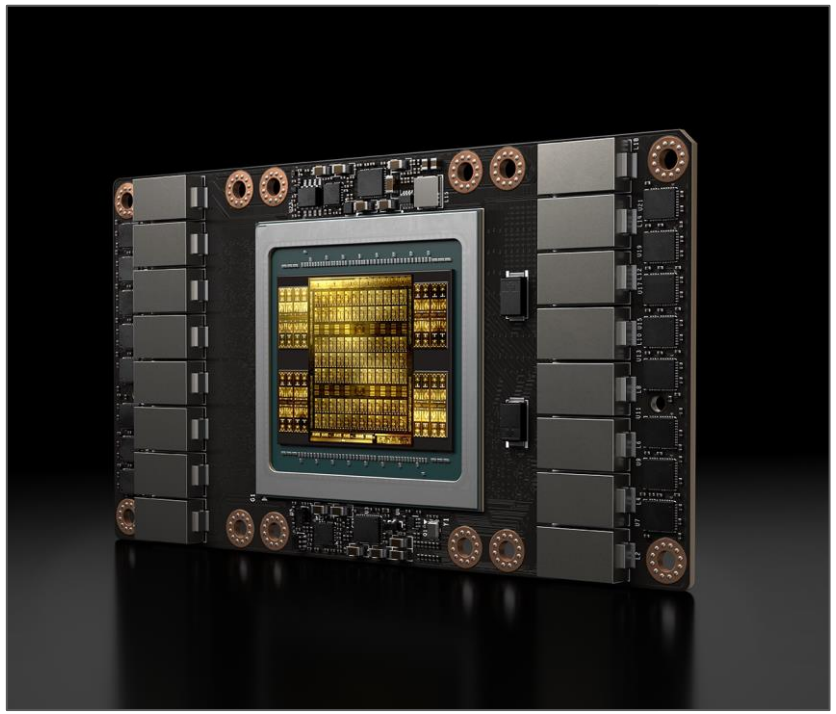


Speedup: 27x faster
ResNet-50 (7ms latency limit)



Speedup: 36x faster
GNMT

VOLTA TENSOR CORE GPUS POWER SUMMIT: WORLD'S FASTEST AI SUPERCOMPUTER

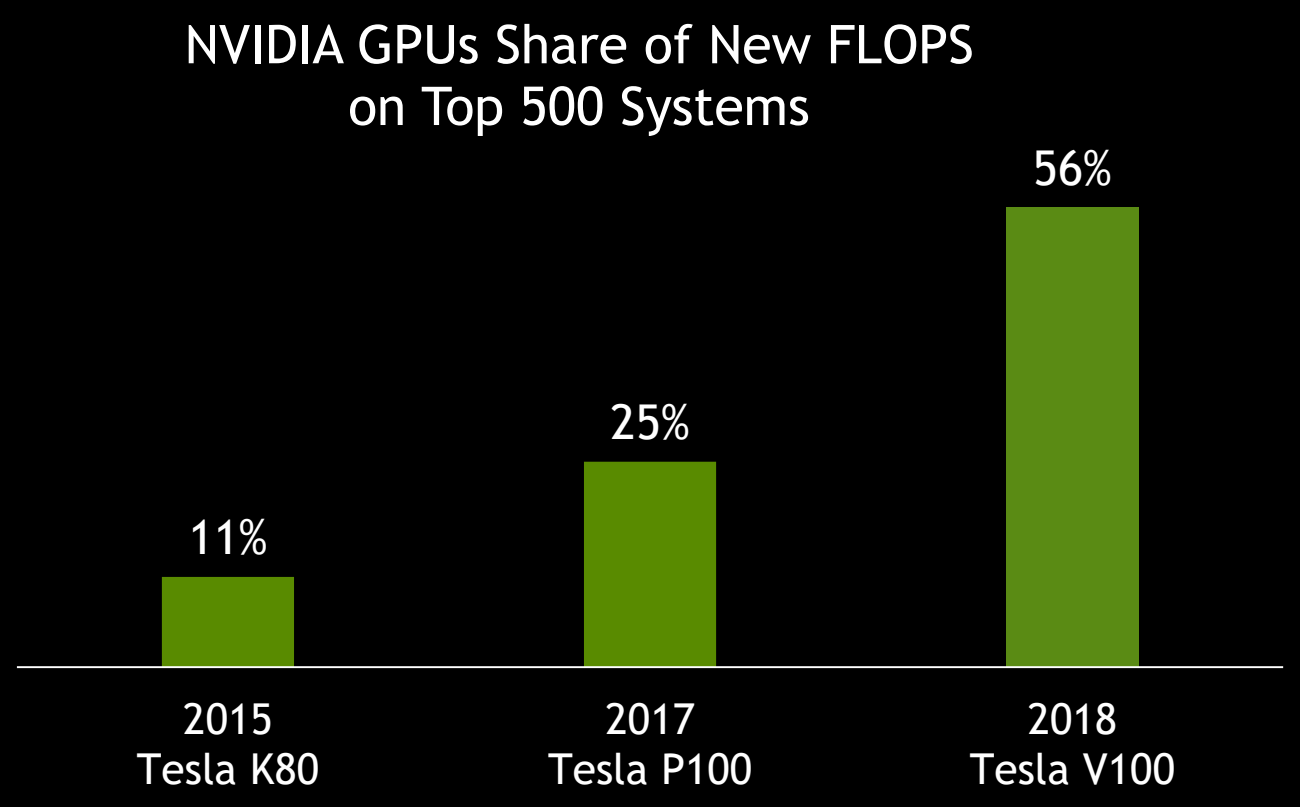


122 PetaFLOPS
HPC

3 ExaFLOPS
AI

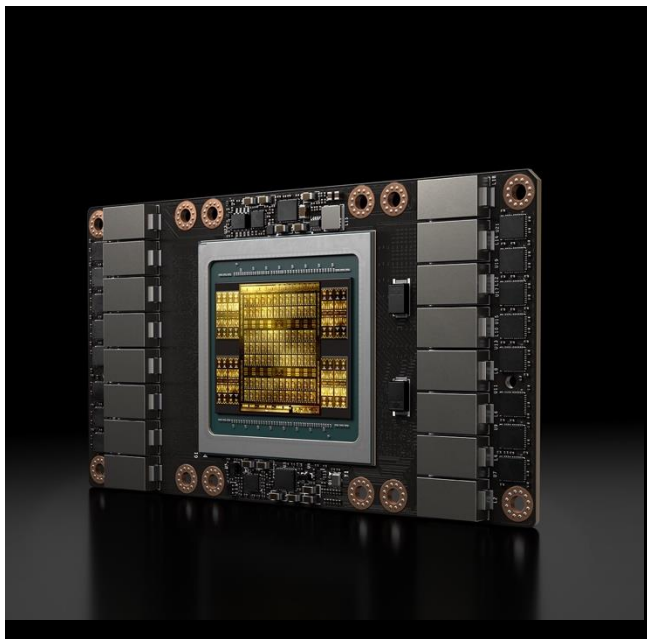
27,648
Volta Tensor Core GPUs

DELIVERING MAJORITY OF THE NEW COMPUTING PERFORMANCE

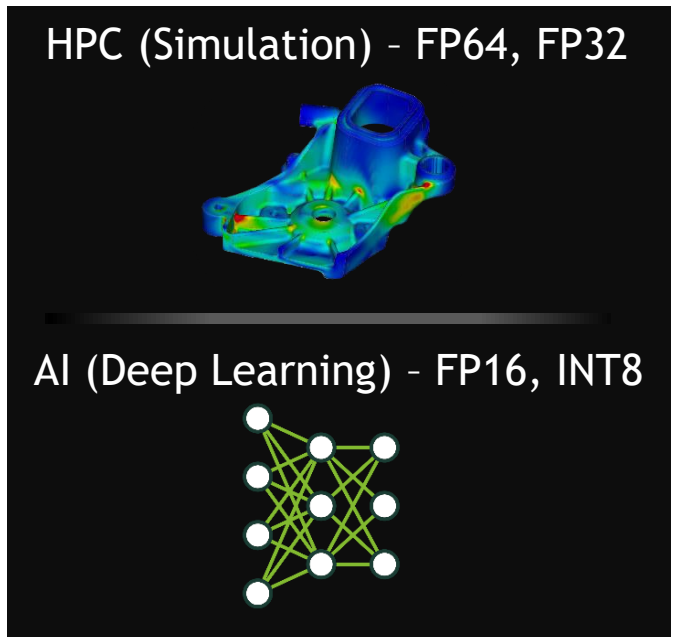


Thanks to NVIDIA innovation!

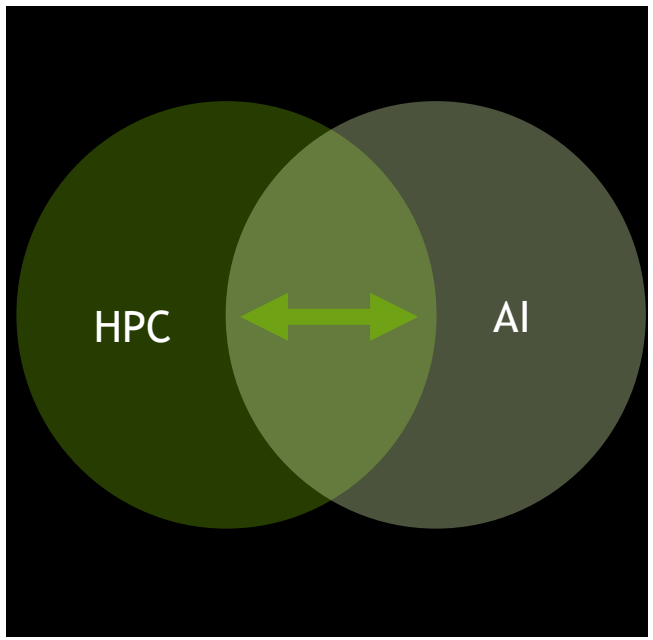
TENSOR CORE GPU FUSES HPC & AI COMPUTING



VOLTA TENSOR CORE GPU



MULTI-PRECISION COMPUTING



FUSION OF HPC & AI

TENSOR CORE

Mixed Precision Matrix Math - 4x4 matrices

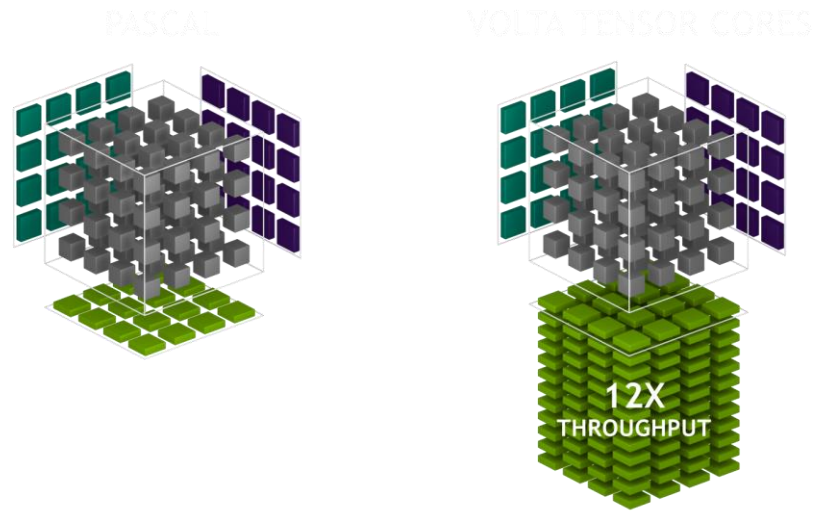
New CUDA TensorOp instructions
& data formats

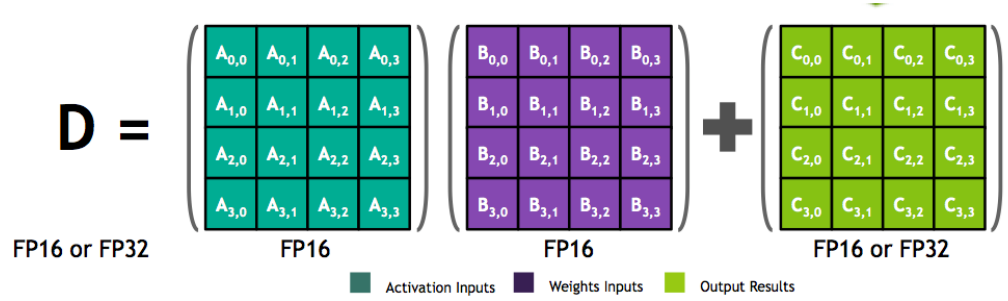
4x4 matrix processing array

D[FP32] = A[FP16] * B[FP16] + C[FP32]

Using Tensor cores via

- Volta optimized frameworks and libraries (cuDNN, CuBLAS, TensorRT, ..)
- CUDA C++ Warp Level Matrix Operations



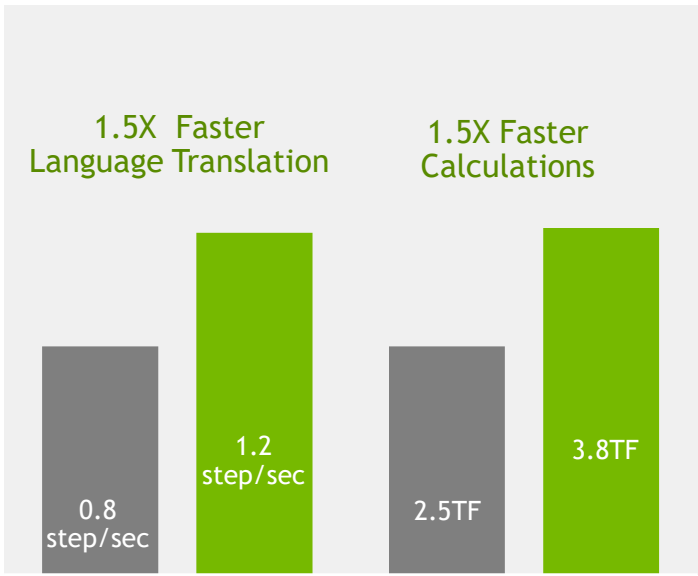


FASTER RESULTS ON COMPLEX DL AND HPC

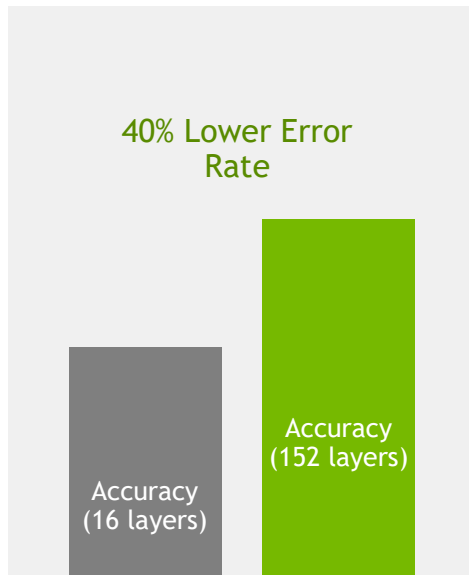
V100: Up to 50% Faster Results With 2x The Memory

■ V100 16GB ■ V100 32GB

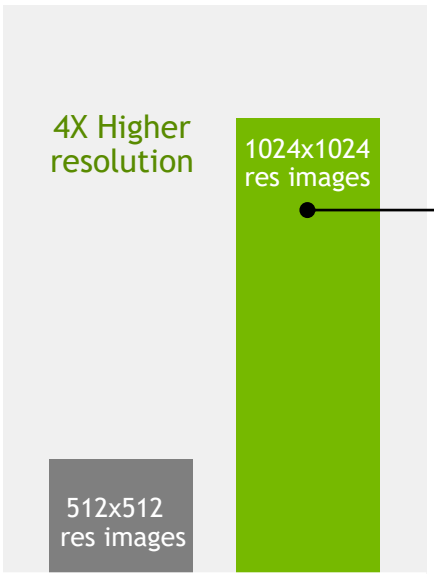
FASTER RESULTS



HIGHER ACCURACY



HIGHER RESOLUTION



Unsupervised Image Translation

Input winter photo

AI converts it to summer

Dual E5-2698v4 server, 512GB DDR4, Ubuntu 16.04, CUDA9, cuDNN7| NMT is GNMT-like and run with TensorFlow NGC Container 18.01 (Batch Size= 128 (for 16GB) and 256 (for 32GB) | FFT is with cufftbench 1k x 1k x 1k and comparing 2 V100 16GB (DGX1V) vs. 2 V100 32GB (DGX1V)

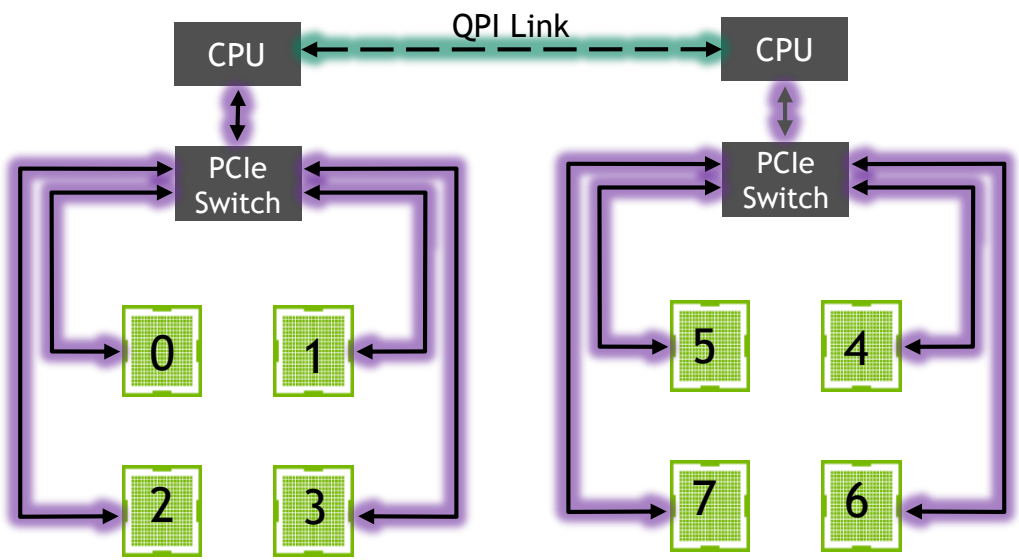
R-CNN for object detection at 1080P with Caffe | V100 16GB uses VGG16| V100 32GB uses Resnet-152

GAN by NVRESEARCH (https://arxiv.org/pdf/1703.00848.pdf) | V100 16GB and V100 32GB with FP32

NVLINK AND MULTI-GPU SCALING

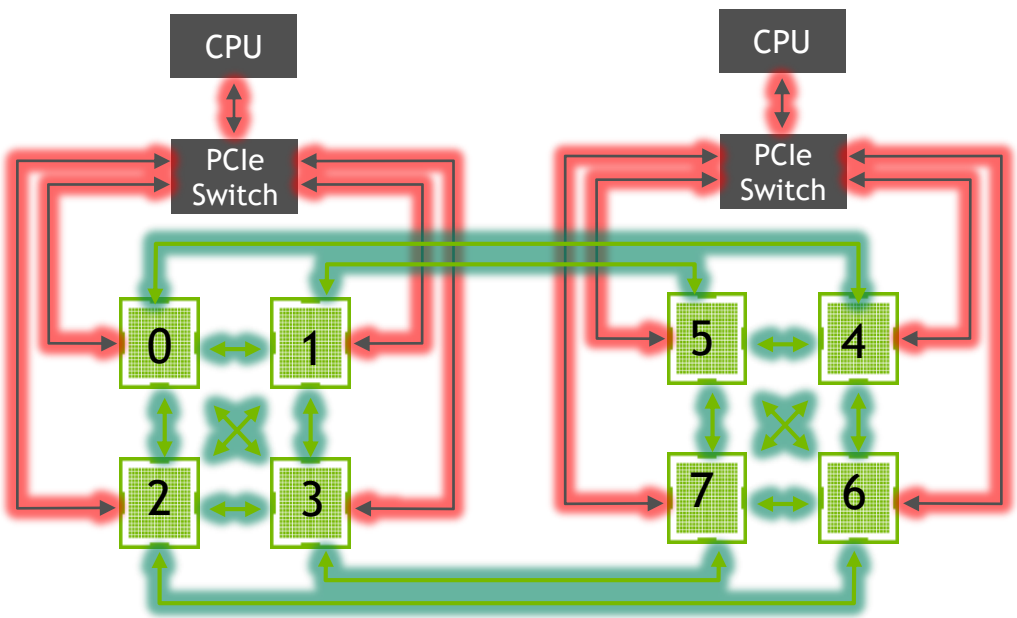
For Data Parallel Training

PCIe based system



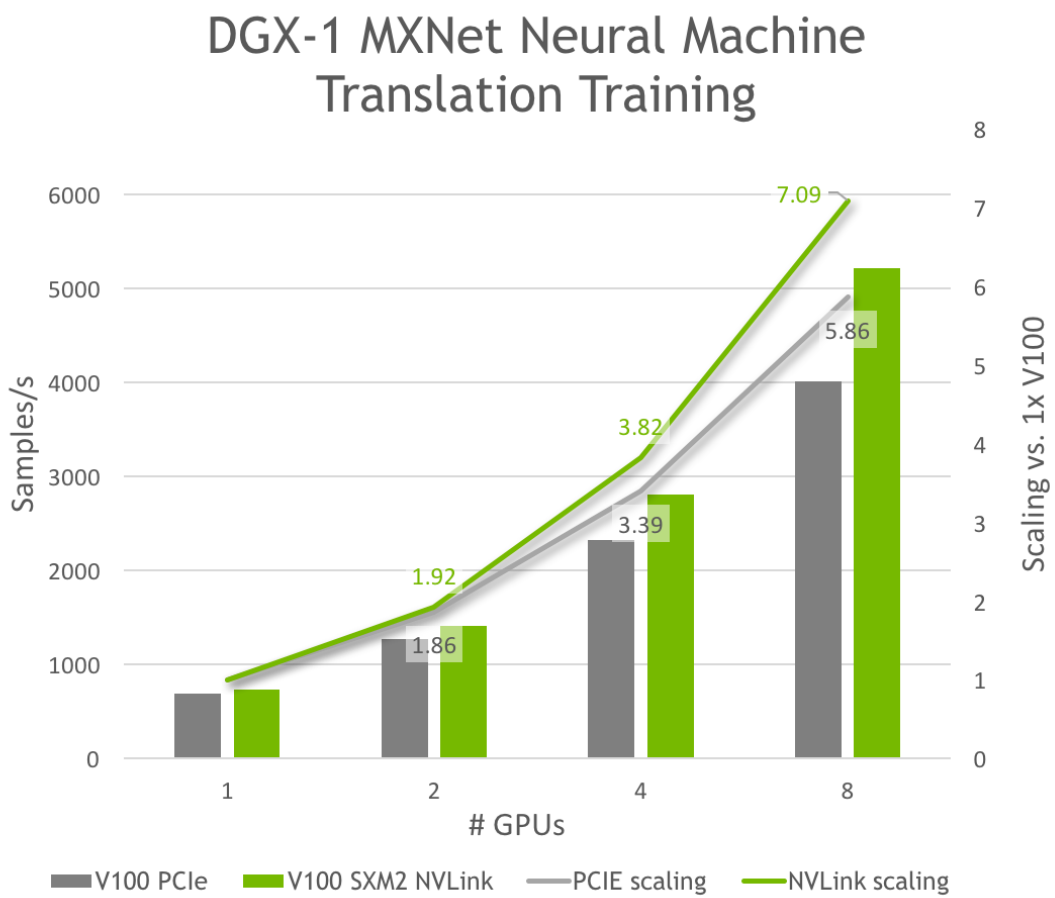
- Data loading over PCIe
- Gradient averaging over PCIe and QPI
- Data loading and gradient averaging share communication resources: Congestion

NVLINK based system



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

30% BETTER PERFORMANCE WITH NVLINK THAN PCIE



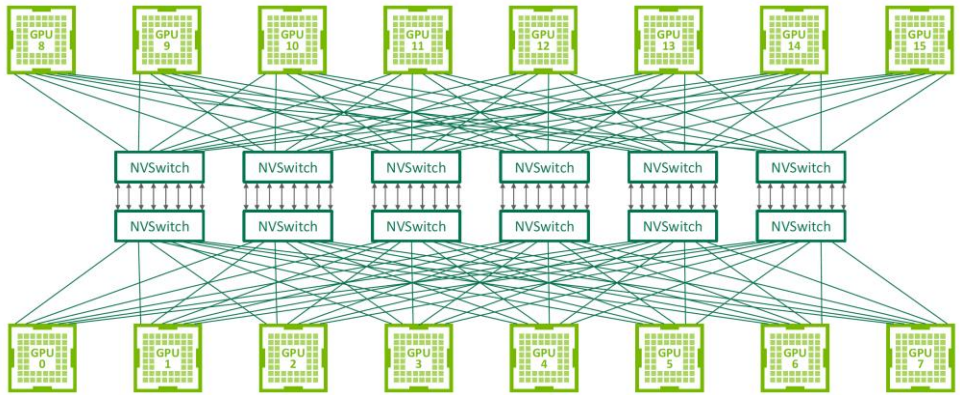
- Encoder and decoder embedding size of 512
- Batch size of 256 per GPU
- NVIDIA DGX containers version 17.11, processing real data with cuDNN 7.0.4, NCCL 2.1.2

NVSWITCH

Features:

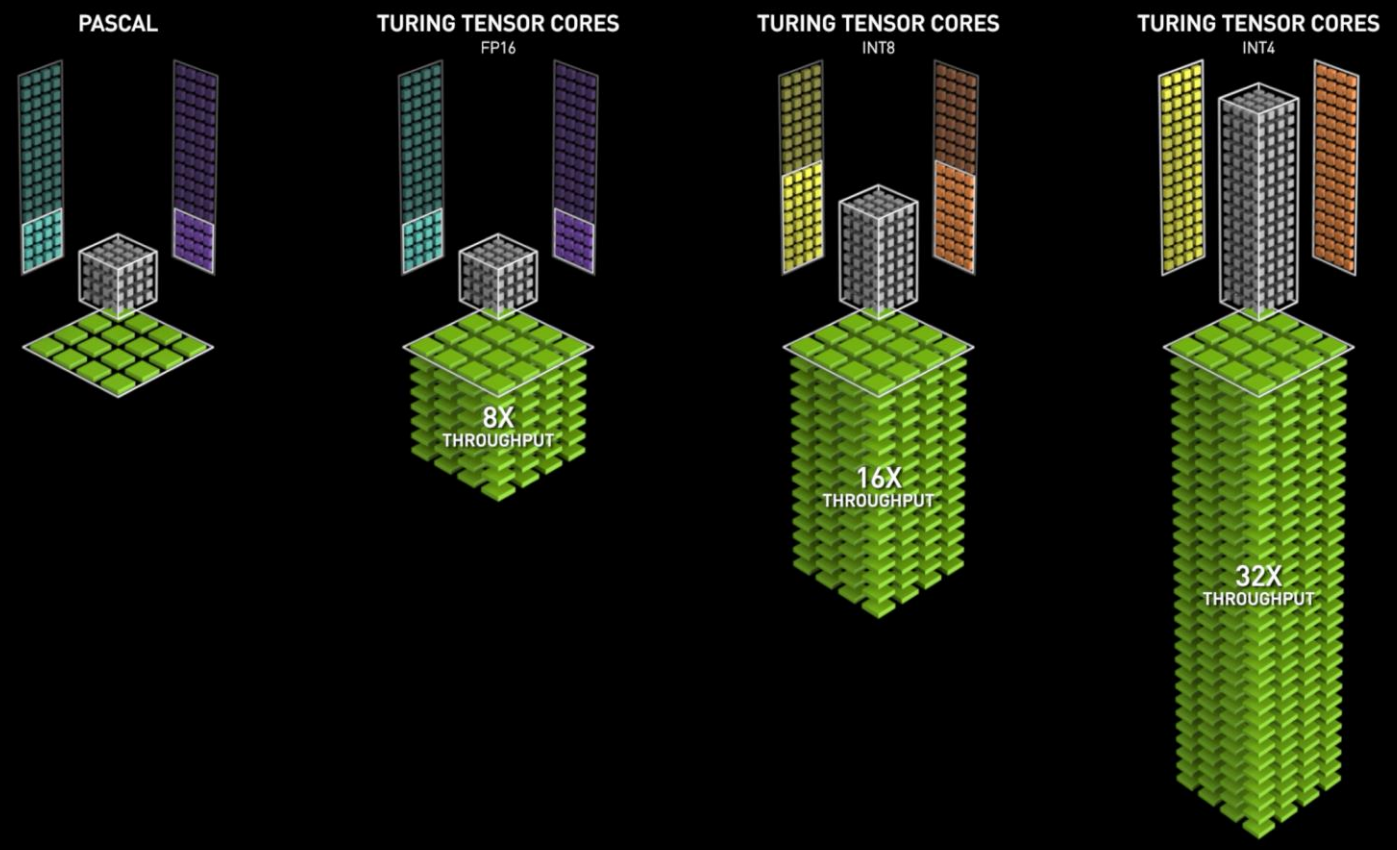
- 18 NVLink ports
  - @ 50GB/s per port
  - 900 GBs total
- Fully connected crossbar
- x4 PCIe Gen2 Management Port
- GPIO
- I2C

Transistor count:
2 billion



NEW TURING TENSOR CORE

MULTI-PRECISION FOR AI INFERENCE
65 TFLOPS FP16 | 130 TeraOPS INT8 | 260 TeraOPS INT4



Announced at GTC Europe

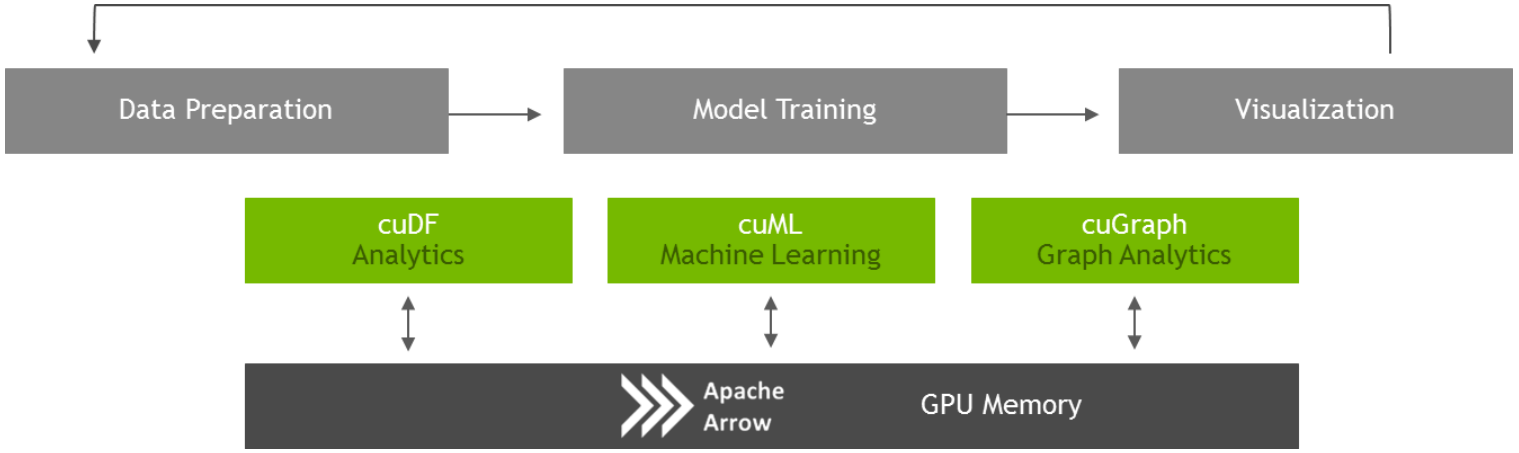
RAPIDS

rapids.ai

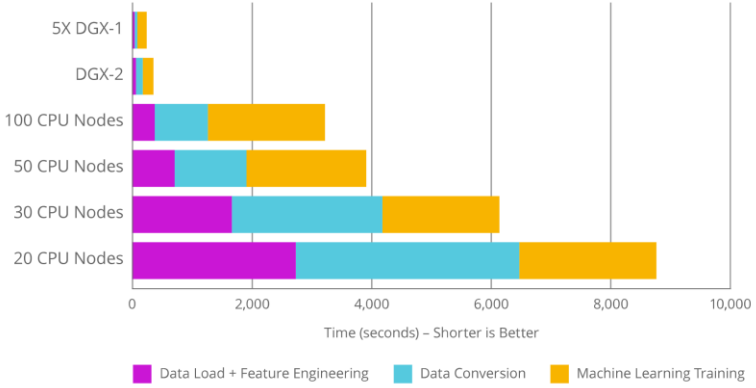
Dramatically increase accuracy, reduce training time and infrastructure costs

GPU Accelerated Data Science

RAPIDS : a set of open source libraries for GPU accelerating **data preparation** and **machine learning.**



End-to-End Faster Speeds on RAPIDS



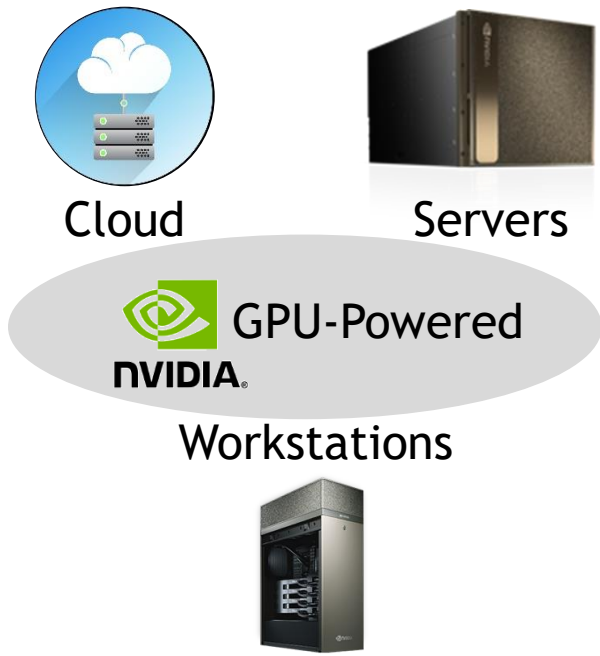
game-changing 50x speedups

NVIDIA GPU CLOUD

NVIDIA GPU CLOUD (NGC)

Simple Access to GPU-Accelerated Software

Discover 35 Optimized Containers



Run Anywhere with Maximum Performance

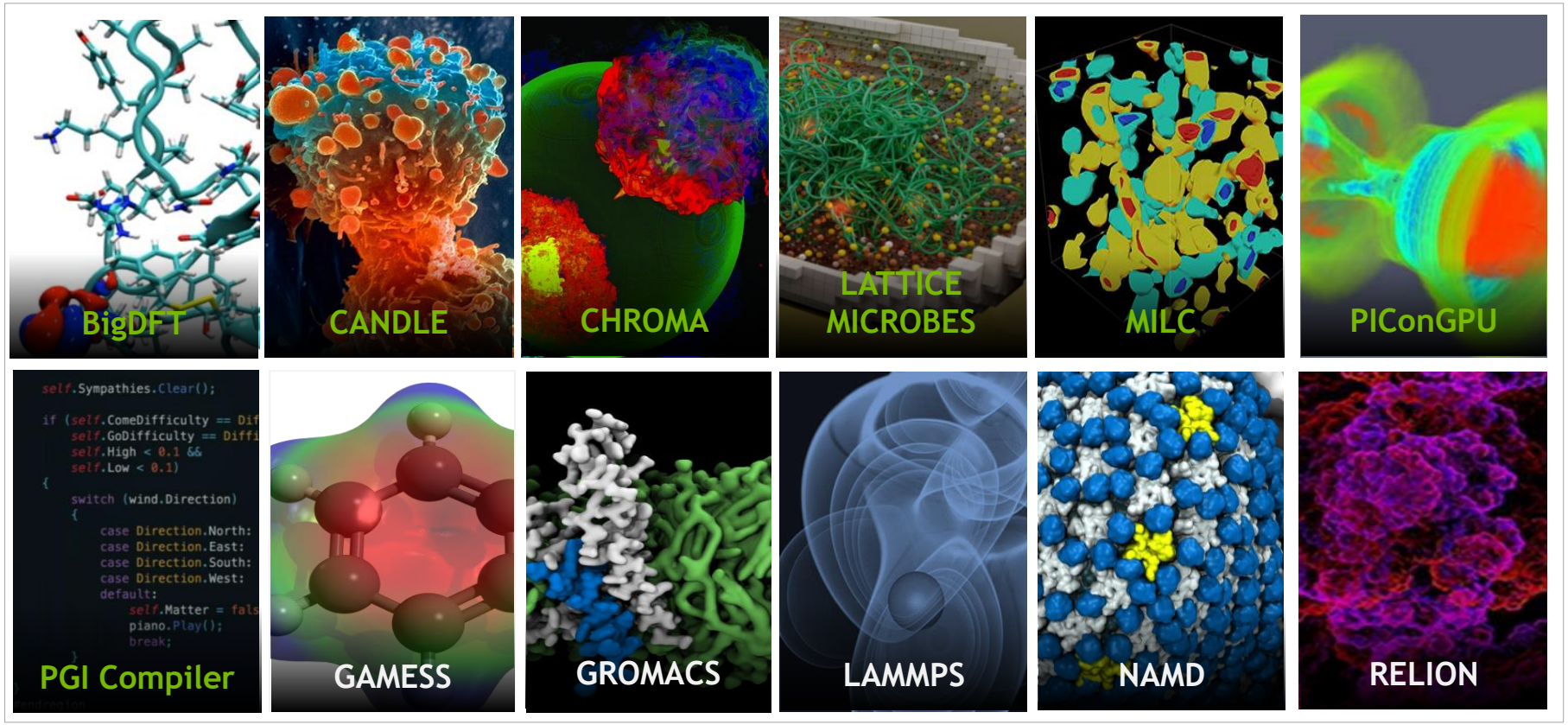
Accelerate
Time to Market

Deploy Applications In Minutes, Not Days

DEEP LEARNING CONTAINERS ON NGC

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

HPC APP CONTAINERS ON NGC



京都大学 KYOTO UNIVERSITY
JOHNS HOPKINS UNIVERSITY
DKRZ DEUTSCHES KLIMARECHENZENTRUM
Pfizer
KAUST
Schlumberger
MONASH University
Stanford University
PennState
UNIVERSITY OF CAMBRIDGE
YONSEI UNIVERSITY

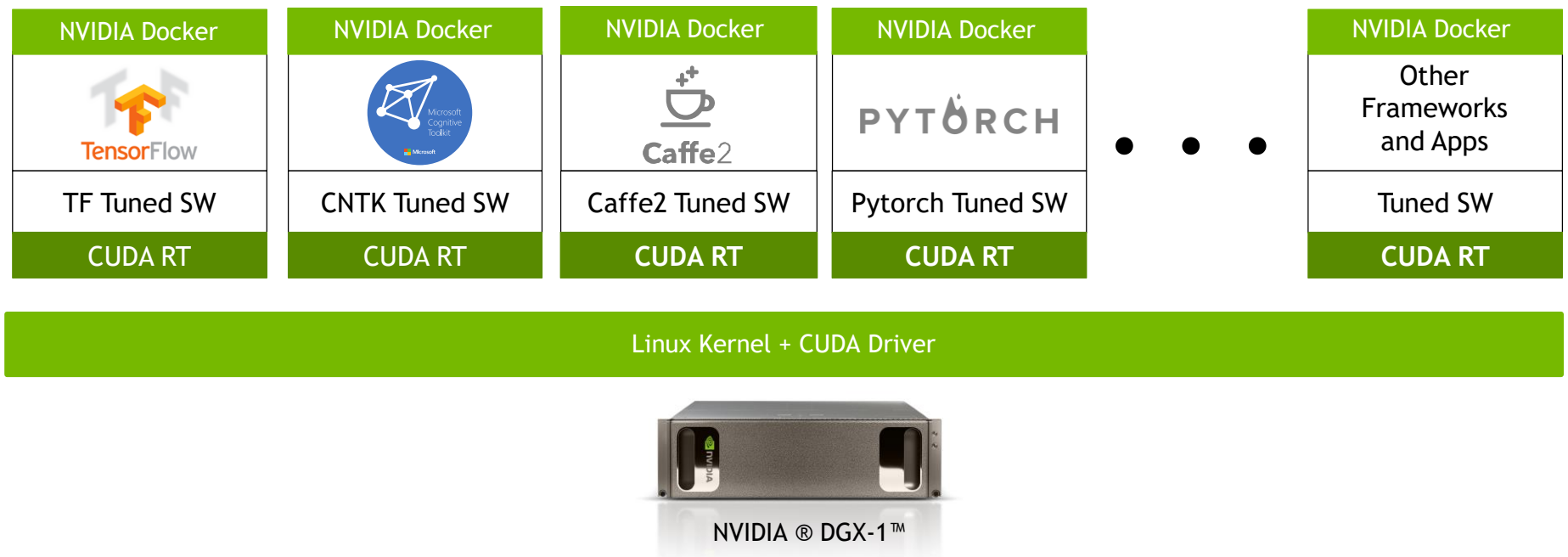
RAPID CONTAINER ADDITION

RAPID USER ADOPTION

THE POWER TO RUN MULTIPLE FRAMEWORKS AT ONCE

Container Images portable across new driver versions

Containerized Applications



TESLA PRODUCT FAMILY

END-TO-END PRODUCT FAMILY

HPC/ TRAINING

NVIDIA GPU CLOUD

INFERENCE

FULLY INTEGRATED AI SYSTEMS

DGX-1

DGX-2

DESKTOP

TITAN

VIRTUAL WS

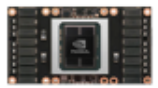
Virtual GPU

WORKSTATION

Quadro

DGX Station

DATA CENTER



Tesla V100

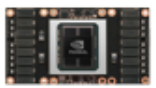
V100 PCIE

SERVER CONFIGS

HGX

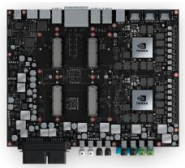
DATA CENTER

Tesla T4



Tesla V100

AUTOMOTIVE



Drive AGX Pegasus

EMBEDDED

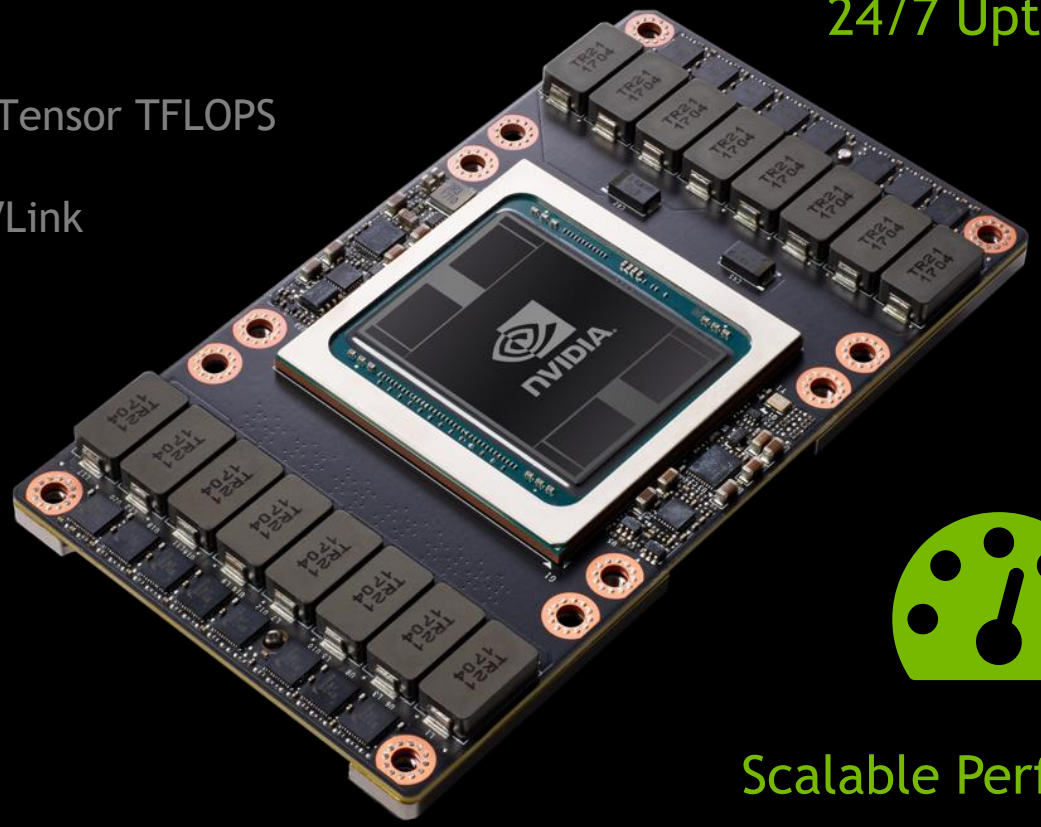
Jetson AGX Xavier

TESLA V100

WORLD'S MOST ADVANCED 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
16GB/ 32GB HBM2 @ 900GB/s | 300GB/s NVLink

24/7 Uptime



Data Center Ready

Scalable Performance

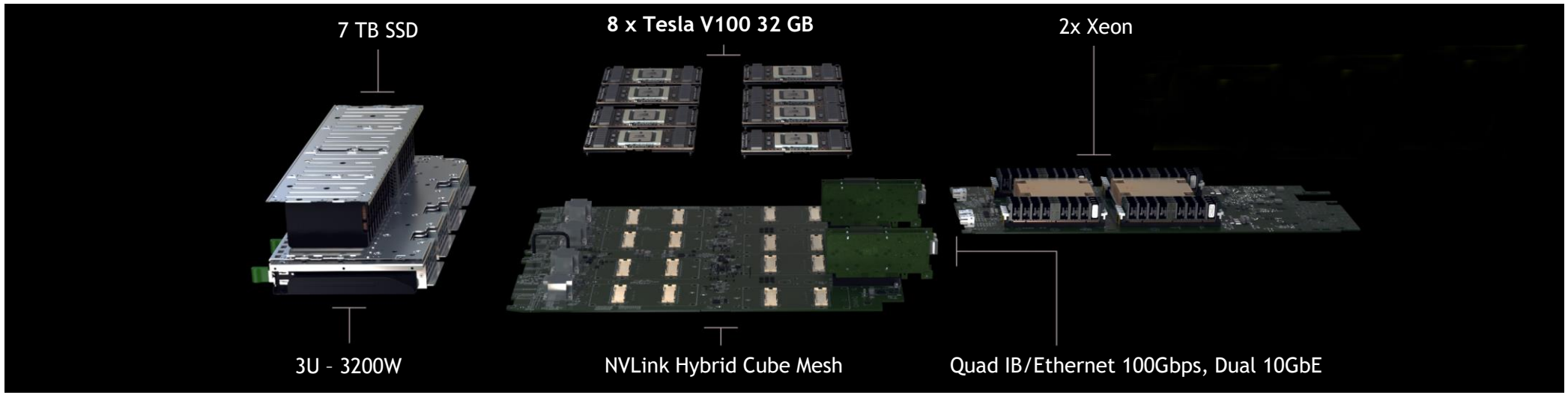
TESLA T4

WORLD'S MOST ADVANCED INFERENCE GPU

Universal Inference Acceleration
320 Turing Tensor cores
2,560 CUDA cores
65 FP16 TFLOPS | 130 INT8 TOPS | 260 INT4 TOPS
16GB | 320GB/s

NVIDIA DGX-1 WITH VOLTA

Highest Performance, Fully Integrated HW System

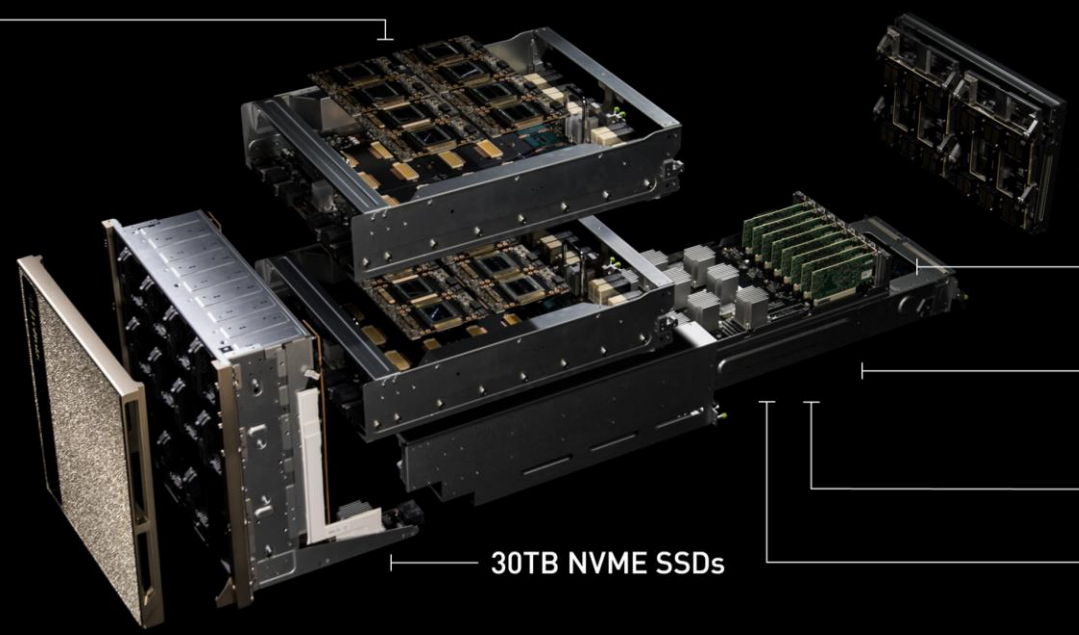


1 PetaFLOPS | 8x Tesla V100 32GB | 300 Gb/s NVLink Hybrid Cube Mesh

2x Xeon | 7 TB RAID 0 | Quad IB/Ethernet 100Gbps, Dual 10GbE | 3U — 3500W

NEW NVIDIA DGX-2

The Largest GPU Ever Created



2 PFLOPS | 512GB HBM2 | 16 TB/sec Memory Bandwidth | 10 kW / 160 kg

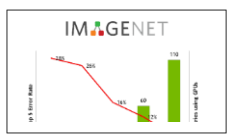
POWERING THE DEEP LEARNING ECOSYSTEM

NVIDIA SDK accelerates every major framework

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION



SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

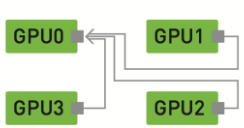
DEEP LEARNING FRAMEWORKS

Caffe | Caffe2 | Chainer | Cognitive Toolkit | KALDI | mxnet | PaddlePaddle | PYTORCH | TensorFlow

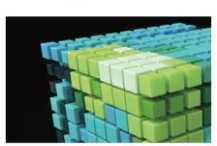
NVIDIA DEEP LEARNING SDK and CUDA

cuDNN

NCCL



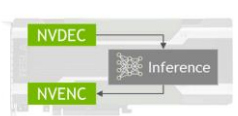
cuBLAS



TensorRT

cuSPARSE

DeepStream SDK

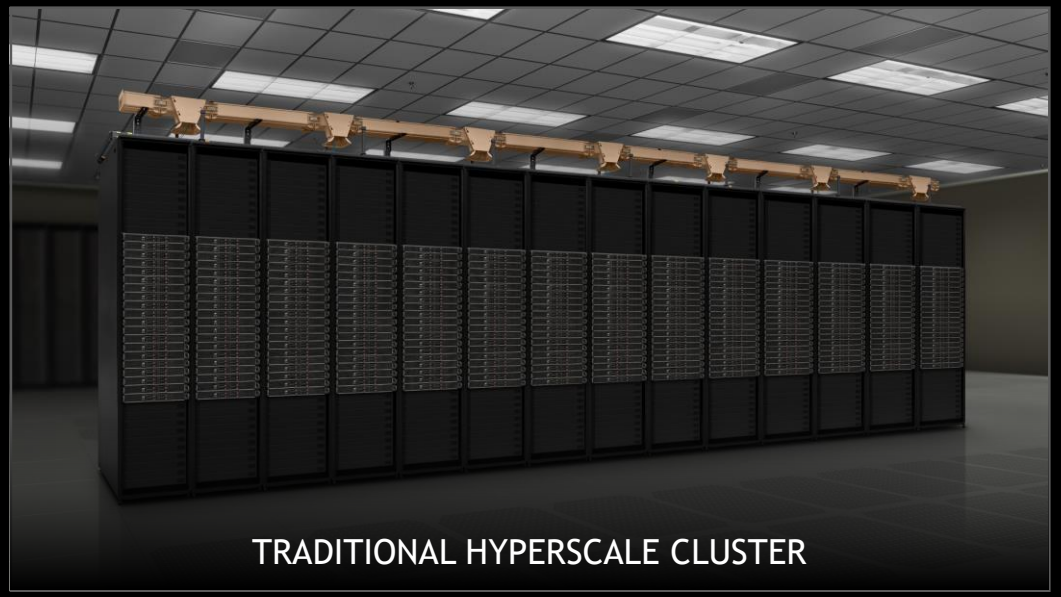


developer.nvidia.com/deep-learning-software

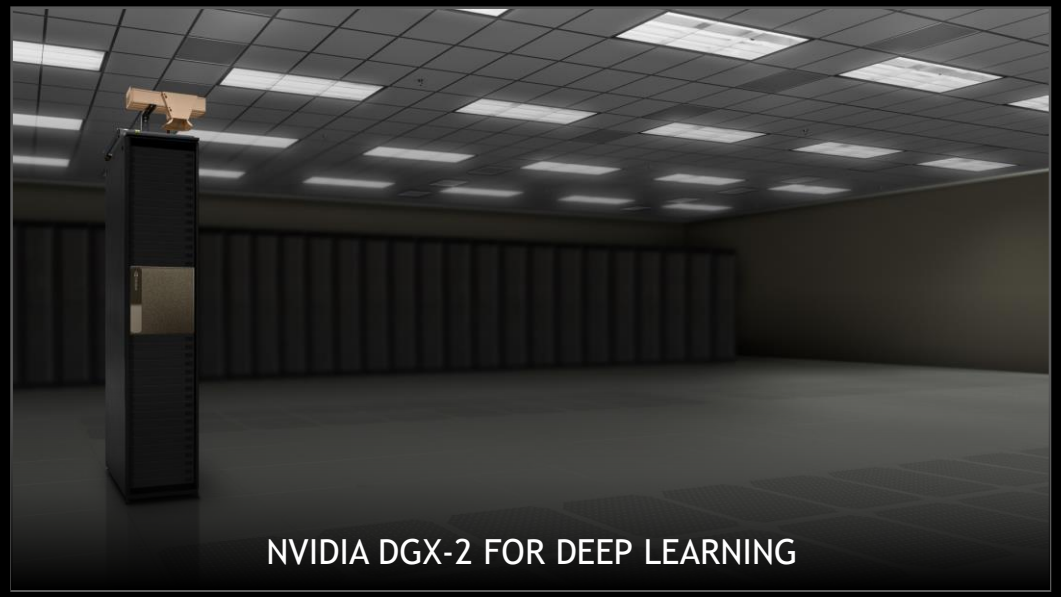
The more you buy, the more you save!

"The More GPUs You Buy, The More You Save"

—Jensen Huang



300 Dual-CPU Servers | $3M | 180 kW



1 DGX-2 | $399K | 10kW

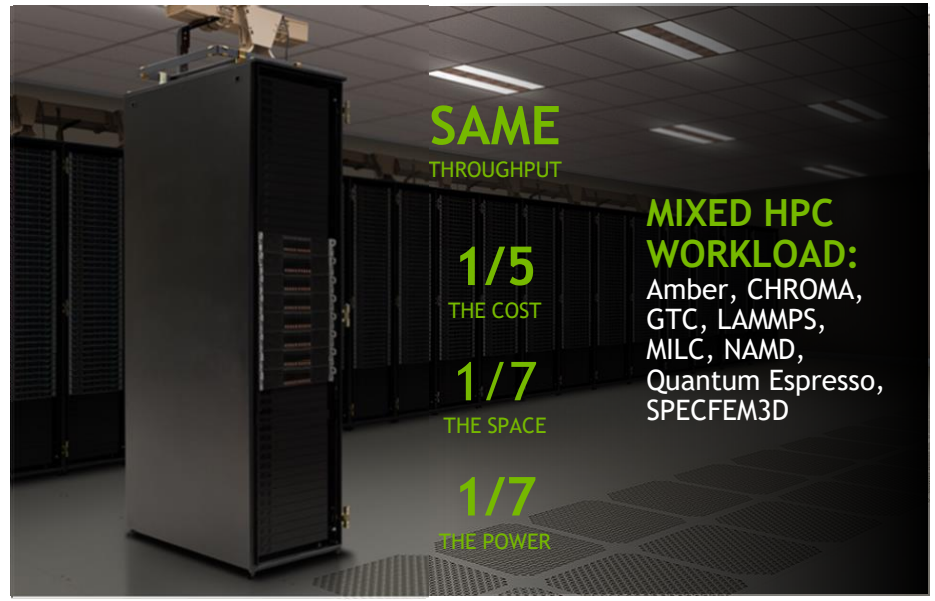
1/8 the Cost 1/60 the Space 1/18 the Power

DRAMATICALLY MORE FOR YOUR MONEY

5X Better HPC TCO for Same Throughput

160 Self-hosted Skylake CPU Servers
96 KWatts

=



8 Accelerated Servers w/4 V100 GPUs
13 KWatts

NVIDIA PROGRAMS

DEEP LEARNING INSTITUTE

DLI Mission: Help the world to solve the most challenging problems using AI and deep learning

We help developers, data scientists and engineers to get started in architecting, optimizing, and deploying neural networks to solve real-world problems in diverse industries such as autonomous vehicles, healthcare, robotics, media & entertainment and game development.

NVIDIA DEEP LEARNING INSTITUTE

HOW TO ACCESS DLI TRAINING

SELF-PACED ONLINE

Get started anywhere, any time with access to a GPU-accelerated workstation in the cloud

All you need is a device with an Internet connection

Courses (8 hrs) are $90

Electives (2 hrs) are $0-30

Take online courses at www.nvidia.com/dli

INSTRUCTOR-LED WORKSHOP

Full-day workshops are available by request

Workshops are delivered by DLI certified instructors through NVIDIA or DLI partners

MSRP: $10K/day for up to 20 attendees (EDU pricing available)

Request a workshop at www.nvidia.com/requestDLI

INDUSTRY CONFERENCES

Training available as instructor-led and self-paced at industry events

Deep learning presentations offered for business & technology leaders

Special training pass available for GTC (NVIDIA's GPU Technology Conference)

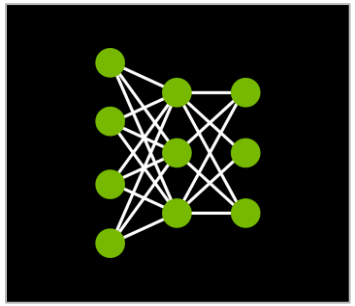
View upcoming DLI workshops at www.nvidia.com/dli

NVIDIA CONFIDENTIAL. DO NOT DISTRIBUTE.

RICH CONTENT PORTFOLIO

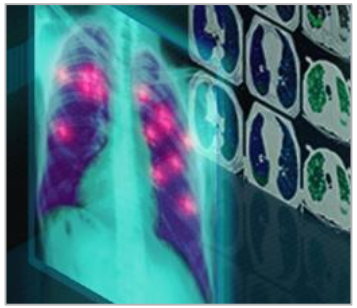
Fundamentals and advanced hands-on training in key technologies and application domains that matter

Caffe2 | Microsoft Cognitive Toolkit | mxnet | TensorFlow | PYTORCH

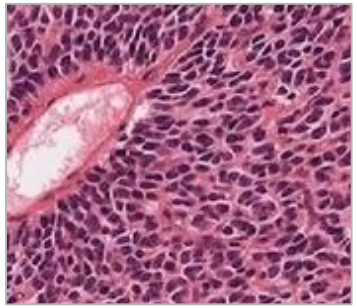


Deep Learning Fundamentals

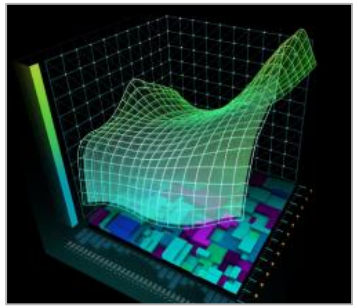
Autonomous Vehicles



Medical Image Analysis



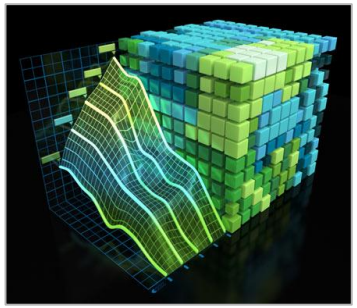
Genomics



Finance

Digital Content Creation

Game Development



Accel. Computing Fundamentals

More industry-specific training coming soon…

UNIVERSITY AMBASSADOR PROGRAM

Training the next generation of AI practitioners

University Ambassador Program enables qualified faculty and researchers to teach DLI courses to their students and academic staff at no cost

40 universities around the world are part of the Ambassador Program

Arizona State University

Columbia University in the City of New York

香港科技大學 The Hong Kong University of Science and Technology

Massachusetts Institute of Technology

NUS National University of Singapore

University of Oxford

NVIDIA